Apprentissage statistique

Chapitre 2 : Enjeux et concepts de l'apprentissage statistique

Lucie Le Briquer

18 février 2018

Table des matières

L'enjeu principal de l'apprentissage statistique est de "prédire" certains évènements. Pour cela on a un certain nombre d'observations $(x_1, \dots, x_n)$. On distingue deux types d'apprentissage :

1. *Apprentissage supervisé.* En plus des observations (X_i), on a une étiquette ou label (Y_i) pour chacune. Dans ce cadre, on va chercher à prédire le prochain label Y d'une nouvelle observation X. Moralement, on veut donc comprendre la dépendance entre X et Y.
2. *Cadre non supervisé.* Dans ce cadre les labels (Y_i) sont soit "cachés" ou inexistants. Cependant, on veut toujours discriminer/trier en un certain sens les (X_i).

Exemples.

1. AS supervisé
   (a) prédiction des températures à partir de la pression, prédiction de trafic routier
   (b) classification de formes
   (c) images de chiffres manuscrits $\rightarrow$ quel chiffre ces images représentent ? (data set : MNIST)
2. AS non supervisé : détection d'anomalie (défauts, d'usinage, tissus sains ou non dans une IRM)

1 Formulation mathématique

Soit deux ensembles mesurables $(\mathbb{X}, \mathcal{X})$ et $(\mathbb{Y}, \mathcal{Y})$.

Deux cas particuliers :

1. $\mathbb{Y} = \mathbb{R}$ et $\mathcal{Y} = \mathcal{B}(\mathbb{R})$. On dit que l'on est dans le cas de la *régression*.
2. $\mathbb{Y} = \{0, 1\}$, $\mathcal{Y} = \mathcal{P}(\mathbb{Y})$. On dit que l'on est dans le cas de la *classification*.

À partir de ces deux espaces, on considère une expérience statistique :

$$\big(\Omega, \mathcal{F}, \mathbb{X} \times \mathbb{Y}, \mathcal{X} \otimes \mathcal{Y}, \mathcal{P}, (X, Y)\big)$$

où X et Y sont des v.a. $\Omega \longrightarrow \mathbb{X} \times \mathbb{Y}$.

La différence maintenant entre les statistiques "classiques" et l'apprentissage est qu'en apprentissage on cherche, à partir de nouvelles observations X, à prédire Y. On ne cherche plus à déterminer le modèle/la probabilité qui a généré X. Contrairement aux statistiques où l'on s'intéresse à la probabilité qui explique le mieux X et Y.

Ce qui peut paraître alors plus intéressant en apprentissage, c'est à partir de $\mathbb{P} \in \mathcal{P}$ de connaître la loi conditionnelle de $Y|X$. En général, ce n'est pas évident, on est plutôt amenés à s'intéresser aux fonctions $\mathbb{X} \longrightarrow \mathbb{Y}$, qui peuvent dépendre de $\mathbb{P}$.

Définition 1 (classifieur/prédicteur)

On appelle classifieur/prédicteur toute application mesurable $f \colon \mathbb{X} \longrightarrow \mathbb{Y}$. On note l'ensemble des classifieurs $\mathcal{F}(\mathbb{X}, \mathbb{Y})$.

Définition 2 (fonction de coût)

On appelle fonction de coût toute fonction $c\colon \mathbb{Y} \times \mathbb{Y} \longrightarrow \mathbb{R}_+$ mesurable et vérifiant :

$$y = y' \;\Rightarrow\; c(y, y') = 0$$

Remarque. Les fonctions de coût vont permettre de mesurer l'efficacité d'un classifieur.

Définition 3 (risque)

On définit le risque de f pour la fonction de coût c par :

$$\mathbb{P} \in \mathcal{P} \longmapsto \mathbb{E}_{\mathbb{P}}[c(Y, f(X))]$$

On la note $R^c_{\mathbb{P}}(f)$ ou $R(f)$ (ou encore $L(f)$).

Remarque. On peut étendre ces définitions à des fonctions de coût qui dépendent de X.

2 Classifieur de Bayes

Pour l'instant on se fixe $\mathbb{P} \in \mathcal{P}$ et on cherche $f \in \mathcal{F}(\mathbb{X}, \mathbb{Y})$ un classifieur qui minimise $R^c_{\mathbb{P}}(f)$.

