Table des matières

Remerciements

Je tiens tout d'abord à remercier Vincent Lemaire qui m'a fait confiance en me procurant ce stage. Je le remercie pour son encadrement rigoureux et son professionnalisme. Je remercie aussi Nicolas Voisine pour ses remarques pertinentes et ses conseils qui m'ont permis de faire des analyses plus pertinentes de mes résultats. Je manifeste aussi ma gratitude à Miguel Vazquez, éric Legal, mes collègues de bureau. Mes remerciements vont aussi à toute l'équipe TSI de France Télécom R&D Lannion et son manager Fabrice Clérot pour leur accueil et leur soutien. Je manifeste également ma reconnaissance au Groupe France Télécom. Enfin je tiens à remercier mes parents et ma famille qui ont su m'accompagner tout au long de ces années d'études.

1

Chapitre 1

Introduction

1.1 Présentation de France Télécom

Le groupe France Telecom, opérateur historique de télécommunications français, emploie 207 000 personnes et a aujourd'hui près de 126 millions de clients à travers le monde. Il intervient dans trois secteurs principaux :

- la téléphonie fixe
- le téléphonie mobile
- Internet

France Télécom est présent sur les cinq continents (220 pays ou territoires), à travers des marques et filiales d'envergure internationale comme Orange, Wanadoo, Equant, GlobeCast et TP. Le chiffre d'affaires du groupe est de 47.2 milliards d'euros, en progression de 4.1% en 2004. France Télécom a défini, depuis le 1er janvier 2005, quatre nouveaux segments d'activité :

1. Le segment " Services de communication Personnels " ("Personal") rassemble les activités de services mobiles en France, au Royaume-Uni, en Pologne et dans le Reste du Monde ;
2. le segment " Services de communication Résidentiels " (" Home ") rassemble les activités de services fixes de télécommunication (téléphonie fixe, services Internet, services aux opérateurs) et les revenus de la distribution et des fonctions supports fournies aux autres segments du groupe France Télécom ;
3. le segment "Entreprises " regroupe les services aux entreprises en France et les services mondiaux ;
4. le segment " Annuaires " consolide les activités de la filiale PagesJaunes Groupe.

1.1.1 Organisation de France Télécom

France Telecom dirigée par Didier Lombard est alors structurée en :

- Cinq divisions opérationnelles orientées vers la demande des clients et les marchés correspondants ;

  - La Division *Service de communication entreprises* a en charge le développement des services de communications aux entreprises dans le monde entier et a en charge la vente des services qu'elle développe pour les grandes entreprises ;
  - la Division *Service de Communication Résidentiels* a en charge le développement de l'ensemble des services de communication à domicile, notamment les services haut-débit à travers le fixe en Europe. Elle regroupe les équipes actuelles de Wanadoo ;
  - la Division *Service de Communication Personnels* a en charge le développement des services de communication destinés aux particuliers au travers des supports mobiles. Elle regroupe l'intégralité de la filiale Orange ;
  - *la division Ventes et Services France* a en charge la distribution de tous les produits du groupe en France pour les marchés Grand Public, Petites et moyennes Entreprises. Elle représente également le groupe auprès des collectivités locales et elle a en charge la responsabilité des bassins d'emplois pour l'ensemble des salariés du Groupe sur son territoire ;
  - *la Division International* a en charge le suivi et le développement du groupe TPSA et des autres filiales du Groupe à l'étranger.
- cinq divisions métier chargées de l'amélioration de la performance opérationnelle du groupe
  - *La Division Réseaux, Opérateurs et Système d'Information* : en cohérence avec les évolutions technologiques récentes et à venir, les métiers réseaux et système d'information sont regroupés. Elle est en charge plus particulièrement du développement et de la gestion des réseaux de France Telecom, toutes technologies confondues, de la vente de services aux opérateurs tiers ainsi que du développement et de la maintenance de l'ensemble des systèmes d'information du groupe ;
  - *La Division Recherche & Développement* a en charge la conduite des programmes de recherche du Groupe et la valorisation de la propriété intellectuelle. Elle joue un rôle moteur dans l'ensemble des processus d'innovation ;
  - *La division Achats* a en charge l'optimisation des achats et des dépenses opérationnelles d'investissement. Ses missions et son périmètre restent inchangés ;
  - *La Division Programme TOP* s'assure de la mise en oeuvre du programme TOP et veille à sa bonne exécution dans le cadre de la nouvelle organisation ;
  - *La Division Intégrations* des Contenus a en charge les partenariats avec les fournisseurs de contenus et la coordination du développement des plates-formes technologiques des terminaux associées.
- cinq fonctions supports au service des divisions opérationnelles et des divisions métier, assurent la cohérence des politiques du Groupe : Finance, Ressources Humaines, Animation des Réseaux et Management et Communication Interne, Secrétariat Général, Communication Externe.

1.1.2 La division Recherche & Développement

Le stage s'est déroulé au centre de Recherche et Développement de France Telecom situé à Lannion. Ce centre R & D est l'un des plus importants de France Telecom. La division Recherche & Développement a la responsabilité de :

– Conduire les programmes de recherche du Groupe ;
– valoriser la propriété intellectuelle.

La R& D joue un rôle moteur dans l'ensemble des processus d'innovation. France Télécom est l'opérateur qui accorde les moyens les plus importants à l'innovation, facteur majeur de différenciation. En 2005, les investissements en Recherche & Développement (R&D) représentaient 1.5% du chiffre d'affaires. Avec 4200 chercheurs, 16 implantations dont 8 à l'étranger et un portefeuille de 7300 brevets (en mars 2005), cette capacité d'innovation intégrée au service de la croissance place le Groupe parmi les cinq premiers centres mondiaux de recherche et développement en télécommunications.

La Division Recherche & Développement est sous la responsabilité de pascal Viginier et est composée de 6 Centres de R&D (CRD) structurés autour de service intégrés et de la convergence des réseaux :

– *CRD Services intégrés, résidentiels et personnels* a pour mission d'aider les divisions opérationnelles du Groupe à tenir leurs roadmaps de services R&P, d'imaginer, de concevoir, de faire développer des services intégrés résidentiels, nomades et personnels générateurs de CA, enfin de parfaire la connaissance des clients R&P, de leurs usages, et des offres des autres acteurs mondiaux du secteur ;
– *CRD Services aux entreprises* est chargé d'aider les divisions opérationnelles du Groupe à développer le CA sur le marché des Entreprises grâce à l'innovation de services et a l'anticipation des usages, notamment en développant des services et communication intégrée d'entreprise, de nomadisme, de relation clients, et des services avancés sur réseaux privés virtuels. Il doit également apporter conseil et intégration dans le cadre d'offres à des grands clients ;
– *CRD Middleware et Plates-formes avancées* est chargé de développer les composants middleware (primitives) et les plates-formes de services de l'opérateur intégré. Il définit un urbanisme de plates-formes de services inter opérables, mutualisant les fonctions et données communes, et en cohérence avec le système d'information et le coeur du réseau. Il est chargé de définir une architecture technique modulable et les outils de création de services associés pour diminuer le time to market des services et les coûts. Enfin, le CRD permet de déployer les services de communication, de production et de consommation de contenus sur tous les terminaux ;
– *CRD Technologies* a pour mission de détecter et d'identifier les ruptures impactant sur le groupe dans les domaines des technologies de la voix et de l'image, des terminaux et des nouvelles interfaces et interaction, des nano et biotechnologies communicantes, ainsi que dans les évolutions d'usages de perception client et de business models ; il veille et assure au Groupe un positionnement compétitif sur les technologies stratégiques et la normalisation associée. Il développe et fournit les briques technologiques des nouveaux services intégrés du groupe et contribue au développement des revenus de FT par la protection intellectuelle et la valorisation de ses travaux et expertises.
– *CRD coeur de réseau* assure les développements du coeur de réseau fixe et mobile en maintenant une cohérence et un urbanisme favorisant l'intégration des services, pour la voix et les données (transport, collecte et longue distance). Il

définit l'évolution de l'architecture des réseaux, en particulier pour la convergence des réseaux coeur, le multiservice et le haut débit, la VoIP, la sécurité et QOS des réseaux support. Il identifie les ruptures potentielles des nouvelles technologies en coeur de réseau, notamment du point de vue économique ;

– *CRD Réseaux d'accès* étudie les réseaux d'accès et domestiques qui utilisent un support de transmission filaire ou radio ainsi que les services de transmission de données pour les entreprises. Il propose des évolutions pour faire de la complémentarité des divers réseaux d'accès un avantage compétitif pour le groupe France Telecom.

La R& D de France Telecom est présente en France sur 8 sites. Lannion est l'un des plus importants avec 1400 techniciens, ingénieurs et chercheurs répartis en 6 centres de R&D. Chaque centre a en son sein plusieurs laboratoires, dont 12 sont actifs à Lannion et 10 représentés par des unités de R&D.

1.1.3 Objectifs du stage

Le stage s'est déroulé dans l'unité *Traitement Statistique de l'Information* (TSI) représentant le laboratoire Sociologie des Usages et Traitement Statistique de l'Information (SUSI) sous la direction de Monsieur Vincent Lemaire, Docteur-Ingénieur à France Télécom R & D. Ce laboratoire fait partie du CRD Technologie que nous avons présenté précédemment. Le stage comporte deux parties :

1. *La sélection de variables* : il s'agit de déterminer le meilleur sous-ensemble de variables qui donnent les meilleurs résultats en prédiction. Une méthode de mesure d'importance des variables a été développée au sein de la R& D de France Télécom. Le premier objectif est de démontrer mathématiquement les résultats de la méthode pour un modèle de régression linéaire et un naïf bayes. Ensuite, on entamera une phase de tests qui permettra d'évaluer la méthode et de comparer ses performances à d'autres méthodes de mesure d'importance des variables ;
2. *Interprétation des résultats* : la deuxième partie traite du domaine de l'interprétation des résultats des modèles "boite noire". On veut répondre à la question : pourquoi un modèle délivre telle valeur en sortie ? La réponse à cette question permettra de savoir quelle action entreprendre pour modifier un score.

1.2 Notations

Nous utiliserons les notations suivantes :

– $f : R^n \rightarrow R^p$ représente le modèle ;
– $I = \{I_1, \ldots, I_N\}$ représente l'ensemble des exemples ;
– $V = \{V_1, \ldots, V_n\}$ représente l'ensemble des variables d'entrée du modèle ;
– $v = \{v_1, \ldots, v_n\}$ est une realisation de $V = \{V_1, \ldots, V_n\}$;
– $I_k = \begin{pmatrix} v_{k1} \ldots v_{kn} \end{pmatrix}$ est la représentation de l'exemple I_k ;
– $S = \{S_1, \ldots, S_p\}$ représente la sortie du modèle.

La figure 1.1 représente un modèle qui prend en entrée un vecteur de taille J et en sortie un vecteur de taille P. Le vecteur d'entrée peut représenter un exemple ou

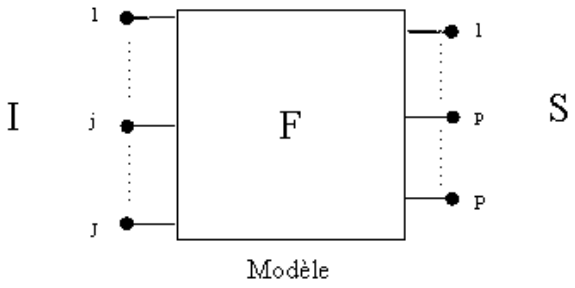

FIG. 1.1 – représentation schématique d'un modèle "boite noire"

objet dont chaque composante représente une caractéristique et chaque composante p du vecteur de sortie peut représenter la probabilité d'appartenir à la classe C_p si le modèle est un classifieur.

Chapitre 2

Sélection de variables

2.1 Etat de l'art des méthodes de séléction de variables

Dans de nombreuses situations, on est amené à construire un modèle prédictif pour modéliser un phénomène. Par exemple, l'éléctricité ne pouvant être stockée, les compagnies d'électricité utilisent des modèles pour prédire la consommation d'électricité. Ce modèle est une fonction f qui prend en entrée un vecteur I et délivre en sortie un vecteur S. Le vecteur I décrit un "objet". Chaque composante de ce vecteur représente une caractéristique (ou une variable) de cet objet. La taille de I peut être très grande. Toutes les variables ne sont pas aussi informatives : elles peuvent correspondre à du bruit, être peu significatives ou encore non pertinentes.

On est donc amené à sélectionner les variables les plus pertinentes. La sélection de variables permet de réduire la taille des informations et facilite la conception du modèle. De plus le modèle peut être plus efficace lorsque les variables non pertinentes ou redondantes sont supprimées. La sélection de variables a pour but de réduire les paramètres du modèle en sélectionnant le sous ensemble de variables de taille minimale qui permette d'avoir les meilleures performances.

Les méthodes de sélection de variables peuvent être regroupées au sein de deux grandes approches : celle des "wrappers"et celle des "filters". Chacune de ces approches comprend généralement les étapes suivantes :

- un algorithme permettant d'explorer l'espace des combinaisons des variables ;
- un critère permettant d'évaluer les différents sous-ensembles de variables ;
- un critère permettant d'arrêter l'algorithme de sélection (voir [3]).

Nous expliciterons ces approches dans les sections suivantes.

2.1.1 L'approche "Wrappers"

La méthode "wrappers" utilise le modèle pour déterminer le meilleur sous-ensemble de variables. Soit un sous-ensemble de variables S dont on souhaite évaluer la pertinence. Soit f_S le modèle qui lui est associé. Supposons qu'on ait un ensemble de N exemples. La méthode "wrappers" consiste à évaluer les performances du modèle f_S.

On dira que le sous-ensemble de variables *S* sera plus pertinent qu'un sous-ensemble de variables *A* si les performances du modèle f_S sont supérieures à celles du modèle f_A associé à *A*.

Par exemple, si le modèle est un classifieur, la performance du modèle sera le taux d'exemples bien classés. *S* sera alors plus pertinent que *A* si le taux d'exemples bien classés par f_S est plus grand que le taux d'exemples bien classés par f_A.

2.1.2 L'approche "Filters"

L'approche "Filters"contrairement à l'approche "Wrappers" n'utilise pas le modèle pour déterminer le "meilleur" ensemble de variables. Elle est donc indépendante du modèle. Elle utilise une fonction *J* (par exemple la corrélation) qui calcule l'indice de pertinence d'un sous-ensemble de variables *S*. Le meilleur sous-ensemble sera celui qui aura le plus grand indice de pertinence. Il existe plusieurs méthodes pour calculer l'indice de pertinence d'un ensemble de variables (voir [6] section 3) .
Elles sont regroupées sous les trois catégories suivantes :

- mesure de la pertinence basée sur la corrélation (coefficient de corrélation linéaire de Fisher, χ^2, le rapport signal sur bruit,....) ;
- mesure de l'indice de pertinence basée sur les distances entre les distributions de probabilité des variables ;
- mesure de l'indice de pertinence issue de la théorie de l'information (par exemple le gain d'information).

2.1.3 Algorithmes *"Backward" et "Forward"*

Pour sélectionner le meilleur sous-ensemble de variables, il existe plusieurs approches :

1. *L'approche exhaustive* :
   - on détermine tous les sous-ensembles de l'ensemble des variables candidates ;
   - on évalue la pertinence de chacun des sous-ensembles ;
   - on choisit le sous-ensemble dont l'indice de pertinence est le plus grand.

   Cette approche permet de trouver le meilleur sous-ensemble. Cependant les calculs deviennent très longs dès que le nombre de variables est supérieur à 5. Par exemple pour un ensemble de composé de 6 variables, on a $2^6 = 64$ sous-ensembles à explorer
2. *L'approche heuristique* : cette approche utilise essentiellement deux algorithmes : l'algorithme "Backward" et l'algorithme "Forward".

L'algorithme *"Forward selection"*

- soit $\{X_1, \ldots, X_n\}$ un ensembles de n variables candidates ;
- soit $A = \emptyset$, l'ensemble vide ;
- soit *J* une fonction qui permet d'évaluer la pertinence d'un sous-ensemble de variables (*J* représente le modèle par exemple) ;

- *Phase 1* : on calcule $J(X_i)$ pour tout $i = 1, \ldots, n$;
- on sélectionne X_k tel que $X_k = \max J(X_i)$;
- on ajoute X_k à A. On considère maintenant $A = \{X_k\}$;
- *Phase 2* :on calcule

$$J(X_k, X_i), \text{ pour } i = 1, \ldots, n \; i \neq k;$$

- on sélectionne X_l tel que $X_l = \max J(X_k, Xi)$;
- on ajoute X_l à A et on considère $A = \{X_k, X_l\}$;
- *Phase 3* : On continue le processus jusqu'à ce que l'un des 3 critères suivants soit satisfait :
  - le nombre de variables pré-défini est atteint ;
  - les performances du modèle sont satisfaisantes ;
  - les performances du modèle n'augmentent plus.

L'algorithme *"Backward selection"*

- soit $\{X_1, \ldots, X_n\}$ un ensembles de n variables candidates ;
- soit $A = \{X_1, \ldots, X_n\}$, l'ensemble de toutes les variables ;
- soit *J* une fonction qui permet d'évaluer la pertinence d'un sous-ensemble de variables (*J* représente le modèle par exemple) ;
- *Phase 1* : on calcule $J(i) = J(\{X_1, \ldots, X_n\} \backslash \{X_i\})$pour tout $i = 1, \ldots, n$;
- on extrait de A X_k tel que $X_k = \min J(i)$;
- on considère maintenant $A = \{X_1, \ldots, X_n\} \backslash \{X_k\})$;
- *Phase 2* : on calcule $J(i) = J(A \backslash \{X_k, X_i\})$, pour tout $X_i \in A$;
- on supprime de A X_l tel que $X_l = \min J(i)$;
- on considère maintenant A tel que $A = \{X_1, \ldots, X_n\} \backslash \{X_k, X_l\}$;
- *Phase 3* : on continue le processus jusqu'à ce que l'un des 3 critères suivants soit satisfait :

  - le nombre de variables pré-défini est atteint ;
  - les performances du modèle sont satisfaisantes ;
  - les performances du modèle n'augmentent plus.

2.1.4 Résumé

Les algorithmes de sélection de variables permettent de trouver un sous-ensemble de variables qui donnent les meilleures performances par rapport aux ensembles explorés. Ils ne garantissent pas que le sous-ensemble de variables sélectionné est le meilleur de tous les sous-ensembles de variables possibles. Le groupe de variables sélectionné donne des résultats satisfaisants.

2.2 Etat de l'art des méthodes de mesure d'importance d'une variable

Dans cette section, nous présentons quelques méthodes de mesure d'importance des variables d'entrée d'un modèle prédictif. Nous verrons des méthodes basées sur l'estimation de la covariance entre une variable et la sortie du modèle comme la méthode ARD (automatic relevance determination), le coefficient de correlation de Pearson, l'indice de pertinence de Gram-Schmidt. Nous verrons aussi une méthode issue de la théorie de l'information qui permet de calculer l'information apportée par une variable par rapport à la sortie désirée.

2.2.1 La méthode ARD (automatic relevance determination)

Cette méthode a été proposée par Mackay et neal (voir [6] pages 407-420).
Soit $D = (I_i, S_i), i = 1, \ldots, N$ où $I_i \in R^n$ est le vecteur d'entrée du modèle et S_i la sortie désirée. Soit f le modèle prédictif et $f(I_i), i = 1, \ldots, N$ les sorties du modèle. La covariance entre les sorties $f(I_i)$ et $f(I_j)$ est définie en utilisant la fonction noyau gaussienne :

$$cov[f(I_i), f(I_j)] = k_0 exp(-\frac{1}{2}\sum_{l=1}^{n} k_{a,l}(I_{i,l} - I_{j,l})^2) + k_b$$

où l est l'indice de la variable, k_0 et k_b sont des constantes strictement positives.

