UNIVERSITÉ DE SHERBROOKE
Faculté des sciences
2500, boul. de l'Université
Sherbrooke (Québec) J1K 2R1
CANADA
Département d'informatique
+1 (819) 821-7030
+1 (819) 821-8200
http://www.usherbrooke.ca/informatique/

Quantitative Structure-Activity Relationship (QSAR)

Présentation générale du laboratoire

Le but du TP1 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. Il devra être complété avant le vendredi 29 février 2024 avant 23h59. Vous devez remettre, sur Moodle, un fichier Ipython notebook (nommé nomEquipe_tp1.ipynb et les données nettoyées – au format souhaité) 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 Excel (QSAR_data.xlsx). Ces données proviennent de l'étude de Lévêque, Tahiri *et al.* (2021) intitule « Quantitative Structure-Activity Relationship (QSAR) modeling to predict the transfer of environmental chemicals across the placenta » [1]. Les données sont un extrait de l'étude d'origine et modifié pour la composition de ce travail pratique. L'ensemble des données contient 154 observations représentées suivant 76 variables. Les données sont segmentées en 5 classes (« class 2 », « class 1 », « class 0 », « class -1 », « class -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 75 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) Vous devez analyser l'ensemble de vos données, par :

- Analyser chaque attribut.
- Proposer un prétraitement de vos données (comme vu en cours et en atelier).
- Sélectionner les 10 meilleurs attributs (donnez une justification statistique).
- Visualiser la distribution des 10 meilleurs attributs.

(b) En visualisant puis en évaluant quantitativement les relations de corrélation entre les 10 meilleurs attributs de représentation (voir le point précédent), déterminez s'il est nécessaire d'appliquer une transformation d'attributs basée sur l'analyse des composantes principales (ACP). Les relations de corrélation entre les variables sont-elles similaires pour toutes les 5 classes ?

(c) En visualisant la séparation entre les 5 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.

[1] Lévêque, L., Tahiri, N., Goldsmith, M.R. and Verner, M.A., 2021. Quantitative Structure-Activity Relationship (QSAR) Modeling to Predict the Transfer of Environmental Chemicals Across the Placenta. Computational Toxicology, p.100211.

2. Mesure de distance :
   (a) D'après les résultats sur l'analyse de corrélation entre les variables de représentation (voir question 1b), 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. *Comme ce modèle à était traité en cours et en laboratoire, je vous laisse donc l'implémenter.*
   (b) On fait l'hypothèse que les objets correspondent à des mélanges de distributions gaussiennes correspondant aux classes. Déterminez si cette hypothèse est vraisemblable 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). Ce modèle se trouve dans le script python (model.py) que vous devez l'ajouter à votre Jupiter Notebook.
4. Application : À l'aide du modèle retenue au Point 3, déterminez la classe de la nouvelle observation suivante : voir test.csv

Remise du travail

Pour soumettre votre travail, connectez-vous, dans un fureteur, au serveur Moodle. Chargez votre fichier nomEquipe_tp1.ipynb, fichier des données nettoyés et soumettez-le. Indiquez bien les noms des deux membres de l'équipe dans le fichier. Ne faites qu'une seule soumission par équipe.

Bon travail ☺

[1] Lévêque, L., Tahiri, N., Goldsmith, M.R. and Verner, M.A., 2021. Quantitative Structure-Activity Relationship (QSAR) Modeling to Predict the Transfer of Environmental Chemicals Across the Placenta. Computational Toxicology, p.100211.