Examen Final

Data Mining et Recherche d'Information

Exercice 1 (12 points)

On considère des classifieurs binaires, c'est à dire qui classent un élément donné parmi les *positifs* ou les *négatifs*.

Pour chaque élément présenté à un tel classifieur, quatre cas peuvent se présenter (Voir Figure 1):

1. L'élément est un *positif* et le classifieur détermine sa classe comme *positif* : c'est un *vrai positif (VP)*.
2. L'élément est un *positif* et le classifieur détermine sa classe comme négat*if* : c'est un *faux négatif* (FN)
3. L'élément est un *négatif* et le classifieur détermine sa classe comme *négatif* : c'est un *vrai négatif (VN)*
4. L'élément est un *négatif* et le classifieur détermine sa classe comme *positif* : c'est un *faux positif (FP)*

	Vraie classe	
Classé	Positif	Négatif
Positif	Vrai Positif (VP)	Faux Positif(FP)
Négatif	Faux Négatif(FN)	Vrai Négatif(VN)
Total colonne	P	N

Fig. 1 Matrice de confusion

Parmi les différentes mesures du comportement d'un classifieur, on peut distinguer :

- La proportion de positifs bien classés par rapport au nombre de positifs (*tp rate* : true positive rate : rapport des vrais positifs) : ***tp rate = VP / P***
- La proportion de négatifs mal classés par rapport au nombre total de négatifs (fp rate : false positive rate : rapport des faux positifs) ***: fp rate = FP / N***

Une façon de juger de la valeur d'un classifieur est de le placer dans un carré de coté 1, indicé en abscisse par *fp rate* et en ordonnée par *tp rate*. Le graphe obtenu est appelé ROC (Receiver Operating Characteristic). La figure 2 représente cinq classifieurs dans cet espace.

Q1 : Sur le graphe, à quelles stratégies de classification correspondent les points (0,0) et (1,1) ?

Q2 : Comment peut-on caractériser en une ou deux phrases le comportement des classifieurs tels que A, i.e. ceux qui se situent vers la gauche du graphe ?

Q3 : Même question pour les classifieurs se situant au dessus de la diagonale, mais à droite du graphe.

Q4 : Que dire des classifieurs se situant sur la diagonale ?

Q5 : Quel est, parmi A,B,C,D,E le meilleur classifieur ?

Q6 : Où se situerait le classifieur idéal ?

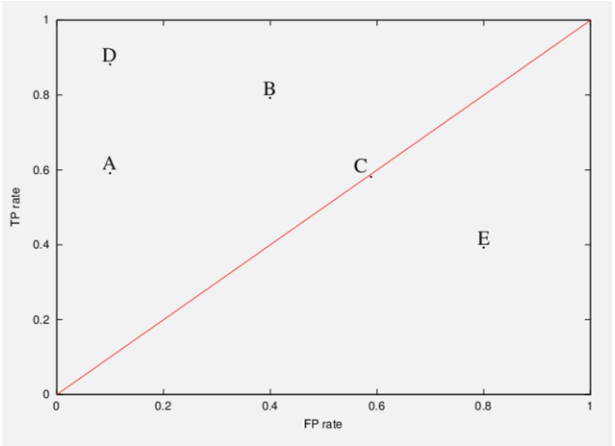


Fig. 2 Un graphe ROC pour des classifieurs discrets

Exercice 2 (08 points)

Le dendrogramme est une représentation graphique d'une classification (clustering) hiérarchique par un arbre.

1- Le dendrogramme d'une classification ascendante (ou descendante) est-il unique ? Si OUI dites Comment ? Si NON dites pourquoi ? **(2 Pts)**

2- Comment déterminer le nombre de classes (clusters) à partir du dendrogramme ? Cette méthode est-elle exacte ou approximative ? Quel est le moyen le plus efficace pour avoir un nombre de classes (clusters) proche de la réalité? **(2 Pts)**

3- Considérer la matrice de similarité suivante de cinq documents d1, d2, d3, d4 et d5. Déterminer le dendrogramme résultant de l'application du text clustering hiérarchique ascendant en utilisant le « **lien maximum** ». **(4 Pts)**

	d1	**d2**	**d3**	**d4**	**d5**
d1	0	0.5	0.5	0.6	0.8
d2	0.5	0	0.7	0.6	0.5
d3	0.5	0.7	0	0.6	0.5
d4	0.6	0.6	0.6	0	0.9
d5	0.8	0.5	0.5	0.9	0

Master 2 Système d'information et Génie Logiciel

Examen Final

Data Mining et Recherche d'Information

Correction

Exo1 : (12 points : 02 points pour chaque réponse)

Q1 : Sur le graphe, à quelles stratégies de classification correspondent les points (0,0) et (1,1) ?