Les paramètres ARD sont les coefficients $k_{a,l}$. Pour tout l, $k_{a,l}$ représente le degré de pertinence de la *l-ième* variable. On peut construire une matrice de covariance Σ où chaque élément de Σ est égale à $cov[f(I_i), f(I_j)]$. L'indice de pertinence de chaque variable est la valeur optimale des vecteurs ARD $(k_a^l)_{l=1}^n$. Pour ordonnancer les variables, on considère l'ensemble $(r^i)_{i=1}^n$ de taille n des indices de pertinence où

$$r^i = \frac{k_{a,i}}{\sum_{j=1}^{n} k_{a,j}}$$

sont les paramètres ARD normalisés.

2.2.2 Le coefficient de correlation de Pearson

Le coefficient de correlation de Pearson est un indice de pertinence qui permet de mesurer l'importance des variables individuellement et de les ordonnancer.Soit X la variable dont on souhaite mesurer la pertinence et x_i la réalisation de X pour l'exemple i. Soit Y la variable de sortie désirée du modèle prédictif et y_i la réalisation de Y pour l'exemple i ; le coefficient de correlation de Pearson pour la variable X est égale à :

$$C(X,Y) = \frac{|\sum_{i=1}^{N}(x_i - E(X))(y_i - E(Y))|}{\sqrt{\sum_{i=1}^{N}(x_i - E(X))^2 \sum_{i=1}^{N}(y_i - E(Y))^2}}$$

où $E(.)$ représente l'espérance mathématique.

Nous introduisons une variante de cette méthode pour calculer l'importance d'une variable étant donné un modèle f. Soit V_j la variable dont on souhaite mesurer la pertinence et V_{ij} la réalisation de V_j pour l'exemple i. Soit f le modèle prédictif ; le coefficient de correlation de Pearson pour la variable V_j est égale à :

$$C(V_j|f) = \frac{|\sum_{i=1}^{N}(V_{ij} - E(V_j))(f(I_i) - E(f)|}{\sqrt{\sum_{i=1}^{N}(V_{ij} - E(V_j))^2 \sum_{i=1}^{N}(f(I_i) - E(f))^2}}$$

où $f(I_i)$ représente la sortie du modèle pour l'exemple I_i et $E(.)$ représente l'espérance mathématique.

2.2.3 Indice de pertinence de Gram-Schmidt

Cette méthode a été développée par G. Dreyfus et I. Guyon (voir [6] chapitre 2). Cet indice de pertinence est aussi basée sur une mesure de correlation entre une variable X d'entrée d'un modèle prédictif et Y de sortie du modèle. Soit X^T le vecteur des réalisations de la variable X pour les N exemples. $X^T = (x_1, \ldots, x_N)$ et $Y^T = (y_1, \ldots, y_N)$ le vecteur des sorties désirées du modèle pour les N exemples.
L'indice de pertinence de X est défini de la manière suivante :

$$S(X, Y) = \cos^2(X^T, Y^T) = \frac{< X^T, Y^T >}{\|X^T\|^2 \|Y^T\|^2}$$

Comme pour le coefficient de correlation de Pearson, nous introduisons une variante de cette méthode pour calculer l'indice de pertinence d'une variable étant donné un modèle. Soit V^j le vecteur des réalisations de la variable V_j pour les N exemples. $V^j = (V_{1j}, \ldots, V_{Nj})$ et $f(S) = (S_1, \ldots, S_N)$ le vecteur de sortie du modèle f pour les N exemples.
L'indice de pertinence de V_j sachant f est défini de la manière suivante :

$$S(V_j|f, S) = \cos^2(V^j, f(S)) = \frac{< V^j, f(S) >}{\|V^j\|^2 \|f(S)\|^2}$$

Cette méthode est utilisée aussi bien pour les méthodes Wrappers que Filters. Elle permet de réaliser une sélection des variables candidates efficace.

2.2.4 Indice de pertinence basé sur la théorie de l'information

Soit V_j une variable aléatoire et V_{ij} sa réalisation pour l'exemple i. On dispose de N exemples I_i de taille n. Soit $S = \{S_1, \ldots, S_p\}$ l'ensemble des sorties désirées..

On définit la quantité d'information liée à la réalisation V_{ij} de la variable V_j de la manière suivante :

$$I(V_j = V_{ij}) = -log_2[p(V_j = V_{ij})]$$

L'entropie de la variable V_j évalue la quantité d'information contenue dans la variable V_j. Elle est définie comme la moyenne des quantités d'information liée à chaque réalisation de la variable V_j . On la note $H(V_j)$.

$$H(V_j) = -\sum_i P(V_j = V_{ij}) log_2[P(V_j = V_{ij})]$$

Si le vecteur S de sortie est de taille p, l'entropie de la sortie S du modèle est égale à :

$$H(S) = -\sum_{k=1}^{p} P(S = S_k) log_2[P(S = S_k)]$$

On définit l'entropie du couple (V_j, S) où V_j est une variable d'entrée et S la sortie désirée du modèle comme :

$$H(S, V_j) = -\sum_{k=1}^{p}\sum_{i=1}^{n} P(S = S_k; V_j = V_{ij}) log_2[P(S = S_k; V_j = V_{ij})]$$

Le Gain d'Information IG que V_j apporte à S est égale à :

$$IG(S, V_j) = H(S) + H(V_j) - H(S, V_j)$$

Le Gain d'Information permet de classer les variables par ordre de pertinence : une variable X est d'autant plus importante que son Gain d'Information $IG(S, X)$ par rapport à la sortie S est grand.

2.2.5 Résumé

Ces méthodes permettent de mesurer individuellement l'importance de chaque variable candidate. On pourra donc les classer par ordre de pertinence. Elles sont le plus souvent utilisées dans le cadre des méthodes Filters. Toutefois, nous avons introduit une adaptation du coefficient de correlation de Pearson et de l'indice de pertinence de Gram-Schimdt pour pouvoir les utilisées dans le cadre des méthodes Wrappers. .

2.3 La méthode *"Robelon"*

La méthode *"Robelon"* [7] s'inscrit dans le cadre des méthodes "wrappers". La méthode *Robelon* permet de mesurer l'importance d'une variable étant donné un modèle. Elle permet de classer les variables selon leur importance étant donné un modèle prédictif.

2.3.1 Introduction

L'objectif est de mesurer l'importance d'une variable d'entrée étant donné un modèle f. Pour évaluer l'influence d'une variable sur la sortie d'un modèle prédictif, Leray et Gallinari [8] introduisent l'approche "perturbation" : si x_i une réalisation d'une variable X est une entrée du modèle, quelle incidence aurait-elle sur la sortie du modèle si on remplaçait x_i par une autre réalisation x_k de X ? La variable serait donc d'autant plus importante que la sortie du modèle serait perturbée.

La définition donnée par Lemaire et Clérot [7] s'inspire de deux approches. La première initiée par R. Féraud ([4]) est inspirée de l'approche "perturbation" La deuxième approche initiée par Breiman ([2]) se fonde sur la distribution de probabilité de la variable dont on veut mesurer l'importance.

En combinant ces deux méthodes, V. Lemaire et F. Clérot ([7]) définissent l'importance d'une variable de la manière suivante :

Définition de *l'importance d'une variable selon Robelon*

L'importance d'une variable d'entrée est une fonction de la distribution de probabilité des exemples et de la distribution de probabilité de la variable dont on veut mesurer l'importance.

Soit :

- V_j la variable dont on veut mesurer l'importance ;
- V_{ij} la réalisation de la variable V_j pour l'exemple i ;
- I_m l'exemple m un vecteur de taille n ;
- f le modèle ;
- $P_{V_j}(u)$ la distribution de probabilité de la variable V_j ;
- $P_I(\nu)$ la distribution de probabilité des exemples I.

et

$$f_j(a;b) = f_j(a_1, ..., a_n; b) = f(a_1, ..., a_{j-1}, b, a_{j+1}, ..., a_n) \tag{2.1}$$

où a_p est la p-ième composante du vecteur a.

L'importance de la variable V_j est la somme des variations de la sortie S du modèle, lorsque les exemples sont "perturbés" suivant la distribution de probabilité de la variable V_j. La sortie "perturbée" du modèle f, pour un exemple I_i , est égale à la sortie du modèle pour cet exemple lorsqu'on remplace la j-ième composante de cet exemple par j-ième composante d'un autre exemple k. La variation de la sortie pour l'exemple I_i, est la différence entre la "vraie sortie" $f_j(I_i; V_{ij})$ et "la sortie perturbée" $f_j(I_i; V_{kj})$ du modèle.

L'importance de la variable V_j est alors la somme des $|f_j(I_i; V_{ij}) - f_j(I_i; V_{kj})|$ sur la distribution de probabilité des exemples et la distribution de probabilité de la variable V_j.

L'importance de la variable V_j pour le modèle f est alors :

$$S(V_j|f) = \iint P_{V_j}(u)du P_I(v)dv \, |f_j\,(I_i; V_{ij}) - f_j\,(I_i; V_{kj})| \tag{2.2}$$

Mesure d'importance d'une variable par Robelon

On peut approximer les distributions de probabilité par leurs distributions empiriques. Pour calculer la moyenne des $S(V_j|f)$ on aura donc besoin d'utiliser toutes les valeurs possibles de la variables V_j pour tous les exemples disponibles. Pour N exemples et donc N valeurs possibles de V_j, le temps de calcul est en $O(N^2)$ et devient donc très long pour une grande base de données.

Il existe au moins deux heuristiques qui permettent de calculer rapidement $S(V_j|f)$:

1. En considérant simultanément I_i et V_{kj} et on calcule une réalisation de $|f_j(I_i; V_{ij}) - f_j(I_i; V_{kj})|$. La mesure de la moyenne de $S(V_j|f)$ est obtenue par les moyennes du filtre de Kalman jusqu'à convergence (voir [9] pour les filtres de Kalman).
2. on peut écrire $S(V_j|f)$ de la manière suivante :

$$S(V_j|f) = \int P_I(v)dv \int P_{V_j}(u)du \, |f_j\,(I_i; V_{ij}) - f_j\,(I_i; V_{kj})| \tag{2.3}$$

On approxime la distribution de probabilité des données par la distribution empirique des exemples :

$$S(V_j|f) = \frac{1}{N^2} \sum_{i \in N} \sum_{k \in N} |f_j\left(I_i; V_{ij}\right) - f_j\left(I_i; V_{kj}\right)| \tag{2.4}$$

Comme la distribution de probabilité des exemples peut être approcher en utilisant les exemples représentatifs (P) d'une statistique d'ordre.
On a donc :

$$S(V_j|f) = \frac{1}{N} \sum_{i \in N} \sum_{p \in P} |f_j\left(I_i; V_{ij}\right) - f_j\left(I_i; v_p\right)| \,\text{Prob}(v_p) \tag{2.5}$$

Cette méthode est spécialement adaptée lorsque V_j est discrète puisque dans ce cas le calcul de la distribution de probabilité de V_j est exact.

Application à la sélection des sous-ensembles de variables

Comme nous l'avons vu dans la section 2.1.3, les méthodes *backward* et *forward* permettent de sélectionner les sous-ensembles de variables selon un certain critère. Dans la suite nous utiliserons l'algorithme *backward*.

Pour une éxécution plus rapide de l'algorithme backward, en plus du critère permettant de sélectionner les sous-ensembles de variables, nous rajoutons un autre paramètre : un seuil que nous fixons à 10^{-6}. Ainsi après chaque exécution de l'algorithme, nous mesurons l'importance de chaque variable par la méthode "robelon" et nous supprimons toutes les variables dont l'importance est inférieure au seuil. Un exemple de l'application de cette méthode avec un MLP est traité dans [7].

2.3.2 Calcul autour de Robelon pour le modèle de régression linéaire

Illustration graphique de l'importance d'une variable pour un modèle de régression linéaire

Considérons un modèle de régression linéaire simple. Son équation est de la forme $y = ax + b$ où a est la pente et b l'ordonnée à l'origine. La pente a de la droite signifie que pour un déplacement dx sur l'axe x correspond un déplacement dy sur l'axe y avec $a = \frac{dy}{dx}$. La pente correspond donc au taux de variation de y quand x varie. Une variable sera d'autant plus importante que sa pente est élevée puisque pour une petite perturbation (ou variation) de la variable x, on a une grande perturbation (ou variation) de y.

Sur la figure 2.1, nous avons représenté la pente d'une droite. Elle est égale à la tangente de l'angle α, l'angle que fait la droite d'équation $y = ax + b$ avec l'axe des abcisses.

Dans le cas d'une regression linéaire multiple, pour 2 variables l'équation de la droite de régression est de la forme $Y = aX + bZ + C$. L'importance de la variable X est égale à la dérivée partielle de Y par rapport à X, i.e le taux de variation de Y

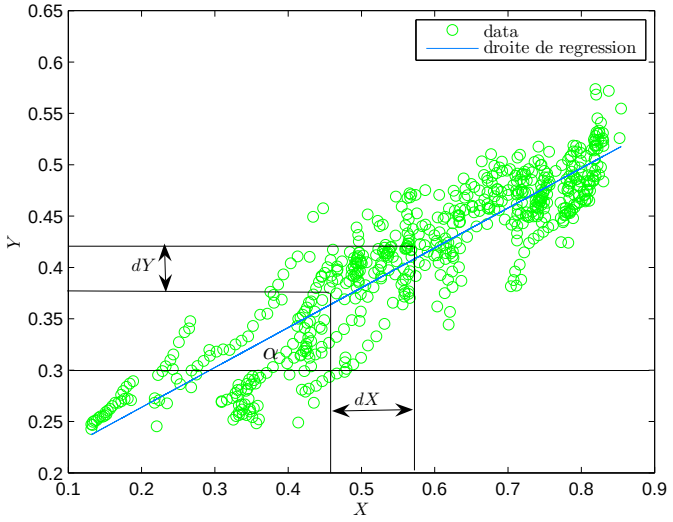

FIG. 2.1 – Illustration graphique de l'importance d'une variable pour un modèle de régression linéaire simple

quand X varie et l'importance de la variable Z est égale à la dérivée partielle de Y par rapport à Z : $a = \frac{dy}{dx}$ et $b = \frac{dy}{dz}$. Sur la figure 2.2, nous illustrons l'importance d'une variable pour une régression linéaire multiple.

Calcul de l'importance d'une variable pour un modèle de regression linéaire

Soit V_j une variable candidate. Selon Lemaire et Clérot dans [7] on a :
$pour\ tout\ k = 1, \ldots, N$:

$$S(V_j|f) = \frac{1}{N} \sum_{i \in N} E\left\{ \left| f_j\left(I_i; V_{ij}\right) - f_j\left(I_i; V_{kj}\right) \right| \right\}$$

Si f est le modèle de régression linéaire alors :

$$\begin{aligned} f_j\left(I_i; V_{ij}\right) &= \beta_0 + \beta_1 V_{i1} + \beta_2 V_{i2} + \ldots + \beta_j V_{ij} + \ldots + \beta_n V_{in} \\ f_j\left(I_i; V_{kj}\right) &= \beta_0 + \beta_1 V_{i1} + \beta_2 V_{i2} + \ldots + \beta_j V_{kj} + \ldots + \beta_n V_{in} \end{aligned}$$

On a donc :

$$\left| f_j\left(I_i; V_{ij}\right) - f_j\left(I_i; V_{kj}\right) \right| = \left|\beta_j\right| \left| V_{ij} - V_{kj} \right|$$

Pour mesurer la dispersion de la sortie du modèle dans l'expression $S(V_j|f)$, on peut utiliser la valeur absolue $\left| f_j\left(I_i; V_{ij}\right) - f_j\left(I_i; V_{kj}\right) \right|$ ou le carré de la différence des sorties $\left(f_j\left(I_i; V_{ij}\right) - f_j\left(I_i; V_{kj}\right)\right)^2$.

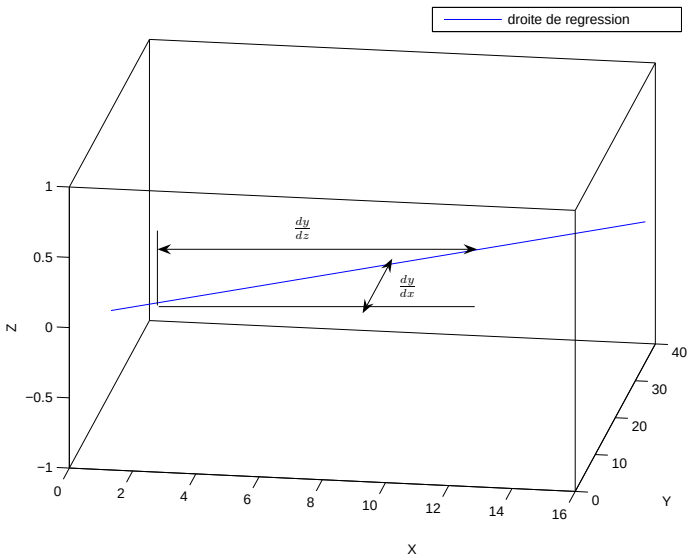

FIG. 2.2 – Illustration graphique de l'importance d'une variable pour un modèle de régression linéaire multiple

Pour calculer l'importance d'une variable selon robelon, nous avons utilisé quatre méthodes :

- la première méthode utilise la définition de l'espérance mathématique et les propriétés d'indépendance des tirages. Dans cette méthode, on mesure le dispersion des tirages en utilisant la valeur absolue. Les données seront normalisées à l'issue des calculs ;
- pour la deuxième méthode, on calcule l'espérance mathématique du bruit de la variable considérée. On peut normaliser les données avant ou à l'issue des calculs pour ordonnancer les variables. On utilise également la valeur absolue de la différence des tirages pour mesurer la dispersion ;
- la troisième méthode utilise les propriétés d'indépendance des tirages et les propriétés de l'espérance mathématique. On mesure la dispersion des tirages en évaluant le carré de la différence des tirages. On utilise les données normalisées pour simplifier les calculs ;
- la quatrième méthode utilise les propriétés de la régression linéaire et les propriétés des variables gaussiennes. On peut normaliser les données avant ou après les calculs. Les calculs ont été faits pour les deux mesures de dispersion. Dans la première partie du calcul, on considère le carré de la différence des tirages et dans la deuxième partie, on considère la valeur absolue de la différence des tirages.

