Data Science

Apprentissage statistique

Exercice I [Détection de fraude]

L'apprentissage automatique peut être utilisé pour détecter les fraudes : l'exercice ci-après en est une illustration très simple.

On dispose d'un dé à 6 faces, parfaitement équilibré. On confie ce dé à des individus en leur demandant de procéder à un certain nombre de lancers et de faire part de leurs résultats. La population est composée de personnes honnêtes (H) qui font exactement ce qu'on leur demande, mais aussi d'un certain nombre de tricheurs (T) qui, chaque fois qu'on leur demande de lancer une fois le dé, le lancent en réalité deux fois et annoncent le plus grand des nombres obtenus. Ainsi, si l'on demande à un tricheur de lancer 5 fois le dé, il pourra obtenir la suite de résultats (2,2), (5,2), (4,1), (5,4), (6,3), et annoncer (2,5, 4, 5, 6).

1. Calculer $p(i|H)$ et $p(i|T)$ pour $i = 1$ à 6.
2. Calculer $p(25456|H)$ et $p(25456|T)$.
3. On suppose que la population contient 10% de tricheurs. Que doit-on décider sur l'honnêteté d'un individu qui annonce 25456 si l'ont suit respectivement :
   (a) la règle majoritaire,
   (b) la règle du maximum de vraisemblance,
   (c) la règle de décision de Bayes ?

Exercice II [Prédire l'issue d'un match]

Le tableau ci-après récapitule les conditions qui ont accompagné les succès et les échecs d'une équipe de football. Est-il possible de prédire l'issue d'un match en fonction des conditions dans lesquelles il se déroule ?

Les conditions d'un match sont modélisées par un élément $\mathbf{x}$ de $\mathcal{X} = \{V, F\}^4$, correspondant aux valeurs des attributs figurant sur la première ligne du tableau. D'après la règle de classification de Bayes, il suffit de connaître $P(V|\mathbf{x})$ pour pouvoir classer $\mathbf{x}$ de manière optimale : $f(\mathbf{x}) = V$ si $P(V|\mathbf{x}) \geq 1/2$ et $f(\mathbf{x}) = F$ sinon.

D'après la formule de Bayes, on a :

$$P(V|\mathbf{x}) = \frac{P(\mathbf{x}|V)P(V)}{P(\mathbf{x})} \text{ et } P(F|\mathbf{x}) = \frac{P(\mathbf{x}|F)P(F)}{P(\mathbf{x})}$$

Match à domicile ?	Balance positive ?	Mauvaises conditions climatiques ?	Match précédent gagné ?	Match gagné
V	V	F	F	V
F	F	V	V	V
V	V	V	F	V
V	V	F	V	V
F	V	V	V	F
F	F	V	F	F
V	F	F	V	F

FIGURE 1 – Jeu de données *FootBall*.

soit encore

$$P(V|\mathbf{x}) \geq 1/2 \text{ ssi } P(\mathbf{x}|V)P(V) \geq P(\mathbf{x}|F)P(F).$$

On peut évaluer $P(V)$ et $P(F)$ en comptant le nombre de matchs gagnés et perdus :

$$\hat{P}(V) = 4/7 \text{ et } P(F) = 3/7.$$

L'évaluation de $P(\mathbf{x}|V)$ et de $P(\mathbf{x}|F)$ est plus délicate. La règle *naïve* de Bayes consiste à faire l'hypothèse que les attributs décrivant $\mathbf{x}$ sont indépendants conditionnellement à chaque classe : si l'on écrit $\mathbf{x} = (x_1, x_2, x_3, x_4)$, on suppose que

$$P(\mathbf{x}|V) = \prod_{i=1}^{4} P(x_i|V) \text{ et } P(\mathbf{x}|F) = \prod_{i=1}^{4} P(x_i|F).$$

Pour estimer $P(\mathbf{x}|V)$ et $P(\mathbf{x}|F)$, il suffit alors d'estimer $P(x_i = V|V)$ et $P(x_i = V|F)$ pour $i = 1, \ldots, 4$.

1. Réaliser ces estimations
2. Classer l'élement (V, F, V, F)

Exercice III

Supposons que vous avez à disposition un jeu de données généré par une fonction polynomiale de degé 3. Caractériser les erreurs d'approximation et d'estimation des modèles dans le tableau ci-dessous en encerclant la réponse correcte.

	erreur d'approximation	erreur d'estimation
régression linéaire	faible/élevée	faible/élevée
régression polynomiale de degré 3	faible/élevée	faible/élevée
régression polynomiale de degré 10	faible/élevée	faible/élevée