Correction du TD N° 2

Exercice 1 (Classifieur naïf de Bayes)

Rappel. L'approche classification naïve bayésienne est *généative* car elle répond à la question "*comment les données que l'on observe auraient elles pu être générées ?*' Elle consiste à déterminer les lois de probabilité $\mathbb{P}[Y = c|X = \vec{x}]$ (c est une classe de la variable cible Y) à partir des observations et des hypothèses, puis utiliser ces lois pour déterminer la classe la plus probable d'une observation.

1. **Hypothèse naïve d'indépendance conditionnelle.** On suppose que les variables (features) $X_1, \ldots, X_p$ sont conditionnellement indépendantes par rapport à la variable cible (target) Y, c'est à dire

$$\mathbb{P}[X_j = x_j, X_m = x_m|Y = y] = \mathbb{P}[X_j = x_j|Y = y]\mathbb{P}[X_m = x_m|Y = y],$$

pour tout $x_j, x_m \in \{0,1\}, y \in \{0,1\}$, et $1 \leq j \neq m \leq p$. Cette hypothèse est équivalente à (énoncé dans le cours (Chapitre 4))

$$\mathbb{P}[X = x_j|Y = y, X_m = x_m] = \mathbb{P}[X = x_j|Y = y],$$

pour tout $x_j, x_m \in \{0,1\}, y \in \{0,1\}$, et $1 \leq j \neq m \leq p$. En effet,

$$\begin{aligned}
\mathbb{P}[X = x_j|Y = y, X_m = x_m] &= \frac{\mathbb{P}[X = x_j, Y = y, X_m = x_m]}{\mathbb{P}[Y = y, X_m = x_m]} \\
&= \frac{\mathbb{P}[X = x_j, X_m = x_m|Y = y]\mathbb{P}[Y = y]}{\mathbb{P}[Y = y, X_m = x_m]} \\
&= \frac{\mathbb{P}[X_j = x_j|Y = y]\mathbb{P}[X_m = x_m|Y = y]\mathbb{P}[Y = y]}{\mathbb{P}[Y = y, X_m = x_m]} \quad \text{(indépendance conditionnelle)} \\
&= \frac{\mathbb{P}[X_j = x_j|Y = y]\mathbb{P}[X_m = x_m|Y = y]\mathbb{P}[Y = y]}{\mathbb{P}[X_m = x_m|Y = y]\mathbb{P}[Y = y]} \\
&= \mathbb{P}[X_j = x_j|Y = y].
\end{aligned}$$

Remarque. Sous cette hypothèse, nous pouvons écrire la distribution à posteriori des étiquettes (labels) Y après avoir observé les features $X_1, \ldots, X_p$ En effet $\forall y \in \{0,1\}$, $\forall x_j \in \{0,1\}$, on a, par la règle de Bayes,

$$\mathbb{P}[Y = y|X_1 = x_1, \ldots, X_p = x_p] = \frac{\mathbb{P}[Y = y, X_1 = x_1, \ldots, X_p = x_p]}{\mathbb{P}[X_1 = x_1, \ldots, X_p = x_p]}.$$

En utilisant la formule des probabilités composées (voir notes du Chapitre 2), on a

$$\begin{aligned}
\mathbb{P}[Y = y, X_1 = x_1, \ldots, X_p = x_p] = \mathbb{P}[Y = y]&\mathbb{P}[X_1 = x_1|Y = y] \\
&\times \mathbb{P}[X_2 = x_2|X_1 = x_1, Y = y] \\
&\times \mathbb{P}[X_3 = x_3|X_1 = x_1, X_2 = x_2, Y = y] \\
&\times \cdots \times \mathbb{P}[X_p = x_p|X_1 = x_1, X_2 = x_2, \ldots, X_{p-1} = x_{p-1}, Y = y].
\end{aligned}$$

Maintenant en utilisant l'indépendance conditionnelle entre X_j et X_m sachant Y pour tout $j \neq m$,

$$\begin{aligned}
&\mathbb{P}[X_2 = x_1|X_1 = x_1, Y = y] = \mathbb{P}[X_2 = x_1|Y = y], \\
&\mathbb{P}[X_3 = x_3|X_1 = x_1, X_2 = x_2, Y = y] = \mathbb{P}[X_3 = x_3|Y = y], \\
&\ldots \\
&\mathbb{P}[X_p = x_p|X_1 = x_1, X_2 = x_2, \ldots, X_{p-1} = x_{p-1}, Y = y] = \mathbb{P}[X_p = x_p|Y = y].
\end{aligned}$$

Donc la probabilité a posteriori devient

$$\mathbb{P}[Y = y | X_1 = x_1, \dots, X_p = x_p] = \frac{\mathbb{P}[Y = y] \prod_{j=1}^{p} \mathbb{P}[X_j = x_j | Y = y]}{\mathbb{P}[X_1 = x_1, \dots, X_p = x_p]}.$$

