Apprentissage statistique

TD 2 : inégalités de concentration et minimisation du risque empirique

Exercice 1 Inégalité de Bennett

Soient $Z_1, \dots, Z_n$ des variables aléatoires réelles, indépendantes, de carré intégrables et telles qu'il existe $b > 0$ telle que

$$Z_i \leq 1, \text{ pour tout } i = 1, \dots, n.$$

1. Soit $f(u) = u^{-2}(e^u - u - 1)$. On admet que la fonction f est croissante sur $\mathbb{R}$. Vérifier que, pour tout $i = 1, \dots, n$, pour tout $\lambda > 0$,
$$e^{\lambda Z_i} - \lambda Z_i - 1 \leq Z_i^2(e^\lambda - \lambda - 1).$$
2. En déduire que
$$\mathbb{E}[e^{\lambda Z_i}] \leq (e^\lambda - \lambda - 1)\mathbb{E}[Z_i^2] + \lambda \mathbb{E}[Z_i] + 1$$
3. En utilisant un argument de type Chernoff, montrer que
$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^n Z_i - \mathbb{E}[Z_1] \geq t\right) \leq e^{-\lambda t + \psi(\lambda/n)},$$
où $\psi(\lambda) = \sum_{i=1}^n \ln(\mathbb{E}[e^{\lambda(Z_i - \mathbb{E}[Z_i]}])$.
4. En utilisant le fait que $\ln(u) \leq u - 1$, montrer que
$$\psi(\lambda) \leq \sum_{i=1}^n \left(\mathbb{E}[e^{\lambda Z_i}] - \lambda \mathbb{E}[Z_i] - 1\right).$$
5. Notons $v = \sum_{i=1}^n \mathbb{E}[Z_i^2]$, montrer que
$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^n Z_i - \mathbb{E}[Z_1] \geq t\right) \leq \exp\left(-v^2 h\left(\frac{nt}{v}\right)\right)$$
où $h(t) = (1+u)\ln(1+u) - u$.
6. Nous supposons maintenant que
$$Z_i \leq b,$$
où b est une constante positive (qui peut être différente de 1). Vérifier que
$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^n Z_i - \mathbb{E}[Z_1] \geq t\right) \leq \exp\left(-\frac{v}{b^2} h\left(\frac{ntb}{v}\right)\right)$$
Cette inégalité de concentration a pour nom *inégalité de Bennett.*

Answer of exercise 1

1. Comme $Z_i \leq 1$ et $\lambda > 0$, et f croissante
$$f(\lambda Z_i) = \frac{e^{\lambda Z_i} - \lambda Z_i - 1}{\lambda^2 Z_i^2} \leq f(\lambda) = \frac{e^\lambda - \lambda - 1}{\lambda^2},$$
d'où le résultat.
2. On prend juste l'espérance de part et d'autre de l'inégalité.

3. En utilisant l'inégalité de Markov et par indépendance des Z_i entre eux

$$\begin{aligned}
\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n} Z_i - \mathbb{E}[Z_1] \geq t\right) &= \mathbb{P}\left(e^{\lambda\left(\frac{1}{n}\sum_{i=1}^{n} Z_i - \mathbb{E}[Z_1]\right)} \geq e^{\lambda t}\right) \\
&\leq e^{-\lambda t}\mathbb{E}\left[e^{\lambda\left(\frac{1}{n}\sum_{i=1}^{n} Z_i - \mathbb{E}[Z_1]\right)}\right] \\
&= e^{-\lambda t}\mathbb{E}\left[e^{\sum_{i=1}^{n}\frac{\lambda}{n}(Z_i - \mathbb{E}[Z_i])}\right] \\
&= e^{-\lambda t}\mathbb{E}\left[\prod_{i=1}^{n} e^{\frac{\lambda}{n}(Z_i - \mathbb{E}[Z_i])}\right] \\
&= e^{-\lambda t}\prod_{i=1}^{n}\mathbb{E}\left[e^{\frac{\lambda}{n}(Z_i - \mathbb{E}[Z_i])}\right]
\end{aligned}$$

d'où le résultat.

