IFT599 - TP1

29 septembre 2023

Carl Gauthier
GAUC2729

Julien Bernat
BERJ0505

Méthode 1

Objectif

L'objectif de la méthode 1 est d'analyser la séparabilité des données entre 2 classes de gènes à l'aide d'un calcul d'overlap. Nous allons utiliser 3 mesures de distances différentes soit euclidienne, de mahalonobis et cosinus pour calculer l'overlap. Nous allons ensuite comparer les résultats de 3 mesures de distances pour effectuer une analyse. Si l'overlap est inférieur a 1 cela signifie que les 2 classes sont bel et bien séparés.

Démarche

Nous avons décider d'utiliser python pour effectuer les calculs sur les données. À l'aide des formules nous avons donc calculer l'overlap entre tous les différentes classes pour chaque mesure de distance différente. Pour ce faire nous avons utiliser les formules de distances inter et intra classes. Pour ce qui est du calcul de distance utiliser dans les calculs de distance inter et intra classe nous avons utiliser la bibliothèque scipy de python (scipy.spatial.distance.euclidean, scipy.spatial.distance.cosine, scipy.spatial.distance.mahalanobis). Pour calculer la matrice de covariance utile dans la distance de mahalonobis nous avons utiliser la bibliothèque numpy de python (np.cov).

Overlap entre différentes paires de classes selon la mesure de distance

Classes	**Euclidienne**	**Mahalonobis**	**Cosinus**
PRAD BRCA	1.34656696191881	1.0154317397759087	1.971079766090517
PRAD KIRC	1.0859141721450414	0.7580035883303471	1.2206658186935642
PRAD LUAD	1.284960459996921	0.8423389848533355	1.754779036184894
PRAD COAD	1.0929855815908998	0.7781863732445895	1.1592267150262086
BRCA KIRC	1.2245241204635569	1.240109299199415	1.5462780430962064
BRCA LUAD	1.5402502815281198	1.3728332087231132	2.5369623171664357
BRCA COAD	1.2968460606188235	1.0477840538357608	1.6674492281054873
KIRC LUAD	1.3210436966091197	1.3332055014778261	1.8168122497375891
KIRC COAD	1.1787644748123138	0.8570595512525796	1.3305447364032923
LUAD COAD	1.4188265653780767	1.4377388427579436	1.9120936739347032

Analyse

On remarque que avec les distances euclidiennes et cosinus pour aucune paire de classe nous pouvons arriver à la conclusion de séparabilité des données. Par contre, lorsqu'on observe la distance de mahalonobis nous avons 4 paires de classes (PRAD KIRC, PRAD LUAD, PRAD COAD et KIRC COAD) pour lesquels on peut affirmer qu'il y a séparabilité des données. Cela nous permet de conclure que la distance de mahalonobis nous permet d'extraire de l'information que les autres mesures de distances ne sont pas capable de détecter. Cela nous permet également de conclure que la classe PRAD est la plus séparé des autres classes et que les autres classes sont tous assez similaires à l'exception de KIRC et COAD. Ont peut observer ces résultats numériques avec la visualisation utiliser dans la méthode 2.

Méthode 2

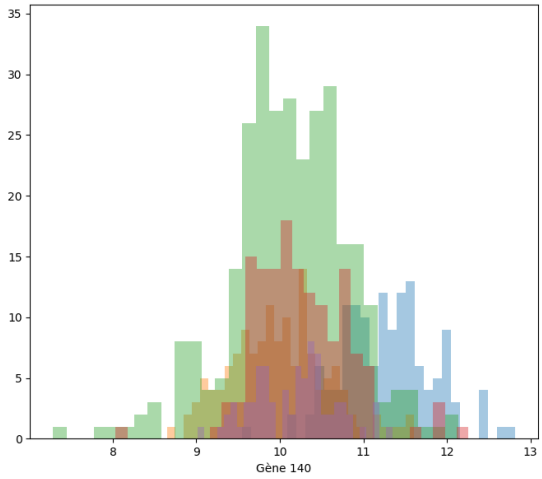
Objectif

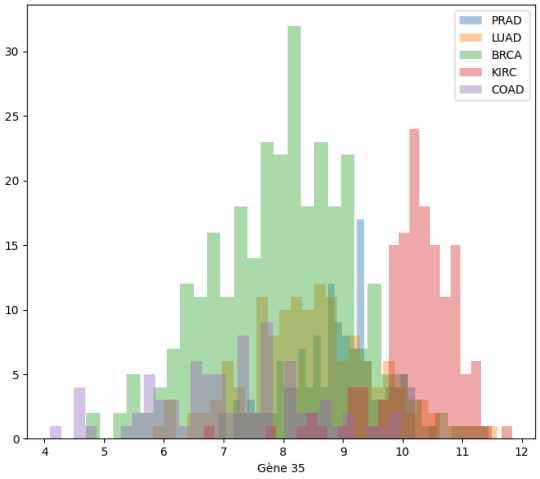
Après avoir obtenu des résultats quantitatifs avec la méthode 1, il serait intéressant d'avoir également accès à des visualisations afin de pouvoir analyser plus intuitivement les relations entre les gènes et les classes. C'est l'objectif de la méthode 2. Nous présentons dans cette section 4 types de visualisations distinctes.

