TD1 : Consistance universelle et classifieur kNN

1 Consistance universelle uniforme

On considère le problème de classification binaire avec

$$(X_i, Y_i) \overset{\text{iid}}{\sim} P, \qquad i = 1, \ldots, n$$

où P est une loi de probabilité sur $\mathcal{X} \times \mathcal{Y}$ avec $\mathcal{Y} = \{0; 1\}$. La qualité de prédiction de la valeur y par la valeur y' est mesurée

$$\ell(y, y') = \mathbb{1}(y \neq y').$$

Le risque d'une fonction de prédiction $g : \mathcal{X} \to \mathcal{Y}$ est alors calculé par

$$R_P(g) = \mathbf{E}_P\big[\ell\big(Y, g(X)\big)\big] = \int_{\mathcal{X}\times\mathcal{Y}} \ell\big(y, g(x)\big)\, dP(x, y).$$

Rappelons que ce risque est minimisé par le classifieur de Bayes défini par $g_P^*(x) = \mathbb{1}(\eta^*(x) > 1/2)$ où $\eta^*(x) = \mathbf{E}_P[Y|X = x]$. Soit $\widehat{g}_n : (\mathcal{X} \times \mathcal{Y})^n \to \mathcal{F}(\mathcal{X}, \mathcal{Y})$ un classifieur. On dit qu'il est uniformément universellement consistant en probabilité, si $\forall \delta > 0$,

$$\sup_P P\Big(|R_P(\widehat{g}_n) - R_P(g_P^*)| > \delta\Big) \xrightarrow[n\to\infty]{} 0. \tag{1}$$

On suppose que $\mathcal{X}$ est fini : $\mathrm{Card}(\mathcal{X}) = K$.

1. Quel est le cardinal de $\mathcal{F}(\mathcal{X}, \mathcal{Y})$?
2. Rappeler la borne de risque obtenue en cours pour le minimiseur du risque empirique $\widehat{g}_{n,\mathcal{G}}$ pour $\mathcal{G} = \mathcal{F}(\mathcal{X}, \mathcal{Y})$. Peut-on en déduire que $\widehat{g}_{n,\mathcal{G}}$ est uniformément universellement consistant ?
3. On suppose maintenant que $K = K_n$ dépend de la taille de l'échantillon. Montrer que si K_n est sous-linéaire en n, alors $\widehat{g}_{n,\mathcal{G}}$ est uniformément universellement consistant.

2 Consistance de l'algorithme kNN

Le but de cet exercice est de montrer que l'algorithme kNN employé avec $k = 1$ n'est pas consistent. Pour cela, nous considérons le problème de classification binaire avec $\mathcal{X} = [0, 1]$ et $\mathcal{Y} = \{0; 1\}$. On note P_X la loi marginale des X_i et suppose que

$$\eta^*(x) = P(Y_1 = 1 | X_1 = x) \equiv \frac{3}{4}, \qquad \forall x \in \mathcal{X}.$$

L'objectif des questions suivantes est de calculer le risque du classifieur oracle g_P^* ainsi que celui du classifieur kNN $\widehat{g}_{n,k}$ avec $k = 1$. On verra que ce dernier ne dépend pas de la taille de l'échantillon et est strictement plus grand que le risque de l'oracle.

1. Montrer que pour tout application déterministe $g : \mathcal{X} \to \{0;1\}$, on a

$$R_P(g) = \mathbf{E}_{P_X}[\eta^*(X)] + \mathbf{E}_{P_X}[g(X)(1-2\eta^*(X))].$$

2. En déduire que si $\eta^* \equiv 3/4$, alors le classifieur oracle (appelé aussi classifieur de Bayes) est donné par $g_P^* \equiv 1$ et son risque vaut $R_P(g_P^*) = 1/4$.
3. Montrer que pour tout $g : \mathcal{X} \to \{0;1\}$, on a

$$R_P(g) = \frac{3}{4} - \frac{1}{2}\int_{\mathcal{X}} g(x)\, P_X(dx).$$

4. Soit $\mathcal{D}_n = \{(X_i, Y_i); i = 1, \ldots, n\}$ et $\widehat{g}_{n,1}(x) = \widehat{g}_{\text{PPV}}(x, \mathcal{D}_n)$ le classifieur du plus proche voisin (PPV). Fixons $x \in \mathcal{X}$ et cherchons à calculer $\mathbf{E}_P[\widehat{g}_{\text{PPV}}(x, \mathcal{D}_n)]$, où l'espérance est par rapport à l'échantillon $\mathcal{D}_n$. Pour tout $i = 1, \ldots, n$, posons

$$Z_i = \begin{cases} 1, & \text{si } X_i \text{ est le PPV de x} \\ 0, & \text{sinon.} \end{cases}$$

   Montrer que

$$\mathbf{E}_P[\widehat{g}_{\text{PPV}}(x, \mathcal{D}_n)] = \sum_{i=1}^{n} \mathbf{E}_P[Y_i Z_i]. \tag{2}$$

5. Vérifier que pour tout i, Y_i est indépendant de $(X_1, \ldots, X_n)$. En déduire que Y_i et Z_i sont indépendantes.
6. En utilisant la question précédente et la relation évidente $\sum_{i=1}^n Z_i = 1$ montrer que

$$\mathbf{E}_P[R_P(\widehat{g}_{\text{PPV}})] = \frac{3}{8}.$$

   Conclure.
7. Considérer le cas des 3 plus proches voisins $\widehat{g}_{3-\text{PPV}}$. Montrer que son risque moyen $\mathbf{E}_P[R_P(\widehat{g}_{3-\text{PPV}})]$ est égal à $21/64$.
8. Passons maintenant au cas général d'un prédicteur kNN $\widehat{g}_{k-\text{PPV}}$. Soient $V_1, \ldots, V_k$ des variables aléatoires i.i.d. de loi de Bernoulli de paramètre $3/4$. Montrer que

$$\mathbf{E}_P[\widehat{g}_{k-\text{PPV}}(x, \mathcal{D}_n)] = \mathbf{P}(\overline{V}_k > 1/2).$$

   En déduire que cette espérance tend vers 1 lorsque $k \to \infty$ et, par conséquent, le risque espéré $\mathbf{E}_P[R_P(\widehat{g}_{k-\text{PPV}})]$ tend vers le risque de l'oracle, c'est à dire vers $1/4$.