Exploration Numérique 3

18 octobre 2023

Nous allons étudier le data set Breast Cancer Wisconsin. Il s'agit d'un problème de classification binaire, la réponse est le diagnotic "M"-malignant et "B"-benign. Il y a $d = 32$ attributs numériques. Le nombre d'observations est $n = 569$.

1. Visualiser les boxplot de chacun des attributs [1 à 10] — Commenter
2. Visualiser la *correlation heatmap* [vous pourrez utiliser `heatmap` de `seaborn`]
3. Commenter.

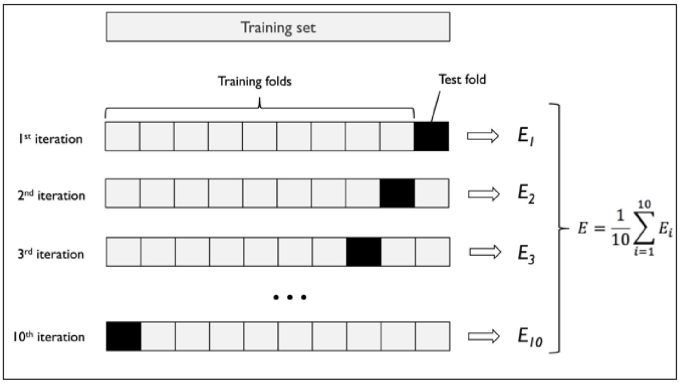
Dans la suite, nous recentrons les régresseurs et les normalisons par leur écart-type empiriques.

Nous allons tout d'abord utiliser une méthode de réduction de dimension afin de réduire le nombre d'attributs. L'approche de base pour réduire la dimension est l'analyse en composantes principales (ACP), qui est conceptuellement assez simple. Nous calculons la matrice de covariance de $d \times d$, $\hat{\mathbf{\Sigma}}$ sur l'ensemble des données. Ensuite, nous calculons les vecteurs propres et les valeurs propres de cette matrice que nous trions par valeur propre décroissante. Nous appelons $\mathbf{e}_i$ le vecteur propre associé à la valeur propre λ_i, Nous choisissons les $\tilde{d}$ plus grands vecteurs propres de ce type. Nous formons une matrice $d \times \tilde{d}$ $\mathbf{A}$ dont les colonnes sont constituées des vecteurs propres $[e_1, \ldots, e_{\tilde{d}}]$. Nous considérons ensuite les données en dimension $\tilde{d}$ définies par

$$\mathbf{x}'_i = \mathbf{A}^\top \mathbf{x}_i \quad i \in \{1, \ldots, n\}$$

Nous prenons tout d'abord $\tilde{d} = 3$.

4. Visualiser les boxplot de chacun des attributs après ACP.
5. Visualiser la *correlation heatmap*
6. Visualiser les données $\{(y_i, \mathbf{x}'_i)\}_{i=1}^n$ [en utilisant des labels différents pour les deux classes]
7. Visualiser les classes en observant les projections sur deux composantes principales $[e_1, e_2]$, $[e_1, e_3]$, $[e_2, e_3]$.
8. Effectuer une analyse linéaire discriminante en considérant les trois composantes.
9. Visualiser la frontière de décision pour chaque paire de composante (trois tracés attendus).
10. Effectuer une analyse quadratique discriminante en considérant les trois composantes.
11. Visualiser la frontière de décision pour chaque paire de composante (trois tracés attendus).



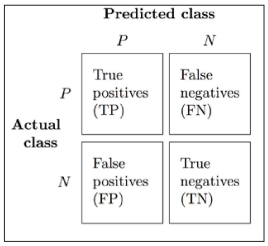


FIGURE 1 – Matrice de confusion

Nous allons maintenant chercher à évaluer les performances de ces deux classifieurs et nous utilisons la méthode de validation croisée. Dans la validation croisée à "k-fold", nous divisons aléatoirement l'ensemble de données en k *fold* sans remplacement, où $k-1$ *fold* sont utilisés pour l'apprentiisage du modèle, et un *fold* est utilisé pour l'évaluation des performances. Cette procédure est répétée k fois afin d'obtenir k modèles et estimations de performance.

Nous calculons ensuite la performance moyenne des modèles basés sur les différents *fold* indépendants afin d'obtenir une estimation de la performance. Étant donné que la validation croisée *k-fold* est une technique de rééchantillonnage sans remplacement, l'avantage de cette approche est que chaque échantillon sera utilisé pour la formation et la validation (dans le cadre d'un *fold*) exactement une fois. La figure ci-dessous illustre le concept de la validation croisée à k *fold* avec $k = 10$.

Une amélioration par rapport à l'approche standard de la validation croisée k-fold est la validation croisée k-fold stratifiée, qui peut donner de meilleures estimations du biais et de la variance, en particulier dans les cas de proportions de classes inégales ; voir [Kohavi et al., 1995]. Dans la validation croisée stratifiée, les proportions de classe sont préservées dans chaque *fold* afin de garantir que chaque *fold* est représentatif des proportions de classe dans l'ensemble de données d'apprentissage. Vous pourrez utiliser `stratifiedKFold` de `scikit-learn`. Nous définissons

— la *précision* (*accuracy*) comme le rapport du nombre de "vrais positifs" et de "vrais négatifs" et du nombre total de cas

— le *rappel* (*recall*) comme le rapport de "vrais positifs" et du nombre de "vrais positifs" et de "faux positifs".

Nous construisons également la matrice de confusion (voir Figure 1).

12. Calculer la précision et le rappel des analyses discriminantes linéaires et quadratiques par validation croisée sur un k-fold avec $k = 10$.
13. Construire la matrice de confusion pour les deux classifieurs.

Références

[Kohavi et al., 1995] Kohavi, R. et al. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Ijcai*, volume 14, pages 1137–1145. Montreal, Canada.