Examen d'Initiation à la Fouille de Donnée et à l'Apprentissage Automatique

Master 2 I2A - Option BDA - 30 octobre 2007

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* (deux tranches : $[18; 35]$ et [36 *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). Le résultat d'une enquête préliminaire sur un panel représentatif de clients donne :

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

1. En utilisant l'algorithme CART avec le gain en information basé sur l'indice de Gini, construisez un arbre de décision sur ces données.
2. Comment votre arbre classe-t-il une cliente de 35 ans propriétaire de son logement ?
3. Une partie des données récoltées n'étant pas complètes, elles ne vous ont pas été communiquées dans un premier temps. En explicitant clairement votre manière 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

Détection de fraude

L'apprentissage automatique peut-être utilisé pour détecter les fraudes : l'exercice suivant en est une illustration très 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, 2, 5, 2, 4, 1, 5, 4, 6, 6 et annoncer 2,5,4,5,6.

1. Calculez $p(i|H)$ et $p(i|T)$ pour i allant de 1 à 6.
2. Calculez $p(25456|H)$ et $p(25456|T)$
3. On suppose que la population contient 20% de tricheurs. Que doit-on décider sur l'honnêteté d'un individu qui annonce 25456 si l'on suit
   (a) la règle majoritaire,
   (b) la règle du maximum de vraisemblance,
   (c) la règle de décision de Bayes
4. On souhaite maintenant fonder la décision sur la moyenne des résultats obtenus. Quelle est l'espérance du résultat annoncé par
   (a) une personne honnête ?
   (b) un tricheur ?
5. Combien faut-il faire faire de tirages à un individu pour être sûr, avec une confiance de 95%, de différencier une personne honnête d'un tricheur ? (Question bonus)

Régression linéaire

On considère l'échantillon suivant de $\mathbb{R}^2 \times \mathbb{R}$:
$S = \{((0,0),0),((0,1),0),((1,0),0),((1,1),1)\}$.
Les exemples sont des réalisations identiquement distribuées de la variable aléatoire Z = (X1,X2, Y).

1. Calculez l'erreur quadratique moyenne $R^S_{emp}(\hat{r}_{1,2})$ de la fonction $\hat{r}_{1,2}(X1, X2) = X1/2 + X2/2 - Y/4$.
2. Calculez les fonctions de régression minimisant le critère des moindres carrés correspondant aux ensembles de variables V suivants : $V = \{X1\}$ et $V = \{X2\}$.
3. Quelle est l'erreur quadratique moyenne de chacune de ces fonctions ?
4. La fonction $\hat{r}_{1,2}$ est la fonction minimisant le critère des moindres carrés pour $V = \{X1, X2\}$. A partir de quelle de valeur de α une pénalisation basée sur la formule $R^S_{emp}(\hat{r}) + \alpha Card(V)/Card(S)$ sélectionne-t-elle une autre fonction que $\hat{r}_{1,2}$?