Réponse : Le point (0,0) correspond *à fprate=VP/P=0* càd *VP=0*, le classifieur n'a trouvé aucun *vrai positif*, et *tprate=FP/N=0* càd *FP=0*, le classifieur n'a trouvé aucun *faux positif*. Dans ce cas le classifieur n'arrive pas à déterminer les *positifs* (ni les *vrais* ni les *faux positifs*).

Le point (1,1) correspond à *fprate=VP/P=1* càd *VP=P* et *FN=0* (car *VP+FN=P*, voir Fig.1), le classifieur n'a trouvé aucun *faux négatif*, et *tprate=FP/N=1* càd *FP=N* et *VN=0* (car *FP+VN=N*, voir Fig.1), le classifieur n'a trouvé aucun *vrai négatif*. Dans ce cas le classifieur n'arrive pas à déterminer les *négatifs* (ni les *vrais* ni les *faux négatifs*).

Q2 : Comment peut-on caractériser en une ou deux phrases le comportement des classifieurs tels que A, i.e. ceux qui se situent vers la gauche du graphe ?

Réponse : généralement, les classifieurs se situant sur la partie gauche du graphe ont un taux *fprate* plus faible et un taux *tprate* plus fort, c'est la catégorie des bons classifieurs.

Q3 : Même question pour les classifieurs se situant au dessus de la diagonale, mais à droite du graphe.

Réponse : les classifieurs se situant sur la partie droite du graphe ont un taux *fprate* plus fort que *tprate*, c'est la catégorie des mauvais classifieurs.

Q4 : Que dire des classifieurs se situant sur la diagonale ?

Réponse : les classifieurs se situant sur la diagonale ont un taux *tprate=fprate*. Ce sont les classifieurs qui donnent de bons résultats aussi bien que de mauvais résultats.

Q5 : Quel est, parmi A,B,C,D,E le meilleur classifieur ?

Réponse : le meilleur classifieur se situe sur la partie gauche du graphe. C'est celui qui donne le moins de faux positifs et le plus de vrais positifs. Le classifieur D est le meilleur classifieur.

Q6 : Où se situerait le classifieur idéal ?

Réponse : le classifieur idéal doit donner zéro faux positif, càd FP=0 (fprate=0) et déterminer tous les vrais positifs, càd VP=P (tprate=1), d'où le meilleur classifieur se situe sur le point (0,1).

Exo2 : (08 points)

1- Le dendrogramme d'un clustering n'est pas unique ; il dépend de la stratégie de regroupement : lien minimum, maximum ou moyen. En plus, si la distance minimale choisie n'est pas unique, le choix aléatoire diversifie le dendrogramme.

2- On détermine le nombre de clusters à partir d'un dendrogramme, en trouvant le nombre de points d'intersection entre la droite *y=d* et le dendrogramme, *d* étant la distance choisie. Cette méthode est exacte. Le moyen le plus efficace pour avoir un nombre de classes (clusters) proche de la réalité se fait généralement à l'aide de l'avis d'un expert dans le domaine.

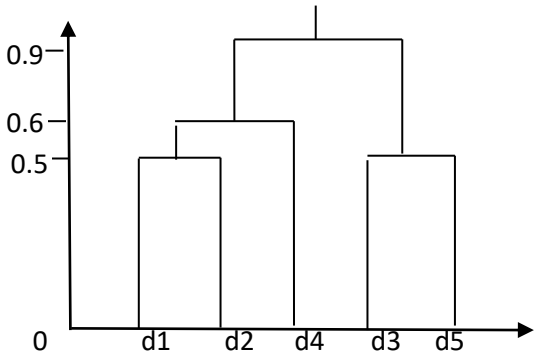
3- Le dendrogramme résultant de l'application du text clustering hiérarchique ascendant de cinq documents d1, d2, d3, d4 et d5, en utilisant le « **lien maximum** ».

	d1	**d2**	**d3**	**d4**	**d5**
d1	0				
d2	**0.5**	0			
d3	0.5	0.7	0		
d4	0.6	0.6	0.6	0	
d5	0.8	0.5	0.5	0.9	0

grouper (d1, d2)

	d1d2	**d3**	**d4**	**d5**
d1d2	0			
d3	0.7	0		
d4	0.6	0.6	0	
d5	0.8	**0.5**	0.9	0

grouper (d3, d5)



Dendrogramme

	d1d2	**d3d5**	**d4**
d1d2	0		
d3d5	0.8	0	
d4	**0.6**	0.9	0

grouper (d1d2, d4)

	d1d2d4	**d3d5**
d1d2d4	0	
d3d5	**0.9**	0

grouper(d1d2d4,d3d5)