2. La règle de décision prend la forme (décision par maximum a posteriori (MAP))

$$\hat{y} = \hat{h}(\vec{x}^{\text{new}}) = \underset{c \in \{0,1\}}{\operatorname{argmax}} \left\{ \mathbb{P}[Y = c] \prod_{j=1}^{p} \mathbb{P}[X_j = x_j | Y = c] \right\}.$$

3. Les paramètres à estimer pour la classification sont :
   - la probabilité des étiquettes $\mathbb{P}[Y = 1]$ (suffira car $\mathbb{P}[Y = 0] = 1 - \mathbb{P}[Y = 1]$).
   - Les probabilités conditionnelles de $X_j | Y$, i.e., $\mathbb{P}[X_j | Y = 1]$ et $\mathbb{P}[X_j | Y = 0]$ pour tout $j = 1, \dots, p$. Donc $2p$ paramètres.
4. D'après la question 2), la règle de décision s'écrit

$$\hat{y} = \hat{h}_{\text{NB}}(\vec{x}) = \begin{cases} 1, & \text{si } \mathbb{P}[Y = 1] \prod_{j=1}^{p} \mathbb{P}[X_j = x_j | Y = 1] > \mathbb{P}[Y = 0] \prod_{j=1}^{p} \mathbb{P}[X_j = x_j | Y = 0], \\ 0, & \text{sinon,} \end{cases}$$

   Ce qui équivalent à

$$\hat{y} = \hat{h}_{\text{NB}}(\vec{x}) = \begin{cases} 1, & \text{si } \mathbb{P}[Y = 1] \mathbb{P}[X_1 = x_1, X_2 = x_2, \dots, X_p = x_p | Y = 1] \\ & \qquad > \mathbb{P}[Y = 0] \prod_{j=1}^{p} \mathbb{P}[X_1 = x_1, X_2 = x_2, \dots, X_p = x_p | Y = 0], \\ 0, & \text{sinon.} \end{cases}$$

   Donc

$$\hat{y} = \hat{h}_{\text{NB}}(\vec{x}) = \begin{cases} 1, & \text{si } \mathbb{P}[Y = 1 | X_1 = x_1, X_2 = x_2, \dots, X_p = x_p] \\ & \qquad > \mathbb{P}[Y = 0 | X_1 = x_1, X_2 = x_2, \dots, X_p = x_p], \\ 0, & \text{sinon.} \end{cases}$$

   Or $\mathbb{P}[Y = 0 | X_1 = x_1, X_2 = x_2, \dots, X_p = x_p] = 1 - \mathbb{P}[Y = 1 | X_1 = x_1, X_2 = x_2, \dots, X_p = x_p]$. Ainsi, nous obtenons

$$\hat{y} = \hat{h}_{\text{NB}}(\vec{x}) = \begin{cases} 1, & \text{si } \mathbb{P}[Y = 1 | X_1 = x_1, X_2 = x_2, \dots, X_p = x_p] \\ & \qquad > 1 - \mathbb{P}[Y = 1 | X_1 = x_1, X_2 = x_2, \dots, X_p = x_p], \\ 0, & \text{sinon.} \end{cases}$$

   alors

$$\hat{y} = \hat{h}_{\text{NB}}(\vec{x}) = \begin{cases} 1, & \text{si } \mathbb{P}[Y = 1 | X_1 = x_1, \dots, X_p = x_p] > \frac{1}{2}, \\ 0, & \text{sinon,} \end{cases}$$

   pour tout $\vec{x} = (x_1, \dots, x_p)^\top$.
5. Soient $\pi = \mathbb{P}[Y = 1], \theta_j = \mathbb{P}[X_j = 1 | Y = 1]$ et $\alpha_j = \mathbb{P}[X_j = 1 | Y = 0]$.
   (a) On a $\mathbb{P}[Y = 0] = 1 - \pi$..
   (b) Remarquons $\forall j \in \{1, \dots, p\}$, la variable aléatoire $X_j \in \{0, 1\}$ est booléenne, alors la variable $X_j | Y = 1$ est aussi booléenne avec $\mathbb{P}[X_j = 1 | Y = 1] = \theta_j$. Ceci implique que la variable $X_j | Y = 1$ suit une loi de Bernoulli de paramètre $\theta_j (\mathcal{B}(\theta_j)$. Alors, $\mathbb{P}[X_j = 0 | Y = 1] = 1 - \theta_j$. On conclut que la de probabilité de masse de $X_j | Y = 1$ est donnée par :

$$\mathbb{P}[X_j = x_j | Y = 1] = \theta_j^{x_j} (1 - \theta_j)^{1 - x_j}, \forall x_j \in \{0, 1\}.$$

   Même raisonnement pour la variable $X_j | Y = 0$, elle suit une loi de Bernoulli $\mathcal{B}(\alpha_j)$. Ainsi, on écrit la probabilité de masse

$$\mathbb{P}[X_j = x_j | Y = 0] = \alpha_j^{x_j} (1 - \alpha_j)^{1 - x_j}, \forall x_j \in \{0, 1\}.$$

