Introduction au Machine learning : devoir maison

Exercice 1

Le but de cet exercice est de montrer l'inégalité d'Hoeffding, selon laquelle si $V_1, \dots, V_n$ sont les variables indépendantes et bornées (pour tout i $V_i \overset{p.s.}{\in} [a_i, b_i]$) alors, pour tout $t > 0$

$$\mathbb{P}\Big(\bar{V}_n - \mathbb{E}(\bar{V}_n) > t\Big) \leq \exp\Big(-\frac{\gamma n^2 t^2}{\sum_{i=1}^n (b_i - a_i)^2}\Big)$$

où $\bar{V}_n = (1/n)\sum_{i=1}^n V_i$, et γ dépend de la démonstration et vaut 2 dans le meilleur cas !

1. On commence par montrer que si W est une variable aléatoire centrée ($\mathbb{E}(W) = 0$) et telle que $W \overset{p.s.}{\in} [\alpha, \beta]$ alors pour tout $s \in \mathbb{R}$

$$\mathbb{E}\Big(\exp(sW)\Big) \leq \exp\Big(s^2(\beta - \alpha)^2/2\Big).$$

   (a) Justifier que $\mathbb{E}\Big(\exp(sW)\Big)$ existe bien puis que pour tout s

$$\mathbb{E}\Big(\exp(sW)\Big) = \mathbb{E}\Big(\exp\Big(s(W - \mathbb{E}(W^{'}))\Big)\Big)$$

   où $W^{'}$ est une copie indépendante de W.

   (b) Montrer que la convexité de la fonction exp et l'inégalité de Jensen entraînent

$$\mathbb{E}\Big(\exp(sW)\Big) \leq \mathbb{E}\Big(\exp\Big(s(W - W^{'})\Big)\Big).$$

   (c) On introduit une variable de Rademacher ε (c'est-à-dire telle que $\mathbb{P}(\varepsilon = 1) = \mathbb{P}(\varepsilon = -1) = 1/2$), indépendante de W et $W^{'}$, montrer que $W - W^{'}$ et $\varepsilon(W - W^{'})$ ont même loi.

   (d) Montrer que

$$\mathbb{E}\Big(\exp\Big(s\varepsilon(W - W^{'})\Big)\Big) = \frac{1}{2}\mathbb{E}\Big(\exp\Big(s(W - W^{'})\Big)\Big) + \frac{1}{2}\mathbb{E}\Big(\exp\Big(-s(W - W^{'})\Big)\Big).$$

   (e) Montrer que $|W - W^{'}| \overset{p.s.}{\leq} \beta - \alpha$ puis que

$$\begin{aligned}\frac{1}{2}\mathbb{E}\Big(\exp\Big(s(W - W^{'})\Big)\Big) + \frac{1}{2}\mathbb{E}\Big(\exp\Big(-s(W - W^{'})\Big)\Big) &\leq \mathbb{E}\Big(\exp\Big(s^2(W - W^{'})^2/2\Big)\Big) \\ &\leq \exp\Big(s^2(\beta - \alpha)^2/2\Big)\end{aligned}$$

   en utilisant les dévélopements en série entière des fonctions cosh et exp.

   (f) Conclure.

2. Montrer successivement que pour tout $\lambda > 0$

$$\begin{aligned}\mathbb{P}\Big(\bar{V}_n - \mathbb{E}(\bar{V}_n) > t\Big) &= \mathbb{P}\Big\{ \exp\Big(\lambda \sum_{i=1}^n \Big(V_i - \mathbb{E}(V_i)\Big) > \exp(\lambda n t)\Big\} \\ &= \mathbb{P}\Big\{ \exp\Big(\lambda \sum_{i=1}^n \Big(V_i - \mathbb{E}(V_i)\Big) - \lambda n t\Big) > 1\Big\} \\ &= \mathbb{E}\Big\{ \mathbb{1}_{\exp\Big(\lambda \sum_{i=1}^n (V_i - \mathbb{E}(V_i)) - \lambda n t\Big) > 1}\Big\}.\end{aligned}$$

