IFT 599/799 – Science des données
TP1 : Visualisation
Date de remise : le vendredi 27 septembre 2019

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, (trouvez et) 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 chacun. Iris contient 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 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)** :
    - **1.a** : Une approche possible serait d'examiner des mesures élémentaires de cohésion et de séparation. Par exemple, la cohésion d'une classe peut être la « variance » de la classe, alors que la séparation entre deux classes peut être la « distance » entre les deux classes. L'interprétation de résultats basés sur des mesures de cohésion et de séparation pose un certain défi cependant car elles ne sont pas toujours très « intuitives ».
    - **1.b** : Une autre approche, plus simple et très intuitive, est de comparer la distance maximale entre un objet quelconque d'une classe, e.g. « setosa », et le centre de cette classe, avec la distance minimale entre un objet quelconque d'une autre classe, e.g. « versicolor », et le centre de la classe « setosa ». Si cette distance maximale est plus petite que la distance minimale sus-mentionnée, alors on peut conclure facilement que les

classes « setosa » et « versicolor » sont bien séparées, mais on ne pourrait pas conclure si ce n'était pas le cas.

- **1.c** : La performance des deux approches précédentes 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. Votre façon de faire doit permettre le plus possible de faciliter la comparaison des résultats.

- **Méthode 2 (avec visualisation)** :
  - **2.a** : Si les classes sont représentées par une seule variable seulement, 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 histogramme 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** : À la place des histogrammes, on peut afficher aussi une distribution par classe. Par exemple, on peut supposer que la distribution est une Gaussienne. Il suffit donc d'estimer les paramètres de la Gaussienne pour chaque classe avant d'afficher la fonction de la Gaussienne. En affichant deux distributions correspondant aux deux classes, on peut montrer l'état de séparation entre les deux classes.
  - **2.c** : 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.d** : Cependant, l'approche par distribution est encore bonne dans le case des données à deux variables. Dans ce cas, nous pouvons supposer que la distribution est encore une Gaussienne (une distribution normale). On estime les paramètres de la distribution et affiche la fonction de distribution de chaque classe comme pour le point **2.b**.
  - **2.e** : Peu importe si on choisit l'affichage d'une variable ou de deux variables, la question clée est de choisir les quels 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.c**, **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.

Méthodes à réaliser pour TP1 : Les combinaisons suivantes sont exigées. Toutes autres combinaisons sont optionnelles et n'apporteront qu'au maximum 10 points sans permettant une note au-delà du 100%.

1. Méthode 1 : **Point 1.b** + **Point 1.c** ;
2. Méthode 2 :
    a. **Point 2.a** avec variables non transformées ;
    b. **Point 2.a** + **Point 2.e** (avec variables transformées) ;
    c. **Point 2.c** avec variables non transformées (nuages des points) ;
    d. **Point 2.c** + **Point 2.e** (avec variables transformées) ;

Toutes les précisions ne sont pas fournies, ce qui veut dire que vous avez beaucoup de libertés de 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 deux classes (dont trois combinaisons possibles). Pour la méthode 2, il est tout à fait possible d'afficher les résultats sur les trois classes dans une seule figure.
Vous devez présenter 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 vendredi 27 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 matricule dans chacun des fichiers que vous soumettez. La remise doit être faite par http://opus.dinf.usherbrooke.ca