M2 MApI3

Big Data
Feuille 1 (Risque empirique/Risque de population)

Le compromis entre complexité des modèles et nombre de données est au centre de la problématique de l'apprentissage statistique. Dans ce TD/TP, nous allons illustrer cette problématique sur un problème jouet de classification.

Le problème est le suivant: on considère un ensemble $A \subset [0,1]^2$. On suppose que X est une variable uniforme sur $[0,1]^2$ et que la classe Y de X est donnée par $Y = \mathbf{1}_A(X)$ où $\mathbf{1}_A(x) = 1$ si $x \in A$ et -1 sinon. On dispose d'un l-échantillon d'apprentissage

$$((x_1, y_1), \dots, (x_l, y_l))$$

de loi (X, Y). Il s'agit au vu de cet échantillon de choisir une fonction $\widehat{h} : [0,1]^2 \to \{-1, 1\}$ pour laquelle la classe Y de X est bien prédite par $\widehat{h}(X)$.

On définit une famille de modèles $(\mathcal{F}_p)_{p\in\mathbb{N}}$ de la façon suivante: pour tout $p \in \mathbb{N}^*$, on considère le découpage en p^2 "carreaux" $(c_{ij})_{1\leq i,j\leq p}$ de $[0,1]^2$ où

$$c_{ij} \doteq [\frac{i-1}{p}, \frac{i}{p}] \times [\frac{j-1}{p}, \frac{j}{p}] \,.$$

La famille $\mathcal{F}_p$ est alors définie comme la famille des classifieurs $g = \mathbf{1}_C$ pour lesquel C est constitué d'une réunion finie (éventuellement vide) de carreaux c_{ij}.

On définit le risque empirique et le risque en population à l'aide de la fonction de perte $L(y, y') = \overline{\mathbf{1}}_{y\neq y'}$, avec

$$\overline{\mathbf{1}}_C = \begin{cases} 1 & \text{, si } C \text{ est vraie} \\ 0 & \text{, sinon.} \end{cases}$$

On a donc, pour tout $g \in \mathcal{F}_p$,

$$R_{\text{emp}}(g) = \frac{1}{l}\sum_{i=1}^{l} \overline{\mathbf{1}}_{g(x_i)\neq y_i} \qquad \text{et} \qquad R(g) = \mathbb{E}(\overline{\mathbf{1}}_{g(X)\neq Y}).$$

On note

$$\widehat{R}_p^* \doteq \min_{g\in\mathcal{F}_p} R_{\text{emp}}(g) \,.$$

Exercice 1.

(1) Calculer le cardinal de $\mathcal{F}_p$.

(2) Pour tout $1 \leq i, j \leq p$, on note

$$\widehat{l}_{ij}^{+} \doteq \sum_{k=1}^{l} \overline{\mathbf{1}}_{x_k\in c_{ij} \text{ et } y_k=1} \qquad \text{et} \qquad \widehat{l}_{ij}^{-} \doteq \sum_{k=1}^{l} \overline{\mathbf{1}}_{x_k\in c_{ij} \text{ et } y_k=-1} \,.$$

(a) Calculer $\mathbb{E}(\widehat{l}_{ij}^{+})$ et $\mathbb{E}(\widehat{l}_{ij}^{-})$.

(b) On note

$$\widehat{C}_p \doteq \bigcup_{i,j \,\mid\, \widehat{l}_{ij}^{+}\geq\widehat{l}_{ij}^{-}} c_{ij} \,.$$

Montrer que $R_{\text{emp}}(\mathbf{1}_{\widehat{C}_p}) = \widehat{R}_p^*$, i.e. $\mathbf{1}_{\widehat{C}_p}$ minimise le risque empirique dans $\mathcal{F}_p$.

(3) Montrer que pour tout $\epsilon > 0$,

$$\mathbb{P}(|R(\mathbf{1}_{\widehat{C}_p}) - \widehat{R}_p^*| > \epsilon) \leq \sum_{g\in\mathcal{F}_p} \mathbb{P}(|R(g) - R_{\text{emp}}(g)| > \epsilon) \,.$$

En déduire que $\mathbb{P}(|R(\mathbf{1}_{\widehat{C}_p}) - \widehat{R}_p^*| > \epsilon) \to 0$ lorsque l tend vers $+\infty$.

(4) Déterminer $\mathbf{1}_{C_p^*} \in \mathcal{F}_p$ tel que $R(\mathbf{1}_{C_p^*}) = R_p^*$ où

$$R_p^* = \inf_{g\in\mathcal{F}_p} R(g) \,.$$

(5) Montrer que $\mathbb{P}(|R(\mathbf{1}_{\widehat{C}_p}) - R^*_p| > \epsilon) \to 0$ lorsque l tend vers $+\infty$. *On montre ainsi la consistance de la minimisation du risque empirique sur la famille $\mathcal{F}_p$.*

(6) On sait (par l'inégalité de Hoeffding) que si $Z_1, \cdots, Z_m$ sont m variables indépendantes à valeurs dans $[0,1]$, alors

$$\mathbb{P}(|\frac{S_m}{m} - \mathbb{E}(\frac{S_m}{m})| > \epsilon) \leq 2\exp(-2m\epsilon^2)\,,$$

où $S_m \doteq \sum_{i=1}^m Z_i$.

