Initiation à la Fouille de Donnée (et à l'Apprentissage Automatique)

Master 2 I2A - Deuxième session - 23 février 2010

Documents autorisés : Une feuille MANUSCRITE - Calculatrice non-programmable

Conseil : vous êtes ici pour gagner le maximum de points dans le temps imparti et le sujet est long. Par conséquent ne perdez pas de temps à recopier l'énoncé ou à justifier des réponses alors que l'énoncé ne demande pas de justification. N'hésitez pas non plus à sauter une question bloquante, les questions les plus faciles ne sont pas forcément les premiéres. Bon courage !

Arbres de décision

Une banque souhaite promouvoir une offre commerciale via les adresses mails de ses clients. Pour cela elle fait appel à vous et à vos connaissances en fouille de donnée pour sélectionner ceux qui sont potentiellement intéressés. Trois attributs descriptifs sont à votre disposition : l'*âge* (trois tranches : $[18; 32]$, $[33; 45]$ et [46 *et plus*]) le *sexe* (H ou F) et le fait d'être *propriétaire* de son logement (O ou N). L'attribut cible prend deux valeurs : O (intéressé) et N (pas intéressé) - Il y a donc 2 classes. Le résultat d'une enquéte préliminaire sur un panel supposé représentatif de clients donne :

Age	Sexe	Propriétaire	Intéressé
20	F	N	N
25	F	N	N
32	H	O	O
34	H	O	O
35	F	O	O
37	H	N	O
41	H	O	N
45	H	O	O
46	F	O	N
52	H	O	O
60	F	O	N

1. En utilisant le gain en information basé sur l'indice de Gini, et faisant clairement apparaître les calculs, dites quel test doit être mis à la racine de l'arbre.
2. Une partie des données récoltées n'étant pas complétes, elles ne vous ont pas été communiquées dans un premier temps. En explicitant clairement votre maniére de gérer les informations manquantes, donnez l'attribut qui est choisi à la racine par l'indice de Gini quand on ajoute au tableau précédent les données suivantes :

Age	Sexe	Propriétaire	Intéressé
?	F	N	N
28	H	?	O
?	F	?	N
36	?	?	N

Détection de fraude

L'apprentissage automatique peut-être utilisé pour détecter les fraudes : l'exercice suivant en est une illustration trés simple.

On dispose d'un dé à 6 faces, parfaitement équilibré. On confie ce dé à des individus en leur demandant de procéder à un certain nombre de lancers et de faire part de leurs résultats. La population est composée de personnes honnêtes (H) qui font exactement ce qu'on leur demande, mais aussi d'un certain nombre de tricheurs (T) qui chaque fois qu'on leur demande de lancer une fois le dé, le lancent deux fois et annoncent le plus grand des nombres obtenus. Ainsi, si l'on demande à un tricheur de lancer 5 fois le dé, il pourra obtenir la suite de résultats 2, 1, 3, 2, 4, 1, 5, 4, 6, 1 et annoncer 2,3,4,5,6.

1. Calculez $p(i|H)$ et $p(i|T)$ pour i allant de 1 à 6.
2. Calculez $p(23456|H)$ et $p(23456|T)$
3. On suppose que la population contient 30% de tricheurs. Que doit-on décider sur l'honnêteté d'un individu qui annonce 23456 si l'on suit
   (a) la régle majoritaire,
   (b) la régle du maximum de vraisemblance,
   (c) la régle de décision de Bayes

Le classifieur naïf de Bayes.

Le tableau ci-dessous récapitule les conditions qui ont accompagné les succés et les échecs d'une équipe de football. Est-il possible de prédire l'issue d'un match en fonction des conditions dans lesquelles il se déroule ?

Match à domicile ?	Balance positive ?	Mauvaises conditions climatiques ?	Match précédent gagné ?	Match gagné
V	V	F	F	V
F	F	V	V	V
V	V	F	F	V
F	V	V	V	F
F	F	V	F	F
V	F	F	V	F
V	F	V	F	F

FIGURE 1 – Jeu de données *FootBall*.

Les conditions d'un match sont modélisées par un élément $\mathbf{x}$ de $X = \{V, F\} \times \{V, F\} \times \{V, F\} \times \{V, F\}$, correspondant aux valeurs des attributs figurant sur la premiére ligne du

tableau. D'aprés la régle de classification de Bayes, il suffit de connaître $P(V|\mathbf{x})$ pour pouvoir classer $\mathbf{x}$ de maniére optimale : $f(\mathbf{x}) = V$ si $P(V|\mathbf{x}) \geq 1/2$ et $f(\mathbf{x}) = F$ sinon.

D'aprés la formule de Bayes, on a :

$$P(V|\mathbf{x}) = \frac{P(\mathbf{x}|V)P(V)}{P(\mathbf{x})} \text{ et } P(F|\mathbf{x}) = \frac{P(\mathbf{x}|F)P(F)}{P(\mathbf{x})}$$

soit encore

$$P(V|\mathbf{x}) \geq 1/2 \text{ ssi } P(\mathbf{x}|V)P(V) \geq P(\mathbf{x}|F)P(F).$$

On peut évaluer $P(V)$ et $P(F)$ en comptant le nombre de matchs gagnés et perdus :

$$\hat{P}(V) = 3/7 \text{ et } \hat{P}(F) = 4/7.$$

L'évaluation de $P(\mathbf{x}|V)$ et de $P(\mathbf{x}|F)$ est plus délicate. La régle *naïve* de Bayes consiste à faire l'hypothése que les attributs décrivant $\mathbf{x}$ sont indépendants conditionnellement à chaque classe : si l'on écrit $\mathbf{x} = (x_1, x_2, x_3, x_4)$, on suppose que

$$P(\mathbf{x}|V) = \prod_{i=1}^{4} P(x_i|V) \text{ et } P(\mathbf{x}|F) = \prod_{i=1}^{4} P(x_i|F).$$

Pour estimer $P(\mathbf{x}|V)$ et $P(\mathbf{x}|F)$, il suffit alors d'estimer $P(x_i = V|V)$ et $P(x_i = V|F)$ pour $i = 1, \ldots, 4$.

1. Réaliser ces estimations.
2. Classer l'élement (V, F, V, F) avec le classifieur naïf de Bayes.

Regression

On dispose de l'échantillon d'apprentissage suivant : $S = \{(0, 0), (1, 2), (2, 3), (3, 3)\}$. On suppose que ce probléme peut être modélisé par une relation afine (une droite donc).

1. Trouver l'équation de la droite de régression estimée qui minimise l'écart moyen quadratique.
2. Sur un graphique, placer les points et dessiner la droite.
3. calculer une estimation de la variance.

Questions de cours

1. Décriver en quelques mots (maximum cinq phrases) un processus de fouille de données.
2. Expliquer, sans forcément donner leur définition formelle, à quoi servent les indices AIC et BIC (2 lignes max.).