4. Nous avons

$$\psi(\lambda) = \sum_{i=1}^{n} \ln(\mathbb{E}[e^{\lambda Z_i}]e^{-\lambda\mathbb{E}[Z_i]}) = \sum_{i=1}^{n}(\ln(\mathbb{E}[e^{\lambda Z_i}]) - \lambda\mathbb{E}[Z_i]) \leq \sum_{i=1}^{n}(\mathbb{E}[e^{\lambda Z_i}] - 1 - \lambda\mathbb{E}[Z_i]).$$

5. Nous avons donc

$$\psi(\lambda) \leq (e^{\lambda} - \lambda - 1)\sum_{i=1}^{n}\mathbb{E}[Z_i^2] \leq v(e^{\lambda} - \lambda - 1).$$

D'où

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n} Z_i - \mathbb{E}[Z_1] \geq t\right) \leq e^{-\lambda(t+v/n)+ve^{\lambda/n}-v},$$

L'inégalité précédente étant vraie quelle que soit la valeur de λ, nous allons chercher la valeur de λ pour laquelle la majoration est la plus petite possible. Notons $r(\lambda) = -\lambda(t + v/n) + ve^{\lambda/n} - v$. Une étude des variations de la fonction r nous indique que le minimum est atteint pour la valeur critique

$$\lambda^* = n\ln\left(\frac{nt}{v} + 1\right)$$

et que

$$r(\lambda^*) = -vh\left(\frac{nt}{v}\right),$$

d'où le résultat.

6. On applique l'inégalité précédente à la variable $\tilde{Z}_i = Z_i/b$ qui vérifie la contrainte $\tilde{Z}_i \leq 1$. Notons

$$\tilde{v} = \sum_{i=1}^{n}\mathbb{E}[\tilde{Z}_i^2] = \frac{v}{b^2},$$

nous avons

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n}\tilde{Z}_i - \mathbb{E}[\tilde{Z}_1] \geq t\right) \leq \exp\left(-\tilde{v}h\left(\frac{nt}{\tilde{v}}\right)\right)$$

ce qui équivaut à

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n}\frac{Z_i}{b} - \frac{\mathbb{E}[Z_1]}{b} \geq t\right) \leq \exp\left(-\frac{v}{b^2}h\left(\frac{ntb^2}{v}\right)\right)$$

en prenant $u = bt$,

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n} Z_i - \mathbb{E}[Z_1] \geq u\right) \leq \exp\left(-\frac{v}{b^2}h\left(\frac{nub}{v}\right)\right).$$

Exercice 2 Dictionnaire fini et inégalité de Hoeffding

Soit $\mathcal{G} = \{g_1, ..., g_p\}$ une famille finie de classifieurs. Soit, avec la notation habituelle, $\mathcal{D}_n = ((X_1, Y_1), \ldots, (X_n, Y_n))$ un famille de couples aléatoires i.i.d. de loi P sur $\mathcal{X} \times \{0, 1\}$. Soit $\widehat{R}_P$ le risque empirique avec la perte de classification $\ell(y, y') = \mathbf{1} - \{y \neq y'\}$. Soit $\widehat{g}$ la règle de minimisation du risque empirique.

1. En utilisant l'inégalité de Markov, montrer que pour tout $t > 0$, pour tout $j \in \{1, ..., p\}$,

$$\mathbb{P}_P\left(|R_P(g_j) - \widehat{R}(g_j)| > t\right) \leq \frac{R_P(g_j)(1 - R_P(g_j))}{nt^2}.$$

2. En déduire que, pour tout $\varepsilon > 0$,

$$\mathbb{P}_P\left(R_P(\widehat{g}) - \min_{1\leq j\leq p} R_P(g_j) \leq \sqrt{\frac{p}{n\varepsilon}}\right) \geq 1 - \varepsilon$$

et

$$\mathbb{E}_P\left[R_P(\widehat{g}) - \min_{1\leq j\leq p} R_P(g_j)\right] \leq \frac{2p}{\sqrt{n}}.$$

3. Comparer avec le résultat obtenu en cours où l'on utilisait l'inégalité de Hoeffding.

Answer of exercise 2

1. Nous avons

$$\widehat{R}(g_j) - R(g_j) = \frac{1}{n}\sum_{i=1}^{n} \ell(Y_i, g_j(X_i)) - \mathbb{E}_P[\ell(Y_i, g_j(X_i))] = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{\{Y_i \neq g_j(X_i)\}} - \mathbb{P}_P(Y \neq g_j(X))$$

d'où, par l'inégalité de Markov,

$$\mathbb{P}_P\left(|R(g_j) - \widehat{R}(g_j)| > t\right) \leq \mathbb{P}_P\left((R(g_j) - \widehat{R}(g_j))^2 > t^2\right) \leq \frac{\mathbb{E}_P\left[(R(g_j) - \widehat{R}(g_j))^2\right]}{t^2} = \frac{\mathrm{Var}_P(\widehat{R}(g_j))}{t^2}.$$

La conclusion vient du fait que $\mathbf{1}_{\{Y_i \neq g_j(X_i)\}} \sim \mathcal{B}(\mathbb{P}(g_j(X) \neq Y))$ et que $\mathbb{P}(g_j(X) \neq Y) = R(g_j)$.