On considère un nouvel échantillon de taille m, $(X'_i, Y'_i)_{1\leq i\leq m}$ indépendant de l'échantillon d'apprentissage. On note alors pour tout $C \subset [0,1]^2$,

$$R_{\text{test}}(\mathbf{1}_C) \doteq \frac{1}{m}\sum_{i=1}^m \overline{\mathbf{1}}_{\mathbf{1}_C(X'_i)\neq Y'_i}\,.$$

(a) Montrer que pour tout $m \geq -\frac{\log(\eta/2)}{2\epsilon^2}$, on a

$$\mathbb{P}(|R_{\text{test}}(\mathbf{1}_C) - R(\mathbf{1}_C)| > \epsilon) \leq \eta\,.$$

(b) On pose $m_0(\eta,\epsilon) = \lceil -\frac{\log(\eta/2)}{2\epsilon^2}\rceil$ où $\lceil x \rceil$ désigne le plus petit entier plus grand ou égal à x. Calculer la valeur de m_0 pour $\eta = 0.05$ et $\epsilon = 0.02$.

(c) Expliquer pourquoi on a

$$\mathbb{P}(|R_{\text{test}}(\mathbf{1}_{\widehat{C}_p}) - R(\mathbf{1}_{\widehat{C}_p})| > \epsilon) \leq \eta\,. \tag{1}$$

(d) On considère la valeur de m_0 définie par $\eta = 0.05$ et $\epsilon = 0.02$. Expliquer dans quelle mesure on peut considérer $R_{\text{test}}(\mathbf{1}_{\widehat{C}_p})$ comme une approximation acceptable de $R(\mathbf{1}_{\widehat{C}_p})$.

Expérience 1. On veut ici implémenter en python le calcul de $\widehat{C}_p$ et mesurer la différence entre le risque $R_{\text{emp}}(\mathbf{1}_{\widehat{C}_p})$ évalué sur l'échantillon et le risque de population $R(\mathbf{1}_{\widehat{C}_p})$.

L'ensemble A de départ est construit comme un ensemble de niveau d'une fonction $f_{g,s}$ définie par

$$f_{g,s}(x) = \sum_{i=1}^n \exp(-\frac{|x-g_i|^2}{2s_i^2})\,,$$

où $g = (g_i)_{1\leq i\leq n}$ est une famille de points tirés au hasard uniformément dans le carré $[0.2, 0.8]^2$ et $s = (s_i)_{1\leq i\leq n}$ est une famille de paramètres positifs tirés au hasard uniformément dans l'intervalle $[0,a]$. On définit alors

$$A \doteq \{\ x \in [0,1]^2 \mid f_{g,s}(x) > \frac{1}{2}\ \}$$

Le jeu sur le nombre n de points et sur la valeur de a permet de construire des ensembles A de formes variées.

(1) Écrire une fonction `Y=intens(X1,X2,g,s)` qui retourne la classe `Y` des points de $[0,1]^2$ dont les deux coordonnées sur chacun des axes sont données respectivement par les tableaux `X1` et `X2`

(2) Écrire une fonction `[g,s]=ensalea(n,a,flag)` qui retourne un tirage de g et s en fonction de n et a. Si la variable `flag` vaut 1 alors un affichage grossier de A est effectué.

Dans la suite, on pourra choisir $n = 4$ et $a = 0.3$ qui donnent des ensembles A raisonnables.

(3) En utilisant la fonction `[g,s]=ensalea(4,0.3,1)`, sélectionner un ensemble A à votre convenance.

(4) Construire la fonction `[X1, X2, Y]=echant(l,g,s)` retournant une réalisation d'un l-échantillon d'apprentissage $(X_i, Y_i)_{1\leq i\leq l}$ où $X_i =$`(X1(i),X2(i))` et $Y_i =$`Y(i)`

(5) Construire la fonction `[B,Re] = estens(X1,X2,Y,p)` qui pour tout échantillon d'apprentissage `[X1, X2, Y]` et toute valeur de $p \in \mathbb{N}^*$ renvoie une matrice `B` de taille $p \times p$ et un scalaire `Re` définis par:

- `B(i,j)=1` si $\widehat{l}_{ij}^{+} \geq \widehat{l}_{ij}^{-}$ et 0 sinon.
- `Re`$=\widehat{R}_p^*$

Notons que `B` code l'ensemble des c_{ij} qui participent à la construction de $\widehat{C}_p$. Faire en sorte d'afficher le résultat afficher votre résultat.

(6) Construire la fonction `R=testens(B,m,g,s)` approximant la valeur de $R(\mathbf{1}_{\widehat{C}_p})$ par $R_{\text{test}}(\mathbf{1}_{\widehat{C}_p})$ sur un échantillon $(X_i', Y_i')_{1 \leq i \leq m}$ indépendant de l'échantillon d'apprentissage $(X_i, Y_i)_{1 \leq i \leq l}$.

(7) Pour chaque valeur de $l \in \{100, 500, 1000, 10000\}$:

- tracer sur un même graphique les courbes $p \rightarrow \widehat{R}_p^*$ et $p \rightarrow R_{\text{test}}(\mathbf{1}_{\widehat{C}_p})$ (on pourra considérer les valeurs de p entre 2 et 60).
- Calculer $\widehat{p}$ minimisant $p \rightarrow R_{\text{test}}(\mathbf{1}_{\widehat{C}_p})$, où $m = m_0$ est donné dans la question (6b) de l'exercice 1.

(8) Commenter :

- l'évolution des deux courbes à l fixé, p variant
- l'évolution de $\widehat{p}$ en fonction de l
- l'évolution de $R_{\text{test}}(\mathbf{1}_{\widehat{C}_{\widehat{p}}})$ en fonction de l, à p fixé.

En particulier comment cette expérience illustre

- le compromis entre complexité des modèles et taille de l'échantillon d'apprentissage.
- la consistance de la minimisation du risque empirique sur la famille $\mathcal{F}_p$.