3. Justifier que pour un réel aă ≥ 0

$$\mathbb{1}_{a>1} = 1\mathbb{1}_{a>1} \leq a\mathbb{1}_{a>1} \leq a.$$

En déduire que

$$\mathbb{E}\Big\{ \mathbb{1}_{\exp\Big(\lambda \sum_{i=1}^n \Big(V_i - \mathbb{E}(V_i)\Big) - \lambda n t\Big) > 1}\Big\} \leq \mathbb{E}\Big\{ \exp\Big(\lambda \sum_{i=1}^n \Big(V_i - \mathbb{E}(V_i)\Big) - \lambda n t\Big)\Big\}$$

4. Justifier que

$$\mathbb{E}\Big\{ \exp\Big(\lambda \sum_{i=1}^n \Big(V_i - \mathbb{E}(V_i)\Big)\Big)\Big\} = \prod_{i=1}^n \mathbb{E}\Big\{ \exp\Big(\lambda (V_i - \mathbb{E}(V_i)\Big)\Big\}.$$

5. Justifier que pour tout $i = 1, \ldots, n$

$$\frac{V_i - \mathbb{E}(V_i)}{b_i - a_i} \overset{p.s.}{\in} [-1, 1]$$

et en déduire une borne pour

$$\mathbb{E}\Big\{ \exp\Big(\lambda \sum_{i=1}^n \Big(V_i - \mathbb{E}(V_i)\Big)\Big)\Big\},$$

et conclure.

Exercice 2

On veut étudier l'algorithme des k plus proches voisins dans le cas où $k = 1$. On se place dans le cas où

- les covariables $X_i \in \mathcal{X} \subset [0,1]^d$
- les $Y_i \in \{0, 1\}$.
- les couples (X_i, Y_i) sont i.i.d. de loi inconnue $\mathcal{L}$
- on note $\mathcal{D}_n = \Big\{(X_i, Y_i), i = 1, \ldots, n\Big\}$.

On considère de plus la perte 0/1 de sorte que le risque d'un classifieur est donné par

$$R(c) = \mathbb{E}_{\mathcal{L}}\Big(\ell(Y_+, c(X_+))|\mathcal{D}_n\Big) = \mathbb{E}_{\mathcal{L}}\Big(\mathbb{1}_{Y_+ \neq c(X_+)}|\mathcal{D}_n\Big)$$

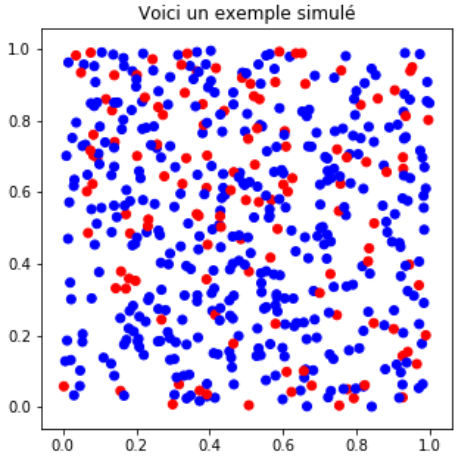
où (X_+, Y_+) est indépendant des (X_i, Y_i) et de même loi.

1. On va étudier des classifieurs non-déterministes, c'est-à-dire qui dépendent des données. On note $\hat{c}$ un tel classifieur. Comme on va intéger par rapport aux lois de (X_+, Y_+) et $\mathcal{D}_n$ alternativement on va noter dans cette question

