UNIVERSITÉ DE SHERBROOKE
DÉPARTEMENT D'INFORMATIQUE

IFT 870 BIN 710 - Forage de données

TP#1 : Exploration de données

Hiver 2020

Le but de ce devoir est de pratiquer l'exploration de données : visualisation de données, analyse de corrélation entre attributs, réduction de dimension, choix d'une mesure de similarité entre objets.

Ce devoir est à faire en équipe de deux obligatoirement. Il devra être complété avant le vendredi 7 février 2029 à 23h59. Vous devez remettre, sur `opus.dinf.usherbrooke.ca`, un fichier Ipython notebook (nommé `tp1.ipynb`) contenant votre rapport et vos scripts Python pour ce devoir.

Description des tâches à réaliser : On vous fournit un ensemble de données stockées dans un fichier au format `.csv` (`TP1_data.csv`). L'ensemble des données contient 59 observations représentées suivant 4 variables (`attribut1,attribut2,attribut3,attribut4`). Les données sont segmentées en 3 classes (0,1,2). La classe d'une observation est représentée par la valeur de la variable `classe` dans le fichier de données. L'objectif du TP est de déterminer si les 4 variables utilisées pour la représentation ont des propriétés discriminantes pour la classification de nouvelles observations. On souhaite utiliser un modèle de classification basée sur la distance : la méthode des $k = 5$ plus proches voisins ou la méthode du plus proche centroïde par exemple.

1. **Représentation des données :**
   (a) En visualisant puis en évaluant quantitativement les relations de corrélation entre les 4 variables de représentation, déterminez s'il est nécessaire d'appliquer une transformation des variables basée sur l'analyse des composantes principales (ACP). Les relations de corrélation entre les variables sont-elles similaires pour toutes les 3 classes ?
   (b) En visualisant la séparation entre les 3 classes après transformation par ACP, déterminez un nombre optimal de composantes principales (CP) à utiliser pour la classification : 2CP ou 3CP. Vérifiez votre réponse en calculant, pour chaque objet, le centroïde dont il est le plus proche par la distance (Euclidienne) dans les cas 2CP et 3CP, puis en comparant avec les classes réelles des objets.
2. **Mesure de distance :**
   (a) D'après les résultats sur l'analyse de corrélation entre les variables de représentation (1.(a)), quelle mesure de distance (Manhattan, Euclidienne, ou Mahalanobis) entre les

objets serait la plus adéquate ? Vérifiez votre réponse en calculant pour chacune des mesures de distance, le centroïde le plus proche de chaque objet, puis en comparant avec les classes réelles des objets.

(b) Pour la distance de Mahalanobis, on peut utiliser une matrice de covariance par classe ou une matrice de covariance pour toutes les données. Laquelle des deux options est la plus adéquate ?

3. **Choix du modèle de classification :**

   (a) En utilisant la meilleure représentation des données retenue au Point 1, et la meilleure mesure de distance retenue au Point 2, tester la méthode des $k = 5$ plus proches voisins ou la méthode du plus proche centroïde, et déterminez la plus adéquate.

   (b) On fait l'hypothèse que les objets correspondent à des mélanges de distributions gaussiennes correspondant aux classes. Déterminez si cette hypothèse est vraissemblable en appliquant une classification par modèle de mélange gaussien ("Gaussian Mixture Model") aux données. Justifiez votre choix parmi les quatre options du modèle pour la covariance des différentes classes (spherical, diag, tied, ou full).

4. **Application :** À l'aide du modèle retenue au Point 3., déterminez la classe de la nouvelle observation suivante : [52.1, 23.0, 6.1, 16.5]

Remise du travail

Pour soumettre votre travail, connectez-vous, dans un fureteur, au serveur `http://opus.dinf.usherbrooke.ca` en utilisant votre CIP, puis choisissez le cours `IFT870 (BIN710)` et le projet `TP1`. Chargez votre fichier `tp1.ipynb` et soumettez-le. Le nom de votre fichier de remise doit être exactement `tp1.ipynb`. Indiquez bien les noms des deux membres de l'équipe dans le fichier. Ne faites qu'une seule soumission par équipe. Ne remettez pas d'autre fichier.