Première méthode

V_{ij} et V_{kj} sont deux réalisations indépendantes de la variable aléatoire V_j.Elles sont issue de la même loi de probabilité p . Notons-les X_i et Y_k.

$$\forall k = 1, \ldots, N \quad S(V_j|f) \quad = \quad \frac{|\beta_j|}{N} \sum_{i \in N} E(|X_i - Y_k|) \tag{2.6}$$

On veut calculer $E(|X_i - Y_k|)$ où X_i et Y_k sont deux réalisations indépendantes de V_j. On a :

$$p_{X_i} = p_{Y_k} = p \quad et \quad F(t) = \int_{-\infty}^{t} p(u)du$$

où F est la fonction de répartition de V_j et p sa distribution de probabilité. Pour le calcul, on prendra $X = X_i$ et $Y = Y_k$.

$$E(|X - Y|) = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} |x - y| p_{(X,Y)}(x, y)dxdy$$

. X et Y étant indépendantes,

$$p_{(X,Y)}(x, y) = p_X(x)p_Y(y) = p(x)p(y)$$

$$E(|X - Y|) = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} |x - y| p(x)p(y)dxdy$$

La fonction $H(x, y) = |x - y|p(x)p(y)$ étant positive, en appliquant le théorème de Tonelli, on a :

$$\begin{aligned} \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} |x - y| p(x)p(y)dxdy \quad &= \quad \int_{-\infty}^{+\infty} p(x)[\int_{-\infty}^{+\infty} |x - y| p(y)dy]dx \\ &= \quad \int_{-\infty}^{+\infty} p(x)[\int_{-\infty}^{x} |x - y| p(y)dy + \\ &\qquad \int_{x}^{+\infty} |x - y| p(y)dy]dx \end{aligned} \tag{2.7}$$

$$\begin{aligned} \text{Sur }]-\infty, x], x \geq y \Rightarrow |x - y| \quad &= \quad x - y \\ \text{Sur } [x, +\infty], x \leq y \Rightarrow |x - y| \quad &= \quad y - x \end{aligned}$$

On a donc :

$$\begin{aligned} E(|X - Y|) \quad &= \quad \int_{-\infty}^{+\infty} p(x)[\int_{-\infty}^{x} (x - y)p(y)dy + \int_{x}^{+\infty} (y - x)p(y)dy]dx \\ &= \quad \int_{-\infty}^{+\infty} p(x)[x\int_{-\infty}^{x} p(y)dy - \int_{-\infty}^{x} yp(y)dy + \int_{x}^{+\infty} yp(y)dy \\ &\qquad - x\int_{x}^{+\infty} p(y)dy]dx \end{aligned} \tag{2.8}$$

$$\begin{aligned} E(|X-Y|) &= \int_{-\infty}^{+\infty} p(x)[x(F(x)-(1-F(x))+\int_{-\infty}^{+\infty} yp(y)dy \\ &\quad -2\int_{-\infty}^{x} yp(y)dy]dx \qquad (2.9) \\ &= \int_{-\infty}^{+\infty} p(x)[x(2F(x)-1)+m-2\int_{-\infty}^{x} yp(y)dy]dx \end{aligned}$$

où

$$m = \int_{-\infty}^{+\infty} yp(y)dy$$

En intégrant par parties on obtient que :

$$\int_{-\infty}^{x} yp(y)dy = xF(x) - \int_{-\infty}^{x} F(y)dy$$

où

$$F(x) = \int_{-\infty}^{x} p(u)du$$

$$\begin{aligned} E(|X-Y|) &= \int_{-\infty}^{+\infty} p(x)[(2xF(x)-x+m-2xF(x)+2\int_{-\infty}^{x} F(y)dy]dx \\ &= \int_{-\infty}^{+\infty} p(x)[-x+m+2\int_{-\infty}^{x} F(y)dy]dx \\ &= -m+m\int_{-\infty}^{+\infty} p(x)dx+2\int_{-\infty}^{+\infty} p(x)[\int_{-\infty}^{x} F(y)dy]dx \end{aligned}$$

$$\begin{aligned} \int_{-\infty}^{+\infty} p(x)[\int_{-\infty}^{x} F(y)dy]dx &= \int_{-\infty}^{+\infty} F'(x)[\int_{-\infty}^{x} F(y)dy]dx \\ &= [F(x)\int_{-\infty}^{x} F(y)dy]_{-\infty}^{+\infty} - \int_{-\infty}^{+\infty} F(x)^2dx \\ &= \int_{-\infty}^{+\infty} F(y)dy - \int_{-\infty}^{+\infty} F(x)^2dx \\ &= \int_{-\infty}^{+\infty} F(t)(1-F(t))dt \end{aligned}$$

D'où

$$E(|X-Y|) = 2\int_{-\infty}^{+\infty} F(t)(1-F(t))dt$$

$$\begin{aligned} S(V_j|f) = &= \frac{|\beta_j|}{N}\sum_{i\in N} E|X_i - X_k| \\ &= 2\frac{|\beta_j|}{N}\sum_{i\in N}\int_{-\infty}^{+\infty} F_i(t)(1-F_i(t))dt \end{aligned}$$

Les X_i étant i.i.d, $\forall i,\ F_i(t) = F(t)$. D'où :

$$\begin{aligned} S(V_j|f) &= 2\frac{|\beta_j|}{N}\sum_{i\in N}\int_{-\infty}^{+\infty} F(t)(1-F(t))dt \\ &= 2\,|\beta_j|\int_{-\infty}^{+\infty} F(t)(1-F(t))dt \end{aligned}$$

Finalement, on a :

$$S(V_j|f) = 2\,|\beta_j|\,l \tag{2.10}$$

où

$l = \int_{-\infty}^{+\infty} F(t)(1-F(t))dt$, (F est la fonction de répartition de la variable V_j)

est une constante. D'après 2.10, pour ordonnancer les $S(V_j|f)$, il suffit d'ordonnancer les $|\beta_j|$.

Deuxième méthode

Supposons que la variable V_j soit continue. Selon Lemaire et Clérot dans [7] on a :

$$\begin{aligned} S(V_j|f) &= \frac{1}{N}\sum_{i\in N} E\left\{\left|f_j\left(I_i; V_{ij}\right) - f_j\left(I_i; V_{kj}\right)\right|\right\} \\ S(V_j|f) &= \frac{1}{N}\sum_{i=1}^{N}\int_{-\infty}^{+\infty}\left|f_j\left(I_i; V_{ij}\right) - f_j\left(I_i; V_{ij}+\epsilon\right)\right| d\epsilon \end{aligned}$$

où ϵ représente la perturbation. On a donc :

$$\begin{aligned} S(V_j|f) &= \frac{1}{N}\sum_{i=1}^{N}\int_{-\infty}^{+\infty}|\beta_j|\,|\epsilon|\,d\epsilon \\ S(V_j|f) &= \frac{|\beta_j|}{N}\sum_{i=1}^{N}\int_{-\infty}^{+\infty}|\epsilon|\,d\epsilon \end{aligned} \tag{2.11}$$

ϵ représente la perturbation sur V_j. C'est une variable aléatoire continue qui varie de $-l\sigma_j$ à $l\sigma_j$ où σ_j est l'écart-type de V_j et l une constante positive.

D'où :

$$S(V_j|f) = \frac{|\beta_j|}{N}\sum_{i=1}^{N}\int_{-l\sigma_j}^{l\sigma_j}|\epsilon|\,d\epsilon \tag{2.12}$$

La fonction $x \mapsto |x|$ étant paire on a donc :

$$S(V_j|f) = 2\frac{|\beta_j|}{N}\sum_{i=1}^{N}\int_{0}^{l\sigma_j}\epsilon\,d\epsilon$$

En intégrant, on trouve :

$$S(V_j|f) = l^2\,|\beta_j|\,var\,(V_j) \tag{2.13}$$

où β_j est le *j-ième* coefficient de la regression linéaire et $var(V_j)$ la variance de la variable V_j. En normalisant les données, on a pour $j = 1, \ldots, n$, $\sigma_j^2 = 1$. L'ordonnancement des variables revient donc à ordonner les $|\beta_j|\,, j = 1, \ldots, n,$ les coefficients de la regression linéaire.

Si la variable aléatoire V_j est discrète alors :

$$S(V_j|f) = \frac{|\beta_j|}{N^2}\sum_{i=1}^{N}\sum_{k=1}^{N}|V_{ij} - V_{kj}| \tag{2.14}$$

où

$$\frac{1}{N^2}\sum_{i=1}^{N}\sum_{k=1}^{N}|V_{ij} - V_{kj}| \tag{2.15}$$

est un coefficient de normalisation.

Troisième méthode

V_{ij} et V_{kj} sont deux réalisations indépendantes de la variable V_j. Notons-les X_i et X_k. La valeur absolue dans le calcul de $S(V_j|f)$ permet de mesurer la dispersion. Pour des commodités de calcul, considérons $(X_i - X_k)^2$ au lieu de $|X_i - X_k|$.

Dans ce cas on a $\forall k = 1, \ldots, N$:

$$\begin{aligned} S(V_j|f) &= \frac{{\beta_j}^2}{N}\sum_{i\in N}E[(X_i - X_k)^2] \\ S(V_j|f) &= \frac{{\beta_j}^2}{N}\sum_{i\in N}E(X_i^2 - 2X_iX_k + X_k^2) \\ &= \frac{{\beta_j}^2}{N}\sum_{i\in N}[E({X_i}^2) + E({X_k}^2)] - 2\frac{{\beta_j}^2}{N}\sum_{i\in N}E(X_i)E(X_k) \end{aligned}$$

X_i et X_k étant issue de la même loi de probabilité, en considérant les variables centrées réduites, on a :

$$E(X_i) = E(X_k) = 0 \quad et \quad E({X_i}^2) = E({X_k}^2) = Var_{X_i} = 1$$

D'où :

$$S(V_j|f) = 2{\beta_j}^2 \tag{2.16}$$

où β_j est le *j-ième* coefficient de la regression linéaire. L'ordonnancement des variables revient donc à ordonner les ${\beta_j}^2, j = 1, \ldots, n,$ les coefficients de la regression linéaire.

Quatrième méthode

1. Dans un premier temps on considère la mesure de dispersion $(-)^2$

   Comme précédemment, V_{ij} et V_{kj} sont des réalisations indépendantes et identiquement distribuées de la variable V_j. Notons-les X_i et X_k. Une fois de plus, pour des commodités de calcul, considérons $(X_i - X_k)^2$ au lieu de $|X_i - X_k|$. On a donc $\forall k = 1, \ldots, N$:

$$S(V_j|f) \quad = \quad \frac{{\beta_j}^2}{N} \sum_{i \in N} E((X_i - X_k)^2) \tag{2.17}$$

   Le modèle considéré dans ce cas étant le modèle de régression linéaire, pour tout i et pour tout k, X_i et X_k sont issues de la même loi gaussienne de paramètres μ et σ^2 $N(\mu, \sigma^2)$ où μ et σ^2 représentent respectivement la moyenne et la variance de V_j. La différence de deux variables gaussiennes indépendantes est une gaussienne. Plus généralement la somme de 2 variables gaussiennes indépendantes est une gaussienne :si on a deux variables indépendantes $X \sim N(\mu_1, {\sigma_1}^2)$ et $Y \sim N(\mu_2, {\sigma_2}^2)$, la somme $X + Y \sim N(\mu_1 + \mu_2, {\sigma_1}^2 + {\sigma_2}^2)$. Pour l'illustrer, sur la figure 2.3 on a représenté la gaussienne $N(-2, 1^2)$, la gaussienne $N(6, 1^2)$ et la somme des 2 gaussiennes $N(-2, 1^2) + N(6, 1^2)$. On a bien que $N(-2, 1^2) + N(6, 1^2)$ est une gaussienne de moyenne $(-2) + 6 = 4$ la somme des moyennes et de variance $1^2 + 1^2 = 2$ la somme des variances.
   $X_i - X_k$ suit donc la loi normale $N(0, 2\sigma^2)$. On a :

$$E((X_i - X_k)^2) \quad = \quad 2\sigma^2, \forall i, k$$

   D'où

$$\frac{{\beta_j}^2}{N} \sum_{i \in N} E((X_i - X_k)^2) \quad = \quad 2{\beta_j}^2\sigma^2$$

   On en déduit que :

$$S(V_j|f) = 2{\beta_j}^2\sigma^2$$

   En normalisant les données on trouve :

$$S(V_j|f) = 2{\beta_j}^2$$

   Il suffit encore d'ordonnancer les ${\beta_j}^2$ pour ordonnancer les variables.

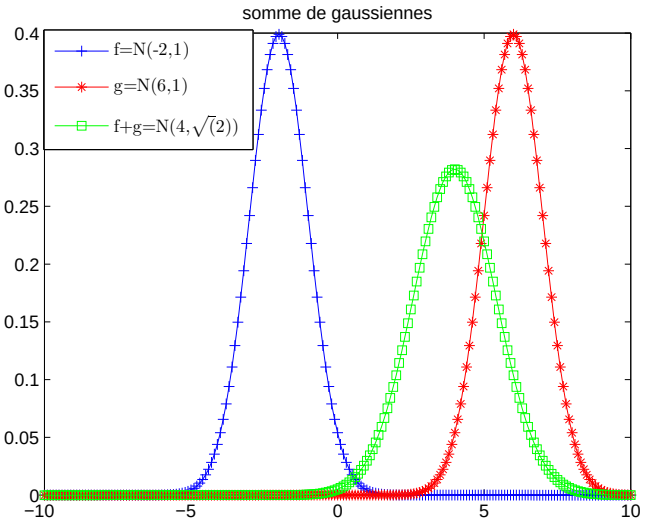

FIG. 2.3 – Illustration graphique de la loi de la somme de 2 variables gaussiennes

2. On considère ensuite la mesure de dispersion $|-|$
   Si on considère maintenant la mesure de dispersion $|-|$, i.e $|X_i - X_k|$, on a :

$$S(V_j|f) = \frac{|\beta_j|}{N} \sum_{i \in N} E(|X_i - X_k|)$$

   $\forall k = 1, \ldots, N$, soit $Y = X_i - X_k$. Si on considère les données normalisées $Y \sim N(0, 2)$.

$$E|Y| = \int_{-\infty}^{+\infty} |y| f(y) dy$$

   où f est la densité de probabilité de Y. La fonction $y \to |y|f(y)$ étant paire on a :

$$E|Y| = 2 \int_0^{+\infty} y f(y) dy$$

$$\begin{aligned} 2 \int_0^{+\infty} y f(y) dy &= \frac{1}{\sqrt{2\pi}} \int_0^{+\infty} y e^{-\frac{1}{4}y^2} dy \\ &= -\frac{1}{2\sqrt{2\pi}} [e^{-\frac{1}{4}y^2}]_0^{+\infty} \\ &= \frac{1}{2\sqrt{2\pi}} \end{aligned}$$

   On a donc :

$$S(V_j|f) = \frac{1}{2\sqrt{2\pi}} |\beta_j|$$

Tableaux récapitulatifs

Outil de mesure de dispersion : $(.)^2$	calcul 3	calcul 4
Pré-normalisation	$\frac{X-\mu}{\sigma}$	$X-\mu$
Post-normalisation	non	$\frac{X}{\sigma}$
$S(V_j\|f)$	${2\beta_j}^2$	${2\beta_j}^2$

TAB. 2.1 – Tableau recapitulatif des prétraitements utilisés et des résultats obtenus en utilisant la mesure de dispersion $(-)^2$

Outil de mesure de dispersion :$\|.\|$	calcul 1	calcul 2	calcul 4
Pré-normalisation	non	$\frac{X}{\sigma}$	$\frac{X-\mu}{\sigma}$
Post-normalisation	$\frac{X}{\sigma}$	$\frac{X}{\sigma}$	$\frac{X-\mu}{\sigma}$
$S(V_j\|f)$	$2\,\|\beta_j\|$	$2\,\|\beta_j\|$	$\frac{1}{2\sqrt{2\pi}}\,\|\beta_j\|$

TAB. 2.2 – Tableau recapitulatif des prétraitements utilisés et des résultats obtenus en utilisant la mesure de dispersion $|-|^2$

Discussion

Pour un modèle de régression linéaire, l'ordonnancement des variables en utilisant la méthode Robelon, revient à ordonnancer les $|\beta_j|$ ou les ${\beta_j}^2$ où les β_j sont les coefficients de la régression linéaire. Dans la méthode Robelon on évalue la dispersion des tirages. Pour mesurer cette dispersion, on peut utiliser la valeur absolue de la différence des tirages ou le carré de la différence des tirages. Les calculs utilisant le carré de la différence des tirages sont plus simples que ceux utilisant la valeur absolue de la différence des tirages. Selon la méthode utilisée, Il sera meilleur de normaliser les données avant (pré-normalisation) ou après (post-normalisation) les calculs pour des commodités de calcul. Les tableaux 2.1 et 2.2 résument les résultats obtenus pour les différents calculs réalisés en fonction de la mesure de dispersion utilisées.

2.3.3 Calcul autour de Robelon pour le modèle Naïf Bayes

Présentation du modèle naïf bayes

Soit $\mathcal{X}$ et Ω deux espaces probabilisés : $\mathcal{X}$ est l'espace des formes et Ω l'espace des classes. Soit $C_1, C_2, \ldots, C_q$ les différentes classes et $P(C_i)$ la fréquence des éléments de la classe i.
$P(X)$ est la densité de probabilité de la forme $X \in \mathcal{X}$ et $P(X/C_i)$ est la densité de probabilité conditionnelle de l'observation X sachant C_i.

On a :

$$\sum_{i=1}^{q} P(C_i) = 1 \quad \text{et} \quad P(X) = \sum_{i=1}^{q} P(X/C_i)P(C_i)$$

$P(C_i)$ est la probabilité *à priori* de la classe C_i.

La règle de Bayes consiste à affecter l'observation X à la classe C_i qui maximise $P(C_i/X)$, la probabilité *à postériori* de C_i sachant X.
Les $P(C_i/X)$ ne sont en général pas facile à estimer. La règle de Bayes permet de calculer les probabilités à postériori :

$$P(C_i/X) = \frac{P(X/C_i)P(C_i)}{P(X)} \quad \text{où} \quad P(X) = \sum_{i=1}^{q} P(X/C_i)P(C_i)$$

Par exemple pour 2 classes C_1 et C_2 la règle d Bayes consiste à affecter X à la classe C_1 si $P(C_1/X) \geq P(C_2/X)$, à la classe C_2 dans le cas contraire.
Pour q classes la règle de Bayes consiste à affecter X à la classe C telle que :

$$C = \max_i P(C_i/X)$$

Considérons le cas où X est un vecteur aléatoire discret, c'est à dire que $X = \{X_1, \ldots, X_n\}$ où les X_j sont des variables aléatoires discrètes. Dans ce cas pour le modèle Naïf Bayes on suppose que les variables X_j sont indépendantes et sous cette hypothèse on :

$$P(C_i/X) = P(C_i) \prod_{j=1}^{n} P(X_j/C_i)$$

on affecte X à la classe C telle que :

$$C(X) = \max_i P(C_i) \prod_{j=1}^{n} P(X_j/C_i)$$

Calcul de Robelon pour un Naïf bayes

Dans la section précédente, nous calculons l'importance d'une variable étant donné un modèle f. Nous considérons maintenant que f est le modèle Naïf Bayes. Notre but ici n'est pas de classer mais de calculer l'importance des variables. C'est pourquoi au lieu de considérer le modèle

$$f(X) = \max_i P(C_i/X)$$

que nous avons présenter au paragraphe précédent dont la sortie est la classe de l'exemple X, nous considérons le modèle f qui nous renvoie toutes les probabilités à posteriori de X. Si on a C classes $C_1, \ldots, C_C$, on a pour un exemple X :

$$f(X) = (P(C_1/X), \ldots, P(C_C/X)) \tag{2.18}$$

Pour la classe $C_l, 1 \leq l \leq C$, la sortie du modèle est égale à :

$$\begin{aligned} f(X, C_l) &= P(C_l/X) \\ f(X, C_l) &= P(C_l) \prod_{j=1}^{n} P(X_j/C_l) \end{aligned} \tag{2.19}$$

Dans le cas général, si un modèle f a plusieurs sorties et si O représente le nombre de sorties de f alors l'importance de la variable V_j selon robelon est égale à :

$$S(V_j|f) = \frac{1}{N}\sum_{o=1}^{O}\sum_{i\in N} E\left\{|f_j\left(I_i; V_{ij}; o\right) - f_j\left(I_i; V_{kj}; o\right)|\right\} \tag{2.20}$$

où $f_j\left(I_i; V_{ij}; o\right)$ représente la sortie o du modèle.

Dans le cas du modèle Naïf Bayes, l'importance de la variable V_j est égale :

$$S(V_j|f) = \frac{1}{N}\sum_{l=1}^{C}\sum_{i\in N} E\left\{|f_j\left(I_i; V_{ij}; C_l\right) - f_j\left(I_i; V_{kj}; C_l\right)|\right\} \tag{2.21}$$

où $f_j\left(I_i; V_{ij}; C_l\right)$ représente la sortie du Naïf Bayes pour la classe C_l.