(c) En utilisant l'hypothèse naïve de Bayes

$$\mathbb{P}[Y=1|X_1=x_1,\dots,X_p=x_p]=\frac{\mathbb{P}[Y=1]\prod_{j=1}^p\mathbb{P}[X_j=x_j|Y=1]}{\mathbb{P}[X_1=x_1,\dots,X_p=x_p]}.$$

Or par la formule des probabilités totales (Chapitre 2), on écrit

$$\begin{aligned}\mathbb{P}[X_1=x_1,\dots,X_p=x_p]&=\sum_{c\in\{0,1\}}\mathbb{P}[X_1=x_1,\dots,X_p=x_p,Y=c]\\&=\sum_{c\in\{0,1\}}\mathbb{P}[Y=c]\mathbb{P}[X_1=x_1,\dots,X_p=x_p|Y=c]\\&=\sum_{c\in\{0,1\}}\mathbb{P}[Y=c]\prod_{j=1}^p\mathbb{P}[X_j=x_j|Y=c]\quad\text{(indépendance conditionnelle)}\\&=\mathbb{P}[Y=0]\prod_{j=1}^p\mathbb{P}[X_j=x_j|Y=0]+\mathbb{P}[Y=1]\prod_{j=1}^p\mathbb{P}[X_j=x_j|Y=1].\end{aligned}$$

Alors,

$$\begin{aligned}&\mathbb{P}[Y=1|X_1=x_1,\dots,X_p=x_p]\\&\quad=\frac{\mathbb{P}[Y=1]\prod_{j=1}^p\mathbb{P}[X_j=x_j|Y=1]}{\mathbb{P}[Y=0]\prod_{j=1}^p\mathbb{P}[X_j=x_j|Y=0]+\mathbb{P}[Y=1]\prod_{j=1}^p\mathbb{P}[X_j=x_j|Y=1]}\\&\quad=\frac{\pi\prod_{j=1}^p\theta_j^{x_j}(1-\theta_j)^{1-x_j}}{(1-\pi)\prod_{j=1}^p\alpha_j^{x_j}(1-\alpha_j)^{1-x_j}+\pi\prod_{j=1}^p\theta_j^{x_j}(1-\theta_j)^{1-x_j}}.\end{aligned}$$

En divisant le numérateur et le dénominateur par le terme $\pi\prod_{j=1}^p\theta_j^{x_j}(1-\theta_j)^{1-x_j}$, on arrive à

$$\mathbb{P}[Y=1|X_1=x_1,\dots,X_p=x_p]=\frac{1}{1+\frac{(1-\pi)\prod_{j=1}^p\alpha_j^{x_j}(1-\alpha_j)^{1-x_j}}{\pi\prod_{j=1}^p\theta_j^{x_j}(1-\theta_j)^{1-x_j}}}.$$

(c) On d'apres la question 4°) b°),

$$\begin{aligned}\mathbb{P}[Y=1|X_1=x_1,\dots,X_p=x_p]&=\frac{1}{1+\frac{(1-\pi)\prod_{j=1}^p\alpha_j^{x_j}(1-\alpha_j)^{1-x_j}}{\pi\prod_{j=1}^p\theta_j^{x_j}(1-\theta_j)^{1-x_j}}}\\&=\frac{1}{1+\exp\left(\log\left(\frac{(1-\pi)\prod_{j=1}^p\alpha_j^{x_j}(1-\alpha_j)^{1-x_j}}{\pi\prod_{j=1}^p\theta_j^{x_j}(1-\theta_j)^{1-x_j}}\right)\right)}.\end{aligned}$$

Nous avons

$$\begin{aligned}
\log\left(\frac{(1-\pi)\prod_{j=1}^{p}\alpha_j^{x_j}(1-\alpha_j)^{1-x_j}}{\pi\prod_{j=1}^{p}\theta_j^{x_j}(1-\theta_j)^{1-x_j}}\right) &= \log\left(\frac{\pi}{1-\pi}\right)+\log\left(\frac{\prod_{j=1}^{p}\alpha_j^{x_j}(1-\alpha_j)^{1-x_j}}{\prod_{j=1}^{p}\theta_j^{x_j}(1-\theta_j)^{1-x_j}}\right)\\
&= \log\left(\frac{\pi}{1-\pi}\right)+\log\left(\prod_{j=1}^{p}\alpha_j^{x_j}(1-\alpha_j)^{1-x_j}\right)-\log\left(\prod_{j=1}^{p}\theta_j^{x_j}(1-\theta_j)^{1-x_j}\right)\\
&= \log\left(\frac{\pi}{1-\pi}\right)+\sum_{j=1}^{p}\log\left(\alpha_j^{x_j}(1-\alpha_j)^{1-x_j}\right)-\sum_{j=1}^{p}\log\left(\theta_j^{x_j}(1-\theta_j)^{1-x_j}\right)\\
&= \log\left(\frac{\pi}{1-\pi}\right)+\sum_{j=1}^{p}\left(x_j\log(\alpha_j)+(1-x_j)\log(1-\alpha_j)\right)-\sum_{j=1}^{p}\left(x_j\log(\theta_j)+(1-x_j)\log(1-\theta_j)\right)\\
&= \log\left(\frac{\pi}{1-\pi}\right)+\sum_{j=1}^{p}\left(x_j\log\left(\frac{\alpha_j}{\theta_j}\right)+(1-x_j)\log\left(\frac{1-\alpha_j}{1-\theta_j}\right)\right)\\
&= \log\left(\frac{\pi}{1-\pi}\right)+\sum_{j=1}^{p}\log\left(\frac{1-\alpha_j}{1-\theta_j}\right)+\sum_{j=1}^{p}\left(\log\left(\frac{\alpha_j}{\theta_j}\right)-\log\left(\frac{1-\alpha_j}{1-\theta_j}\right)\right)x_j\\
&= \beta_0+\sum_{j=1}^{p}\beta_j x_j,
\end{aligned}$$

