Examen – Apprentissage artificiel --- 31IM66

2heures. **Documents non autorisés.**

Prenez soin de lire tous les exercices avant de commencer. La notation est donnée à titre indicatif.

Il y a 2 parties : **Partie 1** (15 pts) et **Partie 2** (5 pts). Chaque partie fera l'objet de **copies séparées**.

PARTIE 1 ---- 1h30mn --- 15 pts

Exercice 1.1 : (7 pts)

1. En apprentissage supervisé, quelle est la différence entre un ensemble de test et un ensemble de validation ? (1 pt)
2. Rappeler les principes des méthodes d'apprentissage : Arbre de décision, Réseaux de neurones, k-PPV et Bayésien Naïf. Illustrer avec l'exemple de l'exercice suivant. (4 pts)
3. Quelles sont leurs ressemblances et leurs différences ? (2 pts)

Exercice 1.2 : (8 pts)

RID	age	income	student	credit	C_i: buy
1	youth	high	no	fair	C_2: no
2	youth	high	no	excellent	C_2: no
3	middle-aged	high	no	fair	C_1: yes
4	senior	medium	no	fair	C_1: yes
5	senior	low	yes	fair	C_1: yes
6	senior	low	yes	excellent	C_2: no
7	middle-aged	low	yes	excellent	C_1: yes
8	youth	medium	no	fair	C_2: no
9	youth	low	yes	fair	C_1: yes
10	senior	medium	yes	fair	C_1: yes
11	youth	medium	yes	excellent	C_1: yes
12	middle-aged	medium	no	excellent	C_1: yes
13	middle-aged	high	yes	fair	C_1: yes
14	senior	medium	no	excellent	C_2: no

On dispose du fichier ci-dessus possédant une variable de classe BUY. On découpe l'ensemble en 2 : D1 et D2. **D1 contient les 6 premiers objets**, et D2 contient les **8 derniers (7 à 14)**.

1- Quel attribut n'est pas pertinent pour construire un modèle de classification ? (0,5 pt)
2- Construire le modèle du plus proche voisin, 3-PPV, en utilisant D1 comme ensemble d'apprentissage. (0,5 pt)
3- L'ensemble D2 va être utilisé pour tester le modèle 3-PPV. Déterminer la classe des 8 objets de D2. (4 pts)
4- Donner la matrice de confusion sur D2; (1 pt)
5- Calculer le taux d'erreur apparente de la méthode avec D2; (0,5 pt)
6- Calculer la précision de la classe 'yes' sur D2; (0,5 pt)
7- Calculer le rappel pour la classe 'no' sur D2; (0,5 pt)
8- On souhaite calculer la précision totale du modèle 3-PPV en tenant compte du poids de chacune des classes dans D2 . (0,5 pt)

PARTIE 2 ---- 30mn --- 5 pts

Licence 3 Informatique
Examen Apprentissage

Olivier Raynaud
raynaud@isima.fr

La durée de cette épreuve est de 30 minutes. Les documents ne sont pas autorisés. Il sera tenu compte de la propreté de la copie lors de l'évaluation.

Exercice 1 (Machines à registres)**.**

Question 1. *Concevoir une machine à registres qui reçoit dans les registres R_1 et R_2 deux booléens a et b et retourne dans R_1 la valeur booléenne de l'expression 'a ou b'.*

Question 2. *Montrer que le prédicat $x =?y$ est calculable par une machine à registres.*

Question 3. *Concevoir une machine à registres qui calcule la partie entière inférieure de $log_2(n)$ si $n \neq 0$ et retourne 0 si $n = 0$.*

Annexes :

Naive Bayes : Estimation des probabilités conditionnelles

Ai : une valeur de l'attribut A

Nic : Nombre d'objets ayant la valeur Ai dans la classe c

Nc : Nombre d'objets de la classe c

k : nombre de valeurs de l'attribut A

p : probabilité apriori

m : paramètre

$$\text{Original: } P(A_i \mid C) = \frac{N_{ic}}{N_c}$$

$$\text{Laplace: } P(A_i \mid C) = \frac{N_{ic}+1}{N_c+k}$$

$$\text{m-estimate: } P(A_i \mid C) = \frac{N_{ic}+mp}{N_c+m}$$

Given a training dataset $\mathcal{D}$ of N labeled examples (assuming complete data)

