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

Dans ce projet, l'objectif final est de dresser un portrait des 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 est la variable `income`.

Toutes vos réponses doivent être **rédigées** et **justifiées**. Les descriptions méthodologiques doivent s'appuyer sur une terminologie statistique adéquate (et non des noms de fonctions R).

1. Questions de cours [4 points]

1. Quel critère est minimisé quand on estime les axes factoriels d'une analyse en composantes principales ? Quel calcul matriciel réalise-t'on pour obtenir les axes factoriels en pratique ?
2. Citer 3 différences importantes entre l'analyse en composantes principales et l'analyse factorielle des correspondances.
3. Qu'est-ce que ces deux méthodes on en commun ?

2. Analyse descriptive [4 points]

1. Décrire et justifier la démarche employée pour réaliser l'analyse des correspondances multiples sur les données du rescensement de 1994 (recodage de variables, variables supplémentaires, choix du nombre d'axes, etc).
2. Commentez les résultats et donner une description synthétique des données.

3. Classification non supervisée [4 points]

1. Décrire et justifier la démarche employée pour obtenir une typologie des individus de la base (différentes étapes, choix des distances, critères de comparaisons, difficultés rencontrées, etc.).
2. Peut-on caractériser les différentes classes ? Y a t'il des classes dans lesquelles les forts revenus sont en forte proportion ou au contraire en faible proportion ?
3. Proposer une méthode pour attribuer une classe à un nouvel individu.

4. Modélisation [8 points]

On construit des modèles pour prédire la variable `income`.

1. Lister les modèles que vous ajustez en indiquant quelle table de données (*) est utilisée ?
   (*) *données d'origine, données recodées (préciser), composantes de l'ACM, centres de classes, ...*
2. Sur quel modèle de loi repose le modèle de l'analyse discriminante linéaire ? Sur quel modèle de loi repose le modèle de l'analyse discriminante quadratique ? Pourquoi (et/ou comment) peut-on utiliser ou non ses modèles dans cette étude ?
3. On ajuste un modèle logistique pour prédire le niveau de salaire en fonction de l'âge, du sexe, de la race, du niveau d'étude et des pertes et gains du capital. Quel estimateur est utilisé pour estimer les paramètres du modèle de régression logistique ?
   On obtient les résultats du Tableau 1. Quelles sont les 3 modalités (ou variables) qui ont l'impact positif (resp. négatif) le plus fort sur le salaire ?

Variable	Modalité ou Moyenne	Estimation	Std. Error	z	Pr(>|z|)
(Intercept)		-2,824e+01	1,758e+02	-0,161	0,872
Age	38,4	4,144e-02	1,942e-03	21,339	<2e-16
Sexe	Male	1,381e+00	6,270e-02	22,018	<2e-16
Race	Asian-Pac-Islander	1,107e-01	3,496e-01	0,317	0,752
	Black	-1,754e-01	3,358e-01	-0,522	0,602
	Other	-7,248e-01	5,404e-01	-1,341	0,180
	White	3,954e-01	3,226e-01	1,226	0,220
Education	1st-4th	2,112e+01	1,758e+02	0,120	0,904
	5th-6th	2,155e+01	1,758e+02	0,123	0,902
	7th-8th	2,160e+01	1,758e+02	0,123	0,902
	9th	2,203e+01	1,758e+02	0,125	0,900
	10th	2,256e+01	1,758e+02	0,128	0,898
	11th	2,220e+01	1,758e+02	0,126	0,899
	12th	2,270e+01	1,758e+02	0,129	0,897
	HS-grad	2,329e+01	1,758e+02	0,133	0,895
	Some-college	2,367e+01	1,758e+02	0,135	0,893
	Assoc-acdm	2,395e+01	1,758e+02	0,136	0,892
	Assoc-voc	2,382e+01	1,758e+02	0,136	0,892
	Bachelors	2,459e+01	1,758e+02	0,140	0,889
	Masters	2,503e+01	1,758e+02	0,142	0,887
	Doctorate	2,555e+01	1,758e+02	0,145	0,884
	Prof-school	2,560e+01	1,758e+02	0,146	0,884
Gain du capital	1058	3,390e-04	1,594e-05	21,265	2e-16
Perte du capital	86	6,979e-04	5,192e-05	13,440	2e-16

TABLE 1 – Résultats de l'ajustement du modèle de régression logistique

4. Quels types de variables explicatives peut-on utiliser quand on construit un arbre de décision ? Sur quels critères sont choisis les divisions des noeuds ?
   A l'aide de la Figure 1, donner une interprétation du modèle d'arbre et prédire le niveau de salaire des individus suivants :
   a. Alfred, agé de 38 ans, né aux État-Unis, divorcé, de niveau d'éducation *HS-grad*, gain sur capital de 0K$, ne vit pas en famille.
   b. Suzie, agée de 31 ans, née aux État-Unis, célibataire, titulaire d'un Master, gain sur capital de 14K$, ne vit pas en famille.

