



Université de
Sherbrooke

IFT 599 / IFT 799 - Science de données

TP1 : Compréhension et visualisation des données

Automne 2024

Enseignant

	Courriel	Local	Téléphone
Shengrui Wang	shengrui.wang@usherbrooke.ca	D4-1018-1	+1 819 821-8000 x62022

FACULTÉ DES SCIENCES,
DÉPARTEMENT D'INFORMATIQUE

9 septembre 2024

Sommaire

Dans le cadre de ce travail pratique (TP) est mis à la disposition des personnes étudiantes un jeu de données. Il est question ici, à partir de ce jeu de données, de mettre en exergue les concepts de prétraitement de données vus dans le Thème 1 du cours. Plus précisément, ce TP consiste à chercher des combinaisons de plusieurs concepts et techniques pour comprendre et visualiser la séparation des données. Il s'agit d'un projet d'exploration des données. À travers ce projet, vous allez maîtriser plusieurs techniques de traitement et acquérir une bonne capacité d'analyse.

Contents

1	Jeu de données et énoncé du problème	1
1.1	Jeu de données	1
1.2	Énoncé du problème	1
2	Travail à faire	3
2.1	Programmation	3
2.2	Travail à réaliser	3
2.3	Présentation des résultats	4
2.4	Remise du TP	4

1 Jeu de données et énoncé du problème

1.1 Jeu de données

Les données soumises à votre étude sont extraites du jeu de données RNA-Seq (HiSeq) PANCAN (<https://archive.ics.uci.edu/dataset/401/gene+expression+cancer+rna+seq>). Il s'agit d'une extraction aléatoire d'expressions génétiques des patients ayant différents (5) types de tumeurs : BRCA, KIRC, COAD, LUAD et PRAD. Le jeu de données est également disponible dans Teams du cours (sous-répertoire ../travaux pratiques/TP1). Le dossier compressé comprend deux fichiers sous format tabulaire. L'un des fichiers, très gros, comprend les profils génomiques des patients. Chaque ligne correspond à un patient tandis que les colonnes sont des gènes. Dans le deuxième fichier, on trouve les différentes tumeurs associées aux différents patients.

1.2 Énoncé du problème

À partir d'une étude exploratoire portée sur ces profils génomiques, on voudrait savoir si les différents types de cancers sont bien séparés.

Méthode 1 (sans visualisation des données) La méthode 1 vise à présenter la séparation entre les classes par des fonctions de séparation. Les choix possibles sont présentés dans les parties 1 et 2 ci-dessous. **Pour cette méthode, 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. : On peut étudier la séparation des données par analyse de deux types de mesures, la **cohésion** et la **séparation**. Pour ce TP, on définit la cohésion par une formule appelée distance intra-classe, et la séparation par une formule appelée distance inter-classe.

- Une distance intra-classe (de la classe BRCA par exemple) est définie comme étant la distance maximale entre un patient quelconque de la classe BRCA et le centre de cette classe. Formellement, pour une classe de données $C = \{x_1, x_2, \dots, x_n\}$ de n patients, la distance intra-classe ($dist_{intra}(C)$) est définie comme suit,

$$dist_{intra}(C) = \max\{dist(x_i, x_C) \mid \forall x_i \in C\}$$

Ici $dist()$ est une mesure choisie (par ex. la distance Euclidienne), x_C le centre de la classe C (généralement calculé comme étant la moyenne des données de C).

- Une distance inter-classe (par ex. entre BRCA et KIRC) est définie comme étant le minimum des deux distances suivantes : la distance minimale entre un objet quelconque de la classe BRCA et le centre de la classe KIRC, et la distance minimale entre un objet quelconque de la classe KIRC et le centre de la classe BRCA.

Formellement, étant donnée deux classes $C_1 = \{x_1, x_2 \dots, x_{n_1}\}$ et $C_2 = \{x_1, x_2 \dots, x_{n_2}\}$ de n_1 et n_2 patients respectivement, la distance inter-classe ($dist_{inter}(C_1, C_2)$) est définie comme suit,

$$dist_{inter}(C_1, C_2) = \min\left(dist(C_1, C_2), dist(C_2, C_1)\right)$$

Avec

$$dist(C_1, C_2) = \min\{dist(x_i, x_{C_2}) \mid \forall x_i \in C_1\}$$

$$dist(C_2, C_1) = \min\{dist(x_j, x_{C_1}) \mid \forall x_j \in C_2\}$$

x_{C_1} et x_{C_2} étant les centres respectifs des classes C_1 et C_2 .

- Dans cette première méthode, le test à faire pour confirmer la séparation entre les deux classes est de regarder à quel point les classes sont distantes entre elles. Pour ce faire, on se donne un indicateur de superposition $Overlap()$ de classes défini comme suit,

$$Overlap(C_1, C_2) = \frac{dist_{intra}(C_1) + dist_{intra}(C_2)}{2 \times dist_{inter}(C_1, C_2)}$$

Si $Overlap(C_1, C_2) < 1$ on pourra dire que les classes C_1 et C_2 sont bien séparées.

