

# SY09

## Théorie de la décision

T. Dencœux

### 1 Introduction

Dans beaucoup d'applications, chaque individu d'une population (objet, entité, état d'un système) peut être décrit par  $p$  variables explicatives (entrées)  $x_1, \dots, x_p$  et une variable explicative (sortie)  $y$ . Le problème considéré consiste alors à prédire la valeur de  $y$  à partir du vecteur  $\mathbf{x}$  des  $p$  variables explicatives  $x_1, \dots, x_p$ . Lorsque la variable  $y$  est quantitative, on dit que l'on a un problème de *régression*. Dans le où  $y$  est une qualitative à  $c$  modalités, on a un problème de *discrimination* en  $c$  classes.

Pour résoudre un problème de régression ou de discrimination, il faut disposer d'un ensemble d'apprentissage  $\mathcal{L} = \{(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_n, y_n)\}$  composé des observations des variables  $\mathbf{x}$  et  $y$  pour  $n$  individus de la population considérée. Les données peuvent être disposées dans un tableau de la forme suivante :

$$\begin{bmatrix} x_{11} & \dots & x_{1p} & y_1 \\ x_{21} & \dots & x_{2p} & y_2 \\ \vdots & \vdots & \vdots & \vdots \\ x_{i1} & \dots & x_{ip} & y_i \\ \vdots & \vdots & \vdots & \vdots \\ x_{n1} & \dots & x_{np} & y_n \end{bmatrix}$$

L'apprentissage supervisé consiste à trouver à partir de  $\mathcal{L}$  une fonction de décision  $g : \mathcal{X} \rightarrow \mathcal{Y}$ ,  $\mathcal{X}$  et  $\mathcal{Y}$  étant les domaines respectifs de  $\mathbf{x}$  et de  $y$ , telle que  $g(\mathbf{x}) \approx y$  pour tout couple  $(\mathbf{x}, y) \in \mathcal{X} \times \mathcal{Y}$  (problème 1).

Une façon de résoudre ce problème est de trouver  $g$  telle que  $g(x_i) \approx y_i$  pour tout  $i \in \{1, \dots, n\}$  (problème 2). Cependant, cette méthode a ses limites, car une fonction  $g$  solution exacte du problème 2 n'est pas forcément une solution du problème 1. Il faut imposer une contrainte de "régularité" à la solution du problème 2 pour espérer qu'elle soit une bonne solution du problème 1.

### 2 Formalisation

On peut formaliser le problème d'apprentissage supervisé de la manière suivante :

- Un *générateur*  $G$  tire au hasard des entrées  $\mathbf{x}$  selon une loi de probabilité  $f(\mathbf{x})$ , constante mais inconnue.
- Un *superviseur*  $S$  tire pour chaque  $\mathbf{x}$  une sortie  $y$  selon une loi de probabilité conditionnelle  $f(y|\mathbf{x})$ , également constante mais inconnue.

– Une *machine*  $M$  fournit pour chaque  $\mathbf{x}$  une décision  $d = g(\mathbf{x})$  à valeurs dans un ensemble  $\mathcal{D}$ , ce qui engendre un coût  $L(d, y)$ .

Chaque couple  $(\mathbf{x}, y)$  est donc une réalisation d'un couple de variables aléatoires  $(\mathbf{X}, Y)$ , de loi jointe  $f(\mathbf{x}, y) = f(\mathbf{x})f(y|\mathbf{x})$ . On appelle *risque conditionnel* de la fonction  $g$  sachant  $\mathbf{X} = \mathbf{x}$  fixé l'espérance conditionnelle du coût :

$$R(g|\mathbf{x}) = \mathbb{E}_Y [L(g(\mathbf{x}), Y)|\mathbf{x}].$$

Le *risque moyen* de  $g$  est

$$R(g) = \mathbb{E}_{\mathbf{X}, Y} [L(g(\mathbf{X}), Y)] = \mathbb{E}_{\mathbf{X}} [R(g|\mathbf{X})].$$

La fonction de décision optimale  $g^*$  est celle qui minimise le risque, pour  $f(\mathbf{x})$ ,  $f(y|\mathbf{x})$  et  $L$  donnés.

### 3 Fonctions de décision optimales

#### 3.1 Coût quadratique

Considérons tout d'abord un problème de régression :  $\mathcal{Y} \subseteq \mathbb{R}$ , et supposons  $\mathcal{D} = \mathcal{Y}$ . Soit la fonction de coût suivante (*coût quadratique*) :

$$L(d, y) = (y - d)^2.$$

On a

$$\begin{aligned} R(g|\mathbf{x}) &= \mathbb{E} [Y^2 - 2Yg(\mathbf{x}) + g(\mathbf{x})^2|\mathbf{x}] \\ &= \text{Var}(Y|\mathbf{x}) + (\mathbb{E}(Y|\mathbf{x}) - g(\mathbf{x}))^2. \end{aligned}$$

Toute fonction  $g$  telle que  $g(\mathbf{x}) = \mathbb{E}(Y|\mathbf{x})$  minimise donc le risque conditionnel sachant  $\mathbf{x}$ . Par conséquent, la fonction  $g^*$  optimale pour ce problème est la *fonction de régression*, définie par

$$g^*(\mathbf{x}) = \mathbb{E}(Y|\mathbf{x}).$$

Supposons maintenant que l'on ait un problème de discrimination à deux classes, avec  $\mathcal{Y} = \{0, 1\}$  et  $\mathcal{D} = [0, 1]$ . La fonction de décision optimale pour la fonction de coût quadratique est alors

$$g^*(\mathbf{x}) = \mathbb{E}(Y|\mathbf{x}) = \mathbb{P}(Y = 1|\mathbf{x}).$$

Dans ce cas, la décision optimale pour une entrée  $\mathbf{x}$  est donc la probabilité a posteriori de la classe 1 sachant  $\mathbf{X} = \mathbf{x}$ .