avec

$$\beta_0 = \log\left(\frac{\pi}{1-\pi}\right)+\sum_{j=1}^{p}\log\left(\frac{1-\alpha_j}{1-\theta_j}\right) \text{ et } \beta_j = \log\left(\frac{\alpha_j}{\theta_j}\right)-\log\left(\frac{1-\alpha_j}{1-\theta_j}\right).$$

Ainsi,

$$\mathbb{P}[Y=1|X_1=x_1,\ldots,X_p=x_p] = \frac{1}{1+\exp\left(\beta_0+\sum_{j=1}^{p}\beta_j x_j\right)},$$

Remarque. Nous pouvons écrire la règle de décision pour le classifieur naïf de Bayes $\mathbb{P}[Y = 1|X_1 = x_1,\ldots,X_p = x_p]$ sous une forme qui correspond à la distribution de classe $Y = 1$ dans une régression logistique (voir chapitre 6).

Exercice 2 (Classification binaire, coût $0/1$)

Soit $\mathcal{X} = [a,b] \subset \mathbb{R}$ et l'ensemble des étiquettes $\mathcal{Y} \subset \{0,1\}$. Supposons que $\mathbb{P}[Y=1] = 0.8$ et les distributions conditionnelles $\mathbb{P}[X|Y=1]$ et $\mathbb{P}[X|Y=0]$ sont uniformes sur $\mathcal{X}$.

1. Le coût $0/1$ est la perte $0/1$ est la fonction

$$\begin{aligned}
\ell_{0/1}\colon \{0,1\}\times\{0,1\} &\to \mathbb{R}_+\\
(y,h(\vec{x})) &\mapsto \begin{cases}1, & \text{si } y \neq h(\vec{x})\\ 0, & \text{sinon.}\end{cases}
\end{aligned}$$

   Autrement $\ell_{0/1}(y,h(\vec{x})) = \mathbb{1}(y \neq h(\vec{x}))$ ($\mathbb{1}(\cdot)$ la fonction indicatrice.) On pénalise de 1 l'erreur de classification.
2. Par définition le classifieur idéal (classifieur de Bayes) $h^\star$ est défini par (voir cours Chapitre 4)

$$h^\star = \underset{h:\mathcal{X}\to\{0,1\}}{\operatorname{argmin}} R(\ell_{0/1}(Y,h(X)))$$

où le risque réel

$$R(h) = R(\ell_{0/1}(Y, h(X))) = \mathbb{E}_{(X,Y)\sim\mathbb{P}[X,Y]}[\ell_{0/1}(Y, h(X))] = \int_{\mathcal{X}\times\mathcal{Y}} \ell_{0/1}(y, h(x)) f_{X,Y}(x, y) dx dy.$$

Notons que la densité conjointe $f_{X,Y}(x, y)$ s'écrit

$$f_{X,Y}(x, y) = f_Y[y] f_{X|Y}(x|y).$$

La fonction $f_Y[y]$ correspond aux probabilités de masse $\mathbb{P}[Y = 0]$ et $\mathbb{P}[Y = 1]$ et $f_{X|Y}(x|y)$ est une densité d'une loi uniforme donnée par $f_{X|Y}(x|y) = \frac{1}{b-a}\mathbb{1}_{[a,b]}(x)$. Donc pour tout h prédicteur, son risque réel vaut

$$\begin{aligned} R(h) = R(\ell_{0/1}(Y, h(X))) &= \int_{\mathcal{X}\times\mathcal{Y}} \ell_{0/1}(y, h(x)) f_{X,Y}(x, y) dx dy \\ &= \int_{\mathcal{X}=[a,b]} \mathbb{1}(h(x) \neq 0)\mathbb{P}[Y = 0] \frac{1}{b-a}\mathbb{1}_{[a,b]}(x) dx \\ &\quad + \int_{\mathcal{X}=[a,b]} \mathbb{1}(h(x) \neq 1)\mathbb{P}[Y = 1] \frac{1}{b-a}\mathbb{1}_{[a,b]}(x) dx \\ &= \frac{0,2}{b-a} \int_{\mathcal{X}=[a,b]} \mathbb{1}(h(x) \neq 0) dx + \frac{0,8}{b-a} \int_{\mathcal{X}=[a,b]} \mathbb{1}(h(x) \neq 1) dx. \end{aligned}$$

