

Epreuve de Rattrapage
Data Mining

Exo1/ (10pts)

Soient les instances tirées de la base d'apprentissage HEART-Statlog et pour lesquelles seuls certains attributs sont pris en compte :

Age	Sex	chest	RestBlood	SerumChol	MaxHeart Rate	Thal	Class
70.0	1	4	130.0	322.0	109.0	3.0	2-
67.0	0	3	115.0	564.0	160.0	7.0	1
57.0	1	2	124.0	261.0	141.0	7.0	2-
46.0	1	4	140.0	311.0	120.0	7.0	2-
43.0	1	4	15.0	303.0	181.0	3.0	1-
41.0	0	2	126.0	306.0	163.0	3.0	1-
58.0	1	4	100.0	234.0	156.0	7.0	2-

1- Calculer le mode, la mediane et la moyenne pour chaque attribut

2- Décrire la classification supervisée par les réseaux bayésiens,

3- A-t-on besoin de discrétiser les attributs numériques lorsqu'on utilise la technique Naive Bayes pour la classification supervisée ? si oui faire la discrétisation des attributs numériques de ces données.

4- Utiliser cette technique pour classer les instances suivantes

54.0	0	2	132.0	288.0	159.0	3	?
57.0	1	3	128.0	229.0	150.0	7	?

Exo2/ (10pts)

Nous comptons mettre en œuvre une méthode hybride pour la classification non supervisée des données suivantes :

Age	Sex	chest	RestBlood	SerumChol	MaxHeart Rate	Thal	Class
67.0	0	3	115.0	564.0	160.0	7.0	1
57.0	1	2	124.0	261.0	141.0	7.0	2
46.0	1	4	140.0	311.0	120.0	7.0	2
43.0	1	4	15.0	303.0	181.0	3.0	1
41.0	0	2	126.0	306.0	163.0	3.0	1

Pour cela nous comptons hybrider un algorithme génétique qui aura en entrée les données et qui doit donner en sortie DEUX INDIVIDUS ayant une distance maximale. En second lieu, l'algorithme K-Mean est appliqué pour répartir le reste des individus sur deux groupes dont les centroides sont formés au départ par les deux individus fournis par l'AG.

Il est demandé :

- de donner une modélisation de cette segmentation (ou clustering) par l'AG puis par la K-Mean
- trouver par le calcul de distances les deux individus les plus éloignés de ce nuage de points
- appliquer la K-Mean pour trouver les clusters en prenant comme centres de gravités de départ, les deux individus les plus éloignés

Bon Courage !