1 Apprentissage machine

1.1 Probabilité

Soit $\mathbb{P}$ une fonction $\mathbb{P}\colon \mathcal{F} \mapsto \mathbb{R}$ sur certains[1] sous-ensembles $A \in \mathcal{F} \subseteq 2^\Omega$. L'ensemble Ω s'appelle *l'univers* des *événements atomiques* $x \in \Omega$. $\mathbb{P}$ est une **probabilité** si et seulement si elle satisfait les **axiomes de Kolmogorov** comme suit :

⋆ la probabilité est toujours entre 0 et 1 :

$$0 \leq \mathbb{P}A \leq 1; \tag{1.1}$$

⋆ l'univers comprenant tous les événements est certain

$$\mathbb{P}\Omega = 1; \tag{1.2}$$

⋆ elle est σ-additive : pour toute suite (infinie) d'ensembles $A_i\colon i = 0, 1, 2, \ldots$ disjoints ($A_i \cap A_j = \emptyset$ pour tout $i \neq j$),

$$\mathbb{P}\bigcup_{i=0}^{\infty} A_i = \sum_{i=0}^{\infty} \mathbb{P}A_i. \tag{1.3}$$

1.2 Variables aléatoires (v.a.)

W(fr)

Une variable aléatoire X est définie par sa probabilité $(\Omega, \mathcal{F}, \mathbb{P}_X)$. Ainsi, la probabilité $\mathbb{P}_X\{X \in A\} = \mathbb{P}A$ est bien définie pour tout $A \in \mathcal{F}$. On écrit $\mathbb{E}X$ pour l'**espérance** de X. Pour tout ensemble de variables aléatoires (définies sur le même univers)

$$\mathbb{E}[X_1 + X_2 + X_3 + \cdots] = (\mathbb{E}X_1) + (\mathbb{E}X_2) + (\mathbb{E}X_3) + \cdots. \tag{1.4}$$

(Notez que l'indépendance des X_i n'est pas important ici.)

Indicateurs. Un **indicateur** pour l'ensemble A est une v.a. $\mathbb{I}_A$ qui prend la valeur 0 ou 1 selon la mesure de A :

$$\mathbb{P}_{\mathbb{I}_a}\{\mathbb{I}_A = 1\} = \mathbb{P}A$$
$$\mathbb{P}_{\mathbb{I}_a}\{\mathbb{I}_A = 0\} = \mathbb{P}\bar{A} = 1 - \mathbb{P}A.$$

[1] La domaine $\mathcal{F}$ de $\mathbb{P}$ est σ-*algèbre* (ou *tribu*) ce qui veut dire que (1) si un sous-ensemble $A \in \mathcal{F}$ est mesurable, alors son complémentaire $\bar{A} = \Omega \setminus A$ l'est aussi, et que (2) si les ensembles $A_i\colon i = 0, 1, 2, \ldots$ sont mesurables, alors leur union $\bigcup_{i=0}^{\infty} A_i$ l'est aussi. En conséquence, $\mathcal{F}$ est stable aussi par intersection et par soustraction.

Pour économie, on écrit l'indicateur de A simplement par A. Comme cela, $\{X = 1\}$ dénote l'événement que $X = 1$, ainsi que l'indicateur X pour ce même événément. Soit X l'indicateur pour A. On a alors $\mathbb{E}X = \mathbb{P}\{X = 1\} = \mathbb{P}A$, ou autrement $\mathbb{E}A = \mathbb{P}A$.

1.3 Conditionnalité

Soit $(\Omega, \mathcal{F}, \mathbb{P})$ un espace probabilisé. La probabilité conditionnée sur un événément $A \in \mathcal{F}$ avec $\mathbb{P}A > 0$ est l'espace $(A, \mathcal{F}, \mathbb{P}_{|A})$, avec $\mathbb{P}_{|A}B = \mathbb{P}\Big\{B \;\Big|\; A\Big\}$ définie par $\mathbb{P}A \cap B = \mathbb{P}\Big\{B \;\Big|\; A\Big\} \cdot \mathbb{P}A$ pour tout $B \in \mathcal{F}$. (Si $\mathbb{P}A = 0$, $\mathbb{P}\Big\{B \;\Big|\; A\Big\}$ est non-determinée.) Par Axiome (1.3),