Pour une meilleure lisibilité des calculs, nous calulerons l'importance de la variable V_j selon robelon pour un modèle Naïf Bayes qui a une seule sortie :

$$f(X) = P(C_l)\prod_{j=1}^{n} P(X_j/C_l)$$

Première méthode

En considérant l'equation 2.4, l'importance de la variable V_j, est égale à :

$$\forall k = 1, \ldots, N$$

$$S(V_j|f) = \frac{1}{N}\sum_{i\in N} E\left\{|f_j\left(I_i; V_{ij}\right) - f_j\left(I_i; V_{kj}\right)|\right\} \tag{2.22}$$

$$\begin{aligned} f_j\left(I_i; V_{ij}\right) - f_j\left(I_i; V_{kj}\right) \quad = \quad & P(C_l)P(V_{i1}/C_l)\ldots P(V_{ij}/C_l)\ldots P(V_{in}/C_l) \\ & - P(C_l)P(V_{i1}/C_l)\ldots P(V_{kj}/C_l)\ldots P(V_{in}/C_l) \end{aligned}$$

où

$$P(V_{ij}/C_l) = P(V_j = V_{ij}/C_l)$$

Le produit des probabilités $P(C_l)P(V_{i1}/C_l)\ldots P(V_{ij}/C_l)\ldots P(V_{in}/C_l)$ et le produit des probabilités $P(C_l)P(V_{i1}/C_l)\ldots P(V_{kj}/C_l)\ldots P(V_{in}/C_l)$ calculent respectivement la vraisemblance d'apparition de l'exemple I_i dans la classe C_l et la vraisemblance d'apparition de l'exemple I_i perturbée dans la classe C_l.
Les fonctions $f_j(I_i, V_{ij})$ et $f_j(I_i, V_{kj})$ représentent donc les fonctions de vraisemblance $FV(V_{i1}, \ldots, Vij, \ldots, V_{in})$ et $FV(V_{i1}, \ldots, Vkj, \ldots, V_{in})$ relativement à la classe C_l.

$f_j\left(I_i; V_{ij}\right) - f_j\left(I_i; V_{kj}\right)$ est donc la différence de 2 fonctions de vraisemblance.

L'importance de la variable V_j selon robelon en considérant les fonctions de vraisemblance, c'est à dire en considérant l'équation 2.22 est donc égale à :

$$\begin{aligned} S(V_j|f) &= \frac{1}{N}\sum_{i\in N} P(C_l)P(V_{i1}/C_l)\dots P(V_{i(j-1)}/C_l)P(V_{i(j+1)}/C_l)\dots P(V_{in}/C_l) * \\ & E\left\{|P(V_{ij}/C_l - P(V_{kj}/C_l|\right\} \end{aligned}$$

Dans le calcul de robelon, au lieu de considérer les fonctions de vraisemblance, pour simplifier les calculs considérons les log-vraisemblance $log_2 f_j\left(I_i; V_{ij}\right)$ et $log_2 f_j\left(I_i; V_{kj}\right)$.

On a :

$$\begin{aligned} |log_2 f_j\left(I_i; V_{ij}\right) - log_2 f_j\left(I_i; V_{kj}\right)| &= |log_2 P(V_j = V_{ij}/C_l) \\ & \quad - log_2 P(V_j = V_{kj}/C_l)|. \end{aligned}$$

D'où la variante de Robelon suivante :

pour $k = 1, \dots, N$

$$S'(V_j|f) = \frac{1}{N}\sum_{i\in N} E\left\{|log_2 P(V_j = V_{ij}/C_l) - log_2 P(V_j = V_{kj}/C_l)|\right\}$$

Considérons la mesure de dispersion $(.)^2$ au lieu de la valeur absolue $|.|$. On obtient une variante de Robelon :

$$S'(V_j|f) = \frac{1}{N}\sum_{i\in N} E[\{log_2 P(V_{ij}/C_l) - log_2 P(V_{kj}/C_l)\}^2] \text{ pour } k = 1, \dots, N \tag{2.23}$$

$$\begin{aligned} & E[\{log_2 P(V_j = V_{ij}/C_l) - log_2 P(V_j = V_{kj}/C_l)\}^2] = \\ & \frac{1}{N}\sum_{k=1}^{N}\{log_2(P(V_j = V_{ij})) - log_2(P(V_j = V_{ik}))\}^2 \end{aligned}$$

D'où :

$$\begin{aligned} S'(V_j|f) &= \frac{1}{N^2}\sum_{i=1}^{N}\sum_{k=1}^{N}\{log_2(P(V_j = V_{ij})) - log_2(P(V_j = V_{ik}))\}^2 \\ &= \frac{1}{N^2}\sum_{i=1}^{N}\sum_{k=1}^{N}\{log_2^2(P(V_i)) - 2log_2(P(V_i))log_2(P(V_k)) \\ & \quad + log_2^2(P(V_k))\} \end{aligned} \tag{2.24}$$

où

$$log_2(P(V_i)) = log_2 P(V_j = V_{ij}/C_l)$$

$$
\begin{aligned}
S'(V_j|f) &= \frac{1}{N^2}\sum_{i=1}^{N}\sum_{k=1}^{N} log_2^2(P(V_i)) - \frac{2}{N^2}\sum_{i=1}^{N}\sum_{k=1}^{N} log_2(P(V_i))log_2(P(V_k)) \\
&\quad + \frac{1}{N^2}\sum_{i=1}^{N}\sum_{k=1}^{N} log_2^2(P(V_k)) \qquad (2.25) \\
&= \frac{1}{N}\sum_{i=1}^{N} log_2^2(P(V_i)) - \frac{2}{N^2}\sum_{i=1}^{N} log_2(P(V_i))\sum_{k=1}^{N} log_2(P(V_k)) \\
&\quad + \frac{1}{N}\sum_{k=1}^{N} log_2^2(P(V_k)) \qquad (2.26) \\
&= \frac{2}{N}\sum_{i=1}^{N} log_2^2(P(V_i)) - 2(\frac{1}{N})\sum_{i=1}^{N} log_2(P(V_i))(\frac{1}{N})\sum_{k=1}^{N} log_2(P(V_k))
\end{aligned}
$$

On a donc :

$$S'(V_j|f) = 2\{\frac{1}{N}\sum_{i=1}^{N} log_2^2(P(V_i)) - (\frac{1}{N})\sum_{i=1}^{N} log_2(P(V_i))(\frac{1}{N})\sum_{k=1}^{N} log_2(P(V_k))\}$$

Si on note H l'entropie, l'entropie de la variable V_j est égale à :

$$H(V_j) = -E(log(P(V_j))$$

où $E(.)$ représente l'espérance mathématique. on a donc :

$$H({V_j}^2) = -\frac{1}{N}\sum_{i=1}^{N} log_2^2(P(V_i)) \quad et \quad H(V_j) = -\frac{1}{N}\sum_{i=1}^{N} log_2(P(V_i))$$

D'où

$$S'(V_j|f) = 2[H^2(V_j) - H({V_j}^2)] \qquad (2.27)$$

avec $H(V_j) = -E(log(P(V_j))$ où H représente l'entropie de la variable V_j et $E(.)$ l'espérance mathématique.

Sur la figure 2.4, nous avons représenté la distribution de l'information contenue dans une variable V_j (on a supposé ici que cette distribution suit une loi gaussienne), l'entropie H de cette distribution qui correspond à la moyenne de la distribution et l'importance selon "robelon" qui correspond à la variance de la distribution.

Deuxième méthode

Considérons maintenant l'équation 2.5 avec la mesure de dispersion $(.)^2$. Cette équation est un calcul approché de l'importance de la variable V_j selon robelon. Si la variable V_j est continue, en considérant les valeurs représentatives v_p d'une statistique d'ordre sur V_j on a :

$$S(V_j|f) = \frac{1}{N}\sum_{i\in N}\sum_{p\in P}\{f_j\,(I_i; V_{ij}) - f_j\,(I_i; v_p)\}^2 \text{Prob}(v_p)$$

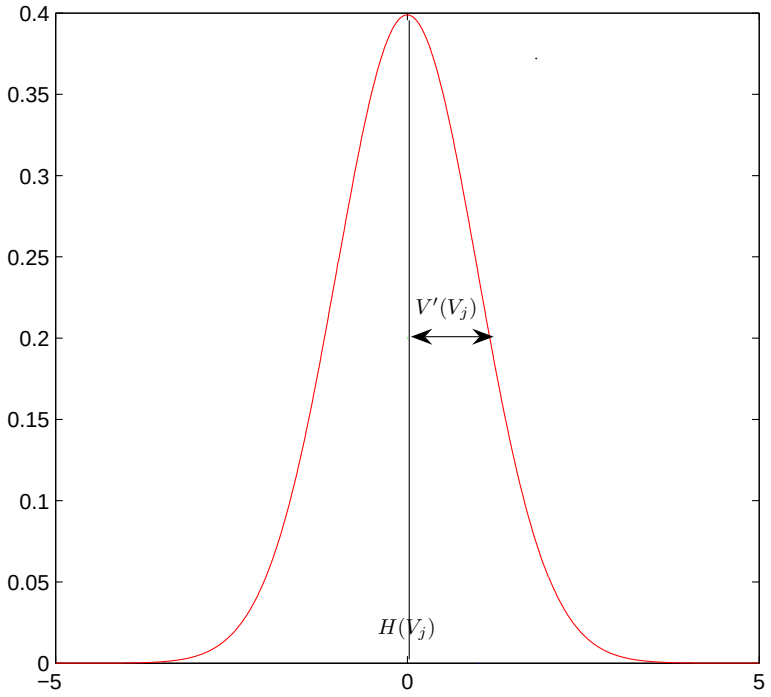

FIG. 2.4 – Représentation graphique de l'entropie d'une variable et de l'importance de cette variable selon robelon

Si la variable V_j est discrète, l'équation 2.5 est un calcul exact de l'importance selon robelon, puisque dans ce cas les v_p sont les valeurs prises par la variables V_j.

Pour le modèle Naïf Bayes, les variables sont discrètes. L'importance de la variable V_j selon robelon en utilisant l'équation 2.5 est donc égale à :

$$S(V_j|f) = \frac{1}{N}\sum_{i\in N} P(C_l)P(V_{i1}/C_l)\dots P(V_{i(j-1)}/C_l)P(V_{i(j+1)}/C_l)\dots P(V_{in}/C_l) * \\ *\{\sum_{p=1}^{P}\{|P(V_{ij}/C_l - P(v_p/C_l|\}\,\text{Prob}(v_p)\}$$

Comme dans la première méthode, en considérant les log-vraisemblance a la place des fonctions de vraisemblance on obtient :

$$\begin{aligned} S'(V_j|f) &= \frac{1}{N}\sum_{i\in N}\sum_{p=1}^{P}[\{log_2 P(V_{ij}/C_l) - log_2 P(v_p/C_l)\}^2]prob(v_p) \\ &= \frac{1}{N}\sum_{i=1}^{N}\sum_{p=1}^{P}\{log_2^2(P(V_i)) - 2log_2(P(V_i))log_2(P(v_p)) + log_2^2(P(v_p))\} \\ &\quad * prob(v_p) \\ &= \frac{1}{N}\sum_{i=1}^{N}\sum_{p=1}^{P} log_2^2(P(V_i))prob(v_p) \\ &\quad - \frac{2}{N}\sum_{i=1}^{N}\sum_{p=1}^{P} log_2(P(V_i))log_2(P(v_p))prob(v_p) + \frac{1}{N}\sum_{i=1}^{N}\sum_{p=1}^{P} log_2^2(P(v_p))prob(v_p) \end{aligned}$$

où

$$log_2(P(V_i)) = log_2 P(V_j = V_{ij}/C_l)$$

$$\begin{aligned} S'(V_j|f) &= \frac{1}{N}\sum_{i\in N} log_2^2 P(V_i) + \frac{2}{N}[\sum_{i\in N} log_2(P(V_i))H(V_j)] + H(V_j) \\ &= -H({V_j}^2) + 2H(V_j)^2 - H({V_j}^2) \end{aligned}$$

$$S'(V_j|f) = 2[H^2(V_j) - H({V_j}^2)] \tag{2.28}$$

Pour une variable aléatoire X, $V(X) = E(X^2) - E^2(X)$ représente la variance. On en déduit que l'expression :

$$V'(V_j) = H^2(V_j) - H({V_j}^2) \tag{2.29}$$

$V'(V_j)$ pourrait représenter la fiabilité de l'information contenue dans V_j.

Discussion commune aux deux calculs

Pour le modèle Naïf Bayes, dans le calcul de Robelon, nous avons utilisé $(.)^2$ pour évaluer la dispersion. La quantité $V'(V_j) = H^2(V_j) - H({V_j}^2)$ pourrait être une mesure de la fiabilité de l'information contenue dans la variable V_j. Plus elle est grande, plus l'information est fiable et plus la variable est pertinente. $V'(V_j)$ pourrait être vu comme un outil de mesure de la diversité de la source d'information :une information issue de plusieurs médias est plus fiable qu'une information issue d'un média. Il serait intéressant de combiner l'entropie d'une variable et la fiabilité d'une variable pour ordonnancer les variables. La sélection serait alors plus efficace puisqu'elle prendrait en compte deux critères : La quantité d'information contenue dans la variable et la fiabilité de cette information.

2.4 Comparaison de Robelon à d'autres méthodes de mesure d'importance

2.4.1 Introduction

Dans cette section, nous comparons les performances du modèle Naïf Bayes obtenues en utilisant le sous-ensemble de variables sélectionné par la méthode Robelon aux performances obtenues par le sous-ensemble sélectionné en utilisant le coefficient de correlation de Pearson (voir 2.2.2). Nous comparons aussi la méthode Robelon à la méthode d'ordonnancement "Enhance Selectif Naïf Bayes" (voir algorithme forward selection section 2.1.3). Nous comparons aussi la mesure d'importance issue du calcul théorique de robelon (voir 2.27) aux 3 méthodes précédentes. Nous utiliserons 3 bases de test : Ada, Hiva et Sylva issues du "Congrès Mondial sur l'Intelligence Numérique". [1]

2.4.2 Description des bases utilisées

Nous avons utilisé les bases ADA, HIVA et SYLVA du WCCI2006. Ces bases sont à l'origine continues et ont été discrétisées par la méthode KHIOPS [2] (voir [1]) : Khiops est une méthode de discrétisation qui vise à déterminer les domaines (ou intervalles) de variation pertinents d'une variables explicatives par rapport à une variable cible. Chaque base est composée d'un ensemble d'apprentissage, d'un ensemble de validation et d'un ensemble de test séparée en deux classes : la classe 1 et la classe 2. Nous avons utilisé l'ensemble d'apprentissage pour apprendre les paramètres du modèle et ordonnancer les variables par ordre de pertinence et l'ensemble de validation pour évaluer les performances du modèle.

La base ADA est composée de 4747 exemples pour l'apprentissage et 415 pour la validation. 70% des exemples appartiennent à la classe 1 et 30% à la classe 2. Chaque exemple est représenté par 48 attributs. Après discrétisation, nous avons obtenu 28

[1] consulter le site `www.WCCI2006.org` pour plus d'information

[2] Méthode de discrétisation de variables continues développée par France Telecom R& D

attributs non constants et 20 attributs constants. Pour une execution plus rapide des calculs, nous avons décidé de supprimer tous les attributs constants puisqu'ils n'apportent aucune information. L'ordonnancement des variables s'est donc effectué sur les 28 variables non constantes.
La base HIVA est constituée de 3845 exemple en apprentissage et 384 en validation. 96.35% des exemples appartiennent à la classe 1 et 3.65% à la classe 2. Chaque exemple est représenté par 1618 variables dont 166 non constantes après discrétisation. Les calculs ont été effectués sur les 166 variables non constantes.
La base SYLVA est composée de 13086 exemples pour l'ensemble d'apprentissage et 1308 exemples l'ensemble de validation.93.85% des exemples appartiennent à la classe 1 et 6.15% à la classe 2. Chaque exemple est représenté par 217 variables dont 65 non constantes après discrétisation.

2.4.3 Méthodologie suivie

Pour chacune des bases nous avons réalisé les étapes suivantes :
- apprentissage des paramètres du Naïf Bayes en utilisant l'ensemble d'apprentissage ;
- ordonnancement des variables de la plus importante à la moins importante ;
    1. en utilisant Robelon sur l'ensemble d'apprentissage (nous avons utilisé l'équation 2.5) ;
    2. en utilisant le coefficient de correlation de Pearson sur l'ensemble d'apprentissage ;
    3. en utilisant la méthode "Enhance Selectif Naïf Bayes" (ESNB) sur l'ensemble d'apprentissage (voir 2.1.3 forward selection) ;
    4. en utilisant la mesure d'importance issue du calcul théorique de "robelon" (voir 2.27) .
- processus forward sur l'ensemble de validation utilisant l'ordonnancement de Pearson (voir 2.2.2). ;
- processus forward sur l'ensemble de validation utilisant l'ordonnancement de robelon ;
- processus forward sur l'ensemble de validation utilisant l'ordonnancement ESNB obtenu sur l'ensemble d'apprentissage ;
- processus forward sur l'ensemble de validation utilisant l'ordonnancement de la mesure d'importance issue du calcul théorique de robelon.

2.4.4 Résultats obtenus

Résultats obtenus sur la base ADA

Nous avons représenté sur le graphe 2.5 le taux d'exemples bien classés par le Naïf Bayes sur l'ensemble de validation pendant le processus forward en fonction de la méthode d'ordonnancement utilisée [3].

[3] Voir le fichier "tableau_recapitulatif_ada.xls" pour les résultats.

Sur le graphe 2.6, nous avons représenté le taux d'erreur de classification dans le cas où la répartition des exemples serait équilibrée, (le BER : Balanced Error Rate) du Naïf Bayes sur l'ensemble de validation pendant le processus forward en fonction de la méthode d'ordonnancement utilisée.

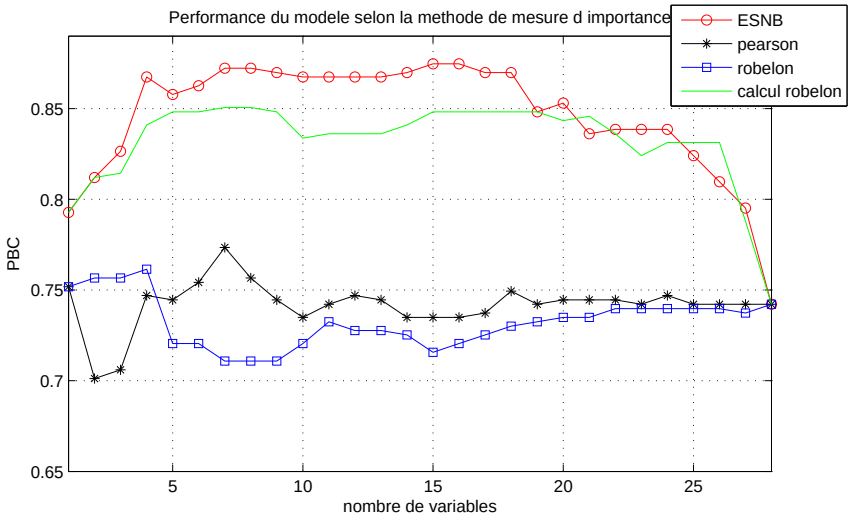

FIG. 2.5 – Taux d'exemples bien classés pendant le processus forward en fonction de différentes méthodes d'ordonnancement des variables.