Maintenant :
- si $h(x) = 1$ alors $R(h) = 0,2\frac{1}{b-a}\int_{\mathcal{X}=[a,b]} dx = 0,2\frac{1}{b-a}\int_a^b dx = 0,2$.
- si $h(x) \neq 1$ alors $R(h) = 0,8\frac{1}{b-a}\int_{\mathcal{X}=[a,b]} dx = 0,8\frac{1}{b-a}\int_a^b dx = 0,8$.

On conclut que $R(h)$ est minimal pour le prédicteur $h(x) = 1$.

3. D'après le développement fait dans la question 2°), le risque de Bayes $h^\star(x) = 1$ vaut $0,2$.
4. Soit $\mathcal{D}_n = (x_1, y_1), \ldots, (x_n, y_n)$ un échantillon d'apprentissage. Par définition,

$$\begin{aligned} R_n(\hat{h}) &= \frac{1}{n}\sum_{i=1}^n \ell_{0/1}(y_i, \hat{h}(x_i)) \\ &= \frac{1}{n}\sum_{i=1}^n \ell_{0/1}(y_i, y_i) \quad (\hat{h}(x_i) = y_i) \\ &= \frac{1}{n}\sum_{i=1}^n 0 \\ &= 0. \end{aligned}$$

Exercice 3 (Consistance de l'estimateur du risque empirique)

1. On l'appelle l'estimateur (prédicteur) de Bayes (estimateur idéal)

$$h^\star = h^\star_{\text{Bayes}} = \underset{h:\mathcal{X}\to\mathcal{Y}}{\operatorname{argmin}} \left\{ R(h) := \mathbb{E}[\ell(h(x), y)] \right\}.$$

2. Le prédicteur (estimateur) par minimisation du risque empirique est

$$\hat{h}_{\mathcal{H}} = \underset{h\in\mathcal{H}}{\operatorname{argmin}} \left\{ R_n(h) := \frac{1}{n}\sum_{i=1}^n \ell(h(x_i), y_i) \right\}$$

3. Le risque $R(\hat{h}_{\mathcal{H}})$ est une variable aléatoire (du fait du caractère aléatoire des données) dont la distribution dépend de $\mathbb{P}$. En effet, l'échantillon $\mathcal{D}_n = \{(x_1, y_1), \ldots, (x_n, y_n)\}$ correpond à une seule réalisation $\omega \in \Omega$ (avec Ω l'univers) des variables (X, Y) i.e., $\mathcal{D}_n = \mathcal{D}_n(\omega)$

$$\mathcal{D}_n(\omega) = \{(X_1(\omega), Y_1(\omega)), \ldots, (X_n(\omega), Y_n(\omega))\},$$

donc le prédicteur du risque empirique est défini

$$\hat{h}_{\mathcal{H}}(\omega) = \underset{h\in\mathcal{H}}{\operatorname{argmin}} \{R_n(h) := \frac{1}{n}\sum_{i=1}^{n} \ell(h(X_i(\omega)), Y_i(\omega))]\}.$$

Autrement, pour chaque tirage d'un échantillon $\mathcal{D}_n \equiv \mathcal{D}_n(\omega)$, on construit un nouveau prédicteur $\hat{h}_{\mathcal{H}}(\omega)$ ce qui explique le caractère aléatoire de ce prédicteur. D'autre part, le risque de $\hat{h}_{\mathcal{H}}(\omega)$ est aussi une variable aléatoire,

$$R(\hat{h}_{\mathcal{H}}(\omega)) = \mathbb{E}_{(X,Y)\sim\mathbb{P}[X,Y]}[\ell(h_{\mathcal{H}}(\omega)(X), Y)].$$

C'est pour cette raison on parle de la consistance statistique de $\hat{h}_{\mathcal{H}}$: le risque $R(\hat{h}_{\mathcal{H}})$ converge en probabilité vers le risque $R(h^\star_{\mathcal{H}})$ du meilleur prédicteur dans $\mathcal{H}$. Soit pour tout $\epsilon > 0$,

$$\lim_{n\to+\infty} \mathbb{P}_{\mathcal{D}_n}[R(\hat{h}_{\mathcal{H}}) - R(h^\star_{\mathcal{H}})] \geq \epsilon] = 0$$

avec $\mathbb{P}_{\mathcal{D}_n} = \mathbb{P}\times\cdots\times\mathbb{P}$ le produit tensoriel [1] (n fois) de la probabilité $\mathbb{P}$. Nous rappelons :
Hypothèse de la loi uniforme des grands nombres (ULLN). L'espace d'hypothèses $\mathcal{H}$ vérifie la loi uniforme des grands nombres si pour tout $\epsilon > 0$ on a:

$$\lim_{n\to+\infty} \mathbb{P}_{\mathcal{D}_n}\Big[\sup_{h\in\mathcal{H}} |R(h) - R_n(h)| \geq \epsilon\Big] = 0.$$

