M2 DataScience **Apprentissage statistique**

APM_5DS27_TP **Stephan Clémençon - Ekhine Irurozki**

TRAVAUX DIRIGÉS N° 1 : Classifieur de Bayes

Stephan CLÉMENÇON <stephan.clemencon@telecom-paris.fr>

Ekhine IRUROZKI <irurozki@telecom-paris.fr>

On se place dans le cadre de la classification binaire. Dans tout le TD, on considère un descripteur aléatoire X à valeurs dans un espace mesurable $\mathcal{X} \subset \mathbb{R}^d$ ($d \in \mathbb{N}^*$) et un label aléatoire Y valant 0 ou 1. La distribution jointe du vecteur (X, Y) est notée P, et la fonction de régression (probabilité a posteriori)

$$\eta : x \in \mathcal{X} \mapsto \mathbb{P}(Y = 1 \mid X = x) \in [0, 1].$$

EXERCICE 1. On considère $\mathcal{X} = [0, 1]$ et P telle que :

- la distribution conditionnelle de X sachant Y = 0 est $P_0 = \mathcal{U}([0, \theta])$ où $\theta \in]0, 1[$,
- la distribution conditionnelle de X sachant Y = 1 est $P_1 = \mathcal{U}([0, 1])$,
- $p = \mathbb{P}(Y = 1) \in]0, 1[$.

Pour $x \in \mathcal{X}$, donner $\eta(x)$ en fonction de p et θ.

EXERCICE 2. On considère $\mathcal{X} = \mathbb{R}_+$ et P telle que :

- la distribution de X est P_X sur $\mathcal{X}$,
- la fonction de régression vaut $\eta(x) = \dfrac{x}{x + \theta}$ pour tout $x \in \mathcal{X}$, où $\theta > 0$ est fixé.

On note h^* le classifieur de Bayes. Expliciter le classifieur de Bayes dans ce modèle. Montrer ensuite que son risque "0-1" (sa probabilité d'erreur) vaut

$$L(h^*) = \int_{\mathcal{X}} \min\Big(\eta(x), 1 - \eta(x)\Big)\, dP_X(x).$$

Calculer le risque de Bayes lorsque $P_X = \mathcal{U}([0, \alpha\theta])$ où $\alpha > 1$.

EXERCICE 3. Soient des poids $\omega(0), \omega(1) \geq 0$ tels que $\omega(0) + \omega(1) = 1$. On considère le risque de classification pondéré :

$$L_\omega(g) = \mathbb{E}\left(2\omega(Y) \cdot \mathbb{1}_{\{Y \neq g(X)\}}\right), \qquad g : \mathcal{X} \to \{0, 1\}.$$

Donner le classifieur de Bayes et le risque de Bayes pour ce critère. Quel est l'intérêt de considérer un tel critère ?

Exercice 4. On considère $X = (T, U, V)$ où T, U, V sont des variables aléatoires réelles i.i.d. de loi exponentielle standard. On pose $Y = \mathbb{1}_{\{T+U+V<\theta\}}$ où $\theta \in \mathbb{R}_+$ est fixé.

1) Calculer le classifieur de Bayes $g^*(T, U)$, lorsque V n'est pas observée. Calculer le risque de Bayes "0-1" associé à ce classifieur.

2) On suppose à présent que seule T est observée. Reprendre les calculs précédents et comparer les risques bayésiens obtenus lorsque $\theta = 9$.

3) Proposer un classifieur lorsque X n'a aucune composante qui soit observée. Calculer son erreur de classification.

Conseils bibliographiques

Vous trouverez ci-dessous quelques points d'entrée utiles pour l'apprentissage automatique :

- Théorique et porté sur les aspects probabilistes : [DGL97]
- Utilitaire et porté sur les aspects pratiques : [HTF13]
- Livre récent porté essentiellement sur l'aspect optimisation : [SSBD14] (et du même auteur sur l'apprentissage en ligne [SS12])
- Méthodes Bayésiennes et modèles graphiques : [MB12]

Références

[DGL97] L. Devroye, L. Györfi, and G. Lugosi. *A Probabilistic Theory of Pattern Recognition.* Stochastic Modelling and Applied Probability. Springer New York, 1997. 2

[HTF13] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning : Data Mining, Inference, and Prediction.* Springer Series in Statistics. Springer New York, 2013. 2

[MB12] K.P. Murphy and F. Bach. *Machine Learning : A Probabilistic Perspective.* Adaptive Computation and Machine Learning Series. MIT Press, 2012. 2

[SS12] S. Shalev-Shwartz. *Online Learning and Online Convex Optimization.* Foundations and Trends in Machine Learning Series. Now Publishers, 2012. 2

[SSBD14] S. Shalev-Shwartz and S. Ben-David. *Understanding Machine Learning : From Theory to Algorithms.* Cambridge University Press, 2014. 2