Discussion

Sur la base ADA, la figure 2.5 montre que les performances du modèle sur l'ensemble de validation utilisant l'ordonnancement de la méthode ENSB sont largement supérieures aux performances utilisant l'ordonnancement robelon. On sélectionnerait pour Robelon les 4 premières variables i.e $\{46, 19, 44, 25\}$ pour un PBC de 76.14% pour robelon contre les 4 premières variables les plus importantes selon ENSB i.e $\{32, 10, 15, 44\}$ pour un PBC de 87.47% . (voir le fichier tableau_récapitulatif_ada.xls pour l'ordonnancement complet). D'autre part Robelon est presque deux fois plus rapide que la méthode ESNB, puisqu'il permet d'ordonnancer toutes les variables non constantes en 7.43 minutes contre 12.87 minutes pour la méthode ESNB.
Sur la figure 2.6 on a représenté le BER [4] en fonction de la méthode d'ordonnancement des variables. On constate que le BER obtenu en utilisant la méthode ESNB est inférieur à celui de la méthode robelon. Pour ENSB, on sélectionnerait les 4 premières variables selon ENSB avec lesquelles on a un BER de 0.2182 . Avec l'ordonnancement de robelon, on sélectionnerait aussi 4 variables mais on obtiendrait un BER légèrement supérieur 0.2367. Sur la figure 2.5 on constate que la méthode utilisant le coefficient de correlation de Pearson donne des performances légèrement inférieures à celles de

[4]BER :taux d'erreur de classification si la répartition des exemples était équilibrée

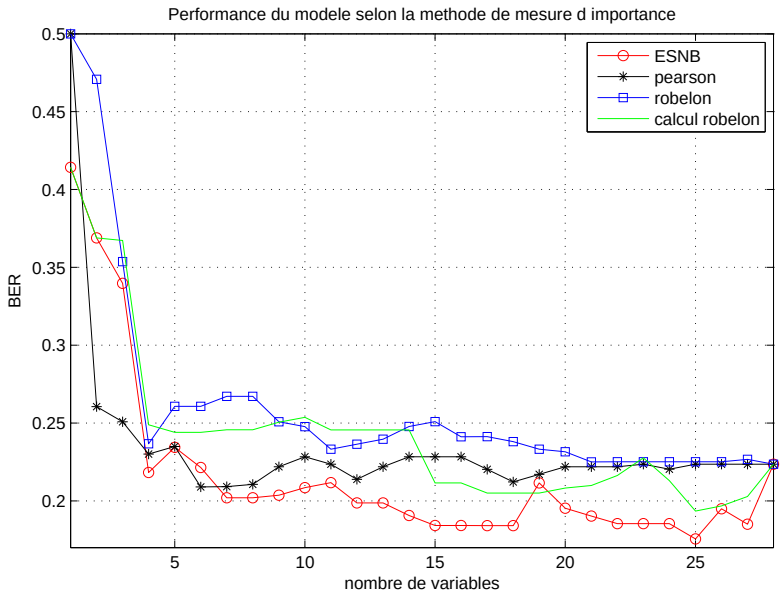

FIG. 2.6 – BER du Naïf Bayes pendant le processus forward en fonction de différentes méthodes d'ordonnancement des variables.

robelon. On sélectionnerait la première variables la plus importante selon la mesure d'importance de Pearson avec laquelle on obtient un PBC de 75.18% sur l'ensemble de validation contre un PBC de 76.14% pour robelon avec 4 variables sélectionnées. Le temps d'éxécution de Pearson est beaucoup plus rapide que celui de robelon puisqu'il est de quelques secondes.

Nous avons aussi représenté sur les graphes 2.5 et 2.6 le PBC et le BER respectivement en fonction de l'ordonnancement des variables utilisant le résultat du calcul théorique de robelon i.e en mesurant pour chaque variable V_j, $S'(V_j)$ et en les rangeant par ordre d'importance en fonction de S'. Nous constatons que les PBC obtenus en utilisant ce critère pour ordonnancer les variables sont largement supérieurs que ceux de "Robelon" et "Pearson" et légèrement inférieurs à ceux de la méthode ENSB. On sélectionnerait avec cette méthode les 8 premières variables les plus importantes et pour ces variables on atteint un PBC de 85.06% contre 76.14% pour robelon avec 4 variables sélectionnées, 75.18% pour Pearson avec une variable sélectionnée et 87.47% pour ENSB avec 4 variable sélectionnée. Par contre pour le BER, les performances de Robelon sont supérieurs à celles de l'ordonnancement utilisant le résultat du calcul théorique de robelon puisqu'on a pour 4 variables sélectionnées un BER de 0.2367 pour robelon contre 5 variables sélectionnées pour le calcul avec un BER de 0.244. Elles sont aussi moins bonnes que celles de ENSB et Pearson puiqu'on a respectivement un BER de 0.2182 pour ENSB avec 4 variables sélectionnées et 0.2301 pour Pearson avec également 4 variables sélectionnées. Sur les tableaux 2.3 et 2.4 nous résumons les ré-

sultats obtenus pour les quatre méthodes d'ordonnancement et sur le tableau 2.5 nous présentons les 5 variables les plus importantes selon chaque méthode.

Tableaux récapitulatifs

	tps execution	# var select	PBC
ESNB	12.87 mn	4	86.75 %
Robelon	7.43 mn	4	76.14%
Pearson	0.39 s	1	75.18%
Calcul Robelon	12 s	8	85.06%

TAB. 2.3 – Résultats obtenus selon la méthode d'ordonnancement

	tps execution	# var select	BER
ESNB	12.87 mn	4	0.2182
Robelon	7.43 mn	4	0.2367
Pearson	0.39 s	4	0.2301
Calcul Robelon	12 s	5	0.244

TAB. 2.4 – Résultats obtenus selon la méthode d'ordonnancement

ESNB	Pearson	Robelon	Calcul robelon
32	44	46	32
10	31	19	10
15	32	44	25
44	15	25	15
40	27	31	24

TAB. 2.5 – Les cinq variables les plus importantes selon la méthode d'ordonnancement

Résultats obtenus sur la base HIVA

Nous avons représenté sur la figure 2.7 le taux d'exemples bien classés par le Naïf Bayes sur l'ensemble de validation pendant le processus forward en fonction de la méthode d'ordonnancement utilisée [5].

[5] Voir le fichier "tableau_recapitulatif_hiva.xls" pour les résultats.

Sur le deuxième, nous avons représenté le taux d'erreur de classification (BER) pondérée par la classe du Naïf Bayes sur l'ensemble de validation pendant le processus forward en fonction de la méthode d'ordonnancement utilisée

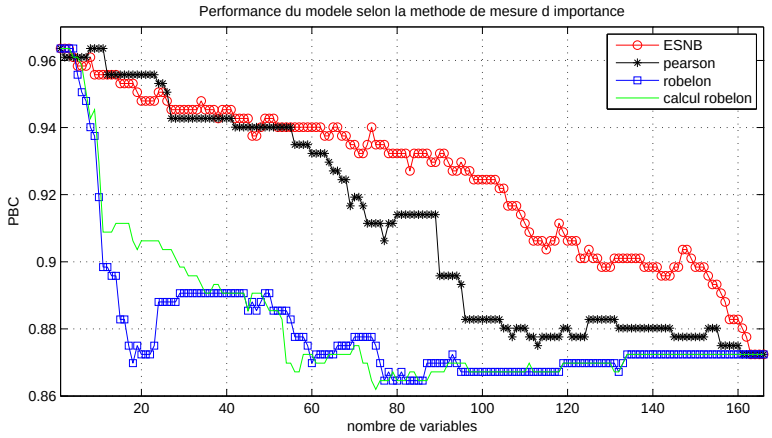

FIG. 2.7 – Taux d'exemples bien classés pendant le processus forward en fonction de différentes méthodes d'ordonnancement des variables.

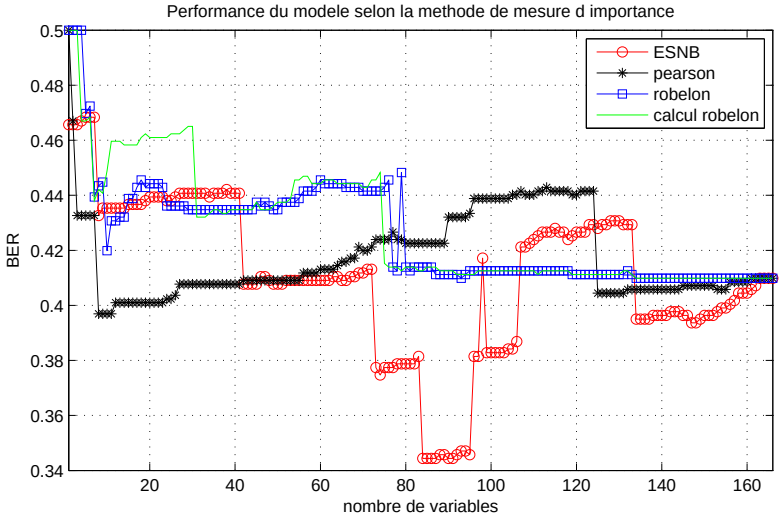

FIG. 2.8 – Taux d'exemples bien classés pendant le processus forward en fonction de différentes méthodes d'ordonnancement des variables.

Discussion

Sur la base HIVA, la figure 2.7 montre que les performances du modèle, pour un ensemble de variables composé de plus de 2 éléments, sur l'ensemble de validation utilisant l'ordonnancement de la méthode ENSB et pearson sont largement supérieures aux performances utilisant l'ordonnancement robelon. Le meilleur PBC est de 96.35% pour les 3 méthodes en considérant respectivement la première variable la plus importante selon la méthode. Cela s'explique par le fait que les exemples de classe 1 constituent 96.35% de l'ensemble des exemples et que le modèle classe les exemples en classe 1 par défaut. Ensuite le PBC s'érode progressivement quelque soit la méthode, avec une baisse du PBC plus importante pour robelon que pour pearson et ESNB. Le PBC de ESNB et pearson sont équivalent jusqu'à la 54 ième variable et vaut 94.01% tandis que le PBC de robelon est de 89.06% à partir de la 29 ième variable.
On sélectionnerait donc pour chacune des 3 méthodes la première variable la plus importante.

En observant le graphe du BER de la figure 2.8 le BER minimum est de 0.3444 pour ESNB et il est obtenu avec les 84 variables les plus importantes selon ESNB. Pour pearson le BER minimum est 0.3969, pour l'ensemble constitué des 8 premières variables les plus importantes selon pearson. Pour robelon, le BER minimum est de 0.4098 et il est obtenu avec les 93 variables les plus importantes selon robelon. On observe donc que robelon donne de moins bon résultats que pearson et ESNB dans l'ensemble. Par contre avec robelon on sélectionnerait les 5 premières variables puisque le BER se dégrade à partir de la sixième variable : il passe de 0.4697 à 0.4724. Pour la méthode ENSB, on sélectionnerait la première variable et pour Pearson on sélectionnerait les 11 premières variables.

Pour la méthode issue du calcul de robelon, on sélectionnerait la première variable si le PBC est le critère à optimiser et les 5 premières variables si on veut optimiser le BER.On obtient respectivement avec cette méthode un PBC et un BER de 96.35% et 0.4670. Sur tableau 2.6 et le tableau 2.7 nous résumons les résultats obtenus pour les quatre méthodes d'ordonnancement et sur le tableau 2.8 nous présentons les 5 variables les plus importantes selon chaque méthode.

Tableaux récapitulatifs

	tps execution	# var select	PBC
ESNB	87.07H	1	96.35 %
Robelon	171 mn	1	96.35 %
Pearson	0.9060 s	1	96.35 %
Calcul Robelon	53 s	1	96.35 %

TAB. 2.6 – Résultats obtenus selon la méthode d'ordonnancement

	tps execution	# var select	BER
ESNB	87.07H	1	0.4656
Robelon	171 mn	5	0.4697
Pearson	0.9060 s	11	0.3969
Calcul Robelon	53 s	5	0.4670

TAB. 2.7 – Résultats obtenus selon la méthode d'ordonnancement

ESNB	Pearson	Robelon	calcul robelon
861	10	413	183
1	963	200	930
299	861	239	1246
136	1068	1368	1297
461	1410	915	1230

TAB. 2.8 – Les cinq variables les plus importantes selon la méthode d'ordonnancement

Résultats obtenus sur la base SYLVA

Sur la figure 2.9, nous avons représenté le taux d'exemples bien classés par le Naïf Bayes sur l'ensemble de validation pendant le processus forward en fonction de la méthode d'ordonnancement utilisée [6].
Sur le deuxième, nous avons représenté le taux d'erreur de classification (BER) pondérée par la classe du Naïf Bayes sur l'ensemble de validation pendant le processus forward en fonction de la méthode d'ordonnancement utilisée

Discussion

Sur la base SYLVA, la figure 2.9 montre que le graphe des PBC de la méthode robelon est légèrement en dessous du graphe ESNB et du graphe pearson. Le graphe "calcul robelon" issue de l'ordonnancement utilisant le résultat du calcul théorique de robelon donne de meilleurs résultats que robelon. En observant le graphe 2.9, la méthode ENSB permet de sélectionner les 6 premières variables (183, 171, 93, 46, 52, 99) ; on atteint avec ces variables un PBC de 99.39%. La méthode de pearson permet de sélectionner les 2 variables les plus importantes selon pearson (183, 171). On atteint avec ces 2 variables un PBC de 98.62%. La méthode robelon quant à elle, nous permet de sélectionner les 2 premières variables (182, 204). Ses performances sont largement inférieures à ENSB et pearson puisqu'on a un PBC de 93.88% en utilisant ces 2 variables. La méthode "calcul robelon" donne des résultats équivalents à pearson et légèrement inférieurs à ENSB. Elle permet de sélectionner les 3 premières variables

[6] Voir le fichier "tableau_recapitulatif_sylva.xls" pour les résultats.

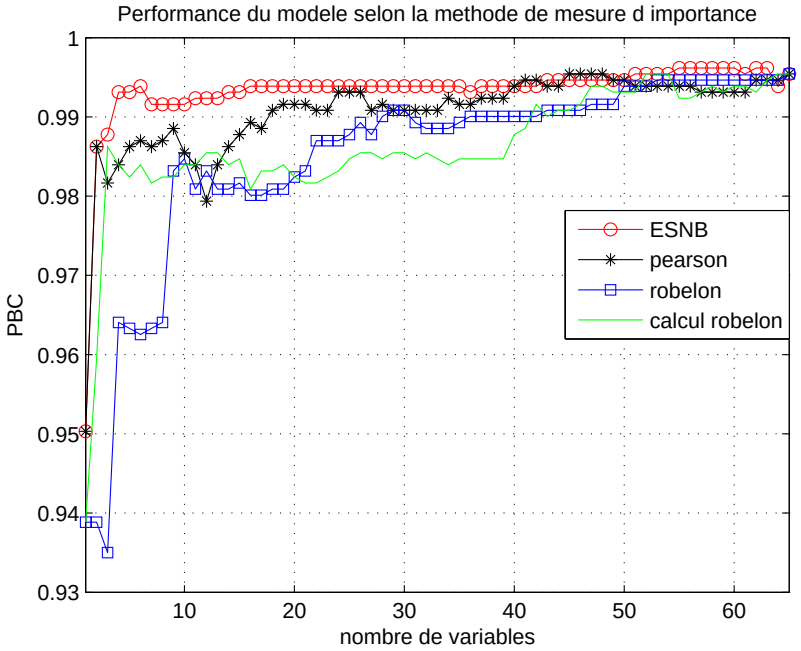

FIG. 2.9 – Taux d'exemples bien classés pendant le processus forward en fonction de différentes méthodes d'ordonnancement des variables.

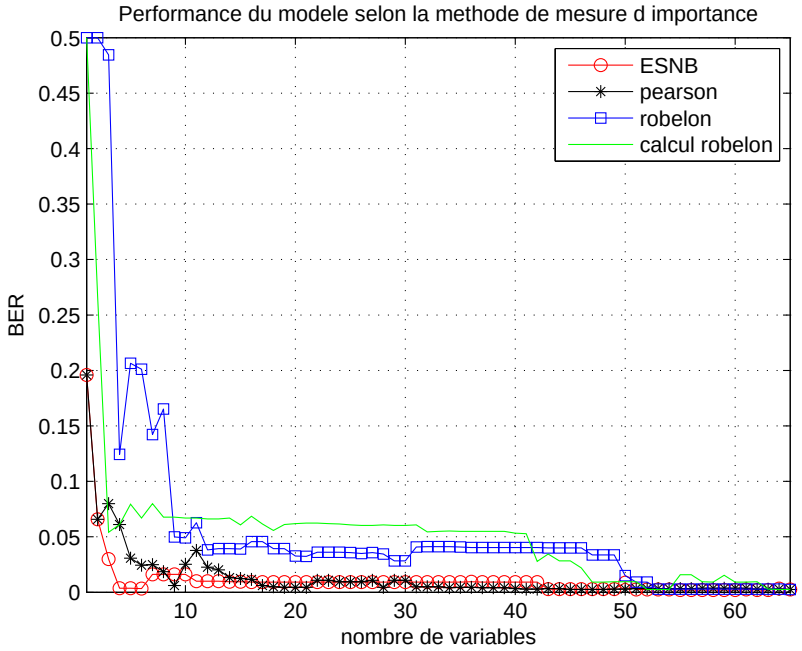

FIG. 2.10 – Taux d'exemples bien classés pendant le processus forward en fonction de différentes méthodes d'ordonnancement des variables.

les plus importantes selon cette mesure d'importance (88, 171, 183) et permet d'atteindre un PBC de 98.62%. Le graphe des BER de la méthode pearson et le graphe des BER de la methode ENSB sont légèrement en dessous et donnent de meilleurs résultats que le graphe des BER de robelon. En observant le graphe des BER, on sélectionnerait pour ENSB toujours les 6 premières variables puisqu'elles donnent un BER 0.0033 et 0.0162 pour les 7 premières variables.Pour pearson, on sélectionnerait également les 2 premières variables comme précédemment. On a avec ces 2 variables un BER de 0.0658 contre 0.0799 pour les 3 premières variables. Par contre avec la méthode robelon, en regardant le BER, on sélectionnerait les 4 premières variables au lieu des 2 premières quand on regarde le PBC. Avec les 4 premières variables, on atteint un BER de 0.1243, contre 0.5 pour les 2 premières variables et 0.2065 pour les 5 premières. Pour la méthode "calcul robelon", on sélectionnerait également les 3 premières variables en regardant le BER. Avec ces 3 variables, on atteint un BER de 0.0541 contre 0.0611 pour les 4 premières et 0.26 pour les 2 premières. Sur le premier tableau nous résumons les résultats obtenus pour les trois méthodes d'ordonnancement et sur le deuxième tableau nous présentons les 5 variables les plus importantes selon chaque méthode.

Tableaux récapitulatifs

	tps execution	# var select	PBC
ESNB	7.69H	6	99.39 %
Robelon	120 mn	2	93.88 %
Pearson	0.86s	2	98.62 %
Calcul Robelon	310 s	3	98.62 %

TAB. 2.9 – Résultats obtenus selon la méthode d'ordonnancement

	tps execution	# var select	BER
ESNB	7.69H	6	0.0033
Robelon	120 mn	2	0.1243
Pearson	0.86s	2	0.0658
Calcul Robelon	310 s	3	0.0541

TAB. 2.10 – Résultats obtenus selon la méthode d'ordonnancement

Discussion générale sur la méthode robelon

La méthode robelon donne pendant le processus forward de moins bons résultats que la méthode ESNB et pearson sur les trois bases ADA, HIVA et SYLVA. Le temps

ESNB	Pearson	Robelon	calcul robelon
183	183	182	88
171	171	204	171
93	21	18	183
46	85	171	182
52	52	88	155

TAB. 2.11 – Les cinq variables les plus importantes selon la méthode d'ordonnancement

de calcul de ESNB pour ordonnancer les variables est en général au moins deux fois plus élevé que celui de robelon pour ENSB.La méthode pearson met quelques secondes pour ordonnancer les variables contre plusieurs minutes pour robelon. La méthode "calcul robelon" quant à elle donne des résultats légèrement supérieurs à pearson. Ses performances sont meilleures que celles de robelon et son temps de calcul est de quelques secondes.