2. D'après le cours

$$R_P(\widehat{g}) - \min_{1\leq j\leq p} R_P(g_j) \leq 2 \max_{1\leq j\leq p} |R_P(g_j) - \widehat{R}(g_j)|$$

Donc, pour tout $t > 0$,

$$\begin{aligned}
\mathbb{P}_P\left(R_P(\widehat{g}) - \min_{1\leq j\leq p} R_P(g_j) > 2t\right) &\leq \mathbb{P}_P\left(\max_{1\leq j\leq p} |R_P(g_j) - \widehat{R}(g_j)| > t\right) \\
&\leq \mathbb{P}_P\left(\bigcup_{j=1}^{p}\left\{|R_P(g_j) - \widehat{R}(g_j)| > t\right\}\right) \\
&\leq \sum_{j=1}^{p} \mathbb{P}_P\left(|R_P(g_j) - \widehat{R}(g_j)| > t\right).
\end{aligned}$$

Par 1., cela nous donne pour tout $\varepsilon > 0$,

$$\mathbb{P}_P\left(R_P(\widehat{g}) - \min_{1\leq j\leq p} R_P(g_j) > \frac{1}{2}\sqrt{\frac{p}{n\varepsilon}}\right) \leq \sum_{j=1}^{p} \frac{R_P(g_j)(1 - R_P(g_j))}{n \times \frac{1}{2^2}\frac{p}{n\varepsilon}} \leq \varepsilon,$$

en remarquant que $R_P(g_j)(1 - R_P(g_j)) \leq \frac{1}{4}$ (faire un tableau de variation de la fonction $x \mapsto x(1 - x)$ sur $[0, 1]$ par exemple). Pour la seconde inégalité nous avons de même

$$\begin{aligned}
\mathbb{E}_P\left[R_P(\widehat{g}) - \min_{1\leq j\leq p} R_P(g_j)\right] &\leq 2\mathbb{E}_P\left[\max_{1\leq j\leq p} |R_P(g_j) - \widehat{R}(g_j)|\right] = 2\int_0^{+\infty} \mathbb{P}_P\left(\max_{1\leq j\leq p} |R_P(g_j) - \widehat{R}(g_j)| > t\right) dt \\
&\leq 2\sum_{j=1}^{p} \int_0^{+\infty} \min\left\{1; \frac{1}{4nt^2}\right\} dt.
\end{aligned}$$

Comme, pour $t > 0$,

$$\min\left\{1; \frac{1}{4nt^2}\right\} = \frac{1}{4nt^2} \iff 4nt^2 \geq 1 \iff t^2 \geq \frac{1}{4n} \iff t \geq \frac{1}{2\sqrt{n}},$$

$$\int_0^{+\infty} \min\left\{1; \frac{1}{4nt^2}\right\} dt = \int_0^{1/(2\sqrt{n})} dt + \int_{1/(2\sqrt{n})}^{+\infty} \frac{dt}{4nt^2} = \frac{1}{2\sqrt{n}} + \left[-\frac{1}{4nt}\right]_{t=1/(2\sqrt{n})}^{+\infty} = \frac{1}{\sqrt{n}}.$$

Exercice 3 Dictionnaire dénombrable et inégalité de Hoeffding

Soit $\mathcal{G} = \{g_j : j \in \mathbb{N}\}$ une famille dénombrable de classifieurs. Soit $\mathcal{D}_n = ((X_1, Y_1), \ldots, (X_n, Y_n))$ une famille de couples aléatoires i.i.d. de loi P sur $\mathcal{X} \times \{0, 1\}$. Enfin, soient $p_j > 0$ des réels tels que

$$\sum_{j=1}^{\infty} p_j = 1.$$

1. En utilisant l'inégalité de Hoeffding, montrer que :

$$\mathbb{P}_P\left(\exists j, |R_P(g_j) - \widehat{R}_P(g_j)| > \sqrt{\frac{\log\left(\frac{2}{p_j \varepsilon}\right)}{2n}}\right) \leq \varepsilon.$$

