TP1 Big Data

Perrot/Deporte

Novembre 2020

1 Rappel Enoncé

On considère $A \subset [0,1]^2$ un ensemble quelconque, et X une V.A. de loi uniforme sur $[0,1]^2$.

La classe Y de X est donnée par $Y = \mathbf{1}_A(X)$, avec $\mathbf{1}_A(x) = 1$ si $x \in A$ et 0 sinon.

On note le training set de taille l : $((x_1^{(1)}, x_2^{(1)}), (x_1^{(2)}, x_2^{(2)}), \ldots, (x_1^{(l)}, x_2^{(l)}))$, de loi (X, Y).

Pour $p \in \mathbb{N}*$, on considère que $[0,1]^2$ est découpé en p^2 carreaux réguliers $(c_{ij}), 1 \leq i, j \leq p$ avec :

$$c_{ij} = \left[\frac{i-1}{p}, \frac{i}{p}\right] \times \left[\frac{j-1}{p}, \frac{j}{p}\right]$$

On considère la famille de modèles $\mathcal{F}_p$ induite par $p \in \mathbb{N}*$, la famille de classifiers $g = \mathbf{1}_C$, où C est une réunion quelconque de c_{ij}.

On définit la loss-function $L(y, y') = \mathbf{1}_{y \neq y'}$, qui permet de calculer, pour tout $g \in \mathbb{F}_p$ le risque empirique (= sur le training set) :

$$\hat{R}(g) = \frac{1}{l} \sum_{i=1}^{l} \mathbf{1}_{g(x^{(i)} \neq y_i}$$

et le risque général (sur la population) :

$$R(g) = \mathbb{E}(\mathbf{1}_{g(X) \neq Y})$$

On note

$$\hat{R}_p^* = \min_{g \in \mathcal{F}_p} \hat{R}(g)$$

2 Questions et Réponses

2.1 Question 1

2.1.1 Enoncé

Calculer le cardinal de $\mathcal{F}_p$

2.1.2 Réponse

On peut voir que $\mathcal{F}_p$ est formé des classifiers qui donnent 0 ou 1 sur chacun des p^2 carreaux de $[0,1]^2$, et donc $|\mathcal{F}_p| = 2^{p^2}$

On peut aussi voir qu'un classifier de $\mathcal{F}_p$ est entièrement défini par la liste des carreaux c_{ij} sur lesquels il indique 1, et que cette liste est un tirage quelconque de k carreaux ($0 \leq k \leq p^2$) parmis les p^2, d'où $|\mathcal{F}_p| = C^1_{p^2} + C^2_{p^2} + \ldots + C^{p^2}_{p^2} = (1+1)^{p^2} = 2^{p^2}$

2.2 Question 2

Pout tout $1 \leq i, j \leq p$, on note :

$$\hat{l}^+_{ij} = \sum_{k=1}^{l} \mathbf{1}_{x^{(k)} \in c_{ij}, y_k = 1}$$

$$\hat{l}^-_{ij} = \sum_{k=1}^{l} \mathbf{1}_{x^{(k)} \in c_{ij}, y_k = -1}$$

qui sont respectivement le nombre de points de l'échantillon présents dans le carreau c_{ij} avec la classe positive $\hat{l}^+_{ij}$, ou négative $\hat{l}^-_{ij}$.

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

2.2.2 Réponse

Pour i, j, k fixés, on a :

$$\begin{aligned}
\mathbb{E}(\mathbf{1}_{x^{(k)} \in c_{ij}, y_k = 1}) &= 1 \times \mathbb{P}(x^{(k)} \in c_{ij}) \times \mathbb{P}(y_k = 1) && (1)\\
&= \mathbb{P}(x^{(k)} \in c_{ij}) \times \mathbb{P}(x^{(k)} \in A) && (2)\\
&= \mathbb{P}(x^{(k)} \in c_{ij} \cap A) && (3)\\
&= |c_{ij} \cap A| && (4)
\end{aligned}$$

Par linéarité de l'espérance :

$$\hat{l}^+_{ij} = l|c_{ij} \cap A|$$

Et

$$\hat{l}^-_{ij} = l|c_{ij} \cap A^c|$$

Avec A^c complémentaire de A : $x^{(k)} \in A^c \Leftrightarrow y_k = -1$

2.2.3 (b) Montrer que $\hat{R}(\mathbf{1}_{\hat{C}_p}) = \hat{R}_p^*$

On note

$$\hat{C}_p = \bigcup_{i,j|\hat{l}_{ij}^+ > \hat{l}_{ij}^-} c_{ij}$$

C'est la réunion des c_{ij} où il y a plus de points de l'échantillon avec une classe positive que de points avec la classe négative.

