Projet - Prediction et sélection

birmele@unistra.fr

Apprentissage - 2021-2022

Instructions pour le rendu du projet

- Le projet est à rendre pour le **mercredi 5 janvier à minuit** sous forme électronique.
- Le rapport sera rendu sous format Markdown (il faut éditer un fichier .Rmd dans R studio). C'est très simple d'utilisation et permet de faire cohabiter du texte avec des commandes et des sorties R. Vous m'enverrez à la fois votre .Rmd et une compilation en .html ou .pdf

Introduction

Nous avons vu en cours des méthodes pour faire de l'apprentissage, notamment pour prévoir des variables binaires. Dans ce projet, le but est de s'intéresser plus précisément à un problème de grande dimension, c'est-à-dire où le nombre de variables est supérieur au nombre d'observations disponibles.

1 Chargement et préparation de données réelles

On considère le jeu de données de cancer du sein *METABRIC* disponible sur Kaggle : https://www.kaggle.com/raghadalharbi/breast-cancer-gene-expression-profiles-metabric

1. Charger le jeu de données, le décrire, et préparer un jeu de données qui sera utilisé pour le projet en utilisant autant que possible les fonctions du package tidyverse et en

a. créant une variable à expliquer Y binaire, qui vaut 1 pour les patientes décédées en raison du cancer et 0 pour les autres.

b. ne gardant que les variables *âge au diagnostic*, *taille de la tumeur* et *nombre de mutations* parmi les variables cliniques (c'est-à-dire autres qu'expressions de gènes et mutations). Les valeurs manquantes de *taille de la tumeur* ou *nombre de mutations* pourront au choix être supprimées ou imputées en utilisant la médiane de la chacune des variables.

c. modifiant les données de mutations en variables binaires comportant un 1 pour chaque mutation, quelque soit le code la décrivant.

2 Prédiction

On s'intéresse d'abord au problème de la prédiction.

2. Choisir trois méthodes de prédiction vues en cours (en expliquant pourquoi vous prenez ces méthodes) ainsi que les méthodes de réduction de dimension suivantes:

a. aucune
b. par ACP (ou PLS si vous connaissez cette méthode)
c. par auto-encodeur

Vous obtenez, suivant le couple *réduction* $\times$ *prédiction* 9 méthodes de prédiction possibles pour votre problème. Discuter de leurs avantages et inconvénients respectifs, et de l'utilité d'une réduction de dimension dans ce cas. Mettre en oeuvre tout ou partie de ces méthodes. Décider de laquelle vous choisiriez et évaluer ses performances.

3 Sélection

On suppose maintenant que l'unité de biologie qui a lancé cette étude est intéressée par la sélection de variables plutôt que par la prédiction.

3. Proposer une ou plusieurs manières d'effectuer cette sélection à l'aide d'une approche de sélection par stabilité. Mettre en oeuvre une telle approche et proposer un ensemble d'ARN/mutations candidats à une étude approfondie.