Chapitre 3

Interprétation des résultats obtenus par un modèle boite noire

3.1 Etat de l'art des méthodes d'interprétation des modèles "boite noire"

3.1.1 Le problème de l'interprétation des modèles "boite noire"

Le domaine de l'apprentissage automatique regorge aujourd'hui de techniques capables de résoudre efficacement des problèmes de régression et de classification. Ces techniques construisent un modèle F à partir d'une base de données d'apprentissage constituée d'un nombre fini d'exemples (des couples de vecteurs entrée-sortie (I, S)).

Le modèle construit est ensuite utilisé pour associer à un exemple positionné en entrée I une sortie $S : S = F(I)$ (voir figure ci-dessous). Dans de nombreuses applications la seule valeur de S est insuffisante à une prise de décision concernant I. Une méthode d'interprétation de la valeur de S connaissant I et le modèle F est nécessaire à une prise de décision. Il est nécessaire de savoir pourquoi le modèle délivre S quand on lui présente I en entrée.

La grande variété des modèles existants dans la littérature impose d'avoir une méthode d'interprétation propre à chaque modèle (connaissance des paramètres internes du modèle). L'interprétation résultante est souvent complexe et inexploitable par une personne extérieure au monde scientifique. De plus tous les types de modèles ne possèdent pas de méthodes d'interprétation. Nous présentons dans les sections suivantes une méthode d'interprétation de modèle.

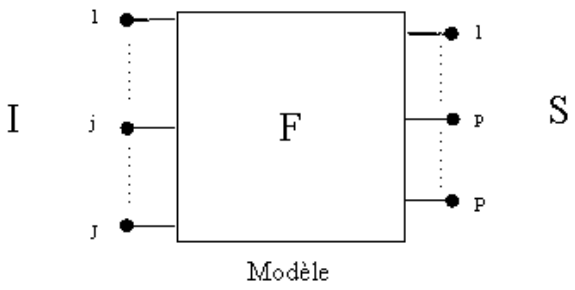

FIG. 3.1 – représentation schématique d'un modèle "boite noire"

3.1.2 Méthode d'interprétation des résultats obtenus par un réseau de neurones

Cette méthode a été développée par Kary FRÄMLING (voir [5]). L'objectif est d'interpréter les résultats obtenus par un réseau de neurones. On ne cherche pas à comprendre comment fonctionne le réseau de neurones mais à dire pourquoi il délivre telle valeur en sortie pour telles valeurs en entrée du modèle. Pour cela, la méthode se base sur deux concepts : l' *Importance Contextuelle* et l' *Utilité Contextuelle* d'une variable d'entrée j du modèle pour une sortie p du modèle d'un exemple I_n.

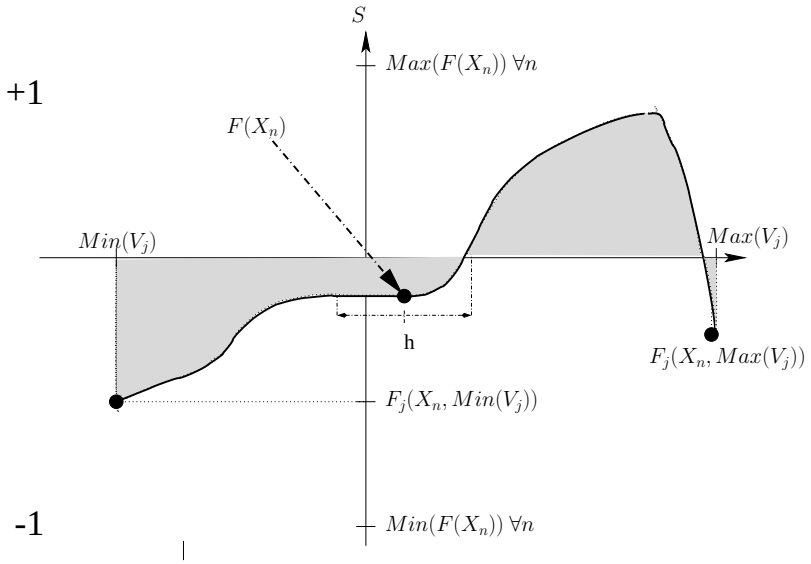

FIG. 3.2 – représentation représentation graphique de la méthode de k. Främling

L'Importance Contextuelle d'une entrée j pour la sortie S_p du modèle pour l'exemple

I_n est définie de la manière suivante :

$$IC(V_j/F, I_n, S_p) = \frac{F_j(I_n, \max V_j) - F_j(I_n, \min V_j)}{\max_n(F(I_n)) - \min_n(F(I_n))} \tag{3.1}$$

où IC est l'importance contextuelle, F le modèle, $F_j(I_n, \max V_j)$ la sortie du modèle pour l'exemple I_n en ayant changé la valeur I_{nj} par la plus grande valeur de la variable V_j, $F_j(I_n, \min V_j)$ la sortie du modèle pour l'exemple I_n en ayant la valeur I_{nj} par la plus petite valeur de la variable V_j, $\max_n(F(I_n))$ le plus grand score réalisé sur tous les exemples d'apprentissage pour la sortie p et $\min_n(F(I_n))$ le plus petit score réalisé sur tous les exemples d'apprentissage pour la sortie p.

L'*Utilité Contextuelle* de la valeur d'une entrée j (la variable V_j peut prendre plusieurs valeurs) pour l'exemple I_n est définie de la manière suivante :

$$UC(V_j/F, I_n, S_p) = \frac{F(I_n) - F_j(I_n, \max V_j)}{F_j(I_n, \max V_j) - F_j(I_n, \min V_j)} \tag{3.2}$$

Cette méthode de mesure d'importance des variables n'est valable que si F est monotone. Ce qui n'est pas le cas pour une grande partie des modèles. La figure 3.2 illustre cet inconvénient de la méthode. Si F n'est pas monotone, elle a tendance à sous-estimer les variables importantes : si la variable V_j est importante et que $F_j(I_n, \max V_j)$ et $F_j(I_n, \min V_j)$ sont proches, IC est petite ; ce qui voudrait dire que la variable V_j n'est pas importante selon la méthode de K. Främling. Or la variable V_j est importante. La méthode est donc érronée pour les modèles non monotones.

Elle permet pour les modèles monotones (les réseaux de neurones par exemple) d'expliquer les résultats de chaque sortie de manière compréhensible par tous, pour chaque exemple en entrée du modèle.

3.2 Une nouvelle méthode d'interprétation : la méthode d'interprétation "Robelon"

3.2.1 Introduction

La méthode d'interprétation présentée ici, est basée sur une approche issue de la méthode "Robelon" (voir [7]), permettant de mesurer l'importance d'une variable étant donné un modèle. C'est pourquoi nous l'avons appelée *la méthode d'interprétation "robelon"*. Elle permet de fournir une explication sous une forme intelligible par tous, de chaque sortie du modèle et ce pour chaque exemple. Elle s'applique à tous les types de modèles, sur des problèmes de régression et de classification. Elle est donc plus générale que la méthode de K. Främling présentée à la section 3.1.2. K. Främling a introduit dans [5] les notions d'Importance Contextuelle (IC) et d'Utilité Contextuelle (UC) pour interpréter chaque sortie du modèle. S'inspirant de ces deux notions, dans la méthode d'interprétation "Robelon" nous introduisons quatre notions : l'"Importance à l'Exemple", l'"Utilité à l'Exemple, l'"Importance Contextuelle à l'Exemple" et l'"Utilité Contextuelle à l'Exemple" notés respectivement IE, UE, ICE, UCE que nous définirons et expliquerons dans les sections suivantes. L'IE,l'UE, l'ICE

et l'UCE sont les outils qui nous permettrons d'expliquer et d'interpréter les résultats d'un modèle.

3.2.2 Définition de l'IE, l'UE, l'ICE, et l'UCE

Définition de l"Importance à l'exemple" (IE)

La méthode permet d'expliquer chaque sortie du modèle pour un exemple donné (pour chaque exemple en entrée du modèle). Pour cela on introduit le concept d'*Importance à l'Exemple* (IE) inspiré par la notion d' *Importance Contextuelle* de K. Främling 3.1.2.

Etant donné le modèle construit F et l'exemple I_n en entrée du modèle. Pour chaque variable V_j en entrée du modèle et chaque variable S_p en sortie du modèle, l'Importance à l'Exemple $IE(V_j/F, I_n, S_p)$ mesure l'importance de la variable étudiée V_j, quand à la délivrance de la valeur de la sortie du modèle S_p étant donné l'exemple I_n et le modèle F : une variable non importante pour l'exemple en question n'influe pas sur le résultat du modèle. On appellera cette importance "importance à l'exemple" (IE)

Définition de l"Utilité à l'exemple" (UE)

L'Utilité à l'Exemple $UE(V_j/F, I_n, S_p)$ mesure l'utilité de la valeur de la variable étudiée V_j, quand à la délivrance de la valeur de la sortie du modèle S_p étant donné l'exemple I_n et le modèle F : une valeur de la variable V_j peut "tirer vers le haut" (valeur de la sortie du modèle forte) ou "tirer vers le bas" (valeur de la sortie du modèle faible) la sortie du modèle. On appellera cette importance "utilité à l'exemple" (noté UE).

Définition de l"Importance Contextuelle à l'Exemple" et de l'"Utilité Contextuelle à l'Exemple"

L'importance contextuelle pour un exemple de chaque variable V_j en entrée du modèle et ce pour chaque sortie S_p du modèle (noté ICE) et l'utilité contextuelle pour un exemple de la valeur de chaque variable V_j en entrée du modèle et ce pour chaque sortie S_p (noté UCE) sont définies comme suit :

Etant donné un modèle F et un exemple I_n. On définit pour chaque variable V_j d'entrée du modèle et chaque variable S_p de sortie du modèle les deux mesures suivantes :

- Pour "l'Importance Contextuelle à l'Exemple" (ICE) les étapes suivantes sont à réaliser :
  1. On calcule l'$IE(V_j/F, I_n, S_p)$ telle que définie précédemment sur l'ensemble des exemples et des variables que l'on possède lors de la construction du modèle ;
  2. on ordonne la distribution des valeurs de l'*IE* calculées à l'étape 1 ;
  3. on établit la statistique d'ordre (fonction de distribution cumilative) de l'*IE* calculées à l'étape 2 ;

4. pour l'exemple dont on désire connaitre l'importance de la variable considérée (V_j) :
   - On calcule l'$IE(V_j/F, I_n, S_p)$ telle que définie précédemment ;
   - on évalue le rang de l'$IE(V_j/F, I_n, S_p)$ dans la fonction de distribution cumulative des *IE* ;
5. le rang de l'*IE* détermine "l'Importance Contextuelle à l'Exemple" (ICE)

– Pour "l'Utilité Contextuelle à l'Exemple" (UCE) les étapes suivantes sont à réaliser :

1. On calcule l'$UE(V_j/F, I_n, S_p)$ telle que définie précédemment sur l'ensemble des exemples et des variables que l'on possède lors de la construction du modèle ;
2. on ordonne la distribution des valeurs de l'*UE* calculées à l'étape 1 ;
3. on établit la statistique d'ordre (fonction de distribution cumilative) de l'*UE* calculées à l'étape 2 ;
4. on conserve cette statistique d'ordre ;
5. pour l'exemple dont on désire connaitre l'importance de la variable considérée (V_j) :
   - On calcule l'$UE(V_j/F, I_n, S_p)$ telle que définie précédemment ;
   - on évalue le rang de l'$UE(V_j/F, I_n, S_p)$ dans la fonction de distribution cumulative des *UE* ;
6. le rang de l'*UE* détermine "l'Importance Contextuelle à l'Exemple" (ICE)

3.2.3 Calcul de l'*IE* et représentation graphique du calcul

Soit j la variable dont on cherche l'importance à l'exemple (IE).
Soit V_{ij} une réalisation de la variable de la variable j.
Soit I un vecteur de dimension J, un exemple ayant servi à la construction du modèle.
Soit I_n le vecteur n.
Soit I_{nl} la composante l du vecteur n.
Soit F le modèle réalisé.
Soit $P_{V_j}(\nu)$ la distribution de probabilité de la variable j.
Soit $P_I(\nu)$ la distribution de probabilité des exemples I.

On pose $F_j(a;b) = F_j(a_1,\ldots,a_n;b) = F_j(a_1,\ldots,a_{j-1},b,a_{j+1},\ldots,a_J)$; a_p étant la p^{ime} composante du vecteur a.

On définit pour le modèle F et l'exemple I_n, l'importance de la variable V_j pour la variable de sortie S_p comme étant la somme des variations mesurées en sortie du modèle lorsqu'on perturbe l'exemple considéré I_n (voir la figure 3.4 pour la représentation graphique de l'importance à l'exemple) , en fonction de la distribution de probabilité de la variable V_j (voir la figure 3.3 pour l'illustration graphique du tirage selon la distribution de probabilité de la variable j).

La sortie perturbée du modèle F, pour un exemple I_n est définie comme étant la sortie du modèle pour cet exemple I_n mais en ayant échangé la composante j de cet exemple par l'une des valeurs k, de la variable V_j. La variation mesurée, pour l'exemple

I_n est donc la différence entre la "vraie sortie" du modèle $F_j(I_n, I_{nj})$ pour l'exemple I_i et la "sortie perturbée du modèle $F_j(I_n, I_{kj})$. L'importance à l'exemple de la variable V_j est alors la somme des $|F_j(I_n, I_{nj}) - F_j(I_n, I_{kj})|$ sur la distribution de probabilité de la variable V_j.

On définit donc l'importance à l'exemple comme étant :

$$IE(V_j|F, I_n, S_k) = \int P_{V_j}(u)\{|F_j(I_n, I_{nj}) - F_j(I_n, I_{kj})|\}$$

La figure 3.4 est une illustration graphique de l'IE de la variable V_j. La courbe de la figure 3.4 représente les perturbation de l'exemple I_i en fonction de la valeur de la variable j. Par définition de l'intégrale, cette quantité correspond à l'aire en dessous de la courbe (l'aire grise). Plus cette aire est grande, plus la variable est importante

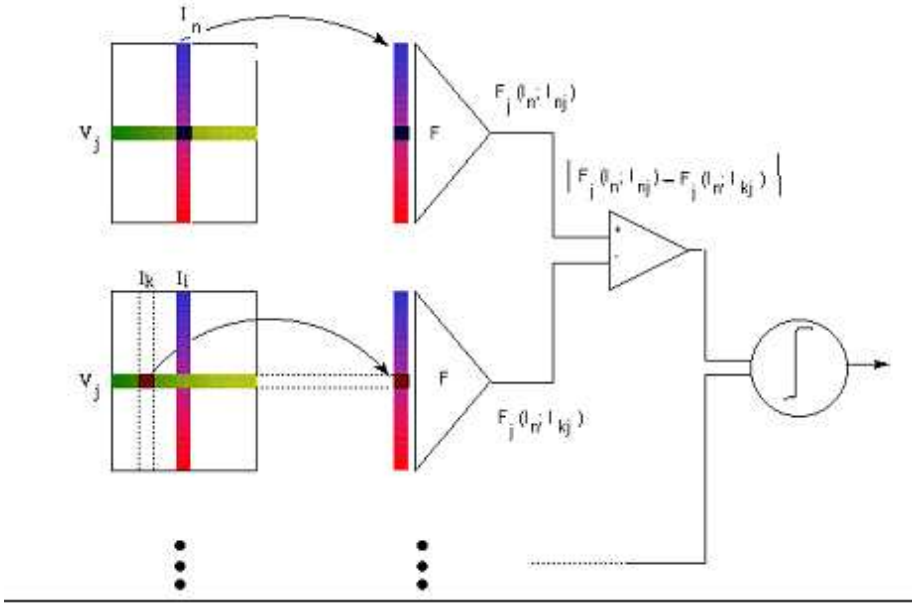

FIG. 3.3 – représentation graphique du tirage selon la distribution de probabilité de la variable j

3.2.4 Calcul de l'*UE*

L'Utilité à l'Exemple (UE) pour le modèle F, l'une des variables S_p en sortie du modèle, l'une des variable V_j en entrée du modèle et un exemple I_n est défini tel que :

$$UE(V_j|F, I_n, S_p) = F(I_n, I_{nj})$$

L'Utilité à l'Exemple (UE) est simplement la valeur de la sortie du modèle S_p étant donné l'exemple I_n.

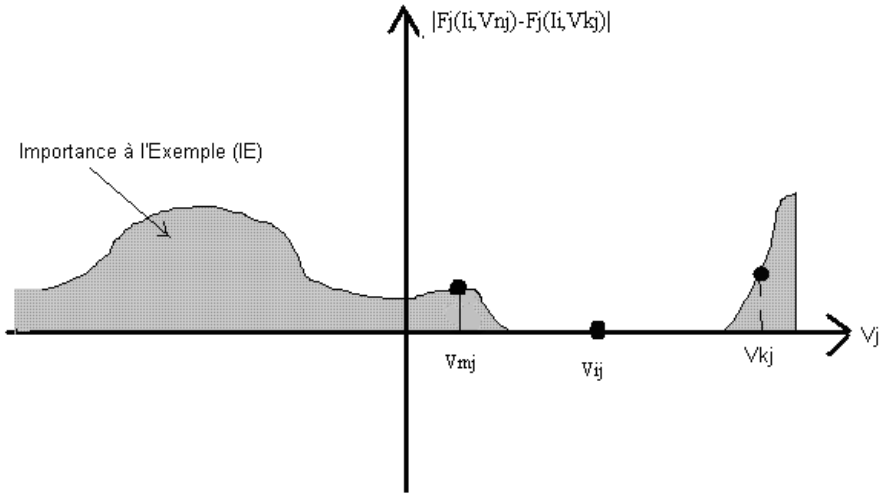

FIG. 3.4 – représentation graphique : l'aire sous la courbe représente l'importance à l'exemple

L'Utilité à l'Exemple "perturbée" ($UE_{perturb}$) pour le modèle F, l'une des variables S_p en sortie du modèle, l'une des variables V_j en entrée du modèle et un exemple I_n est défini tel que :

$$UE_{perturb}(V_j|F, I_n, S_p) = F(I_n, I_{kj})$$

L'Utilité à l'Exemple "perturbé" ($UE_{perturb}$) est simplement la valeur de la sortie du modèle S_p étant donné l'exemple I_n mais pour lequel on a remplacé la valeur de sa j^{ime} composante par la valeur d'un autre exemple k.

3.2.5 Illustration graphique de la *notion d'Importance à l'Exemple* (IE)

La méthode "robelon" permet d'évaluer pour un exemple donné I_i, l'importance de chaque variable étant donné un modèle F. Cette mesure est basée sur les perturbations de la sortie du modèle pour cet exemple quand on fait varier la valeur de la variable étudiée suivant la distribution de probabilité de cette variable. Cela veut dire que si on étudie la variable V_j par exemple, et que cette variable prend n valeurs $\{V_1, \ldots, V_n\}$. Si on note V_{ij} la valeur de la variable j pour l'exemple i, on fait varier la variable j en changeant successivement la valeur V_{ij} par les autres valeurs que peut prendre la variable j.

La figure 3.5 illustre garphiquement l'influence d'une variable sur la sortie du modèle. Elle donne une image du raisonnement utilisée pour mesurer l'importance d'une

variable.