Les gènes 140 et 35 ont été séléctionner pour les visualisations à 2 variables en **a)**, **b)** et **c)**

a)

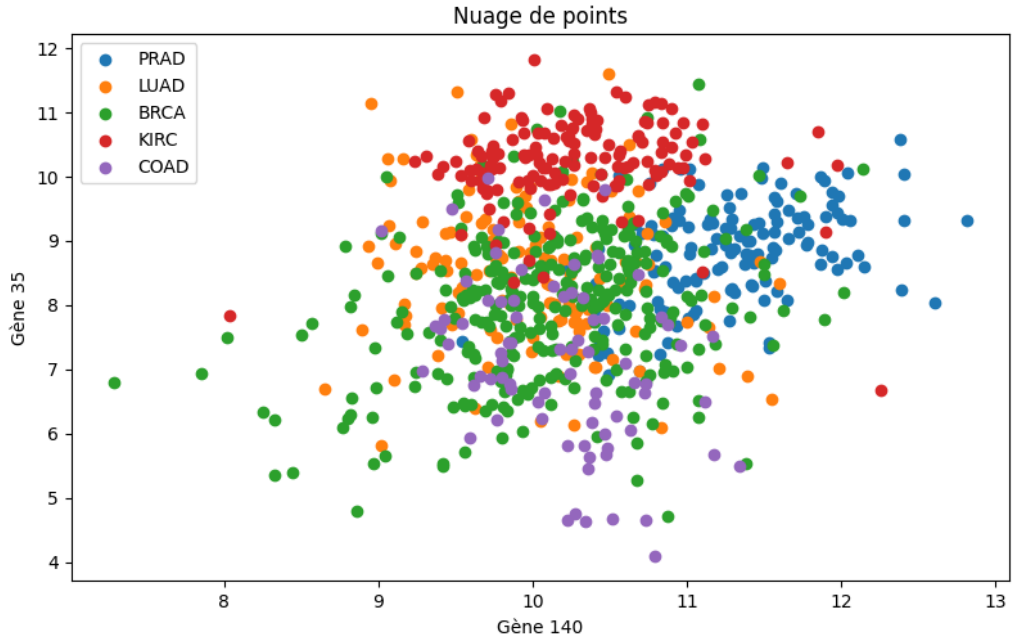
Voici deux histogrammes qui montre la distribution de chaque classe pour les variables concernées. Nous avons décidé d'inclure chaque classe dans le même graphique avec une légende afin de rendre les différences plus apparentes en conservant les axes.





b)

La deuxième visualisation est un nuage de points. Les informations présentées ici sont les mêmes qu'en **a)**. Ce format est mieux approprié aux données puisqu'on observe 2 variables à la fois. Aussi la division des classes est plus facile à observer car il y a moins d'obstruction entre les distributions.



On remarque ici que la valeur associée au gène 35 a tendance à être plus élevé pour la classe **KIRC**. La même remarque s'applique pour le gène 140 et la classe **PRAD**.

c)

Nous devions réaliser une visualisation pour les distributions jointes entre chaque pair de classe. L'histogramme 2D convient bien à cette tâche. Par contre ce type de visualisation exige d'avoir le même nombre d'échantillons pour les deux dimensions. Puisque ce n'est pas le cas on doit réechantillonner l'une des variables. Nous avons choisi de suréchantillonner, c'est-à-dire qu'on ajoute de nouveaux échantillons dans la dimension la plus petite en respectant la distribution initiale afin de faire correspondre les quantités d'échantillons.

