Initiation à la fouille de données - TD 2 - 09/2011

Le classifieur naïf de Bayes.

Le tableau ci-dessous récapitule les conditions qui ont accompagné les succès 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	V	F	V
V	V	F	V	V
F	V	V	V	F
F	F	V	F	F
V	F	F	V	F

FIGURE 1 – Jeu de données *FootBall*.

Les conditions d'un match sont modélisées par un élément $\mathbf{x}$ de $X = \{V, F\}^4$, correspondant aux valeurs des attributs figurant sur la première ligne du tableau. D'après la règle de classification de Bayes, il suffit de connaître $P(V|\mathbf{x})$ pour pouvoir classer $\mathbf{x}$ de manière optimale : $f(\mathbf{x}) = V$ si $P(V|\mathbf{x}) \geq 1/2$ et $f(\mathbf{x}) = F$ sinon.

D'après 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 :

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

L'évaluation de $P(\mathbf{x}|V)$ et de $P(\mathbf{x}|F)$ est plus délicate. La règle *naïve* de Bayes consiste à faire l'hypothèse 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).

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 30% 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

Régression

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

1. Trouver l'équation de la droite de régression estimée qui minimise l'écart quadratique moyen.
2. Sur un graphique, placer les points et dessiner la droite.
3. Calculer le risque empirique.