#### 3.2 Coûts 0/1

Restons dans le cadre de la discrimination à deux classes, et supposons maintenant que l'espace des décisions est  $\mathcal{D} = \{0, 1\}$ . La fonction de coût quadratique précédente se réduit à

$$L(d, y) = \begin{cases} 0 & \text{si } d = y, \\ 1 & \text{si } d \neq y \end{cases}$$

et le risque conditionnel s'écrit

$$\begin{aligned} R(g|\mathbf{x}) &= L(g(\mathbf{x}), 0)\mathbb{P}(Y = 0|\mathbf{x}) + L(g(\mathbf{x}), 1)\mathbb{P}(Y = 1|\mathbf{x}) \\ &= \begin{cases} \mathbb{P}(Y = 1|\mathbf{x}) & \text{si } g(\mathbf{x}) = 0, \\ \mathbb{P}(Y = 0|\mathbf{x}) & \text{si } g(\mathbf{x}) = 1. \end{cases} \end{aligned}$$

On voit que le risque conditionnel n'est autre dans ce cas que la probabilité d'erreur sachant  $\mathbf{x}$ . Pour  $\mathbf{x}$  fixé, le risque est minimisé pour  $g^*(\mathbf{x})$  défini par :

$$g^*(\mathbf{x}) = \begin{cases} 0 & \text{si } \mathbb{P}(Y = 0|\mathbf{x}) \geq \mathbb{P}(Y = 1|\mathbf{x}), \\ 1 & \text{sinon.} \end{cases}$$

La fonction de décision  $g^*$  minimisant le risque (c'est-à-dire, dans ce cas, la probabilité d'erreur), consiste donc à choisir la classe de plus grande probabilité a posteriori. Cette règle est appelée règle de Bayes minimisant la probabilité d'erreur. Cette règle peut également s'exprimer à l'aide du rapport de vraisemblance  $f_0(\mathbf{x})/f_1(\mathbf{x})$ ,  $f_k(\mathbf{x})$  désignant la densité conditionnelle de  $\mathbf{x}$  sachant  $Y = k$ ,  $k \in \{0, 1\}$ . En effet, en notant  $\pi_k = \mathbb{P}(Y = k)$ , on peut écrire :

$$\begin{aligned} g^*(\mathbf{x}) = 0 &\Leftrightarrow \mathbb{P}(Y = 0|\mathbf{x}) > \mathbb{P}(Y = 1|\mathbf{x}) \\ &\Leftrightarrow \frac{f_0(\mathbf{x})\pi_0}{f(\mathbf{x})} > \frac{f_1(\mathbf{x})\pi_1}{f(\mathbf{x})} \\ &\Leftrightarrow \frac{f_0(\mathbf{x})}{f_1(\mathbf{x})} > \frac{\pi_1}{\pi_0}. \end{aligned}$$

Le risque de la règle de Bayes, appelé dans ce cas *probabilité d'erreur de Bayes*, est égal à :

$$\begin{aligned} R(g^*) &= \int R(g^*|\mathbf{x})f(\mathbf{x})d\mathbf{x} \\ &= \int \min(\mathbb{P}(Y = 0|\mathbf{x}), \mathbb{P}(Y = 1|\mathbf{x})) f(\mathbf{x})d\mathbf{x} \\ &= \int \min(f_0(\mathbf{x})\pi_0, f_1(\mathbf{x})\pi_1) d\mathbf{x}. \end{aligned}$$

### 3.3 Discrimination avec coûts quelconques

Généralisons maintenant le cadre précédent, en supposant que la fonction de coût est définie par la matrice suivante :

$d/y$	0	1
0	$L_{00}$	$L_{01}$
1	$L_{10}$	$L_{11}$

Dans ce cas, la règle décision optimale s'écrit :

$$\begin{aligned} g^*(\mathbf{x}) = 0 &\Leftrightarrow L_{00}\mathbb{P}(Y = 0|\mathbf{x}) + L_{01}\mathbb{P}(Y = 1|\mathbf{x}) < L_{10}\mathbb{P}(Y = 0|\mathbf{x}) + L_{11}\mathbb{P}(Y = 1|\mathbf{x}) \\ &\Leftrightarrow (L_{00} - L_{10})\frac{f_0(\mathbf{x})\pi_0}{f(\mathbf{x})} < (L_{11} - L_{01})\frac{f_1(\mathbf{x})\pi_1}{f(\mathbf{x})} \\ &\Leftrightarrow \frac{f_0(\mathbf{x})}{f_1(\mathbf{x})} > \frac{L_{01} - L_{11}}{L_{10} - L_{00}} \frac{\pi_1}{\pi_0}. \end{aligned}$$