IFT 599 / IFT 799 - Science de données

TP1 : Compréhension et visualisation des données

Automne 2023

Enseignants

	Courriel	Local	Téléphone
Shengrui Wang	shengrui.wang@usherbrooke.ca	D4-1018-1	+1 819 821-8000 x62022
Etienne G. Tajeuna	etienne.gael.tajeuna@usherbrooke.ca		+1 819 821-8000 x

Faculté des Sciences,
Département d'informatique

September 15, 2023

Sommaire

Dans le cadre de ce travail pratique (TP) est mis à la disposition des personnes étudiantes un (01) jeu de données. Il est question ici, à partir de ce jeu de données de mettre en exergue les conceptes de prétraitement de données vu 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 mini projet d'exploration des données. À travers ce projet, vous allez acquérir une bonne capacité d'analyse et maîtriser quelques techniques de base.

Contents

1 Jeu de données et énoncé du problème

1.1 Jeu de données

Les données soumis 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 de patients ayant différents types de tumeurs : BRCA, KIRC, COAD, LUAD et PRAD. Le jeu de données est également disponible dans Teams (`https://usherbrooke.sharepoint.com/:u:/r/sites/Cours-IFT599-01-IFT799-51-A23/Documents%20partages/General/Travaux%20pratiques/TP1/gene+expression+cancer+rna+seq.zip?csf=1&web=1&e=T0Rohd`). Le dossier compressé comprends deux fichiers sous format tabulaire. L'un des fichiers comprend les profils génomiques des patients. Chaque ligne est patient tandis que les colonnes sont des gènes. Dans le deuxième fichier, vous avez les différentes tumeurs associées aux différents patients.

1.2 Énoncé du problème

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

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. (a) and 1. (b) 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. (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-classe.
   - Une distance intra-classe (de la classe BRCA par exemple) peut être définie comme étant la distance maximale entre un patient quelconque de la classe BRCA et le centre de cette classe.
     Formellement, étant donnée une classe $C_1 = \{x_1, x_2 \ldots, x_{n_1}\}$ de n_1 patients, ladistance intra-classe ($dist_{intra}(C_1)$) est définie comme suit,

     $$dist_{intra}(C_1) = max\{mes(x_i, x_{c_1}) \,|\, \forall x_i \in C_1\}$$

     Avec $mes()$ une métrique donnée (exemple distance Euclidienne), x_{c_1} le centre de la classe C_1 (généralement représentée comme étant la moyenne).
   - Une distance inter-classe (exemple entre BRCA et KIRC) est définie comme étant la distance minimale entre un objet quelconque de la classe BRCA ou KIRC et du centre de la classe BRCA ou KIRC.

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

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

Avec

$$dist(C_1, C_2) = min\{mes(x_i, x_{c_2}) \,|\, \forall x_i \in C_1\}$$
$$dist(C_2, C_1) = min\{mes(x_j, x_{c_1}) \,|\, \forall x_j \in C_2\}$$

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. (b): 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,
   - 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 les tableaux.

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

   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 reduire le nombre de dimensions en utilisant chacune des méthodes suivantes :
   - ACP
   - TSNE
   - UMAP

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 1. (b). Pour la distance Mahalanobis, vous devez déterminer les sous-ensembles de données à utiliser pour construire les fonctions de distance (En fait, 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.
2. Méthode 2 :
   (a) Avec quelques variables de votre choix (max 2), pour chacune d'elle, afficher les distributions des différentes classes.
   (b) Afficher le nuage des points. Les points doivent être coloriés suivant chaque classe.
   (c) Pour chacune des variables que vous avez choisis, afficher conjointement la distribution des paires de classes.
   (d) Effectuez les transformation ACP, TSNE et UMAP données en 2. (b) et affichez le nuage de points. Les points doivent être coloriés suivant chaque classe.

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 deux ou trois;
- La date de remise du TP est le 29 septembre 2023 23h59, aucun TP ne sera accepté à partir de 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 http://turnin.dinf.usherbrooke.ca