IFT 599/799 – Science des données
TP1 : Compréhension et visualisation des données
Date de remise : le mercredi 30 septembre 2020

Ce TP porte sur l'analyse des données pour comprendre et visualiser la répartition des données. De façon plutôt informelle, ce TP pourrait être décrit comme suit :

Étant donnée un ensemble de données contenant un certain nombre de classes, appliquez deux méthodes pour montrer, le plus fidèlement possible, l'état de la séparation entre les différentes classes. L'une de ces méthodes doit comporter un aspect visuel.

Évidemment, cette description est un peu trop générale. Plus concrètement, pour le TP1, les éléments suivants doivent être considérés :

- **Données** : Iris est un ensemble de données très connu du domaine de la science des données. Il contient 150 observations (ou objets) réparties en 3 classes (appelées respectivement setosa, versicolor, virginica) de 50 observations chacune. Chaque objet d'Iris est décrit par 4 variables (sepal_length, sepal_width, petal_length, petal_width). Pour plus d'informations concernant l'ensemble d'iris, vous pouvez consulter https://en.wikipedia.org/wiki/Iris_flower_data_set et https://archive.ics.uci.edu/ml/datasets/iris. L'ensemble des données Iris est placé dans le répertoire publique et le Moodle du cours en trois versions : .csv, .txt, et .ARFF. Ce dernier est le format du progiciel Weka.

- **Séparation entre les classes** : Pour le TP1, nous considérons plutôt un cas simplifié de la séparation. Les trois classes d'Iris sont relativement très bien séparées les unes des autres dans l'espace de R^4. Donc, pour ce TP, vous pouvez vous donner comme objective de montrer que les trois classes sont réellement bien séparées. C'est évidemment plus simple que de « montrer le plus fidèlement possible l'état de la séparation entre les différente classes », qui est un défi plus grand.

- **Méthode 1 (pas de visualisation)** : Pour la méthode 1, vous présentez vos résultats d'analyse **par des tableaux** car ce sont des mesures quantitatives que vous calculez. Cette méthode est intuitive. Elle sert à se familiariser avec des mesures de distance et des mesures de qualité de classe.
    - **1.a** : On peut étudier la séparation des données par analyse de deux types de mesures *cohésion* et *séparation*. Pour ce TP, à la place de *cohésion* on utilise le terme *distance intra-classe*, et à la place de *séparation* on utilise le terme *distance inter-clase*.
        - Une distance intra-classe (de la classe *setosa* par exemple) peut être définie comme la distance maximale entre un objet quelconque de la classe *setosa* et le centre de cette classe.
        - Une distance inter-classe « directionnelle » d'une classe, e.g. *versicolor*, vers une autre classe, e.g. *setosa*, peut être définie

comme la distance minimale entre un objet quelconque de la classe *versicolor* et le centre de la classe *setosa*. (On pourrait évidemment définir une distance inter-classe non-directionnelle (comment?), mais pour ce TP, elle n'est pas nécessaire.)

- Le test à faire pour confirmer la séparation entre les deux classes *setosa* et *versicolor* peut se faire comme suit. Si la distance intra-classe de *setosa* est plus petite que la distance inter-classe « directionnelle » de *versicolor* vers *setosa*, alors on peut conclure que les classes *setosa* et *versicolor* sont bien séparées, mais on ne pourrait pas conclure si ce n'était pas le cas. De même, Si la distance intra-classe de *versicolor* est plus petite que la distance inter-classe « directionnelle » de *setosa* vers *versicolor*, alors on peut conclure que les classes *versicolor* et *setosa* sont bien séparées.

- **1.b** : La performance de l'approche précédente dépend de la mesure de distance utilisée.
  - Vous devez tester la distance Euclidienne et la distance Mahalanobis.
  - Il y a de différentes façons d'appliquer la distance Mahalanobis. **Trouver une bonne façon d'utiliser cette distance fait partie des tâches à compléter pour TP1**.
  - Il y a aussi de différentes façons d'appliquer la distance Euclidienne car la distance sera différente en fonction des variables choisies. Par exemple, vous n'êtes pas obligés d'utiliser toutes les variables dans votre calcul de distances. Mais, trouver une bonne combinaison de variables pour la distance Euclidienne n'est pas demandé pour ce TP.

- **Méthode 2 (avec visualisation)** : Pour la méthode 2, vous présentez vos résultats d'analyse **par des figures** de nuages de points ou de histogrammes. On cherche donc la présentation visuelle. Il n'est pas nécessaire de fournir des résultats quantitatifs (en utilisant les tableaux).
  - **2.a** : Si les objets sont représentés par une seule variable, alors, on peut utiliser l'histogramme pour représenter la distribution de chaque classe. Pour visualiser l'état de la séparation entre deux classes, on peut tout simplement afficher deux histogrammes sur une seule figure à deux dimensions (en utilisant une différente couleur pour chaque histogramme) : x représente l'axe des données et y représente les fréquences.
  - **2.b** : Maintenant, si les classes sont représentées par deux variables, on pourrait encore utiliser l'approche par l'histogramme, mais on ne génère pas de très belles figures de cette façon. Une méthode plus simple serait de tout simplement afficher les nuages de points pour chaque classe (*scatter plot* en anglais).
  - **2.c** : Peu importe si on choisit l'affichage d'une variable ou de deux variables, la question clée est de choisir quelles variables à utiliser.