Sous l'hypothèse ULLN, montrons la consistance de l'estimateur du risque empirique $\hat{h}_{\mathcal{H}}$.
Pour tout $\epsilon > 0$

$$R(\hat{h}_{\mathcal{H}}) - R(h^\star_{\mathcal{H}}) = \big[R(\hat{h}_{\mathcal{H}}) - R_n(\hat{h}_{\mathcal{H}})\big] + \big[R_n(\hat{h}_{\mathcal{H}}) - R(h^\star_{\mathcal{H}})\big].$$

Or $h^\star_{\mathcal{H}} \in \mathcal{H}$ alors $R_n(\hat{h}_{\mathcal{H}}) \leq R_n(\hat{h}^\star_{\mathcal{H}})$. Ainsi

$$\begin{aligned} R(\hat{h}_{\mathcal{H}}) - R(h^\star_{\mathcal{H}}) &= \big[R(\hat{h}_{\mathcal{H}}) - R_n(\hat{h}_{\mathcal{H}})\big] + \big[R_n(\hat{h}^\star_{\mathcal{H}}) - R(h^\star_{\mathcal{H}})\big] \\ &\leq |R(\hat{h}_{\mathcal{H}}) - R_n(\hat{h}_{\mathcal{H}})| + |R(h^\star_{\mathcal{H}}) - R_n(\hat{h}^\star_{\mathcal{H}})| \\ &\leq \sup_{h\in\mathcal{H}} |R(h) - R_n(h)| + \sup_{h\in\mathcal{H}} |R(h) - R_n(h)| \\ &\leq 2\sup_{h\in\mathcal{H}} |R(h) - R_n(h)|. \end{aligned}$$

Ceci implique l'inclusion de l'événement

$$\big\{R(\hat{h}_{\mathcal{H}}) - R(h^\star_{\mathcal{H}}) \geq \epsilon\big\} \subset \big\{2\sup_{h\in\mathcal{H}} |R(h) - R_n(h)| \geq \epsilon\big\} \subset \big\{\sup_{h\in\mathcal{H}} |R(h) - R_n(h)| \geq \frac{\epsilon}{2}\big\}.$$

Donc

$$\mathbb{P}_{\mathcal{D}_n}[R(\hat{h}_{\mathcal{H}}) - R(h^\star_{\mathcal{H}}) \geq \epsilon] \leq \mathbb{P}_{\mathcal{D}_n}[\sup_{h\in\mathcal{H}} |R(h) - R_n(h)| \geq \frac{\epsilon}{2}].$$

Appliquons l'hypothèse ULLN

$$\lim_{n\to\infty} \mathbb{P}_{\mathcal{D}_n}[R(\hat{h}_{\mathcal{H}}) - R(h^\star_{\mathcal{H}}) \geq \epsilon] \leq \lim_{n\to\infty} \mathbb{P}_{\mathcal{D}_n}[\sup_{h\in\mathcal{H}} |R(h) - R_n(h)| \geq \frac{\epsilon}{2}] = 0,$$

On conclut $\forall\epsilon > 0$, $\lim_{n\to\infty} \mathbb{P}_{\mathcal{D}_n}[R(\hat{h}_{\mathcal{H}}) - R(h^\star_{\mathcal{H}}) \geq \epsilon] = 0$.

[1] https://fr.wikipedia.org/wiki/Produit_tensoriel

Exercice 4 (Excès de risque, compromis estimation, approximation et optimisation)

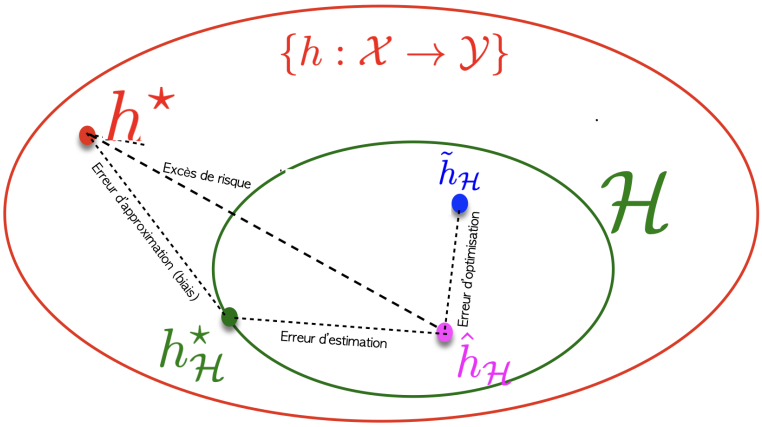
Le principe d'apprentissage que nous avons étudié en cours consiste à sélectionner d'abord un espace d'hypothèses $\mathcal{H}$ pour des estimateurs (prédicteurs), puis définir l'estimateur qui minimise le risque empirique $\hat{h}_{\mathcal{H}} = \text{argmin}_{h\in\mathcal{H}} R_n(h)$, avec $R_n(h) = \frac{1}{n}\sum_{i=1}^{n} \ell(h(\vec{x}_i, y_i)]$. Comme l'estimateur optimal (estimateur idéal de Bayes) $h^\star = \text{argmin}_{h:\mathcal{X}\to\mathcal{Y}} R(h)$, où $R(h) = \mathbb{E}[\ell(h(\vec{x}, y)]$, n'est généralement pas supposé appartenir à l'espace d'hypothèses $\mathcal{H}$, nous avons défini $h^\star_{\mathcal{H}} = argmin_{h\in\mathcal{H}} R(h)$. Nous supposons l'existence et l'unicité de $h^\star$, $h^\star_{\mathcal{H}}$, et $\hat{h}_{\mathcal{H}}$.

