|  |  |  |
| --- | --- | --- |
| BdeB_noir |  | **Année : 2015-2016**  **cours : 420 – BD8-BB**  **Enseignants :** Nesrine Zemirli & Hafed Benteftifa  **Groupe : G 208** |

**Devoir**

**Apprentissage Supervisé et Non Supervisé**

**Exercice 1 : Calcul des distances [10 pts]**

**Exercice 2 : Classification basé sur le kNN [6 pts]**

**Exercice 3 : Clustering avec le k-means et HC [10 pts]**

**Exercice 4 : Naïve Bayes [4 pts]**

**Total des points 30 pts**

**Préparé par :** Nesrine Zemirli

© Hafed Benteftifa et Nesrine Zemirli 2015-201

Ce document ne peut être utilisé dans le cadre d’une formation, publication papier, site internet ou tout support sans mon accord express.

Aucune reproduction, même partielle, ne peut être faite de ce document et de l'ensemble de son contenu : textes, images, etc. sans mon autorisation express.

Pour toutes informations, communiquer avec moi sur [infor@degenio.com](mailto:infor@degenio.com)

Contexte de réalisation

1. Format : individuel

**Livrable**

1. Un document contenant la solution détaillée de chacun des exercices et les explications demandées le cas échéant,
2. Le scripte R ou Python des solutions demandées le cas échéant.
3. Mette votre nom et prénom au début de votre document avec une mise en forme adéquate.
4. Formaliser les réponses de manière claire et commenter vos solutions.

**Exercice 1 : Calcul des distances [10 pts]**

**Q 1.1** [**4 pt**]

Pour chacun des vecteurs suivants, calculer la mesure de similarité ou distance selon Jaccard, cosinus, euclidienne et corrélation.

A : [ x = (1 1 0 0 0), y = (0 0 0 1 1) ]

B : [x = (0 1 0 1 1), y = (1 0 1 0 0) ]

C : [x = (0 1 2 4 5 3), y = (5 6 7 9 10 8) ]

**Q 1.2** [**2 pts**]

Donner le code python ou R qui vous permet d'obtenir ces mesures

**Q 1.3** [**4 pts**]

Étant donné un nouveau vecteur D : [x = (0 1 1 0 0), y=(1 1 0 1 1)].

Calculer la similarité du nouveau vecteur D avec les vecteurs A, B et C selon Jaccard, cosinus, euclidienne

Ordonner le résultat par le plus similaire, au moins similaire pour chacune des mesures de similarité utilisées

***Dérouler l’algorithme manuellement sur papier, aucune programmation avec un langage n’est exigée***

**Exercice 2 : Classification basée sur le kNN [6 pts]**

Dans cet exercice, vous allez appliquer l’algorithme de classification du k-plus proche voisin en utilisant la distance euclidienne comme métrique pour cette classification binaire. On considère la classe de test, celle ayant la majorité des k voisins les plus proches. Notez qu'un point peut être son propre voisin.

**Q 2.1 [4 pts]**

Considérant l’ensemble de données suivant représenté dans la figure 1, délimiter les classes sur la base du 1-plus proche voisin pour cet ensemble de données.

Utiliser la figure suivante pour tracer les lignes séparatrices des classes



Figure : Classification k-nn, un plus indique un exemple positif et une étoile un exemple négatif.

**Q 2.2 [2pts]**

Comment le point (8,1) serait classé en utilisant 1-nn ?

***Dérouler l’algorithme manuellement sur papier. Noter qu'aucune programmation avec un langage n’est exigée***

**Exercice 3 : Clustering avec le k-means et HC [10 pts]**

**Q 3.1 [4 pts]**

Effectuer une classification hiérarchique des données 2-D du tableau ci-dessous.

On utilisera la distance euclidienne en tant que mesure de similarité.

Puis, dessinez un arbre dendrogramme avec la hauteur du nœud, la distance à laquelle les grappes ont été fusionnées.

Astuce : Dessinez les premiers points sur 2-D puis effectuer la simulation manuelle.

**Q 3.2 [4 pts]**

Sur la base des même données 2-D du tableau ci-dessous, simuler l’algorithme du K-means. Utiliser comme centres initiaux les points (2,6), (2,8), (5,8).

Indiquer les étapes lors du déroulement de l'algorithme.

Ensuite, utilisez les mêmes données et simuler algorithme des k-means à partir des centres de cluster des points D, E et H.

**Q 3.3 [2 pts]**

Comparer le résultat des deux solutions ci-dessus.

****

Tableau 1 Données pour le clustering

***Dérouler l’algorithme manuellement sur papier. Noter qu'aucune programmation avec un langage n’est exigée***

**Exercice 4 : Naïve Bayes [ 4 pts ]**

Soit le modèle obtenu sur la base de Naïve Bayes et décrit dans le tableau suivant:

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Total | Age | | | | | | Genre | |
| Salaire |  | 1 à 20 | 20 à 30 | 30 à 40 | 40 à 50 | 50 à 60 | 60 à 100 | Femelle | Male |
| < 50 K | .75 | .1 | .3 | .25 | .17 | .1 | .08 | .39 | .61 |
| > 50 K | .25 | .03 | .08 | .3 | .32 | .2 | .07 | .15 | .85 |
| Total |  | .08 | .24 | .26 | .21 | .12 | .08 | .33 | .67 |

Pour chaque valeur du prédicteur, on a la probabilité conditionnelle correspondante de la cible.

**Q 4.1 [8pts]**

Trouver la probabilité que le salaire soit inférieur à 50K sachant que la personne a 25 ans

***Dérouler l’algorithme manuellement sur papier, aucune programmation avec un langage n’est exigée***