- On pourrait utiliser les variables dans les données (d'Iris).
- On pourrait transformer les variables. C'est par la transformation, on pourra obtenir de « meilleurs » variables permettant de mieux illustrer la séparation entre les classes. **Dans votre démarche pour la Méthode 2, vous devez inclure l'utilisation des variables originales et la transformation des variables**.
- La technique pour la transformation des variables doit être celle basée sur l'analyse des composantes principales (ACP). Comme pour **Point 1.b**, **trouver la bonne façon d'appliquer l'ACP fait partie des tâches à compléter pour TP1**.

- **Programmation** : vous êtes libres d'utiliser le langage de votre choix pour faire ce TP. Vous n'avez pas à programmer les analyses comme ACP car vous pouvez facilement trouver des programmes de ces analyses sur l'Internet. Vous devez citer clairement les sources cependant quand vous utilisez les programmes des autres. **Ne pas citer les sources sera considéré comme une acte de plagiat et pourrait conduire à une note de zéro en plus de s'exposer à des mesures disciplinaires**. Vous pouvez faire les citations soit dans vos programmes par des commentaires soit dans une section ou un paragraphe de votre rapport du TP1 avec une liste des sources.

Travaux à réaliser pour TP1 : Les combinaisons suivantes sont exigées. La présentation des résultats vise toujours à montrer la séparabilité entre deux classes de toutes paires possibles.

1. **Méthode 1** : **Point 1.a** + **Point 1.b**. Attention, vous avez deux distances pour **1.b**. Pour la distance Mahalanobis, vous devez déterminer les sous-ensembles de données à utiliser pour construire les fonctions de distance. Vous avez donc potentiellement plusieurs façons de le faire. Pour la remise, vous présentez la meilleure façon que vous avez trouvée.

2. **Méthode 2** : Les combinaisons suivantes sont exigées.
    a. **Point 2.a** avec variables non transformées : Pour chaque paire de classes, montrer le meilleur résultat suffit. C'est-à-dire, le résultat avec l'une des 4 variables.
    b. **Point 2.a** + **Point 2.c** avec variables transformées : Attention, vous aurez toujours 4 variables à choisir, de la plus significative à la moins significative.
    c. **Point 2.b** : nuages des points avec variables non transformées ;
    d. **Point 2.b** + **Point 2.c** : nuages des points avec variables transformées. Attention, vous aurez toujours 4 variables à choisir, de la plus significative à la moins significative.

Toutes les précisions ne sont pas fournies, ce qui veut dire que vous avez de la liberté des choix. Vous n'avez pas à explorer de façon exhaustive toutes les possibilités. Le plus important est de tester au moins une « configuration » par chaque méthode-combinaison.

Présentation des résultats : Dans votre rapport, vous devez décrire, brièvement, l'objectif et votre démarche pour chaque méthode. Vous pouvez rapporter seulement les meilleurs résultats pour chaque méthode-combinaison. Rappelons que vous avez 3 classes dans l'ensemble. Pour des mesures mutuelles (comme pour la méthode 1), vous devez les appliquer sur des paires de classe (dont trois combinaisons possibles). Pour la méthode 2, il est possible d'afficher les résultats sur les trois classes dans une seule figure. Mais, afficher les résultats pour chaque paire de classes est préféré.

Vous devez fournir quelques commentaires sur les résultats de chaque méthode-combinaison pour faciliter la compréhension de votre présentation et des résultats. Si vous utilisez des ressources Internet, il faut absolument citer les sources aussi. **Ne pas citer les sources sera considéré comme une acte de plagiat et pourrait conduire à une note de zéro en plus de s'exposer à des mesures disciplinaires**. Il est fortement déconseillé d'utiliser des ressources Internet pour la partie de l'analyse des résultats.

Concernant l'équipe et la remise :

1. Le TP **doit** être fait seul ou en équipe de deux personnes ;
2. La date de remise est : le mercredi 30 septembre 23h59, aucun TP ne sera accepté à partir de cette date ;
3. Soignez votre rapport, une pénalité de 5% pourrait être appliquée pour un rapport mal rédigé; Normalement, votre rapport ne devrait pas dépasser 5 ou 6 pages + la page de couverture (pas indispensable) ;
4. Les fichiers à soumettre sont le rapport (en Word ou pdf) et l'ensemble de vos programmes. **Ne pas soumettre les données !**
5. N'oubliez pas de vous identifier. Indiquez votre nom et cip (ou matricule) dans chacun des fichiers que vous soumettez. La remise doit être faite par http://opus.dinf.usherbrooke.ca

Informations supplémentaires : Pour faire un histogramme, vous devez choisir vous-même la largeur de chaque « bin », alors que la largeur affecte la qualité visuelle d'un histogramme. Quand vous affichez deux histogrammes sur une même figure, il vaut mieux afficher un histogramme « par-dessus » l'autre.