Crises cardiaques

November 6, 2023

1 Crises cardiaques

Jeu de données de nature médicale, ou la variable à prédire `output` est 1 dans le cas d'une crise cardiaque, 0 dans le cas d'un individu sain.

1.1 Corrélation

Calculer la corrélation entre la variable `age` et la variable `thalachh` (fréquence cardiaque maximale atteinte). Tracer le scatterplot de ces deux variables et comparer à la valeur de corrélation. Comment interpréter cette valeur ?

1.2 ACP

En excluant la variable `output` (dont la valeur 0 correspond à un individu sain et la valeur 1 à un individu faisant une crise cardiaque), effectuer une ACP sur les variables restantes. Tracer les graphes suivants :

- Variance expliquée par les 15 premières composantes
- Courbe cumulative de la variance expliquée (par toutes les composantes). Interpréter : a-t-on une bonne représentation ?
- Nuage des individus projétés sur les 2 premières composantes principales, colorés par classe (avec ou sans crise cardiaque). Interpréter : peut-on obtenir une bonne classification en n'utilisant que ces deux composantes ?
- Même nuage, colorés selon la qualité de la représentation ($\cos^2$). Interpréter : comment lier cela à la variance expliquée ?
- Premier cercle de corrélation. Interpréter : quelles variables sont bien expliquées dans le premier plan principal ?

1.3 AFD

- Appliquer une AFD linéaire aux mêmes données, en utilisant `output` come variable de référence. Combien d'axes factoriels obtient-on ? Pourquoi ?
- Tracer un graphe de la distribution des projections sur le(s) axe(s) factoriel(s). A-t-on une bonne séparation des deux classes ?
- Calculer l'accuracy de cette méthode de classification et la matrice de confusion. Est-ce satisfaisant ? Si on devait l'utiliser dans la pratique pour prédire si un patient aux urgences est malade, comment pourrait-on le modifier pour réduire le risque pour les patients ?

1.4 SVM

- Entraîner une SVM linéaire pour séparer les deux classes, en utilisant toutes les données
- Calculer l'accuracy et la matrice de confusion. Est-ce meilleur que l'AFD?
- Répéter l'entraînement en n'utilisant que les deux premières composantes principales. Comment change l'accuracy ?
- Tracer la droite de séparation qu'on obtient ainsi dans le plot sur les deux composantes principales.
- Répéter l'entraînement en utilisant, cette fois, une SVM non-linéaire (par exemple, noyau "rbf") sur toutes les données. Est-ce meilleur ?

1.5 Réseaux de neurones

- Entraîner un réseau de neurones avec les paramètres standards. Peut-on dire si c'est mieux des méthodes ci-dessus ?
- Répéter l'entraînement avec beaucoup plus de neurones ; comment change le résultat ?
- Séparer le jeu de données en 80% d'entraînement et 20% de test. Répéter l'entraînement de AFD, SVM (linéaire et non) et les deux réseaux de neurones, en n'entraînant que sur le premier 80% et évaluant sur le reste. Comment changent les résultats ?