Oversampling

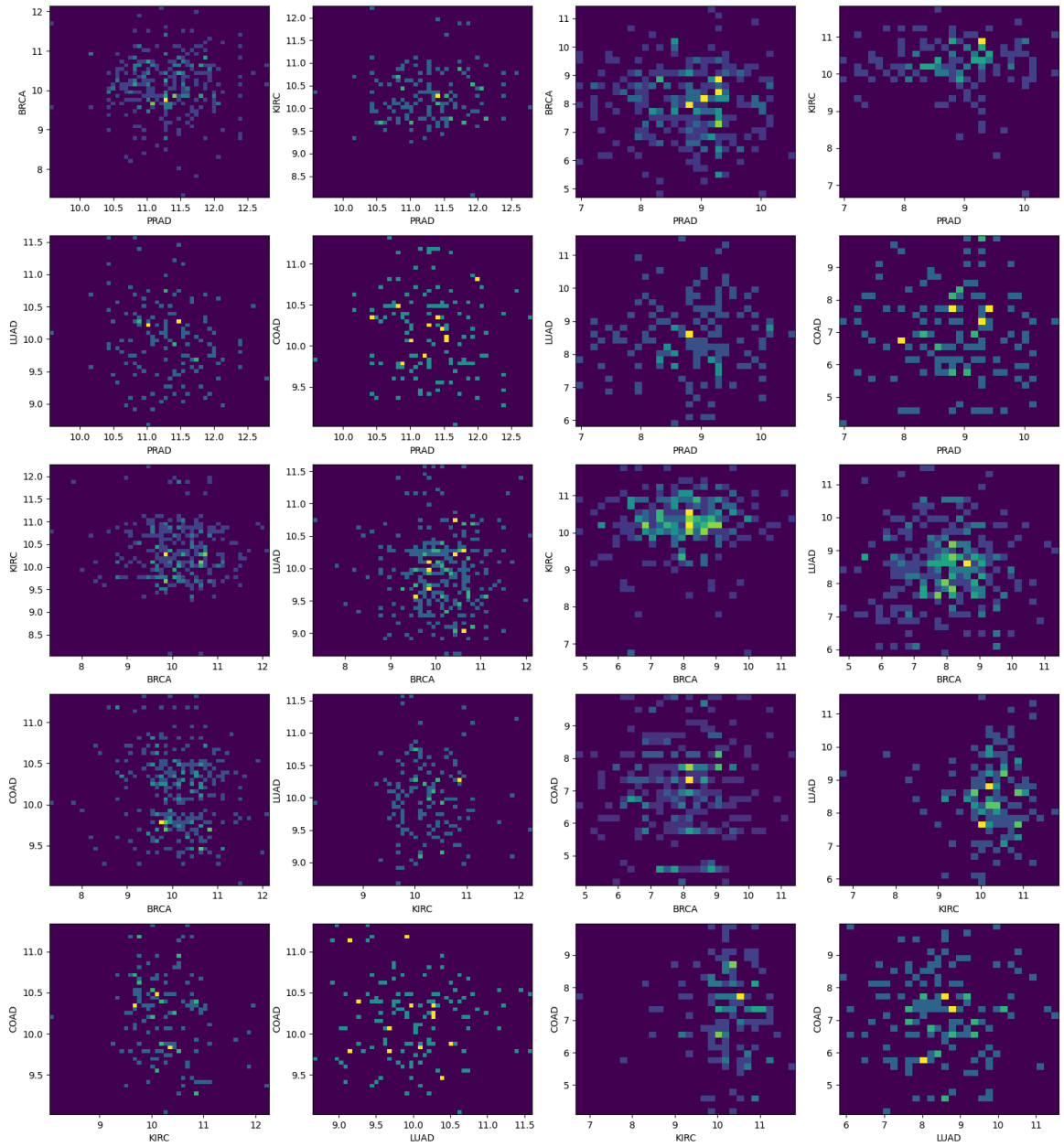
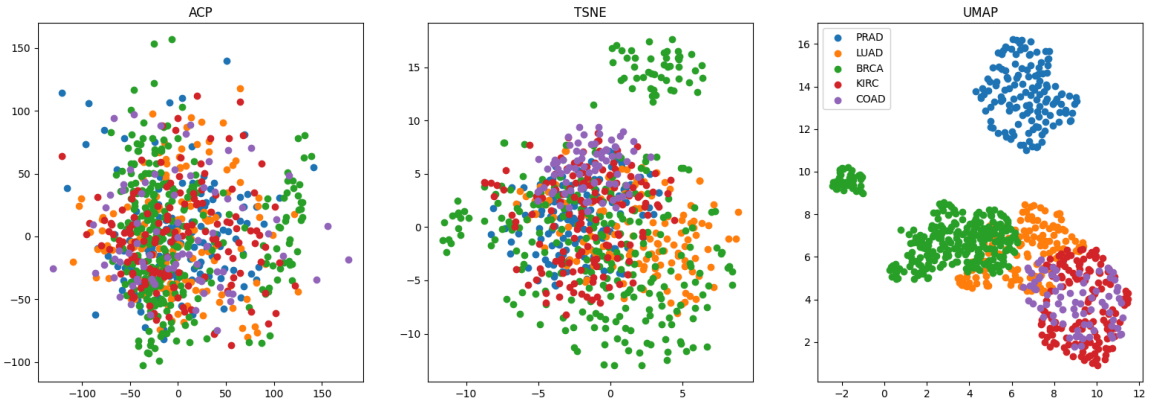


Figure 3: À gauche, dans les deux premières colonnes se trouvent les distributions jointes pour le gène 140. À droite se trouvent celles pour le gène 35.

d)

Pour conclure, on propose trois dernières visualisations qui ont en commun de réduire le nombre de dimensions. Contrairement au visualisation à deux variables celle-ci sont un amalgame de tous les gènes. Les trois méthodes de réduction sont **ACP**, **TSNE** et **UMAP**. Chaque méthodes



Les méthodes TSNE et UMAP sont non déterministes, les résultats dépendent d'un facteur aléatoire et sont différents à chaque exécution.

On remarque dans les graphiques TSNE et UMAP qu'une portion des échantillons classifiés BRCA est détachée du reste. Cela pourrait indiquer qu'on est en présence d'une distribution bimodale. Cette caractéristique est toujours vraie même si les résultats sont aléatoires.

Sources

Numpy: https://numpy.org/

Scipy: https://scipy.org/

Une liste des libraires tierces utilisées est disponible en annexe, voir `requirements.txt`