   - $\mathbb{E}_{(X_+,Y_+)\sim\mathcal{L}}$ quand on intègre par rapport à la loi de (X_+, Y_+)
   - $\mathbb{E}_{\mathcal{D}_n\sim\mathcal{L}}$ quand on intègre par rapport à la loi de $\mathcal{D}_n$

   avec ces notations
   $$R(\hat{c}) = \mathbb{E}_{(X_+,Y_+)\sim\mathcal{L}}(\ell(Y_+, \hat{c}(X_+)).$$
   Vérifier que
   $$\begin{aligned}\mathbb{E}_{\mathcal{D}_n\sim\mathcal{L}}\Big(R(\hat{c})\Big) &= \mathbb{E}_{\mathcal{D}_n\sim\mathcal{L}}\Big(\mathbb{E}_{(X_+,Y_+)\sim\mathcal{L}}\Big(\ell(Y_+, \hat{c}(X_+))|\mathcal{D}_n\Big)\Big)\\ &= \mathbb{E}_{(X_+,Y_+)\sim\mathcal{L}}\Big(\mathbb{E}_{\mathcal{D}_n\sim\mathcal{L}}\Big(\ell(Y_+, \hat{c}(X_+))|(X_+, Y_+)\Big)\Big)\end{aligned}$$

2. Dans toute la suite, chaque Y_i suit la loi de Bernoulli de paramètre $\pi^\star = 3/4$ pour toute valeur valeur de X_i, c'est-à-dire
   $$\mathbb{P}_{\mathcal{L}}\Big(Y_i = 1|X_i = x\Big) = \pi^\star = 3/4 \text{ pour tout } \quad x \in \mathcal{X}.$$
   Montrer que Y_i est indépendant de X_i. Proposez alors un classifieur déterministe (qui ne dépend pas de $\mathcal{D}_n$) c_0, c'est-à-dire une fonction de $\mathcal{X}$ dans $\{0, 1\}$, qui vous paraît convenir au problème.



3. Montrer que le risque d'un classifeur déterministe c peut s'écrire successivement
   $$R(c) = \mathbb{E}_{\mathcal{L}}\Big(\mathbb{1}_{Y_+\neq c(X_+)}\Big) = \pi^\star + (1 - 2\pi^\star)\mathbb{E}_{\mathcal{L}}(c(X_+)).$$

   (a) Montrer que, pour tout classifieur déterministe c,
   $$0 \leq \mathbb{E}_{\mathcal{L}}(c(X_+)) \leq 1.$$
   En déduire une borne inférieure de $R(c)$ et la valeur de $\mathbb{E}_{\mathcal{L}}(c(X_+))$ pour laquelle elle est atteinte.

(b) En déduire que l'oracle $c^\star$ qui vérifie $c^\star = \text{argmin}_c R(c)$ vaut c_0.

(c) Calculer le risque de l'oracle.

4. On considère le classifieur du plus proche voisin (1 plus proche voisin) noté $\hat{c}_1$. Pour chaque $x \in \mathcal{X}$, on introduit les v.a. $Z_i(x)$ $(i = 1, \ldots, n)$ définies par

$$Z_i(x) = \begin{cases} 1 & \text{si } \ i \ \text{ est le plus proche voisin de } \ x \\ 0 & \text{sinon.} \end{cases}$$

(a) Vérifier que $\sum_{i=1}^n Z_i(x) = 1$.

(b) Montrer que $\hat{c}_1(x) = \sum_{i=1}^n Z_i(x) Y_i$.

(c) Déduire de la question 1 que

$$\mathbb{E}_{\mathcal{D}_n \sim \mathcal{L}}\Big(\hat{c}_1(x)\Big) = \sum_{i=1}^n \mathbb{E}_{\mathcal{D}_n \sim \mathcal{L}}\Big(Z_i(x) Y_i\Big) = \pi^\star.$$

(d) En déduire que

$$\mathbb{E}_{\mathcal{D}_n \sim \mathcal{L}}\Big(\ell(Y_+, \hat{c}_1(X_+)) | (X_+, Y_+)\Big) = 3/8.$$

5. Déduire des questions 1 et 4 que $\mathbb{E}_{\mathcal{D}_n \sim \mathcal{L}}\Big(R(\hat{c}_1)\Big) = 3/8$.

6. En déduire que le risque intégré de $\hat{c}_1$ ne tend pas vers celui de l'oracle $c^\star$.

7. Expliquer pourquoi.