1. Estimate $P(c_j)$ for each class c_j

$$\hat{P}(c_j) = \frac{N_j}{N}$$

N_j - the number of examples of the class c_j

2. Estimate $P(X_i=x_k \mid c_j)$ for each value x_k of the attribute X_i and for each class c_j

- X_i discrete

$$\hat{P}(X_i = x_k \mid c_j) = \frac{N_{ijk}}{N_j}$$

N_{ijk} - number of examples of the class c_j having the value x_k for the attribute X_i

- X_i continuous

Two options:

- The attribute is discretized and then treats as a discrete attribute
- A Normal distribution is usually assumed

$$P(X_i = x_k \mid c_j) = g(x_k; \mu_{ij}, \sigma_{ij}) \quad \text{onde} \quad g(x;\mu,\sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{(x-\mu)^2}{2\sigma^2}}$$

The mean μ_{ij} e the standard deviation σ_{ij} are estimated from $\mathcal{D}$

---

2. Estimate $P(X_i=x_k \mid c_j)$ for a value of the attribute X_i and for each class c_j

- A Normal distribution is usually assumed

$$P(X_i = x_k \mid c_j) = g(x_k; \mu_{ij}, \sigma_{ij}) \Rightarrow g(x;\mu,\sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{(x-\mu)^2}{2\sigma^2}}$$

$X_i \mid c_j \sim N(\mu_{ij}, \sigma^2_{ij})$ - the mean μ_{ij} e the standard deviation σ_{ij} are estimated from $\mathcal{D}$

For a variable $X \sim N(74, 36)$, the probability of observing the value 66 is given by:

$$f(x) = g(66; 74, 6) = 0.0273$$

---

k-PPV: Proximité (Similarité, Dissimilarité), Distances

Attribute Type	Dissimilarity	Similarity
Nominal	$d = \begin{cases} 0 & \text{if } p = q \\ 1 & \text{if } p \neq q \end{cases}$	$s = \begin{cases} 1 & \text{if } p = q \\ 0 & \text{if } p \neq q \end{cases}$
Ordinal	$d = \frac{\lvert p-q \rvert}{n-1}$ (values mapped to integers 0 to $n-1$, where n is the number of values)	$s = 1 - \frac{\lvert p-q \rvert}{n-1}$
Interval or Ratio	$d = \lvert p - q \rvert$	$s = -d$, $s = \frac{1}{1+d}$ or $s = 1 - \frac{d - min\_d}{max\_d - min\_d}$

Distance de Minkowski :

$$dist = \left(\sum_{k=1}^{n} | p_k - q_k |^r \right)^{\frac{1}{r}}$$

où r est un paramètre, n est le nombre de dimensions (attributs) et p_k et q_k sont, respectivement, les $k^{èmes}$ attributs (composants) des objets p et q.

r = 1 : City block (Manhattan, taxicab, L_1 norm) distance. Aussi appelée distance de Hamming pour des vecteurs binaires.

r = 2 : distance euclidienne

- Common situation is that objects, *p* and *q*, have only binary attributes
- Compute similarities using the following quantities
  M_{01} = the number of attributes where p was 0 and q was 1
  M_{10} = the number of attributes where p was 1 and q was 0
  M_{00} = the number of attributes where p was 0 and q was 0
  M_{11} = the number of attributes where p was 1 and q was 1
- Simple Matching and Jaccard Coefficients
  SMC = number of matches / number of attributes
  = $(M_{11} + M_{00}) / (M_{01} + M_{10} + M_{11} + M_{00})$

  J = number of 11 matches / number of not-both-zero attributes values
  = $(M_{11}) / (M_{01} + M_{10} + M_{11})$