Analyse de données
M1 Statistique et économétrie - 2011
C. Herzet, V. Monbet
Mini Projet

Dans ce projet, l'objectif est de discriminer des vins blancs portugais. La variable de groupe correspond à des préférences gustatives variant de 0 (très mauvais) à 10 (excellent). Cette variable est mesurée par des tests sensoriels réalisés par des oenologues. Chaque vin est testé par au moins 3 spécialistes et on attribut le vin au groupe médian des groupes déterminés par les experts. Les variables explicatives sont des variables physico-chimiques liées à la constitution des vins. Certaines variables explicatives peuvent être corrélées et/ou non discriminantes de telle sorte qu'il sera utile de procéder à une sélection de variables.

Les données `winequality-white.csv` sont disponibles sur la page `http://perso.univ-rennes1.fr/valerie.monbet/Cours_AD/AD.html`. Importez les données dans le logiciel que vous avez choisi d'utiliser. Pensez à définir la variable à prédire comme une variable ordinale.

Vous vous attacherez à répondre aux questions suivantes. Vos réponses doivent être rédigées, argumentées clairement et permettre de suivre votre démarche statistique, comme si on lisait un compte rendu d'étude. Par exemple, si vous réalisez des tests statistiques les hypothèses doivent être clairement posées et la statistique de test donnée, ainsi que sa loi. Vos choix (méthodologiques, algorithmiques, etc.) doivent être justifiés. Si vous introduisez des notations, il faut que vous les définissiez.

1 Analyse descriptive

Faîtes une analyse descriptive du jeu de données afin de mettre en évidence des individus aberrants éventuels, la nature, la distribution des variables, des liaisons entre les variables, ou tout autre caractère qui vous semble intéressant. Faites des prétraitements si nécessaire.

2 Classification : représentation synthétique des individus

Proposer une classification (typologie) des vins basée sur les variables physico-chimiques. Décrire la démarche : choix de (ou des) algorithme(s), initialisation, sélection du nombre de classes, etc. Interpréter les classes. Discuter la distribution de la qualité des vins dans les différentes classes. La classification obtenue est-elle cohérente avec celle des experts ? Pouvez-vous expliquer ce résultat ?

3 Préparation à la sélection de variables

Une analyse en composante principale et une classification sur les variables vont permettre de mettre en évidence des corrélations entre les variables physico-chimiques et des sous groupes de variables. Mener ces analyses puis interpreter et discuter les résultats.

4 Analyse discriminante

On va maintenant construire des règles de décision permettant de discriminer les vins selon la variable de préférence.

4.1 Préanalyse

1. Rappeler ce qu'est la MANOVA. En quoi cette analyse peut-être intéressante ici ? Mener une MANOVA et discuter les résultats.
   Sous `R` on peut utiliser la fonction `manova` et sous SAS la `proc discrim` avec l'option `MANOVA`.
2. utiliser l'analyse factorielle discriminante pour proposer une représentation graphique des individus mettant en évidence les classes de préférence. Discuter le (ou les) graphique(s) obtenus.

4.2 Modélisation, première étape

1. Générer une base d'apprentissage et une base de validation.
2. Proposer un modèle basé sur l'analyse de Fisher (c'est à dire en supposant que le vecteur des variables explicatives suit une loi de Gauss). Si c'est opportun, vous ferez une sélection de variables (en expliquant votre démarche). Le modèle est-il linéaire ou quadratique ? Pourquoi ?
   Valider le modèle ; vous donnerez notamment la matrice de contingence, vous estimerez les probabilités de mauvais classement pour chacune des classes etle risque de Bayes.
   Remarque : sous `SAS` la fonction `STEPDISC` permet de faire de la sélection forward ou backward pour le modèle d'analyse discriminante linéaire (analyse discriminante de Fisher). Sous `R`, le package `dprep` propose un analogue par la fonction `sfs` pour l'analyse discriminante de Fisher et la méthode des plus proches voisins.
3. Proposer un modèle non paramétrique basé sur les plus proches voisins. Comment choisissez-vous le nombre optimal de voisins ? Valider le modèle et le comparer au modèle de la question précédente. Quel est le meilleur modèle. Pourquoi ?
4. D'après les questions précédentes, pouvez-vous expliquer quelles sont les propriétés physico-chimique des très bons vins ?

4.3 Modélisation, deuxième étape

On a constaté dans la question précédente que certaines classes sont difficiles à prédire car elles contiennent peu de vins. Si on ne cherche pas à faire une analyse fine des très bons vins (respectivement très mauvais vins), on peut penser à regrouper certaines classes.

1. Proposer un regroupement de classes de façon à n'en conserver que 4.
2. Répondre à nouveau aux questions de la section 4.2.
3. Cas particulier.
   Dans le cas particulier de 4 classes et de deux variables (choisir les plus discriminantes), tracer le nuage de points (une couleur par classe), ajouter en niveaux de couleur la probabilité a posteriori d'être dans le groupe 3 puis ajouter les frontières du modèle linéaire de Fisher (resp. du modèle quadratique). Discuter le graphique obtenu.

4.4 Modélisation, une alternative ?

Comme les classes de préférences sont ordonnées, on aurait pu choisir d'ajuster un modèle de régression linéaire multiple. Qu'en pensez-vous, critiquez cette idée ? Ce modèle peut-il nous aider à répondre à certaines questions précédentes ? Si oui, précisez les réponses. Les résultats obtenus ainsi sont-ils cohérents avec les précédents ?

4.5 Test

Classer, le mieux possible, les 10 vins du fichier `winequality-test.csv`.