$$\mathbb{P}B = \mathbb{P}B \cap A + \mathbb{P}B \cap \bar{A} = \mathbb{P}\Big\{B \;\Big|\; A\Big\} \cdot \mathbb{P}A + \mathbb{P}\Big\{B \;\Big|\; \bar{A}\Big\} \cdot \mathbb{P}\bar{A}, \tag{1.5a}$$

et, en général, pour tout $A_i \colon i = 0, 1, 2, \ldots$ disjoints avec $A_i \cap A_j = \emptyset$ pour $i \neq j$ et $\bigcup_i A_i = \Omega$, on obtient la **formule des probabilités totales** :

$$\mathbb{P}B = \sum_i \mathbb{P}\Big\{B \;\Big|\; A_i\Big\} \cdot \mathbb{P}A_i. \tag{1.5b}$$

L'espérance conditionnelle s'écrit par $\mathbb{E}\Big[X \;\Big|\; A\Big]$. Si X, Y sont des v.a. réelles, $\mathbb{E}\Big[X \;\Big|\; Y\Big]$ est une variable aléatoire : elle prend la valeur $\mathbb{E}\Big[X \;\Big|\; Y = y\Big]$ quand $Y = y$. On a alors $\mathbb{E}\Big[\mathbb{E}\Big[X \;\Big|\; Y\Big]\Big] = \mathbb{E}X$.

W(fr)

Théorème de Bayes. Soit $A_i \colon 0, 1, 2, \ldots$ une famille exhaustive d'événements disjoints : $A_i \cap A_j = \emptyset$ et $\cup_i A_i = \Omega$. Par (1.5b),

$$\mathbb{P}\Big\{A_i \;\Big|\; B\Big\} = \frac{\mathbb{P}A_i \cap B}{\mathbb{P}B} = \frac{\mathbb{P}\Big\{B \;\Big|\; A_i\Big\} \cdot \mathbb{P}A_i}{\mathbb{P}B} = \frac{\mathbb{P}\Big\{B \;\Big|\; A_i\Big\} \cdot \mathbb{P}A_i}{\sum_j \mathbb{P}\Big\{B \;\Big|\; A_j\Big\} \cdot \mathbb{P}A_j}. \tag{1.6}$$

Équation (1.6) s'appelle le théorème de Bayes.

1.4 Prédiction

On veut prédire une v.a. Y à partir d'une autre v.a. X. Dans d'autres mots, à partir d'une obsérvation $X = x$, on prédit $\hat{y} = f(x)$, en se servant d'une fonction de prédiction f.

Classification. Si la valeur prédite prend des valeurs finies ou dénombrables, alors on parle d'un problème de classification. En particulier, on a k classes si $\mathbb{P}\big\{Y \in \{0,1,\ldots,k-1\}\big\} = 1$. Le plus commun est d'avoir $k = 2$, ou classification binaire : Y est l'indicateur d'un ensemble A. La prédiction s'évalue par la probabilité d'erreur

$$\mathsf{erreur}(f) = \mathbb{P}\{f(X) \neq Y\}.$$

Si $k = 2$, on a

$$\begin{aligned}\mathsf{erreur}(f) = \mathbb{P}\Big\{Y = 1 \;\Big|\; f(X) = 0\Big\} \cdot \mathbb{P}\{f(X) = 0\} \\ + \mathbb{P}\Big\{Y = 0 \;\Big|\; f(X) = 1\Big\} \cdot \mathbb{P}\{f(X) = 1\} \quad (1.7)\end{aligned}$$

Erreur de Bayes. Soit $g_i(x) = \mathbb{P}\Big\{Y = i \;\Big|\; X = x\Big\}$ pour $i = 0, 1$ et $h(x) = g_1(x) - 1/2$. Soit f^* l'indicateur

$$f^*(x) = \big\{h(x) > 0\big\} = \big\{g_1(x) > g_0(x)\big\}.$$

La fonction f^*, appellée le **prédicteur Bayes** fournit la meilleure prédiction possible. Lors d'une obsérvation $X = x$, on devine f^*. Si $f^*(x) = 1$, $h(x) > 0$, et donc $Y = 1$ est plus probable que $Y = 0$. On se trompe avec une probabilité $\mathbb{P}\Big\{Y = 0 \;\Big|\; X = x\Big\} = g_0(x) = 1/2 - h(x)$. Si, par contre, $f^*(x) = 0$, alors $h(x) \leq 0$ et on se trompe avec une probabilité $\mathbb{P}\Big\{Y = 1 \;\Big|\; X = x\Big\} = g_1(x) = 1/2 + h(x)$. Dans les deux cas, l'erreur est $1/2 - |h(x)|$. Par conséquence, l'**erreur Bayes** s'écrit par

