**MET8130 – Automne 2021 – Jean-Marie Bourjolly**

**Examen final à la maison**

1. Lecture du fichier ***Examen-Take\_home-data.csv***

Assurez-vous d’avoir lu toutes les données.

2. Description des données du fichier.

3. Représentation graphique des données du fichier. Que peut-on en dire à première vue?

4. Appliquez la méthode k-means aux données, avec k=3. Étudiez la structure du résultat et analysez-le. Que concluez-vous?

5. Le fichier de données est maintenant segmenté. Regroupez la classe la moins volumineuse dans un fichier appelé C1. Regroupez la moins volumineuse des deux classes restantes dans un fichier appelé C2. Regroupez la classe la plus volumineuse dans un fichier appelé C3.

6. Ajoutez une colonne de 1 à C1, une colonne de 2 à C2 et une colonne de 3 à C3. C1, C2 et C3 contiennent maintenant deux variables explicatives numériques et une variable expliquée de type catégoriel : 1, 2 ou 3. Empilez C1 au-dessus de C2 puis de C3 pour créer xy. (Il se peut que vous ayez d’abord à nommer les colonnes de C1, C2 et C3 de la même manière, disons, x1, x2 et y.)

7. Représentation des 3 classes dans un même graphique à deux dimensions (les deux premières colonnes de xy), avec une couleur différente pour chaque classe, une légende permettant de les distinguer, et en indiquant les limites de chacun des deux axes de coordonnées.

8. Créez un *data frame* de 30 x 2, appelé new; la 1ère colonne contient des nombres aléatoires entre 0.5 et 2.5; la 2e contient des nombres aléatoires entre -2.5 et 3. Nommez les colonnes de new de la même façon que les deux premières colonnes de xy (par exemple, x1 et x2). Nommez les lignes de new de 1001 à 1030.

9. Utilisez kNN, avec k=7, pour classer les éléments de new (en utilisant xy). Étudiez la structure du résultat et analysez-le.

10. Portez sur un même graphique C1, C2, C3 et new. Les éléments de new se distinguent de ceux de C1, C2, C3 par leur forme. Qu’ils appartiennent à C1, à C2, à C3 ou à new, les éléments classés différemment se distinguent les uns des autres par leur couleur. Indiquez aussi sur le graphique les centres des 3 classes avec la couleur appropriée et une forme ainsi qu’une grosseur qui les distinguent des autres points.

11. En supposant que les 12 premiers éléments de new soient en fait des « 1 », les 10 suivants, des « 2 » et les 8 derniers des « 3 », comparez la classification réelle à celle qui est prédite pour les « 1 » d’un côté, et les « non 1 » de l’autre (les « 1 » sont la classe dite « positive ») : utilisez des méthodes présentées en classe pour effectuer cette comparaison. Analysez les résultats obtenus.

**Question bonus de 5/100.**

**À faire seulement si vous avez terminé tout le travail qui précède.**

Bâtissez votre propre programme, avec k=3.

1. Choisissez au hasard les deux coordonnées de 3 points, la 1ère coordonnée entre -1 et 4 et la 2e entre -4 et 4. Les 3 points sont nommés P1, P2, P3. Ils sont destinés à être les centres des 3 classes que l’on construit.

2. Définissez une fonction dist() qui calcule la distance (euclidienne) entre deux points dont on connaît les coordonnées.

3. Calculez, pour chacun des points du fichier ***Examen-Take\_home-data.csv***, sa distance à P1, P2, P3 et classez-le, provisoirement, en accord avec la plus petite de ces trois distances.

4. Faites une mise à jour de P1, P2, P3. Les nouvelles coordonnées de P1 sont les moyennes des coordonnées des points temporairement classés « 1 ». Faites-en de même pour P2 et P3. Calculez e, la distance totale entre les 3 nouveaux points P1, P2, P3 et les anciens (ceux qui existaient juste avant la mise à jour).

5. Répétez 2., 3., 4., jusqu’à ce que e soit inférieur à 0.1.