A p donné, c'est cette construction qui donne le classifier $\mathbf{1}_{\hat{C}_p}$ minimisant le risque empirique.

Preuve :

On considère un carrelage quelconque $C = \underset{i,j\in I_C,J_C}{c_{ij}}$, qui définit un classifier de $\mathcal{F}_p$: $g = \mathbf{1}_C$

Le risque empirique de ce classifier s'écrit :

$$\hat{R}(g) = \frac{1}{l}\sum_{k=1}^{l} \mathbf{1}_{g(x^{(k)})\neq y_k} \tag{5}$$

On remarque que :

$$\mathbf{1}_{g(x^{(k)})\neq y_k} = \mathbf{1}_{g(x^{(k)})=1, y_k=-1} + \mathbf{1}_{g(x^{(k)})=-1, y_k=+1}$$

D'où :

$$
\begin{aligned}
\hat{R}(g) &= \frac{1}{l}\sum_{k=1}^{l} \mathbf{1}_{g(x^{(k)})\neq y_k} && (6)\\
&= \frac{1}{l}\sum_{k=1}^{l} \left(\mathbf{1}_{g(x^{(k)})=1, y_k=-1} + \mathbf{1}_{g(x^{(k)})=-1, y_k=+1}\right) && (7)\\
&= \frac{1}{l}\sum_{k=1}^{l} \left(\mathbf{1}_{x^{(k)}\in C, y_k=-1} + \mathbf{1}_{x^{(k)}\notin C, y_k=+1}\right) && (8)\\
&= \frac{1}{l}\sum_{k=1}^{l} \left(\sum_{i,j\in I_C,J_C} \mathbf{1}_{x^{(k)}\in c_{ij}, y_k=-1} + \sum_{i,j\notin I_C,J_C} \mathbf{1}_{x^{(k)}\in c_{ij}, y_k=+1}\right) && (9)\\
&= \frac{1}{l}\sum_{k=1}^{l} \left(\sum_{i,j\in I_C,J_C} \hat{l}_{ij}^{-} + \sum_{i,j\notin I_C,J_C} \hat{l}_{ij}^{+}\right) && (10)\\
&\geq \frac{1}{l}\sum_{k=1}^{l} \sum_{i,j} \min\left(\hat{l}_{ij}^{-}, \hat{l}_{ij}^{+}\right) && (11)\\
&= \frac{1}{l}\sum_{k=1}^{l} \left(\sum_{i,j|\hat{l}_{ij}^{+}>\hat{l}_{ij}^{-}} \hat{l}_{ij}^{-} + \sum_{i,j|\hat{l}_{ij}^{+}<=\hat{l}_{ij}^{-}} \hat{l}_{ij}^{+}\right) && (12)\\
&= \hat{R}(\mathbf{1}_{\hat{C}_p}) = \hat{R}_p^* && (13)
\end{aligned}
$$

Donc $g = \mathbf{1}_{\hat{C}_p}$ est le classifier qui minimise le risque empirique. $\hat{C}_p$ est le meilleur carrelage que l'on puisse construire.

2.3 Question 3

2.3.1 Enoncé

Montrer que $\forall \epsilon > 0$, on a :

$$P(|R(\mathbf{1}_{\hat{C}_p}) - \hat{R}_p^*| > \epsilon) \leq \sum_{g\in\mathcal{F}_p} P(|R(g) - \hat{R}(g)| > \epsilon)$$

En déduire : $P(|R(\mathbf{1}_{\hat{C}_p}) - \hat{R}_p^*| > \epsilon) \to 0$ quand $l \to \infty$

2.3.2 Réponse

On sait que $\hat{R}_p^* = \hat{R}(\mathbf{1}_{\hat{C}_p})$, donc

$$P(|R(\mathbf{1}_{\hat{C}_p}) - \hat{R}_p^*| > \epsilon) = P(|R(\mathbf{1}_{\hat{C}_p}) - \hat{R}(\mathbf{1}_{\hat{C}_p})| > \epsilon) \tag{14}$$
$$\leq \sum_{g \in \mathcal{F}_p} P(|R(g) - \hat{R}(g)| > \epsilon) \tag{15}$$

Pour g donné dans $\mathcal{F}_p$, on considère la V.A. $Z = \mathbf{1}_{g(X) \neq Y}$, avec l observations I.I.D. $z_i = \mathbf{1}_{g(x^{(i)}) \neq y_i}$

