Projet d'étude

4 ModIA
UFs "Analyse de données" & "Eléments de modélisation statistique"
Intervenante : Cathy Maugis-Rabusseau

8 septembre 2022

Organisation du projet et documents à rendre

- Le projet sera réalisé par groupe de 4 (ou 3) étudiant-e-s. La constitution des groupes sera faite lors de la première séance. Tous les groupes doivent être mixtes / l'Ecole d'origine (INSA, ENSEEIHT) et si possible / genre (Femme, Homme).
- 8 séances de 2h30 sont dédiées dans votre emploi du temps au travail du projet. Je serai présente lors de ces séances pour répondre à vos questions.
- Livrables : vous devrez rendre (en déposant sous Moodle) au plus tard le **mardi 31 janvier 2023 minuit** les 2 documents suivants :
  1. un fichier Rmarkdown (*nom1-nom2-nom3-Rapport.Rmd*) contenant les codes R, les codes python et générant le rapport au format pdf. Si difficultés avec les codes python, ils pourront être rendus séparément dans un notebook.
  2. un rapport au format .pdf (*nom1-nom2-nom3-Rapport.pdf*) généré par la compilation du fichier .Rmd précédent.
     Attention : le rapport est limité à 25 pages, figures incluses.
- Un dossier "ModeleRapport", disponible sur Moodle, vous donne un exemple avec des consignes pour la rédaction de votre rapport. Il est important d'en prendre connaissance dès la première séance !

Evaluation du projet

Pour chaque UF, la note de projet compte pour un tiers de la note finale de l'UF. Elle sera issue de l'évaluation des critères suivants :

Critère	**UF EMS**	**UF AD**
Utilisation pertinente des méthodes d'exploration de données		x
Utilisation pertinente de méthodes de clustering adaptées à la question traitée		x
Utilisation pertinente de l'analyse discriminante linéaire		x
Choix des modélisations ML et MLG adaptées à la question traitée	x	
Ecriture mathématique des modèles considérés	x	
Utilisation de méthodes de sélection de variables	x	
Aller-retour exploration $\leftrightarrow$ modélisation	x	x
Analyse ($\neq$ lecture !) des résultats obtenus	x	x
Choix et rendu des graphiques illustratifs	x	x
Rédaction d'un document en Rmarkdown	x	
Programmation en R	x	
Couverture du code Python / code R		x
Rédaction générale du document	x	x
Bonus pour des choix originaux adaptés	x	x

Jeu de données étudié

On observe pour $G = 1615$ gènes d'une plante modèle les valeurs suivantes :

$$Y_{gtsr} = \log_2(X_{gtsr} + 1) - \log_2(X_{gt_0} + 1)$$

où

- X_{gtsr} est la mesure d'expression du gène $g \in \{G1, \ldots, G1615\}$ pour le traitement $t \in \{T1, T2, T3\}$ pour le réplicat $r \in \{R1, R2\}$ et au temps $s \in \{1h, 2h, 3h, 4h, 5h, 6h\}$
- X_{gt_0} est l'expression du gène g pour un traitement de référence t_0

Dans la suite, on dira qu'un gène g est sur-exprimé si $Y_{gtsr} > 1$, sous-exprimé si $Y_{gtsr} < -1$ et non-exprimé sinon. A noter que le traitement $T3$ est une combinaison des traitements $T1$ et $T2$. Des explications plus précises du contexte seront données lors de la première séance.

Questions à aborder

Dans votre rapport final, vous devez avoir abordé par une/des méthodes adaptées les questions suivantes :

- Analyse du jeu de données
  - Statistiques descriptives et préparation du jeu de données (variables redondantes ? transformations ? création de features ? outliers ?...)
  - Visualisation dans un espace de faible dimension (voir en particulier l'aspect réplicat biologique, l'effet traitement et l'effet temps)
- Clustering :
  - Obtention d'une classification des variables "Tx_xh_Rx"
  - Obtention d'une classification des gènes ayant des profils d'expression similaires (co-exprimés) dans les différentes conditions
- Etude de l'expression des gènes pour le traitement T3 à 6h :
  - Quels sont parmi les différents temps ceux affectant réellement l'expression des gènes à 6h pour le traitement T3 fixé ?
  - Même question avec tous les traitements et tous les temps.
  - Quelles sont les variables prédictives qui permettent de discriminer les gènes sur-exprimés des gènes sous-exprimés à 6h pour le traitement T3 ?
- Etude de l'expression des gènes pour le traitement T1 à 6h :
  - Quels sont parmi les différents temps ceux affectant réellement l'expression des gènes à 6h pour le traitement T1 fixé ?
  - Même question avec tous les traitements et tous les temps.
  - Quelles sont les variables prédictives qui permettent de discriminer entre les gènes sur- / sous-/non-exprimés à 6h pour le traitement T1 ?

Pour les deux derniers points, vous pouvez vous restreindre au réplicat R2.