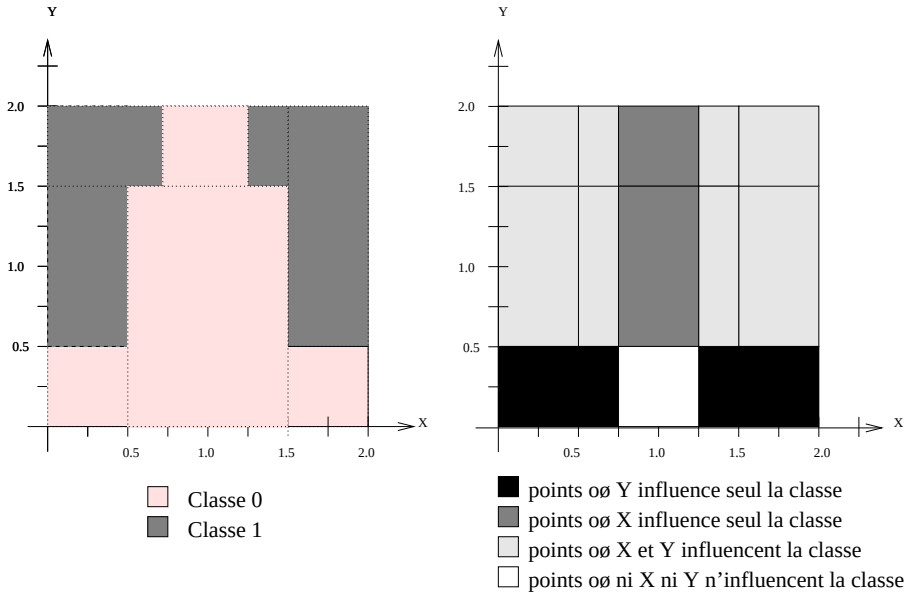

FIG. 3.5 – Illustration graphique de la méthode robelon

La figure 3.5 représente le plan. Chaque point P du plan représente un exemple. Chaque coordonnée X et Y est une caractéristique (variable) de l'exemple P. le plan est séparée en deux parties représentant chacune une classe.

Considérons l'exemple (le point) $I(0.25, 0.25)$ qui appartient à la classe 0. Pour mesurer l'influence de la variable Y sur la sortie du modèle (la classe à laquelle appartient P), on fait varier Y. On se pose donc la question de savoir quelle serait la classe de I si on changeait la valeur de $Y = 0.25$ par une autre valeur de $Y \in [0, 2]$? En regardant la figure 3.5, on remarque que si Y prend ses valeurs dans l'intervalle $[0.25, 0.5]$, la classe de l'exemple ne change pas ; I appartient toujours à la classe 0. Par contre, si Y prend ses valeurs dans l'intervalle $[0.5, 2]$, I appartient à la classe 1 : le résultat change donc et la variable Y a donc une influence sur le résultat. en perturbant la variable Y de l'exemple I selon la distribution de probabilité de Y, I passe de la classe 0 à la classe 1.

Par contre, la variable X n'a aucune influence sur le résultat. En effet si on fait varier la variable X de l'exemple I selon la distribution de probabilité de X, i.e $X \in [0, 2]$, on remarque en observant la figure 3.5 que la classe ne change pas : les exemples I perturbés appartiennent toujours à la classe 0.

Cette illustration grahique permet de comprendre comment fonctionne la méthode robelon et le raisonnement qui permet de mesurer l'importance d'une variable pour un exemple étant donné un modèle.

3.2.6 Calcul de l'*ICE* et représentation graphique

Définition d'un partile

Pour une population infinie d'une distribution d'une variable aléatoire X, le o^{ieme} q-quantile est la valeur des données où la fonction de distribution cumulative vaut $\frac{o}{q}$.

De manière générale, le o^{ieme} q-quantile d'une distribution d'une variable aléatoire X, où la fonction de distribution cumulative a été établie, peut être définit par la valeur x tel que : $P(X \leq x) \geq \frac{o}{q}$.

Pour un nombre fini de N tirages, il faut calculer ($N \quad \frac{o}{q}$) et arrondir à l'entier supérieur.

Application à la définition du rang de la valeur de l'*IE* : l'importance Contextuelle à l'exemple *ICE*

Pour le modèle F, une variable V_j, l'une des variables S_p en sortie du modèle, un exemple I_n et le choix d'une valeur o, par exemple ici $o = 100$ si l'on utilise le rang au sein des "centiles", on définit l'*ICE* tel que :

- ETAPE 1 : sur l'ensemble des J variables :
  on calcule l'$IE(V_j|F, I_i, S_p)$ telle que définie précédemment sur l'ensemble des N exemples I_i qui ont servi à la construction du modèle et l'ensemble des J variables : $IE(V_j|F, I_i, S_p) \quad \forall i, j$ (on possède donc N réalisations de la variable aléatoire $IE(V_j|F, I_i, S_p)$) ;
- ETAPE 2 : on ordonne la distribution des valeurs de l'$IE(V_j|F, I_i, S_p) \quad \forall i, j$ calculées à l'étape précédente de manière à obtenir la fonction de distribution cumulative de $IE(V_j|F, I_i, S_p) \quad \forall i, j$;
- ETAPE 3 : on définit l'*ICE* de l'exemple I_n comme étant le rang o (ou l'appartenance au rang) de son $IE(V_j|F, I_n, S_p)$ au sein de la fonction de distribution cumulative des $IE(V_j|F, I_i, S_p) \quad \forall i, j$ tel que :
$$P[(IE_{cumul}(V_j|F, I_i, S_p) \forall i, j) \leq (IE(V_j|F, I_n, S_p)] \geq \frac{o}{100}$$

La figure 3.6 représente la distribution cumulative des *IE*.

3.2.7 Calcul de l'*UCE*

Pour le modèle F, une variable V_j, l'une des variables S_p en sortie du modèle, un exemple I_n et le choix d'une valeur o, par exemple ici $o = 100$ si l'on utilise le rang au sein des "centiles", on définit l'*UCE* tel que :

- ETAPE 1 :on calcule les N valeurs l'$UE_{perturb}(V_j|F, I_i, S_p)$ telle que définie précédemment à l'aide de l'ensemble des N exemples I_i qui ont servi à la construction du modèle (on possède donc N réalisations de la variable aléatoire $UE_{perturb}(V_j|F, I_i, S_p)$) ;
- ETAPE 2 : on ordonne la distribution des valeurs de l'$UE(V_j|F, I_i, S_p) \quad \forall i, j$ calculées à l'étape précédente de manière à obtenir la fonction de distribution cumulative de $UE_{perturb_{cumul}}(V_j|F, I_i, S_p) \quad \forall i, j$;

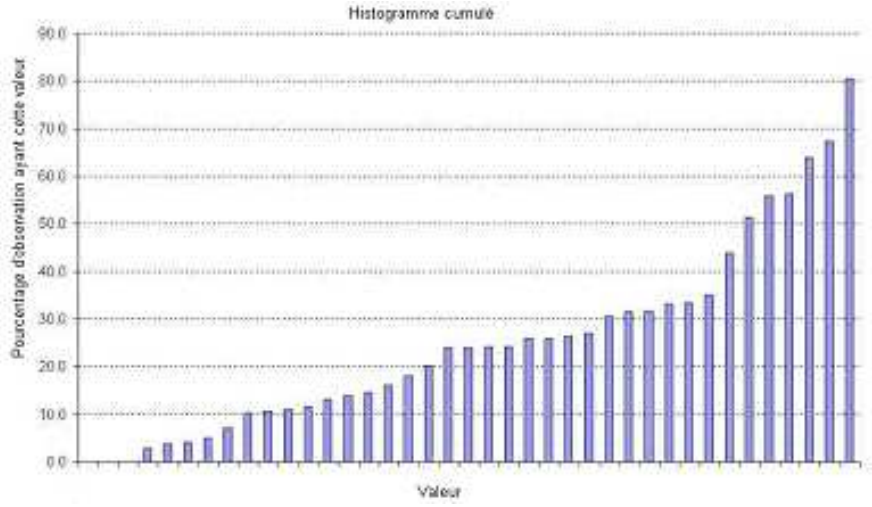

FIG. 3.6 – représentation graphique de la distribution cumulative des *IE*

– ETAPE 3 : on définit l'*UCE* de l'exemple I_n comme étant le rang o (ou l'appartenance au rang) de son $UE(V_j|F, I_n, S_p)$ au sein de la fonction de distribution cumulative des $UE(V_j|F, I_i, S_p) \quad \forall i, j$ tel que :
$P[(UE_{perturb_{cumul}}(V_j|F, I_i, S_p) \forall i, j) \leq (UE(V_j|F, I_n, S_p)] \geq \frac{o}{100}$

3.3 Application de la méthode d'interprétation "Robelon" à un exemple jouet

3.3.1 Description de la base utilisée

Dans cette section, nous appliquons la méthode d'interprétation à un cas concret. Nous voulons interpréter les résultats de sortie d'un modèle naïf bayes. Pour cela, nous disposons d'une base de 2006 exemples séparés en 2 classes. Nous avons utilisé les 1000 premiers exemples de la base pour apprendre les paramètres du modèle (du naïf bayes) et les 6 derniers pour tester la méthode d'interprétation. Chaque exemple est représenté par deux variables X et Y continues et qui ont été discrétisées à l'aide de KHIOPS [1] (voir [1]). La figure 3.5 représente à gauche l'ensemble des points des deux classes et la figure 3.7 représente les 2006 exemples. Sur la figure 3.7, les points encadrés A(0.25,1.5), B(1,1.5), C(1.75,1.5), D(0.25,0.25), E(1,0.25) et F(1.75,0.25) sont les exemples de test. Ils appartiennent chacun à une des zones d'influence représentée sur la figure 3.5.

[1] Méthode de discrétisation de variables continues développée par France Telecom R& D

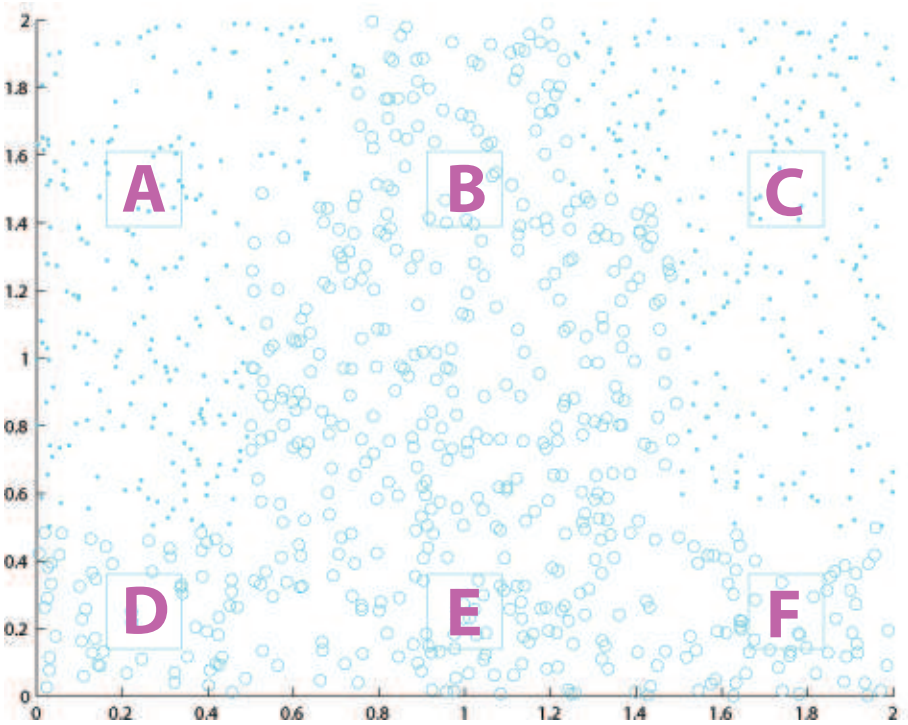

FIG. 3.7 – Illustration des 2006 exemples issue du tirage, dont 1000 ont servi à apprendre et 6 exemples encadrés à tester la méthode

3.3.2 Méthodologie suivie

Le but est d'interpréter les sorties (dans notre cas le modèle a 2 sorties) du naïf bayes. Nous appelons sortie 1, la sortie relative à la classe 1 et sortie 2, la sortie relative à la classe 0. Nous avons réalisé les étapes suivantes :

- Apprentissage des paramètres du naïf bayes ;
- calcul et mémorisation des IE des 1000 exemples d'apprentissage et ce pour chaque variable et chaque sortie du modèle ;
- calcul de l'IE, l'UE, l'ICE et l'UCE pour chaque exemple de l'ensemble de test ;
- interprétation phrasée des résultats pour l'exemple F.

3.3.3 Construction des éléments de l'interprétation et discussion

Le tableau 3.1 récapitule l'IE, l'UE, l'ICE (le rang de l'IE sur 1000) et l'UCE (le rang de l'UE sur 1000) de X et Y et ce pour chaque exemple de l'ensemble de test pour la sortie 1 (sortie relative à la classe 1) du modèle et le tableau 3.2 donne l'IE, l'UE, l'ICE (le rang de l'IE sur 1000) et l'UCE (le rang de l'UE sur 1000) de X et Y pour chaque exemple de l'ensemble de test pour la sortie 2 (sortie relative à la classe 0) du modèle.

V_i, I_n	IE	UE	ICE	UCE	Variables influentes
$V_i = X, I_n = I_A$	44.5903	0.0941	417	876	X et Y influencent la classe
$V_i = Y, I_n = I_A$	24.7726	0.0941	116	524	X et Y influencent la classe
$V_i = X, I_n = I_B$	49.5373	0	852	124	seul X influence la classe
$V_i = Y, I_n = I_B$	71.8888	0	789	121	seul X influence la classe
$V_i = X, I_n = I_C$	43.1347	0.0912	339	628	X et Y influencent la classe
$V_i = Y, I_n = I_C$	24.6387	0.0912	38	383	X et Y influencent la classe
$V_i = X, I_n = I_D$	0	0	121	501	seul Y influence la classe
$V_i = Y, I_n = I_D$	71.8888	0	789	121	seul Y influence la classe
$V_i = X, I_n = I_E$	0	0	121	501	aucune n'influence la classe
$V_i = Y, I_n = I_E$	71.8888	0	789	121	aucune n'influence la classe
$V_i = X, I_n = I_F$	0	0	121	501	seul Y influence la classe
$V_i = Y, I_n = I_F$	71.8888	0	789	121	seul Y influence la classe

TAB. 3.1 – Tableau des IE, UE, ICE, UCE de X et Y et les variables influentes déterminées empiriquement d'après la figure 3.5 pour les exemples A, B, C, D, E, F et pour la sortie 1 (sortie relative à la classe 1)

La figure 3.8 représente la distribution des IE calculés sur les exemples d'apprentissage de la variable X pour la sortie 1 du Naïf bayes . Les valeurs en abcisse représentent les IE des exemples d'apprentissage et les valeurs en ordonnée le rang des IE. De même, la figure 3.9 représente la distribution des IE calculés sur les exemples d'apprentissage de la variable Y pour la sortie 1. Au regard de la figure 3.8, on constate que la distribution des IE de la variable X a quelques modalités importantes. Pour simplifier l'interprétation, on peut subdiviser l'ensemble des rangs en 5 plages de rangs par exemple, pour qualifier l'IE d'un exemple. Par exemple pour la variable X, on peut qualifier l'importance de cette variable X pour un exemple I dont le rang de l'IE

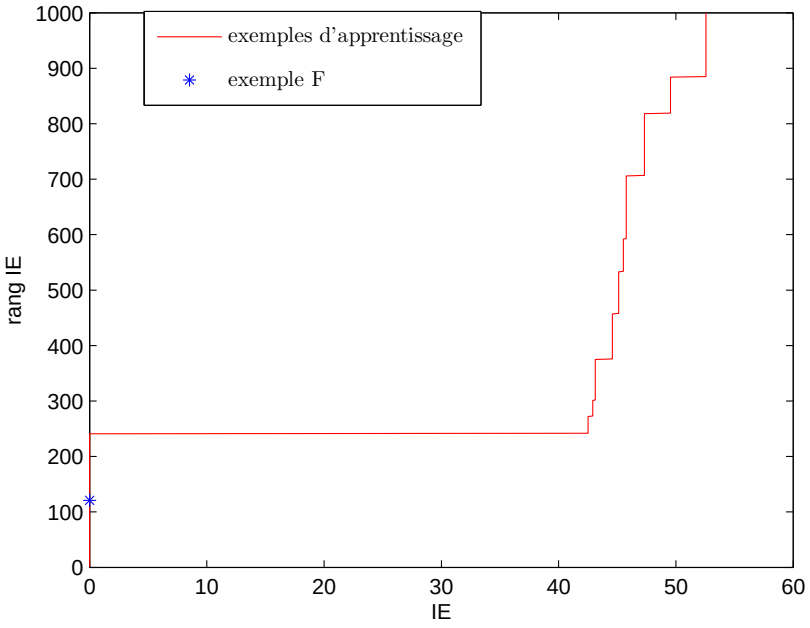

FIG. 3.8 – Distribution ordonnée des IE de la variable X pour la sortie 1 du naïf bayes

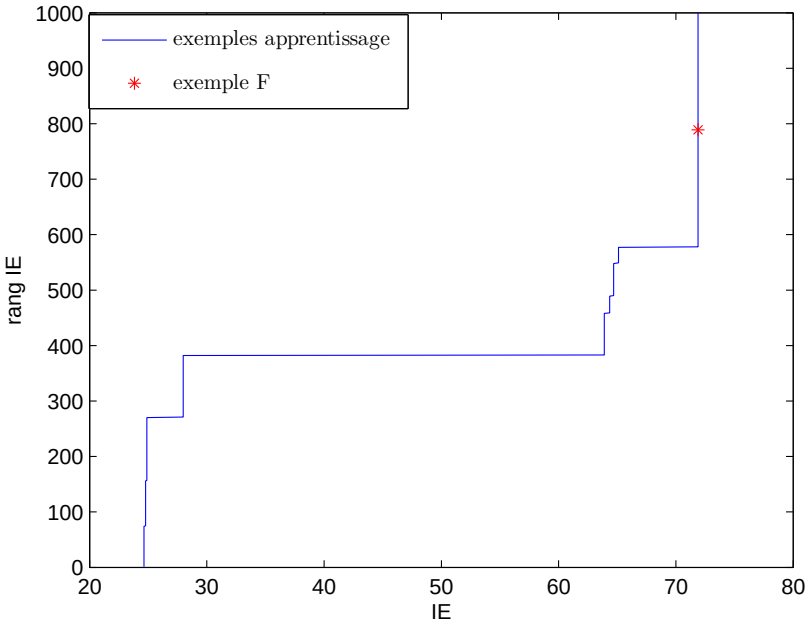

FIG. 3.9 – Distribution ordonnée des IE de la variable Y pour la sortie 1 du naïf bayes

V_i, I_n	IE	UE	ICE	UCE	Variables influentes
$V_i = X, I_n = I_A$	42.5564	0.0069	820	1	X et Y influencent la classe
$V_i = Y, I_n = I_A$	13.5070	0.0069	582	1	X et Y influencent la classe
$V_i = X, I_n = I_B$	24.9138	0.0304	315	497	seul X influence la classe
$V_i = Y, I_n = I_B$	9.9850	0.0304	374	1000	seul X influence la classe
$V_i = X, I_n = I_C$	42.3104	0.0071	820	1	X et Y influencent la classe
$V_i = Y, I_n = I_C$	13.2610	0.0071	504	142	X et Y influencent la classe
$V_i = X, I_n = I_D$	25.3871	0.0241	401	126	seul Y influence la classe
$V_i = Y, I_n = I_D$	5.8824	0.0241	313	283	seul Y influence la classe
$V_i = X, I_n = I_E$	56.6646	0.1061	852	877	aucune n'influence la classe
$V_i = Y, I_n = I_E$	85.7140	0.1061	852	1000	aucune n'influence la classe
$V_i = X, I_n = I_F$	24.9575	0.0249	344	374	seul Y influence la classe
$V_i = Y, I_n = I_F$	5.5078	0.0249	30	403	seul Y influence la classe

TAB. 3.2 – Tableau des IE, UE, ICE, UCE de X et Y et les variables influentes déterminées empiriquement d'après la figure 3.5 pour les exemples A, B, C, D, E, F et pour la sortie 2 (sortie relative à la classe 0)

(l'ICE) serait compris entre 0 et 241 de "TRES FAIBLE", ou de "FAIBLE" si il est compris entre 241 et 375, de "MOYENNE" s'il est compris entre 375 et 706, de "FORTE" si il est compris entre 706 et 882 et de "TRES FORTE" si il est compris entre 882 et 1000. La figure 3.9 montre que la distribution des IE de la variable Y pour la sortie 1 a aussi quelques modalités importantes. Comme pour la variable X, en observant cette figure, pour faciliter l'interprétation, on pourrait subdiviser de même l'ensemble des rangs en plage de rangs.