Définition 4 (risque de Bayes)

Soit $R^c_{\mathbb{P}}\colon \mathcal{F}(\mathbb{X}, \mathbb{Y}) \longrightarrow \mathbb{R}_+ \cup \{+\infty\}$. On appelle risque de Bayes :

$$(R^c_{\mathbb{P}})^* = \inf_{f \in \mathcal{F}(\mathbb{X}, \mathbb{Y})} R^c_{\mathbb{P}}(f)$$

si $f \in \underset{\tilde{f} \neq 0}{\operatorname{argmin}}(R^c_{\mathbb{P}}(\tilde{f}))$, on dit que f est un *classifieur de Bayes*. On le note en général $f^*_{\mathbb{P}}$.

On appelle *excès de risque* noté $e(f, f^*)$ la quantité $R^c_{\mathbb{P}}(f) - (R^c_{\mathbb{P}})^*$.

3 Règles d'apprentissage

On se place dans le cas de n observations $(X_i, Y_i)_{1 \leqslant i \leqslant n}$ i.i.d. (cadre d'une expérience statistique répétée).

Définition 5 (règle d'apprentissage)

Une règle d'apprentissage est fonction mesurable :

$$\hat{f}\colon \bigcup_{n \geqslant 1} \{\mathbb{X} \times \mathbb{Y}\}^n \longrightarrow \mathcal{F}(\mathbb{X}, \mathbb{Y})$$

Notation. $(X_i, Y_i)_{1 \leqslant i \leqslant n} = D_n$, $\hat{f}(D_n, .) \longrightarrow \hat{f}_n(.)$

Définition 6

Soit $\hat{f}$ une règle d'apprentissage. On définit le risque de $\hat{f}$ par rapport aux données D_n par :

$$\mathbb{P} \longmapsto \hat{R}^c_{\mathbb{P}}(\hat{f}, D_n) = \mathbb{E}_{\mathbb{P}}[c(\hat{f}_n(X), Y)|D_n]$$

pour une fonction de coût c.

Remarque. Si D_n est une v.a. alors $\hat{R}^c_{\mathbb{P}}(\hat{f}, D_n)$ est $\sigma(D_n)-$mesurable.

Définition 7 (risque moyen)

On appelle risque moyen de la règle d'apprentissage $\hat{f}$ pour le coût c :

$$\mathbb{P} \longmapsto \mathbb{E}_{\mathbb{P}}\big[\mathbb{E}_{\mathbb{P}}[c(\hat{f}_n(X), Y)|D_n]\big] = \mathbb{E}_{\mathbb{P}}[c(\hat{f}_n(X), Y)]$$

pour un $n \geqslant 1$. On le note $R^{\mathbb{P}}_n(\hat{f})$.

4 Consistance

Une règle d'apprentissage va être considérée comme *efficace* si le risque moyen $R^{\mathbb{P}}_n(\hat{f}) \xrightarrow[n\to+\infty]{} R^*_{\mathbb{P}}$ le risque de Bayes associé à $\mathbb{P}$, et cela pour tout $\mathbb{P} \in \mathcal{P}$ ($\mathcal{P}$ est suffisamment riche).

Définition 8 (faiblement consistant)

Soit $\hat{f}$ une règle d'apprentissage et $\mathbb{P} \in \mathcal{P}$. On dit que $\hat{f}$ est faiblement consistante si $\lim_{n\to+\infty} R^{\mathbb{P}}_n(\hat{f}) = R^*_{\mathbb{P}}$

Définition 9 (fortement consistant)

On dit que $\hat{f}$ est fortement consistant pour $\mathbb{P}$ si :

$$\hat{R}^c_{\mathbb{P}}(\hat{f}, D_n) \xrightarrow[n\to+\infty]{} R^*_{\mathbb{P}}$$

Remarque. Si c est borné et $\hat{f}$ est fortement consistante pour $\mathbb{P}$ alors par T.C.D. on a $\hat{f}$ faiblement consistante.

Définition 10 (universellement consistant)

Si $\hat{f}$ est faiblement/fortement consistante pour tout $\mathbb{P} \in \mathcal{P}$, alors $\hat{f}$ est dite faiblement/fortement universellement consistante.