2. : La performance de l'approche précédente dépend de la mesure de distance utilisée.

Vous devez tester avec chacune des métriques ci-dessous

- Distance Euclidienne,
- Distance Mahalanobis : pour le calcul de la distance Mahalanobis, si votre ordinateur n'a pas la puissance nécessaire pour tenir compte des 16382 dimensions, vous pouvez choisir en utiliser un sous-ensemble. Dans ce cas, indiquez dans votre remise les dimensions utilisées,
- Distance cosinus.

Méthode 2 (avec visualisation) La méthode 2 vise à visualiser la séparation des données par des figures de nuages de points ou de histogrammes. Il n'est pas nécessaire de fournir des résultats quantitatifs en utilisant des tableaux.

1. : 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 pourrait regarder conjointement la distribution des deux classes (une illustration est donnée ici <https://seaborn.pydata.org/generated/seaborn.jointplot.html>)
2. : 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).

Dans le jeu données fourni, nous avons beaucoup trop de variables (attributs) qu'il serait fastidieux de trouver une bonne combinaison de deux ou trois variables permettant de bien visualiser la séparation des différents types de cancers. Pour cette raison, vous devez réduire le nombre de dimensions en utilisant chacune des méthodes suivantes :

- ACP : Pour la même raison que pour le calcul de la distance Mahalanobis, vous pouvez utiliser un sous-ensemble des dimensions. Si vous optez pour un sous-ensemble des dimensions, utilisez le même que celui pour la distance Mahalanobis
- T-SNE

2 Travail à faire

2.1 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.

2.2 Travail à réaliser

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 : vous devez réaliser la première approche en utilisant les trois métriques données en partie 2. Pour la distance Mahalanobis, vous devez déterminer les sous-ensembles de données à utiliser pour construire les fonctions de distance. D'autre part, vous pourriez choisir aussi entre plusieurs combinaisons possibles des variables. ~~Vous avez donc potentiellement plusieurs façons de le faire. Pour la remise, vous présentez la meilleure façon que vous avez trouvée.~~ **Il est conseillé, mais non obligatoire, d'utiliser un même ensemble de variables pour les trois distances.** Le choix de l'ensemble de variables peut se faire de beaucoup de façons. Plus qu'il permet de séparer les classes, mieux c'est. Cependant, on ne vous demande pas de faire une recherche exhaustive pour un ensemble optimal des variables. Il suffit d'avoir une bonne compréhension de ce problème et d'effectuer quelque recherches à la place de prendre un sous-ensemble quelconque.
2. Méthode 2 : **Le mode d'affichage principal est l'affichage des nuages de points à deux variables. L'ajout des histogrammes ou courbes de distribution de chacune des variables dans ces affichages est optionnel. En principe, vous devez faire une figure par paire de classes par paire de variables sauf pour le résultat de T-SNE. Voir plus de précisions ci-dessous.**
 - (a) ~~Avec quelques variables de votre choix (max 2), pour chacune d'elle, afficher les distributions des différentes classes.~~

Pour l’affichage des classes selon des variables originales, vous choisissez parmi les variables utilisées dans la Méthode 1 soit une ”bonne” paire de variables pour toutes les paires de classes, ou une ”bonne” paire de variables par paire de classes.

- (b) Afficher les nuages des points déterminés dans le point (a). Les points doivent être coloriés suivant chaque classe. Il est permis d’afficher plus de deux classes par affichage surtout si les classes ne se recouvrent pas beaucoup. Évidemment, ce dernier suppose que la paire de variables choisie pour ces classes soit la même.
- (c) **Optionnellement**, pour chacune des paires variables que vous avez choisies pour affichage, afficher conjointement aussi la distribution des paires de classes, sous forme d’histogramme, selon chaque variable impliquée. En autres termes, il s’agit d’illustrer la séparation des classe par l’affichage 1D.
- (d) Effectuez les transformations ACP et T-SNE données en partie 2 et affichez le nuage de points. Les points doivent être coloriés suivant chaque classe. L’affichage des résultats obtenus de l’ACP doit suivre les mêmes consignes de (a), (b) et (c) ci-dessous. L’affichage des résultats obtenus de t-SNE peut se faire sur une seule figure.

2.3 Présentation des résultats

Dans votre rapport, vous devez décrire, brièvement, l’objectif et votre démarche pour chaque méthode. 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.

2.4 Remise du TP

- Le TP doit être fait en équipe de trois personnes ou moins;
- La date de remise du TP est le lundi 30 septembre 2024 23h59, **AUCUN TP NE SERA ACCEPTÉ APRÈS CETTE DATE**;
- Soignez votre rapport, une pénalité de 5% pourrait être appliquée pour un rapport mal rédigé;
- Les fichiers à soumettre sont le rapport (en Word ou pdf) et l’ensemble de vos programmes. Ne pas soumettre les données!
- N’oubliez pas d’identifier les membres du groupe de travail. Indiquez les noms et Cips (ou matricules) des membres du groupe dans chacun des fichiers que vous soumettez. La remise doit être faite par Turnin : <http://turnin.dinf.usherbrooke.ca>