L'excès de risque $\mathcal{E}(\hat{h}_{\mathcal{H}})$ se décompose

$$\mathcal{E}(\hat{h}_{\mathcal{H}}) = \underbrace{\mathcal{R}(\hat{h}_{\mathcal{H}}) - \mathcal{R}(h^\star)}_{\text{Excès de risque}} = \underbrace{\mathcal{R}(\hat{h}_{\mathcal{H}}) - \mathcal{R}(h^\star_{\mathcal{H}})}_{\text{Erreur d'estimation}} + \underbrace{\mathcal{R}(h^\star_{\mathcal{H}}) - \mathcal{R}(h^\star)}_{\text{Erreur d'approximation}}$$

L'excès du risque $\mathcal{E}(\hat{h}_{\mathcal{H}})$ est une variable aléatoire (du fait du caractère aléatoire des données) dont la distribution dépend de $\mathbb{P}$. Notons $\varepsilon = \mathbb{E}_{\mathcal{D}_n}[\mathcal{E}(\hat{h}_{\mathcal{H}})]$, $\varepsilon_{\text{estim}} = \mathbb{E}_{\mathcal{D}_n}[\mathcal{R}(\hat{h}_{\mathcal{H}}) - \mathcal{R}(h^\star_{\mathcal{H}})]$, et $\varepsilon_{\text{approx}} = \mathbb{E}_{\mathcal{D}_n}[\mathcal{R}(h^\star_{\mathcal{H}}) - \mathcal{R}(h^\star)]$, où l'espérance est prise par rapport à l'échantillon d'apprentissage $\mathcal{D}_n$. Ainsi, on obtient $\varepsilon = \varepsilon_{\text{estim}} + \varepsilon_{\text{approx}}$. Nous illustrons ces prédicteurs dans la figure suivante:

$$h^\star = \arg\min_{h:\mathcal{X}\to\mathcal{Y}} R(h) \qquad h^\star_{\mathcal{H}} = \arg\min_{h\in\mathcal{H}} R(h)$$



$$\hat{h}_{\mathcal{H}} = \arg\min_{h\in\mathcal{H}} R_n(h) \qquad \tilde{h}_{\mathcal{H}} = \arg\min_{h\in\mathcal{H}} R_n(h)$$

Supposons que l'algorithme de minimisation pour calculer $\hat{h}_{\mathcal{H}}$ retourne une solution approchée $\tilde{h}_{\mathcal{H}}$ qui minimise la fonction objective avec une tolérance prédéfinie $\delta \geq 0$, c'est à dire

$$R_n(\tilde{h}_{\mathcal{H}}) \leq R_n(\hat{h}_{\mathcal{H}}) + \delta.$$

Rappelons :

- Échantillon d'apprentissage $\mathcal{D}_n = \{(X_i, Y_i)\}_{i=1,\dots,n}$ i.i.d avec $(X_i, Y_i) \sim \mathbb{P}[X, Y]$.
- Prédicteur de Bayes $h^\star_{\text{Bayes}} = \text{argmin}_{h:\mathcal{X}\to\mathcal{Y}} \left\{R(h) := \mathbb{E}[\ell(h(x), y)]\right\}$
- Meilleur prédicteur dans $\mathcal{H}$: $h^\star_{\mathcal{H}} = \text{argmin}_{h\in\mathcal{H}} R(h)$ (meilleur prédicteur minimisant dans $\mathcal{H}$ le risque réel).
- Prédicteur du risque empirique $\hat{h}_{\mathcal{H}} = \text{argmin}_{h\in\mathcal{H}} \left\{R_n(h) := \frac{1}{n}\sum_{i=1}^n \ell(h(x_i), y_i)]\right\}$.
- l'excès de risque est définie par $R(\hat{h}_{\mathcal{H}}) - R(h^\star)$.
- On décompose l'excès de risque sous la forme

$$R(\hat{h}_{\mathcal{H}}) - R(h^\star) = \underbrace{\left[R(\hat{h}_{\mathcal{H}}) - R(h^\star_{\mathcal{H}})\right]}_{\text{Erreur d'estimation}} + \underbrace{\left[R(h^\star_{\mathcal{H}}) - R(h^\star)\right]}_{\text{Erreur d'approximation}}.$$

