Module Fouille de données

2h. Seule une feuille A4 **manuscrite** est autorisée.

NB : La notation tient compte de la présentation. Elle est donnée à titre indicatif.

Exercice 1 : (8 pts)

Soient O un ensemble fini d'exemples ou de transactions ; A un ensemble fini d'attributs ou d'items, X et Y sont des sous-ensembles de A.

- Un ensemble d'items X est dit fréquent si le nombre de transactions contenant les items de X est supérieur au seuil de support minimum (minsup).
- Une règle associative (ou d'association) est une implication de la forme « si X alors Y » (notée X → Y, X est la prémisse, et Y est la conclusion) et l'intersection de X et Y est vide, traduisant le fait que si les items X sont présents dans une transaction, alors les items Y le sont avec une certaine probabilité (confiance).
- Le support d'une règle est la mesure indiquant le pourcentage de transactions qui vérifient une règle associative.
- La confiance d'une règle est la mesure indiquant le pourcentage de transactions qui vérifient la conclusion d'une règle associative parmi celles qui vérifient la prémisse.

Soit l'ensemble des transactions : {P, L}, {P, C, B, O}, {L, C, B, K}, {P, L, C, B}, {P, L, C, K}.

Questions :

1- Calculer les itemsets candidats et fréquents pour un support minimum égal à 2. (2,5 pts)
2- Donnez :
   a. la liste des itemsets fréquents fermés (1 pt)
   b. la liste des générateurs minimaux fréquents (1,5 pt)
3- Etant donné un seuil de confiance minimum égal à 60%, lister **au moins 2** règles associatives valides, à partir d'un 3-itemset. (0,5 pt)
4- A partir de la bordure négative (BN) :
   a. Donner un pseudo-code qui permet de lister tous les itemsets fréquents. (1,5 pt)
   b. Peut-on dériver le support de ces itemsets fréquents ? Justifiez. (0,5 pt)
   c. Que peut-on en déduire entre la BN et la liste des itemsets fréquents ? (0,5 pt)

Exercice 2 : (12 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 3 : D_1, D_2 et D_3. **D_1 contient les 6 premiers objets**, **D_2** contient les **4 suivants (7 à 10)**, et **D_3** contient les **4 derniers (11 à 14)**.

La *précision* pour une classe donnée mesure le taux d'exemples corrects parmi les exemples prédits dans cette classe. Le *rappel* mesure le taux d'exemples corrects parmi les exemples de la classe. Le taux de *faux positifs* d'une classe mesure le nombre d'objets positifs parmi ceux n'appartenant pas à la classe. Le taux de *vrais positifs* d'une classe mesure le nombre d'objets positifs parmi les vrais objets de la classe.

Pour les questions 3) à 7), vous devez **indiquer la formule de calcul**.

Questions :

1- Quelle est la différence entre un ensemble de test et un ensemble de validation ? 1 pt
2- L'ensemble D_2 va être utilisé pour tester la méthode des k-plus proches voisins : k-PPV. Déterminer la classe des 4 objets de D_2. 4 pts
3- Donner la matrice de confusion sur D_2; 0,5 pt
4- Calculer le taux d'erreur apparente de la méthode avec D_2; 0,5 pt
5- Calculer le taux de faux positifs (FP rate) pour la classe C_1 ; 0,5 pt
6- Calculer le taux de vrais positifs (TP rate) pour la classe C_2 ; 0,5 pt
7- Calculer la précision de la classe C_1 sur D_2; 0,5 pt
8- Calculer le rappel pour la classe C_2 sur D_2; 0,5 pt
9- On souhaite calculer la précision totale du modèle k-PPV en tenant compte du poids de chacune des classes dans D_2 . 0,5 pt
10- Ecrire un algorithme (pseudo-code) qui donne comme résultat le modèle à choisir par l'utilisateur pour la résolution de son problème. En entrée on a 2 matrices de confusion T_1 et T_2, obtenues respectivement sur D_2 et D_3, par application respective des modèles M_1 et M_2, générés à partir de D_1. 1 pt
11- Comment peut t-on s'assurer que ce mode de comparaison de modèles est robuste ? 0,5 pt
12- Citer une autre méthode de classification supervisée autre que la méthode k-PPV. Rappeler son principe. 1 pt
13- Comparer ces deux méthodes. 1 pt

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
  = $(M_{11}) / (M_{01} + M_{10} + M_{11})$