$$\begin{aligned}\mathsf{erreur}(f^*) = \mathbb{P}\Big\{Y = 1 \;\Big|\; h(X) \leq 0\Big\} \cdot \mathbb{P}\{h(X) \leq 0\} \\ + \mathbb{P}\Big\{Y = 0 \;\Big|\; h(X) > 0\Big\} \cdot \mathbb{P}\{h(X) > 0\} \\ = \mathbb{E}\Big[1/2 - \big|h(x)\big|\Big]\end{aligned}$$

Pour voir que f^* est la meilleure prédiction possible, soit f un prédicteur quelconque. Si $f^*(x) = 1$ et $f(x) = 0$, alors f se trompe avec une probabilité $\mathbb{P}\Big\{Y = 1 \;\Big|\; X = x\Big\} = 1/2 + h(x)$. Si $f^*(x) = 0$ et $f(x) = 1$, alors f se trompe avec une probabilité $\mathbb{P}\Big\{Y = 0 \;\Big|\; X = 1\Big\} = 1/2 - h(x)$. Dans les deux cas, la probabilité d'erreur est $1/2 + \big|h(x)\big|$. On a donc

$$\mathsf{erreur}(f) = \mathsf{erreur}(f^*) + 2\mathbb{E}\Big[\big|h(X)\big| \;\Big|\; f(X) \neq f^*(X)\Big] \cdot \mathbb{P}\big\{f(X) \neq f^*(X)\big\}.$$

Régression. Si Y est réelle ($Y \in \mathbb{R}$) on parle d'un problème de régression. L'erreur d'un prédicteur f se mesure par la distance Δ entre la prédiction et la valeur observée : $\Delta(f(X), Y)$. Typiquement, on vise à minimiser soit l'erreur L_1 : $\Delta(f(X), Y) = \big|f(X) - Y\big|$, ou l'erreur L_2 avec $\Delta(f(X), Y) = \big(f(X) - Y\big)^2$.

1.5 Apprentissage machine

En pratique, la distribution jointe de X et Y est rarement connue, et donc le prédicteur Bayes n'est pas disponible. On construit un prédicteur plutôt à partir d'un échantillon d'apprentissage (*learning sample*) D. L'échantillon comprend n paires observées : $D = \big\{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\big\}$. Typiquement, on assume que les paires sont indépendant et identiquement distribuées (iid). En apprentissage machine, on cherche des algorithmes de construction de prédicteurs.

Apprentissage non-supervisée. Quand on n'a q'un échantillon $D = \{x_1, x_2, \ldots, x_n\}$ sans étiquettes, on veut «découvrir» des classes à partir des similarités observés parmi les points x_i. On cherche un *groupage* (*clustering*) des données : il s'agit d'une fonction $f\colon \mathcal{R}^d \mapsto \{0, 1, 2, \ldots, k-1\}$ pour classifier les points x_i (de d dimensions) en k classes différentes. (Comment définir l'erreur d'un groupage ? On retournera à cette question plus tard.)

1.6 Modèles paramétriques

Les méthodes d'apprentissage machine peuvent être paramétriques ou non : tombent en une des classes suivantes

- ⋆ **méthodes non-paramétriques** : la prédiction se fait à partir de l'échantillon directement. Exemples : plus proche voisin, histogramme.
- ⋆ **méthodes paramétriques** : la prédiction se fait en choisissant le meilleur prédicteur à partir d'une classe fixée. Exemples : machine à vecteurs de support, réseau de neurones.

En bio-informatique, on se sert souvent de modèles paramétriques qui capturent les aspects spécifiques du problème biologique analysé. Un tel modèle impose des dépendances entre composantes de l'observation X et la sortie Y. En particulier, le modèle donne corps à un hypothèse, permettant sa validation et quantification. Exemple : modèles d'évolution de séquences moléculaires.