5. Citer deux méthodes de combinaison de modèles et expliquer leur principe (en deux ou trois phrases).
6. Parmi les modèles que vous avez ajustés, quel est le meilleur ? Selon quel critère ? Pensez-vous que vous auriez pu améliorer les résultats ? Si non, pourquoi ? Si oui, comment ?
7. D'après les résultats de l'étude, dresser le profil typique d'un individu dont le revenu est supérieur à 50K$.

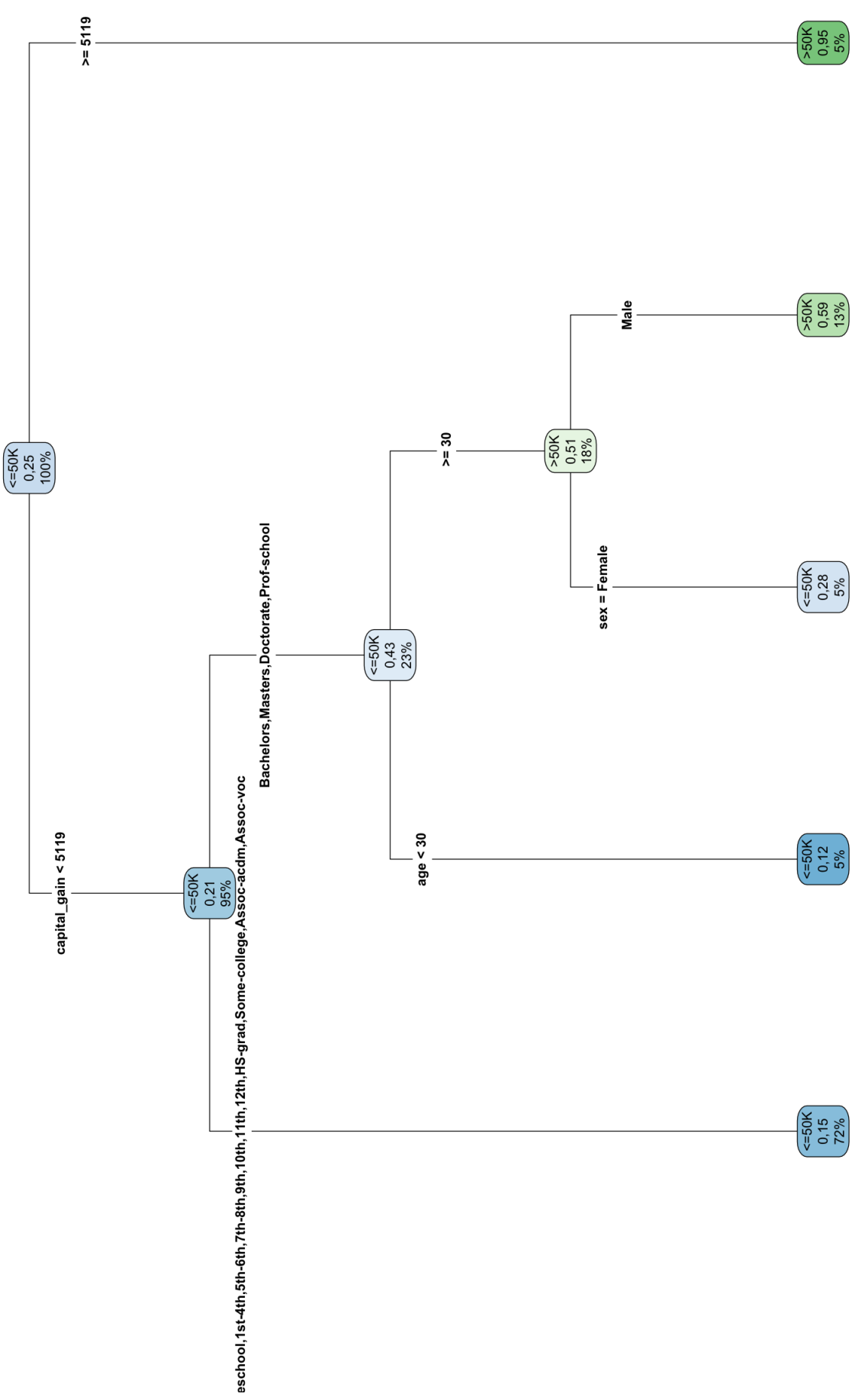

FIGURE 1 – Arbre de décision