Université Pierre et Marie Curie
Examen Reconnaissance des formes et introduction à la décision (RFIDEC)
Master 1

18-12-2008
Durée 2 heures - Documents autorisés

Exercice 1

1. Dans le cours, nous avons traité le problème de la classification en présentant 2 familles d'approches, la classification Bayesienne et la discrimination linéaire.
   Donner succinctement (en ½ page) une description de chacune de ces approches.
2. *Classifieur Gaussien*
   2.1. Rappeler les principes du classifieur gaussien.
   2.2. Dans quel cas est ce que le classifieur gaussien devient il un classifieur linéaire ?
   2.3. Quel lien cela illustre-t-il entre les approches Bayesienne et la discrimination linéaire ?

Exercice 2

Nous allons examiner un ensemble de modèles de discrimination simples couramment utilisés pour effectuer de la classification dans les textes.
Pour affecter un document noté x à une classe C^*, on utilisera la règle de décision Bayesienne :

$$C^* = \arg\max_C P(C|\ x) = \arg\max_C \frac{p(x|\ C)P(C)}{p(x)} \qquad (1)$$

Pour représenter un document, on utilise un ensemble de termes correspondant à un index ou dictionnaire de taille finie n. Un document sera représenté par un vecteur noté $x_1^n = (x_1,...,x_n)$, où x_i et n sont respectivement le i^{ème} élément et la taille du vecteur. La i^{ème} coordonnée x_i, est ainsi associée au i^{ème} terme d'index.
Nous allons étudier deux représentations des documents et les classifieurs qui leurs sont associés.

1. Représentation vectorielle binaire

Un document est représenté sous la forme d'un vecteur x, dont la i^{ème} coordonnée indique la présence ou l'absence du i^{ème} termes d'index, i.e.

$$x_i = \begin{cases} 1 & \text{si le terme i est présent dans le texte} \\ 0 & \text{sinon} \end{cases} \qquad \text{pour i} = 1, ..., \text{n}$$

Les x_i sont supposés indépendants étant donnée la classe du document.

1.1 On note $p_{ik} = p(x_i = 1|\ C_k)$ la probabilité d'apparition du terme i dans un document de la classe C_k montrer $p(x_i|\ C_k) = (p_{ik})^{x_i}(1-p_{ik})^{(1-x_i)}$

1.2 Montrer en utilisant l'hypothèse d'indépendance des x_i :

$$\log(P(C_k|x) = \log(P(C_k)) + \sum_{i=1}^{n} x_i \log\left(\frac{p_{ik}}{1-p_{ik}}\right) + \sum_{i=1}^{n} \log(1-p_{ik}) - \log(p(x))$$

1.3 De nombreux problèmes de classification se ramènent au cas à deux classes (e.g. document pertinent ou pas pour une requête).

1.3.1 Montrer que dans ce cas :

$$\log\frac{P(C_1/x)}{P(C_2/x)} = \sum_{i=1}^{n} x_i \frac{p_{i1}(1-p_{i2})}{p_{i2}(1-p_{i1})} + \sum_{j=1}^{n} \log\frac{(1-p_{i1})}{(1-p_{i2})} + \log\frac{P(C_1)}{P(C_2)} \qquad (2)$$

1.3.2 On veut affecter x à une des deux classes selon la règle de décision Bayesienne. Montrer que l'on peut utiliser le rapport $\log\frac{P(C_1/x)}{P(C_2/x)}$ pour cette affectation et donner la règle de décision correspondante.

2. *Représentation vectorielle fréquentielle*.

Un document sera représenté par un vecteur de taille *n* qui donne pour chaque terme de l'index sa fréquence d'apparition dans le document, x_i sera donc le nombre d'occurrences du terme d'index *i* dans le document. La probabilité d'apparition du terme *i* dans la classe C_k est notée p_{ik}. On notera *q* la taille en nombre de mots du document *x*.

On considère la classe C_k. On suppose que les documents de cette classe sont générés par un tirage avec remise parmi les *n* mots de l'index. On commence par examiner la probabilité d'apparition du terme d'index *i* dans un document *x* comportant *q* mots. On note *y* l'évènement « apparition du terme d'index *i* lors d'un tirage», i.e. *y* = 1 si le terme *i* est tiré et *0* sinon. Soit $(y_1,\ldots, y_q)$ la séquence correspondant aux *q* tirages pour la génération du document x. y_j vaudra 1 si le j^{eme} terme du document est le terme d'index *i*. On remarque que les y_j sont indépendants.

2.1 Montrer

$p(y_1,...,y_q) = p_{ik}^{q_i}(1-p_{ik})^{q-q_i}$ où $q_i = \sum_{j=1}^{q} y_j$ est le nombre d'apparitions du terme i dans le document.[1]

2.2 Combien peut il y avoir de séquences $(y_1\ldots y_q)$ vérifiant $\sum_{j=1}^{q} y_j = q_i$?

2.3 Montrer alors $p(x_i = q_i / C_k) = C_q^{q_i} p_{ik}^{q_i}(1-p_{ik})^{q-q_i}$.[2]

2.4 On considère le cas simple d'un index comprenant uniquement 2 mots, i.e. *n* = 2. Le document *x* sera une séquence de *q* mots, chacun pouvant prendre l'une de ces deux valeurs.

2.4.1 Montrer $p(x_1, x_2 \mid C_k) = p(x_1 \mid C_k)\, p(x_2 \mid x_1, C_k)$

2.4.2 Montrer $p(x_2 = q_2 \mid x_1 = q_1, C_k) = 1$.

2.4.3 En déduire que $p(x = (q_1, q_2) \mid C_k) = \frac{q!}{q_1! q_2!} p_{1k}^{q_1} p_{2k}^{q_2}$ (3)

2.5 Comme en 1., on considère le cas à deux classes. On admettra que dans le cas d'un index de n mots,

l'expression (3) se généralise à $p(x = (q_1,\ldots,q_n) \mid C_k) = \frac{q!}{q_1!\ldots q_n!}\prod_{i=1}^{n} p_{ik}^{q_i}$

Donner l'expression de $\log\frac{P(C_1/x)}{P(C_2/x)}$ et en déduire une règle de décision sous la forme d'une fonction discriminante linéaire.

[1] On a alors $q = q_1 + \ldots + q_n$

[2] Cette loi est la loi binomiale $B(q, p_{ik})$