1. On décompose l'excès de risque de l'estimateur approché $\tilde{h}_{\mathcal{H}}$ sous la forme

$$\mathcal{E}(\tilde{h}_{\mathcal{H}}) = R(\tilde{h}_{\mathcal{H}}) - R(h^\star) = \underbrace{\left[R(\tilde{h}_{\mathcal{H}}) - R(\hat{h}_{\mathcal{H}})\right]}_{\text{Erreur d'optimisation}} + \underbrace{\left[R(\hat{h}_{\mathcal{H}}) - R(h^\star_{\mathcal{H}})\right]}_{\text{Erreur d'estimation}} + \underbrace{\left[R(h^\star_{\mathcal{H}}) - R(h^\star)\right]}_{\text{Erreur d'approximation}}.$$

   Appliquons l'espérance par rapport à la loi de l'échantillon $\mathcal{D}_n$,

$$\begin{aligned}\tilde{\varepsilon} &= \mathbb{E}_{\mathcal{D}_n}[\mathcal{E}(\tilde{h}_{\mathcal{H}})] \\ &= \mathbb{E}_{\mathcal{D}_n}[R(\tilde{h}_{\mathcal{H}}) - R(\hat{h}_{\mathcal{H}})] + \mathbb{E}_{\mathcal{D}_n}[R(\hat{h}_{\mathcal{H}}) - R(h^\star_{\mathcal{H}})] + \mathbb{E}_{\mathcal{D}_n}[R(h^\star_{\mathcal{H}}) - R(h^\star)] \\ &:= \varepsilon_{\text{optim}} + \varepsilon_{\text{estim}} + \varepsilon_{\text{approx}}\end{aligned}$$

2. Étudier l'effet de croître l'espace d'hypothèses $\mathcal{H}$, c'est à dire considérer un nouveau espace plus grand $\mathscr{H}$, tel que $\mathcal{H} \subset \mathscr{H}$, sur les erreurs $\varepsilon_{\text{estim}}$ et $\varepsilon_{\text{approx}}$.
   Quand on augmente l'espace d'hypothèses $\mathcal{H}$, c'est à dire on cherche des prédicteurs dans un espace plus large alors le risque $R(h^\star_{\mathcal{H})}$ diminue et ainsi $\varepsilon_{\text{approx}}$ diminue (donc le biais diminue). Toutefois, augmentant l'espace d'hypothèses fait croître la complexité de modèle et donc $\varepsilon_{\text{estim}}$ qui est la variance de $\hat{h}_{\mathcal{H}}$ croît.
3. Étudier l'effet de croître la taille de l'échantillon d'apprentissage n sur les erreurs $\varepsilon_{\text{estim}}$ et $\varepsilon_{\text{approx}}$.
   Augmenter la taille de l'échantillon d'apprentissage n :
   - sur $\varepsilon_{\text{approx}}$: il n' y aucun effet car le biais est indépendant de n.
   - sur $\varepsilon_{\text{estim}}$: quand n devient plus grand, la variance devient plus petite et ainsi $\varepsilon_{\text{estim}}$ décroît.
4. Étudier l'effet de croître la tolérance δ sur l'erreur d'optimisation $\varepsilon_{\text{optim}}$.
   Augmenter la tolérance δ induit une solution approchée moins précise donc $\varepsilon_{\text{optim}}$ augmente.
5. L'estimateur approché $\tilde{h}_{\mathcal{H}}$ par un algorithme de minimisation nécessite un nombre d'itérations T. Étudier l'effet de croître $\{\mathcal{H}, n, \delta\}$ sur T.
   - quand $\mathcal{H}$ devient plus large, l'espace de recherche de $h^\star_{\mathcal{H}}$ devient large et donc le temps nécessaire pour calculer $h^\star_{\mathcal{H}}$ augmente, donc plus d'itérations.
   - Quand n croît, le calcul du risque empirique devient plus lent et ainsi T augmente.
   - Quand δ augmente, une solutions moins précise est acceptable, et ainsi le temps nécessaire pour calculer $\tilde{h}$ devient court, c'est à dire T diminue.

   On résume ces variations dans le tableau suivant :

Erreurs \ Paramètres	$\mathcal{H}(\uparrow)$	$n(\uparrow)$	$\delta(\uparrow)$
$\varepsilon_{\text{approx}}$	$\downarrow$	$\times$	$\times$
$\varepsilon_{\text{estim}}$	$\uparrow$	$\downarrow$	$\times$
$\varepsilon_{\text{optim}}$	$\times$	$\times$	$\uparrow$
T	$\uparrow$	$\uparrow$	$\downarrow$

$(\uparrow)$ croît

$(\times)$ pas d'effet