Analyse de données
M1 Statistique et économétrie
Projet 2017
V. Monbet

Dans ce projet, l'objectif final est de proposer un modèle permettant de prédire quelles sont les personnes dont les revenus sont supérieurs à 50K$ par an à partir de variables socio-démographiques. Les données issues du rescensement de 1994 sont disponibles dans la table `Census94_salaires.Rdata`. La variable à prédire des la variable `income`.

Cette base de données contient un nombre n important d'individus. Dans un premier temps, vous pouvez travailler sur un échantillon de la base de façon à ne pas perdre trop de temps en temps de calcul.

```
set.seed("111")
subs = sample(1:n,round(n/10))
Dsub = D[subs,]
```

Une alternative consiste à travailler avec Python.

Analyse descriptive

Faire une analyse descriptive des données en utilisant, notamment, l'analyse des correspondances (ACM). Vous avez le choix de discrétiser les variables continues ou de les considérer comme variables supplémentaires.
Commentez les résultats.

Classification non supervisée

Proposer une ou plusieurs typologies des individus de la base.

Vous devrez ici utiliser une méthode adaptée au grand nombre d'individus et aux types de variables. Par exemple

1. Réaliser une ACM sur les variables catégorielles.
2. Construire une nouvelle base de données incluant les variables continues et les composantes de l'ACM.
3. Obtenir une partition en 100 classes de cette nouvelle base par la méthode des k-moyennes.

4. Faire une classification hiérarchique ascendante des centres des classes obtenues au points précédent.

Est-ce qu'on peut caractériser les différentes classes ? Y a t'il des classes dans lesquelles les forts revenus sont en forte proportion ? Comment peut-on répartir des individus dans les différentes classes ?

Modélisation

Proposer plusieurs modèles pour prédire la variable `income`. Quelques suggestions sont données ci-dessous, mais on peut bien sûr imaginer d'autres solutions.

— Analyse discriminante linéaire et/ou quadratique appliquée sur les composantes d'une analyse en facteurs
— Classification par la méthode des plus proches voisins
— Régression logistique
— Arbres de décision et forêts aléatoires

Les algorithmes/modèles devront être validés en validation croisée.
Quel est le meilleur modèle au sens du risque de Bayes ?

Prédiction

Le fichier `Census94_salaires_test.Rdata` contient des individus pour lesquels on ne connait pas le niveau de salaire.

1. Prédire le niveau de salaire `<=50K` ou `>50K` avec votre meilleur modèle.
2. Stocker les résultats dans un fichier au format .Rdata (utiliser la commande `save`) dont les noms des lignes sont les numéros d'individus et avec une seule colonne `income`. Le nom du fichier sera composé de votre prénom et de votre nom. Exemple : `valerie_monbet.Rdata`.
3. Envoyer le fichier par mail à `valerie.monbet@univ-rennes1.fr` avant le 4 avril à 8 :00.