La figure 3.10 représente la distribution des UE de l'exemple F pour la sortie 1 en ayant perturbé la variable X. Elle est unimodale (voir 3.10). La variable X n'a donc aucune influence pour cet exemple, sur la sortie 1, car quelle que soit la valeur de la variable X pour cet exemple, le score reste constant. Ce résultat est cohérent puisque d'après le tableau 3.1 et le graphe 3.8, l'IE de X pour l'exemple F (pour la sortie 1) est nulle. D'après le tableau 3.1 et le graphe 3.9, la variable Y est importante pour F puisque le rang de son IE est égal à 789 pour la sortie 1. Son UE est égale à zéro et son UCE est mal classée (son UCE est égale à 121). Ces résultats sont logiques et correspondent à nos attentes puisque le point F a été choisi dans la zone noire du graphe3.5. Dans cette zone, la variable X n'a aucune influence. Seule la variable Y est importante et peut influencer le résultat.

L'IE, l'UE, l'ICE, l'UCE et les graphes 3.8, 3.9, 3.10, 3.11 sont les éléments qui peuvent nous permettre d'améliorer le score. Par exemple, dans notre cas, on peut améliorer le score de la sortie 1 du point F. Au regard de la distribution des UE de F par rapport à la variable Y (voir le graphe 3.11),on peut améliorer l' UE (le score de la sortie 1) de F en agissant sur la variable Y, i.e en donnant une autre valeur à Y. On peut donc améliorer le score de la sortie 1 de F (le score de F est égale à zéro initialement) pour l'amener à 0.09 ou 0.1 par exemple.

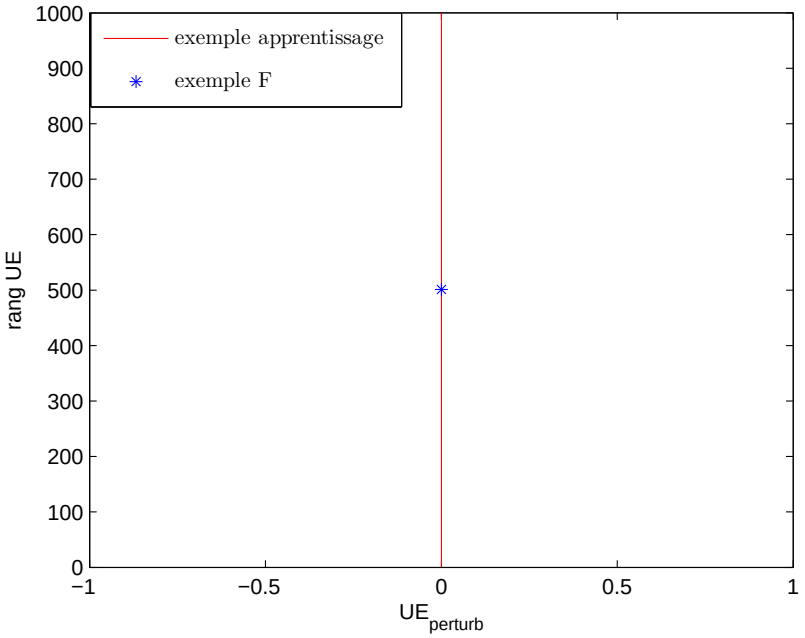

FIG. 3.10 – Distribution ordonnée des UE de la variable X du point F

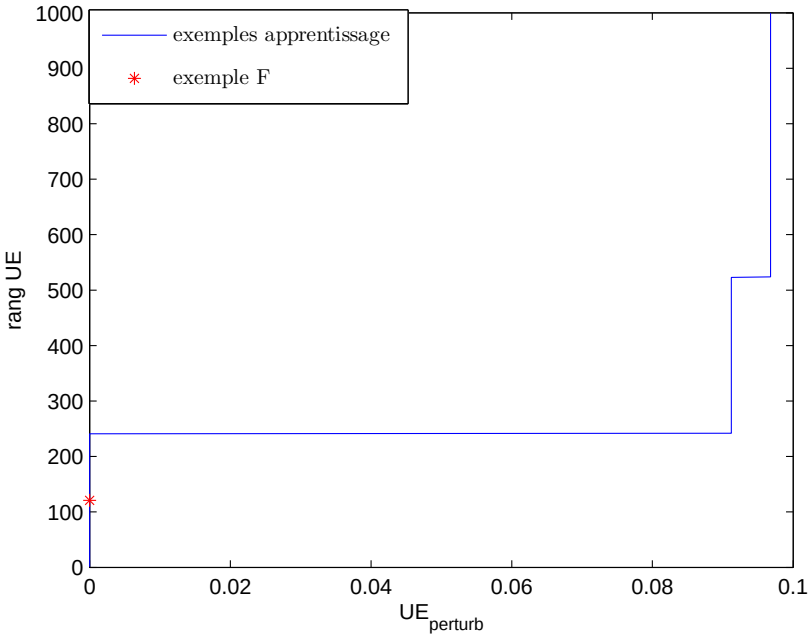

FIG. 3.11 – Distribution ordonnée des UE pour la variable Y du point F

Chapitre 4

Conclusion

Dans ce rapport, nous avons étudié les grandes familles de méthodes de sélection de variables. Ces méthodes sont essentielles pour améliorer les performances des modèles et réduire le nombre de variables à prendre en compte pour la conception de modèles prédictifs. On a besoin, pour sélectionner les variables, de mesurer leurs pertinences pour la tâche à réaliser. La méthode "Robelon", que nous avons étudié dans ce rapport, est une méthode de mesure d'importance des variables. Elle permet d'évaluer la pertinence d'une variable étant donné un modèle quel qu'il soit. Nous avons comparé les performances de la méthode "Robelon" à d'autres méthodes de mesure d'importance de variables, pour un modèle Naïf Bayes : ses résultats sont satisfaisants. Pour un modèle de régression linéaire, on retrouve bien avec la méthode "robelon" que ordonnancer les variables revient à ordonancer les valeurs absolues des coefficients de de la régression linéaire.

Ensuite, nous avons étudié et testé une méthode d'interprétation des scores délivrés par les modèles "boîte noire". Cette méthode, permet de savoir pourquoi le modèle délivre telle score en sortie. Elle indique par exemple, sur quelles variables agir et comment agir sur ces variables pour modifier un score. Elle permet de fournir aux équipes de marketing des outils pour faciliter la personnalisation des offres et répondre plus efficacement aux attentes des clients ou prospects et donc fidéliser la clientèle. Nous avons, dans la deuxième partie de ce rapport, étudié et testé cette méthode pour un modèle Naïf bayes sur un exemple jouet.

Pour étayer le fait que ces méthodes s'appliquent à tous les modèles, il faudrait tester, la méthode de sélection de variable "robelon" et la méthode d'interprétation sur d'autres modèles "boite noire" tels que les SVM par exemple.

Chapitre 5

Annexe

5.1 Pseudo-codes des fonctions utilisées

5.1.1 Fonction d'apprentissage des paramètres du modèle

$$function \quad learning\_fonction$$

- Entrées : tableau des données $mat = \{V_1, \ldots, V_n, classe\}$
- variables globales :
  1. P_posteriori (table des probas à posteriori)
  2. P_priori (table des probas à priori)
  3. realisation (table des valeurs prises par les variables)
  4. classe (table des classes)
  5. position

DEBUT
$N =$ nombre de lignes de la table mat
$n =$ nombre de colonnes de la table mat
$label = mat_{.(n+1)}$ vecteur des étiquettes de chaque exemple
$mat_{.(n+1)} = \emptyset$ on supprime les étiquettes de la matrices des données

on repère les variables constantes
pour $i = 1, \ldots, n$

$$\begin{aligned} var\_const &= \{V_i;\ \sigma(V_i) = 0\} \\ var\_non\_const &= \{V_i;\ \sigma(V_i) \neq 0\} \end{aligned}$$

fin *pour* i
on supprime les variables constantes de la table des données

$$mat = mat[var\_non\_const] = \{V_i;\ \sigma(V_i) \neq 0\}$$

$E = \emptyset$

$classe = \{C_i, i = 1, \ldots, C\}$, la table des classe

pour $i = 1, \ldots, \#(classe)$

$$\begin{aligned} Q1 &= \emptyset \\ Q2 &= \emptyset \\ Q3 &= \emptyset \end{aligned}$$

on repère les exemples de classe i

$$\begin{aligned} I_i &= \{C_l \in label / C_l = i\} \\ C &= \{V_{ij}, i \in I_i, 1 \leq j \leq n\} \end{aligned}$$

on calcule la table des probas à priori

$$P\_priori = \#(C)/N$$

$G_i = \emptyset$

pour $j = 1, \ldots, n$

on sélectionne la variable j

$$D_j = \{V_{ij}, i \in I_i, 1 \leq j \leq n\} = V_j$$

On trouve toutes les valeurs prises par la variable j

$$D1_j = \{v \in C_j\} \, avec \, C_j = \{V_{ij}, \ i \in I_i, \ 1 \leq j \leq n\}$$

on trouve toutes les valeurs prises par la variables j

$$E_j = \{v \in D1_j\}$$

$$F_j = \emptyset$$

pour $k = 1, \ldots, \#(E_j)$

on calcule les occurences de chaque valeur de E_j

$$J1_k = \{i \ / \ V_{ij} = v_k\}$$

on calcule la frequence de chaque realisation de la variable j

$$P_k = \#(J1_k)/\#(D1_j)$$

on crée la table F_j des probas d'apparition de chaque valeur de la variable j

$$F_j = \{P_k, \ k = 1, \ldots, \#(E_j)\}$$

fin *pour* k

on crée la table $Q1_i$ de toutes les fréquences prises par toutes les variables

$$Q1_i = \{F_j,\ j = 1, \ldots, n\}$$

on calcule la table $Q2_i$ de toutes les valeurs prises par toutes les variables aléatoires

$$Q2_i = \{E_j,\ j = 1, \ldots, n\}$$

on calcule la table $Q3$ qui contient le nombre de valeurs prises par chaque variable j

$$Q3 = \{\#(E_j), æ = 1, \ldots, n\}$$

fin pour j

on calcule la table des probas à postériori de toutes les valeurs prises par chaque variable

$$P\_posteriori = \{Q1_i,\ i = 1, \ldots, C\}$$

on calcule la table de toutes les valeurs prises par chaque variable

$$realisation = \{Q2_i,\ i = 1, \ldots, C\}$$

on calcule le cardinal de $Q1_i$

$$a = \#(Q1_i)$$

on calcule le cardinal de $Q2_i$

$$a1 = \#(Q2_i)$$

on calcule le cardinal de $Q3$

$$a = \#(Q3)$$

fin pour i

on réorganise la table $P\_posteriori$ en une table à C colonnes et a lignes
on réorganise la table $realisation$ en une table à C colonnes et $a1$ lignes
on calcule la table $position$ des index qui permettent de reconstituer les tables de chaque variable
on sauvegarde les tables $var\_non\_const\ et\ var\_const$
FIN

5.1.2 Fonction naïf bayes

$$function \quad naif\_bayes3$$

- Entrées :
  1. X une table $1 * N$ ou $N * 1$;
  2. num_var : une table $1 * N$ ou $N * 1$ qui représente les index des variables considérées
- Sortie : table c $1 * 3$ ou $3 * 1$
- variables globales :
  1. P_posteriori (table des probas à posteriori)
  2. P_priori (table des probas à priori)
  3. realisation (table des valeurs prises par les variables)
  4. classe (table des classes)
  5. position

DEBUT

$$\begin{aligned} b &= \emptyset \\ n &= \#(X) \\ C &= nbre\ de\ classe \end{aligned}$$

pour $i = 1, \ldots, C$

$$\begin{aligned} E_i &= realisation_{.i} \\ F_i &= P\_posteriori_{.i} \\ G_i &= \emptyset \end{aligned}$$

pour $k = 1, \ldots, n$

$$\begin{aligned} j &= num\_var(k) \\ E1_k &= \emptyset \\ F1_k &= \emptyset \\ F1_k &= F_i[position(j)\ \ position(j+1) - 1] \\ E1_k &= E_i[position(j)\ \ position(j+1) - 1] \\ J2_k &= \{y \in E1_k\ / y = X(k)\} \\ Pro_k &= F1_k[J2_k] \\ G_i &= \{Pro_k,\ k = 1, \ldots, n\} \end{aligned}$$

fin pour k

$$\begin{aligned} b1_i &= P\_priori[i] \prod_{k=1}^{n} Pro_k^i \\ b &= \{b1_i,\ i = 1 \ldots C\} \end{aligned}$$

fin pour i

$$\begin{aligned} C &= \max_i \{b1_i,\ i = 1 \ldots C\} \\ I &= i\ tel\ que C = \max_i \{b1_i,\ i = 1 \ldots C\} \end{aligned}$$

$si \quad c = 0$

$$\begin{aligned} a &= \max_i \{P\_priori(i),\ i = 1 \ldots C\} \\ j &= i\ tel\ que a = \max_i \{P_priori(i),\ i = 1 \ldots C\} \\ CLASSE &= j \end{aligned}$$

$sinon$

$$CLASSE = I$$

fin pour si

$$c = [CLASSE, b1_i,\ i = 1, \ldots, C]$$

5.1.3 Fonction robelon

$$function \quad robelon$$

- Entrées :
  1. mat la table des données ;
  2. f le modèle ;
  3. var_const les indices des variables constantes
  4. var_non_const les indices des variables non constantes
- Sorties :

1. $s1$ le vecteur des mesures d'importance triées par ordre décroissant ;
2. I_x le vecteur des indices triés par ordre décroissant ;

DEBUT

$$\begin{aligned} I_1 &= var\_const \\ I_2 &= var\_non\_const \end{aligned}$$

on supprime les variables constantes

$$\{mat_{.i} = \emptyset,\ i \in var\_const\}$$

on supprime les labels
N =le nombre d'exemples (lignes) de la table mat
n =le nombre de variables (colonnes) de la table mat

$$\begin{aligned} S &= \emptyset \\ d &= \emptyset \end{aligned}$$

pour $j = 1, \ldots, n$
on sélectionne la variable j

$$V_j = mat_{.j}$$

on trouve l'ensemble des valeurs prises par la variable j

$$valeur\_V_j = \{x \in V_j\}$$

on calcule le nombre d'éléments de l'ensemble des valeurs de la variable j

$$\begin{aligned} P = \#\{valeur\_V_j\} \\ B = \emptyset \end{aligned}$$

pour $i = 1, \ldots, N$
on selectionne l'exemple i

$$I_i = mat_{i.}$$

on copie I_i dans J_i
on evalue la sortie du du modèle pour l'exemple i

$$\begin{aligned} a_i &= f(I_i, [1, 2, \ldots, n]) \\ B1 &= \emptyset \end{aligned}$$

pour $p = 1, \ldots, P$
on calcule la frequence de chaque valeur prise par la variable j

$$\begin{aligned} v_p &= valeur\_V_j(p) \\ nb\_occurence\_v &= \#\{x \in V_j \ / x = v_p\} \\ frequence\_v &= nb\_occurence\_v/N \end{aligned}$$

on perturbe l'exemple i

$$J_i(j) = v_p$$

$$\begin{aligned} b_p &= f(J_i(j), [1, 2, \ldots, n]) \\ A_p &= [(a_1 - b_1)^2 + (a_2 - b_2)^2] frequence\_v(p) \\ B1 &= \{A_p, \ p = 1, \ldots, P\} \end{aligned}$$

fin pour p

$$B_2^i = \sum_{p=1}^{P} A_p$$

fin pour i

$$B_j = \{B_2^i, \ i = 1, \ldots, N\}$$

on calcule l'importance de la variable j

$$C_j = \frac{1}{N} \sum_{j=1}^{n} B_j$$

on calcule l'ensemble des mesures d'importance des n variables

$$s = \{c_j, \ j = 1, \ldots, n\}$$

fin pour j
on range l'ensemble s du plus grand élément au plus petit
$s1 = s$ trié dans l'ordre décroissant
$I_x =$ ensemble des indices des variables triées
FIN

5.1.4 Fonction forward_driven

$$function \quad forward\_driven$$

– Entrées :

1. val la table des données de validation ;
2. f le modèle ;
3. var_const les indices des variables constantes
4. var_non_const les indices des variables non constantes
5. I_x la table des indices des variables rangées par ordre d'importance

DEBUT

n le nombre de variables (colonnes de val)
N le nombre d'exemples (lignes de val)
on commence par calculer des etiquettes des exemples (chaque elements i de cette table représente la classe de l'élément i)

$$\begin{aligned} classe\_vrai &= \{val_{i(n+1)}, \quad 1 \leq i \leq N\} \\ num_var &= \emptyset \\ I1 &= var\_const \\ I2 &= var\_non\_const \end{aligned}$$

on supprime les variables constantes

$$\begin{aligned} \{val_{ij}, \quad 1 \leq i \leq N, j \in I1\} &= \emptyset \\ PBC &= \emptyset \\ BER &= \emptyset \\ H_i &= \emptyset \end{aligned}$$

$n =$ le nombre de variables constantes

pour $i = 1, \ldots, n$

$$\begin{aligned} j_i &= I_x(i) \\ num\_var_i &= \{j_i\} \\ H_i &= \{val_{ik}, \quad 1 \leq i \leq N, \quad k \in num\_var_i\} \end{aligned}$$

on calcule le pbc_i et le ber_i en considérant les variables $j \in num\_var$

$$\begin{aligned} [pbc_i, ber_i] &= classification(H_i, f, num\_var_i, classe\_vrai) \\ PBC &= \{pbc_i, \quad i = 1, \ldots, n\} \\ BER &= \{ber_i, \quad i = 1, \ldots, n\} \end{aligned}$$

fin pour i

FIN

Bibliographie

[1] M. Boulle and A.M. Faure. A statistical discretization method of continuous attribute. *Machine learning*, pages 53–69, Avril 2004.

[2] Leo Breiman. Random forest. *Machine Learning*, 45, 2001.

[3] P. Dangauthier. Sélection automatique de variables pertinentes. Master's thesis, Institut National Polytechnique de Grenoble, Juin 2003.

[4] Raphael Féraud and Fabrice Clérot. A methodology to explain neural network classification. *Neural Networks*, 15 :237–246, 2002.

[5] K Främling. Explaining results of neural networks by contextual importance and utility. In *The AISB'96 conference*, 1996.

[6] I. Guyon, A. Eliseeff, G. Dreyfus, W Duch, and J Reunanen. *Feature Extraction Fundamentals*. Journal of machine learning research, 2005.

[7] V. Lemaire and F. Clérot. An input variable importance definition based on empirical data probability and its use in variable selection. In *International Joint conference on Neural Networks (IJCNN)*, 2004.

[8] P. Leray and P. Gallinari. Variable selection. Technical Report ENV4-CT96-0314, University Paris 6, 1998.

[9] Greg Welch and Gary Bishop. SCAAT : Incremental tracking with incomplete information. In *SIGGRAPH*, Los Angeles, August 12-17 2001.