SY09 – P18
Second projet

1 Discrimination

Dans cette première partie, on cherchera à évaluer les performances des principales méthodes de discrimination étudiées dans le cours — analyse discriminante quadratique et linéaire, classifieur bayésien naïf, régression logistique, arbres de décision — sur différents jeux de données binaires : breastcancer, ionosphere, sonar, spambase, et spambase2.

On ne se contentera pas d'afficher des taux d'erreur ou des figures : on cherchera à analyser les résultats obtenus en les expliquant, à la lumière des caractéristiques des jeux de données considérés (géométrie des classes, dimension, quantité de données disponibles, etc). On pourra pour cela s'aider des méthodes étudiées dans la première partie de l'UV.

Si vous rencontrez des difficultés à appliquer certaines méthodes, vous êtes libres d'effectuer des prétraitements à condition de les décrire rigoureusement (et de les justifier dans la mesure du possible) dans votre compte-rendu.

2 Analyse discriminante de données binaires

2.1 Modèle

On cherche à développer un classifieur spécifique aux données « simplifiées » spambase2. Les données spambase correspondent aux données de spams déjà étudiées lors du premier projet. Les variables $X^1, \ldots, X^p$ correspondent aux caractéristiques suivantes :

1. 48 mesures de fréquences de mots pré-définis dans le mail,
2. 6 mesures de fréquences de caractères pré-définis,
3. la longueur moyenne des séquences ininterrompues de lettres majuscules dans le message,
4. la longueur de la plus longue séquence ininterrompue de lettres majuscules,
5. le nombre total de majuscules dans le message.

Les données spambase2 correspondent à une « simplification » de spambase ; les variables ont été transformées en variables binaires $Y^1, \ldots, Y^p$ de la manière suivante :

1. les 48+6 = 54 mesures de fréquence de mots ont été converties en mesures de présence/absence du mot :
$$x_{ij} > 0 \quad \Rightarrow \quad y_{ij} \leftarrow 1;$$
2. les longueurs moyenne et maximale des séquences de lettres majuscules et le nombre total de lettres majuscules dans le message ont été remplacées par 1 si la réalisation originale était supérieure à la médiane des réalisations dans le jeu de données, et 0 sinon :
$$x_{ij} > \underset{i=1,\ldots,n}{\text{médiane}}\,(x_{ij}) \quad \Rightarrow \quad y_{ij} \leftarrow 1.$$

On pourra s'appuyer sur les questions suivantes pour développer un modèle spécifique aux données spambase2.

1. On suppose que pour chaque variable X^j, la proportion (théorique) p_{kj} de valeurs égales à 1 dépend de la classe :
$$p_{kj} = \Pr(X^j = 1 | Z = \omega_k).$$
Quelle est la distribution conditionnelle de X^j étant donné Z ? En déduire la probabilité $\Pr(X^j = x_j | Z = \omega_k)$.
2. En supposant l'indépendance des variables $X^1, \ldots, X^p$ conditionnellement à la classe $Z = \omega_k$, en déduire la probabilité jointe du vecteur aléatoire $\boldsymbol{X}$ conditionnellement à la classe ω_k : $\Pr(\boldsymbol{X} = \boldsymbol{x} | Z = \omega_k)$, où $\boldsymbol{x} = (x_1, \ldots, x_p)^T$ est une réalisation du vecteur aléatoire $\boldsymbol{X}$.
3. En considérant que le i$^{\text{e}}$ exemple d'apprentissage consiste en un couple $(\boldsymbol{x}_i, \boldsymbol{z}_i)$ où $\boldsymbol{x}_i$ est une réalisation de $\boldsymbol{X}$ et $\boldsymbol{z}_i$ une réalisation de $\boldsymbol{Z}$ du vecteur de classe (avec $\boldsymbol{z}_i = (z_{i1}, \ldots, z_{ig})^T$, et $z_{ik} = 1$ si $\boldsymbol{x}_i \in \omega_k$ et $z_{ik} = 0$ sinon), écrire la probabilité jointe $\Pr(\boldsymbol{X} = \boldsymbol{x}_i, \boldsymbol{Z} = \boldsymbol{z}_i)$.
4. En déduire la vraisemblance jointe des paramètres du modèle p_{kj} et $\pi_k = \Pr(Z = \omega_k)$ $(k = 1, \ldots, g,\ j = 1, \ldots, p)$ étant donné l'ensemble d'apprentissage $\{(\boldsymbol{x}_1, \boldsymbol{z}_1), \ldots, (\boldsymbol{x}_n, \boldsymbol{z}_n)\}$.
5. Montrer que l'EMV de chaque paramètre p_{kj} $(k = 1, \ldots, g,\ j = 1, \ldots, p)$ est
$$\widehat{p}_{kj} = \frac{1}{n_k} \sum_{i=1}^{n} z_{ik} x_{ij},$$
avec $n_k = \sum_{i=1}^{n} z_{ik}$ (on supposera pour cela que $0 < p_{kj} < 1$ pour tout $k = 1, \ldots, g$ et $j = 1, \ldots, p$), et que l'EMV de chaque probabilité a priori est
$$\widehat{\pi}_k = \frac{1}{n} \sum_{i=1}^{n} z_{ik} = \frac{n_k}{n}.$$

2.2 Programmation

On pourra programmer deux fonctions permettant de faire l'implémentation du modèle décrit ci-dessus. La fonction binaryNBCfit permet de faire l'apprentissage : il prend comme arguments d'entrée un tableau individus-variables X et un vecteur d'indicateurs de classe z, et doit retourner les paramètres du modèle : le vecteur des probabilités a priori pik et le vecteur des paramètres pkj.

La fonction binaryNBCval évalue un ensemble de données de test au moyen des paramètres du modèle : elle doit prendre en arguments d'entrée le vecteur des probabilités a priori pik et les paramètres pkj de même que le tableau individus-variables des individus de test Xtst, et doit retourner le vecteur prob des probabilités a posteriori calculées et les prédictions associées pred.

2.3 Test

Utiliser la procédure habituelle pour évaluer les performances du modèle sur les données spambase2. Comparer avec les autres modèles utilisés. Que peut-on remarquer ? Quelles performances obtient-on en comparaison de celles obtenues sur les données spambase ?