Par la loi des grands nombres, on a $\frac{1}{l} \sum_{k=1}^{l} z_i \to \mathbb{E}(Z)$, au sens des probabilités, cad, pour tout g :

$$P(|R(g) - \hat{R}(g)| > \epsilon)|) \underset{l \to \infty}{\longrightarrow} 0$$

Comme $\mathcal{F}_p$ est de cardinal fini, on en déduit, pour p fixé,

$$P(|R(\mathbf{1}_{\hat{C}_p}) - \hat{R}_p^*| > \epsilon) \underset{l \to \infty}{\longrightarrow} 0$$

C'est un résultat important : à p fixé, l'erreur empirique de notre meilleur classifier $\mathbf{1}_{\hat{C}_p}$ tend vers son erreur intrinsèque de généralisation quand la taille du training set augmente.

2.4 Question 4

2.4.1 Enoncé

On cherche maintenant l'oracle, cad un $g = \mathbf{1}_{C_p^*}$ qui va minimiser l'erreur de généralisation sur toute la famille de modèles :

$$R_p^* = R(\mathbf{1}_{C_p^*}) = \inf_{g \in \mathcal{F}_p} R(g)$$

2.4.2 Réponse

On suit la même démarche que la construction de $\hat{C}_p$, cette fois non plus en comptant les points de l'échantillon de classe positive (resp négative) dans chacun des carreaux c_{ij}, mais en regardant la mesure de $c_{ij} \cap A$ (resp. $c_{ij} \cap A^c$)

On construit ainsi :

$$C_p^* = \bigcup_{i,j : |c_{ij} \cap A| \geq |c_{ij} \cap A^c|} c_{ij}$$

Puis on prend un $g = \mathbf{1}_{C_p} \in \mathcal{F}_p$ quelconque, construit sur un carrelage $C_p = \bigcup_{i,j} c_{ij}$

L'erreur de généralisation $R(g)$ vérifie :

$$
\begin{aligned}
R(g) &= \mathbb{E}(\mathbf{1}_{g(X)\neq Y}) && (16)\\
&= \int_{[0,1]^2} \mathbf{1}_{g(x)\neq y} dP(x) && (17)\\
&= \int_{[0,1]^2} \mathbf{1}_{g(x)\neq y} dx \;\; (\text{loi} - \text{uniforme}) && (18)\\
&= \int_{[0,1]^2} \mathbf{1}_{g(x)=1; y=-1} + \mathbf{1}_{g(x)=-1; y=1} dx && (19)\\
&= \sum_{i,j \in I_C, J_C} \int_{c_{ij}} \mathbf{1}_{y=-1} dx + \sum_{i,j \notin I_C, J_C} \int_{c_{ij}} \mathbf{1}_{y=1} dx && (20)\\
&= \sum_{i,j \in I_C, J_C} |c_{ij} \cap A^c| + \sum_{i,j \notin I_C, J_C} |c_{ij} \cap A| && (21)\\
&\geq \sum_{i,j \in I_{C^*}, J_{C^*}} |c_{ij} \cap A^c| + \sum_{i,j \notin I_{C^*}, J_{C^*}} |c_{ij} \cap A| && (22)\\
&= R(\mathbf{1}_{C_p^*}) = R_p^* && (23)
\end{aligned}
$$

Donc l'oracle est bien $\mathbf{1}_{C_p^*}$

2.5 Question 5

2.5.1 Enoncé

On veut montrer ici que $P(|R(\mathbf{1}_{\hat{\mathbf{C}}_\mathbf{p}}) - R_p^*| > \epsilon) \to 0$ quand $l \to \infty$

Cad que l'erreur de généralisation de notre classifier $\mathbf{1}_{\hat{C}_p}$ tend vers la meilleure erreur possible.

Donc que $\mathbf{1}_{\hat{C}_p}$ tend, au sens de l'erreur de généralisation, vers l'oracle.

2.5.2 Réponse

On donne ici une réponse un peu différente de celle donnée en TP.

On voit d'abord que $0 \leq R(\mathbf{1}_{\hat{C}_p}) - R_p^*$ par définition de R_p^*.