2. On suppose qu'il existe $j^* \in \mathbb{N}^*$ tel que

$$\inf_{j \geq 1} R(g_j) = R(g_{j^*}).$$

Montrer que la règle de classification

$$\widehat{g}_n(\cdot) = g_{\widehat{j}}(\cdot) \text{ où } \widehat{j} \in \arg\min_{j \in \mathbb{N}}\left(\widehat{R}(g_j) + \sqrt{\frac{\log\left(\frac{2}{p_j \varepsilon}\right)}{2n}}\right)$$

satisfait pour tout $\varepsilon > 0$:

$$\mathbb{P}_P\left(R(\widehat{g}_n) - R(g_{j^*}) \leq 2\sqrt{\frac{\log\left(\frac{2}{p_{j^*} \varepsilon}\right)}{2n}}\right) \geq 1 - \varepsilon.$$

Answer of exercise 3

1.

$$\begin{aligned}
\mathbb{P}_P\left(\exists j, |R_P(g_j) - \widehat{R}_P(g_j)| > \sqrt{\frac{\log\left(\frac{2}{p_j \varepsilon}\right)}{2n}}\right) &\leq \sum_{j \geq 1} \mathbb{P}_P\left(|R_P(g_j) - \widehat{R}_P(g_j)| > \sqrt{\frac{\log\left(\frac{2}{p_j \varepsilon}\right)}{2n}}\right) \\
&\leq \sum_{j \geq 1} \mathbb{P}_P\left(\frac{1}{n}\sum_{i=1}^{n} |\mathbf{1}_{\{g_j(X_i) \neq Y_i\}} - \mathbb{E}_P[\mathbf{1}_{\{g_j(X) \neq Y\}}]| > \sqrt{\frac{\log\left(\frac{2}{p_j \varepsilon}\right)}{2n}}\right)
\end{aligned}$$

Or $0 \leq \mathbf{1}_{\{g_j(X_i) \neq Y_i\}} \leq 1$ pour tout i donc, d'après l'inégalité de Hoeffding,

$$\begin{aligned}
\mathbb{P}_P\left(\left|\frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{\{g_j(X_i) \neq Y_i\}} - \mathbb{E}_P[\mathbf{1}_{\{g_j(X) \neq Y\}}]\right| > \sqrt{\frac{\log\left(\frac{2}{p_j \varepsilon}\right)}{2n}}\right) &\leq \exp\left(-\frac{2n^2\left(\sqrt{\frac{\log\left(\frac{2}{p_j \varepsilon}\right)}{2n}}\right)^2}{\sum_{i=1}^{n}(0-(-1))^2}\right) \\
&= 2e^{-2\frac{\log\left(\frac{2}{p_j \varepsilon}\right)}{2}} = p_j \varepsilon,
\end{aligned}$$

d'où le résultat.

2. On a

$$R(\widehat{g}_n) - R(g_{j^*}) = R(\widehat{g}_n) - \widehat{R}(\widehat{g}_n) + \widehat{R}(\widehat{g}_n) - R(g_{j^*}).$$

Posons $t_j = \sqrt{\frac{\log\left(\frac{2}{p_j \varepsilon}\right)}{2n}}$. Par définition de $\widehat{g}_n$,

$$\widehat{R}(\widehat{g}_n) + t_{\widehat{j}} \leq \widehat{R}(g_{j^*}) + t_{j^*},$$

d'où

$$R(\widehat{g}_n) - R(g_{j^*}) \quad \leq \quad R(\widehat{g}_n) - \widehat{R}(\widehat{g}_n) + \widehat{R}(g_{j^*}) + t_{j^*} - t_{\widehat{j}} - R(g_{j^*}).$$

Soit

$$\mathcal{A} = \left\{ \forall j, |\widehat{R}(g_j) - R(g_j)| \geq t_j \right\},$$

d'après 1., $\mathbb{P}(\mathcal{A}^c) \leq \varepsilon$. Or si $\mathcal{A}$ est réalisé, cela implique,

$$R(\widehat{g}_n) - R(g_{j^*}) \quad \leq \quad t_{\widehat{j}} + t_{j^*} + t_{j^*} - t_{\widehat{j}} = 2t_j^*.$$

d'où la résultat.