On décompose :

$$
\begin{aligned}
0 \leq R(\mathbf{1}_{\hat{C}_p}) - R_p^* &= R(\mathbf{1}_{\hat{C}_p}) - \hat{R}(\mathbf{1}_{\hat{C}_p}) && (24)\\
&+ \hat{R}(\mathbf{1}_{\hat{C}_p}) - \hat{R}(\mathbf{1}_{C_p^*}) && (25)\\
&+ \hat{R}(\mathbf{1}_{C_p^*}) - R(\mathbf{1}_{C_p^*}) && (26)
\end{aligned}
$$

La deuxième ligne vérifie $0 \leq \hat{R}(\mathbf{1}_{\hat{C}_p}) - \hat{R}(\mathbf{1}_{C_p^*})$ par définition de $\hat{R}(\mathbf{1}_{\hat{C}_p})$.

Donc :

$$0 \leq R(\mathbf{1}_{\hat{C}_p}) - R_p^* \leq 2 \times \sup_{g \in \mathcal{F}_p} |R(g) - \hat{R}(g)|$$

Avec la loss-function $l(y,y') = \mathbf{1}_{y \neq y'}$, on ré-écrit :

$$R(g) = \mathbb{E}(l(g(X), Y))$$

$$\hat{R}(g) = \frac{1}{l} \sum_{i=1}^{l} l(y_i, g(x^{(i)}))$$

On utilise Chebychev :

$$P((X - \mathbb{E}(X))^2 \geq \epsilon) \leq \frac{Var(X)}{\epsilon^2}$$

pour X V.A.

On pose ici $Z = \frac{1}{l} \sum_{i=1}^{l} l(g(X^{(i)}), X^{(i)})$, avec $\mathbb{E}(Z) = \mathbb{E}l(g(X), X) = R(g)$, et :

$$P(|Z - \mathbb{E}(Z)| \geq \epsilon) = P(|Z - \mathbb{E}(Z)|^2 \geq \epsilon^2|) \leq \frac{Var(Z)}{\epsilon^2}$$

Ici $\mathrm{Var}(Z) = \frac{1}{l^2} \times l \times \mathrm{Var}(l(g(X), X)) = \frac{\sigma^2}{l}$

Au final :

$$P(|\hat{R}(g) - R(g)| \geq \epsilon) \leq \frac{\sigma^2}{l}$$

Le majorant est indépendant de g, donc c'est encore un majorant du sup, et donc :

$$P(|R(\mathbf{1}_{\hat{\mathbf{C}}_\mathbf{p}}) - R_p^*| > \epsilon) \leq \frac{2\sigma^2}{l}$$

Cad que l'erreur de généralisation de notre classifier $\hat{\mathbf{1}}_{\hat{C}_p}$ tend vers l'erreur minimale de généralisation sur l'ensemble $\mathcal{F}_p$ de la classe de modèles, quand la taille de l'échantillon tend vers l'infini.

A p fixé, quand la taille de l'échantillon augmente, l'erreur empirique du classifier tend vers son erreur de généralisation ("convergence simple" à modèle fixé), qui elle-même tend vers l'erreur minimale de généralisation sur $\mathcal{F}_p$ ("convergence uniforme" sur toute la classe de modèles).

La borne donnée par Chebychev est grossière, d'où Hoeffding pour aborder numériquement le problème.

2.6 Question 5

On introduit maintenant un test-set $(x_1^{'(i)}, x_2^{'(i)}, y_i'))$, de taille m, distribué suivant les mêmes V.A. X, Y, I.I.D. et indépendant du training-set.

Le risque empirique sur le test set est noté :

$$\hat{R}'(\mathbf{1}_C) = \frac{1}{m} \sum_{i=1}^{m} \mathbf{1}_{C(x'^{(i)}) \neq y_i}$$

2.6.1 Réponse

On note $Z_i = \mathbf{1}_{C(X^{(i)},Y_i)}$, d'où $\hat{R}'(\mathbf{1}_C) = \frac{S_m}{m}$ avec les notations plus haut.

On applique Hoeffding, avec $m \geq -\frac{\log(\eta/2)}{2\epsilon^2}$, la borne supérieure devient :

$$2\exp(-2m\epsilon^2) \leq \eta$$

Donc :

$$P(|\hat{R}'(\mathbf{1}_C) - R(\mathbf{1}_C)| > \epsilon) \leq \eta \tag{27}$$

Cette dernière égalité est valable pour tout C, donc en particulier pour notre $\hat{C}_p$, qui est le carrelage avec lequel nous construisons notre estimateur.

Pour $\eta = 0.05$ et $\epsilon = 0.02$, on trouve $m_0 = 4611$

Au delà de 4611 points dans le test-set, l'écart entre l'erreur commise sur le test-set par notre classifier et son erreur de généralisation, est donc inférieur à 2% avec une probabilité de 95% .