SY19
Minimisation du risque empirique et modèles linéaires

T. Denœux

1 Introduction

Nous avons calculé dans le chapitre précédent les fonctions de décision optimales pour différents problèmes d'apprentissage supervisé. Cependant, ces fonctions de décision optimales ne peuvent être calculées dans la pratique car les distributions de probabilité $f(\mathbf{x})$ et $f(y|\mathbf{x})$ sont inconnues. En revanche, on dispose en général d'un ensemble d'apprentissage $\mathcal{L} = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)\}$. Une approche peut alors consister à minimiser le *risque empirique*

$$\widehat{R}(g) = \frac{1}{n} \sum_{i=1}^{n} L\left(g(\mathbf{x}_i), y_i\right).$$

Si l'on n'impose aucune contrainte sur la fonction $\widehat{g}$ minimisant $\widehat{R}(g)$, il est clair que n'importe quelle fonction $\widehat{g}$ vérifiant $\widehat{g}(\mathbf{x}_i) = y_i$ pour tout i est solution du problème. Cette solution correspond à un « apprentissage par cœur » des données et ne présente pas d'intérêt. En pratique, on cherche à minimiser le risque empirique au sein d'une famille paramétrée de fonctions : $\mathcal{G} = \{g(\cdot; \mathbf{w}), \mathbf{w} \in \mathcal{W}\}$.

Dans le paragraphe suivant, nous appliquons ce principe à la régression, en définition $\mathcal{G}$ comme l'ensemble des fonctions linéaires par rapport à $\mathbf{x}$.

2 Régression linéaire

2.1 Moindres carrés

Supposons dans ce paragraphe que $\mathbf{x}$ est défini comme $\mathbf{x} = (x_0, x_1, \ldots, x_p)'$ avec $x_0 = 1$ par convention, et posons

$$g(\mathbf{x}) = \mathbf{w}'\mathbf{x} = w_0 + \sum_{i=1}^{p} w_j x_j.$$

En considérant une fonction de coût quadratique, le risque empirique s'écrit dans ce cas

$$\widehat{R}(g(\cdot; \mathbf{w})) = \widehat{R}(\mathbf{w}) = \frac{1}{n} \sum_{i=1}^{n} \left(\mathbf{w}'\mathbf{x}_i - y_i\right)^2.$$

Notons

$$X = \begin{pmatrix} 1 & x_{11} & \dots & x_{1p} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & \dots & x_{np} \end{pmatrix}$$

la matrice $(n, p+1)$ contenant les valeurs des variables explicatives, et $\mathbf{y} = (y_1, \dots, y_n)'$ le vecteur des observations de la variable y. Le risque empirique $\widehat{R}(\mathbf{w})$ peut alors s'écrire matriciellement :

$$\begin{aligned} \widehat{R}(\mathbf{w}) &= \frac{1}{n}(X\mathbf{w} - \mathbf{y})'(X\mathbf{w} - \mathbf{y}) \\ &= \frac{1}{n}\left(\mathbf{w}'X'X\mathbf{w} - 2\mathbf{w}'X'\mathbf{y} + \mathbf{y}'\mathbf{y}\right). \end{aligned}$$

On a

$$\frac{d\widehat{R}(\mathbf{w})}{d\mathbf{w}} = \frac{1}{n}\left(2X'X\mathbf{w} - 2X'\mathbf{y}\right),$$

d'où

$$\frac{d\widehat{R}(\mathbf{w})}{d\mathbf{w}} = 0 \Leftrightarrow X'X\mathbf{w} = X'\mathbf{y}. \tag{1}$$

Si la matrice $X'X$ est inversible, le minimum du risque empirique est donc obtenu pour $\widehat{\mathbf{w}} = (X'X)^{-1}X'\mathbf{y} = X^+\mathbf{y}$, où $X^+ = (X'X)^{-1}X'$ est appelée *pseudo-inverse* de X. Dans le cas contraire, on montre que la solution de (1) est encore de la forme $\widehat{\mathbf{w}} = X^+\mathbf{y}$ avec

$$X^+ = \lim_{\lambda \to 0}(X'X + \lambda I)^{-1}X'.$$

En pratique, si la matrice $X'X$ est mal conditionnée, on pose $X^+ = (X'X + \lambda I)^{-1}X'$ pour une petite valeur de λ.

2.2 Lien avec la fonction de régression

Soit $\widehat{g}(\mathbf{x}) = \widehat{\mathbf{w}}'\mathbf{x}$. Quel lien peut-on établir entre les fonctions $\widehat{g}$ et $g^* = \mathbb{E}(Y|\mathbf{x})$? Pour répondre à cette question, on peut tout d'abord remarquer que le risque empirique converge en probabilité vers le risque théorique lorsque la taille de l'ensemble d'apprentissage tend vers l'infini :

$$\widehat{R}(\mathbf{w}) = \frac{1}{n}\sum_{i=1}^{n}\left(\mathbf{w}'\mathbf{x}_i - y_i\right)^2 \xrightarrow{P} R(\mathbf{w}) = \mathbb{E}_{X,Y}\left[(\mathbf{w}'\mathbf{X} - Y)^2\right].$$

Or, on a $R(\mathbf{w}) = \mathbb{E}_X\left[R(\mathbf{w}|\mathbf{X})\right]$, avec

$$R(\mathbf{w}|\mathbf{x}) = \mathbb{E}_Y\left[(\mathbf{w}'\mathbf{x} - Y)^2|\mathbf{x}\right] = \left(\mathbf{w}'\mathbf{x} - \mathbb{E}(Y|\mathbf{x})\right)^2 + \mathrm{Var}(Y|\mathbf{x}),$$

d'où l'on déduit

$$R(\mathbf{w}) = \mathbb{E}\left[(\mathbf{w}'\mathbf{X} - g^*(\mathbf{X}))^2\right] + \text{cste}.$$

Asymptotiquement, minimiser $\widehat{R}(\mathbf{w})$ revient donc à minimiser l'espérance de l'écart quadratique par rapport à la fonction de régression. Pour n grand, $\widehat{g}(\mathbf{x})$ est en ce sens la *meilleure approximation linéaire* de la fonction de régression.

3 Discrimination linéaire

La méthode précédente peut être appliquée à la régression en deux classes, avec $y \in \{0, 1\}$. Dans ce cas, $\widehat{g}(\mathbf{x}) = \widehat{\mathbf{w}}'\mathbf{x}$ est une approximation de $g^*(\mathbf{x}) = \mathbb{P}(Y = 1|\mathbf{x})$.

Si on souhaite une décision binaire, on choisira dans ce cas la classe 0 si $\widehat{g}(\mathbf{x}) \leq 0.5$ et la classe 1 sinon. Ceci revient à partitionner l'espace d'entrée $\mathcal{X}$ en deux régions séparé par un hyperplan d'équation $\widehat{\mathbf{w}}'(\mathbf{x}) = 0.5$, appelé *frontière de décision*. La règle de décision correspondante est appelée un *classifieur linéaire*.

Si l'accent est mis sur l'approximation de la probabilité a posteriori $\mathbb{P}(Y = 1|\mathbf{x})$, on peut remarquer que le choix d'une fonction linéaire de la forme $g(\mathbf{x}; \mathbf{w}) = \mathbf{w}'\mathbf{x}$ nest pas très naturel car la contrainte $0 \leq \mathbb{P}(Y = 1|\mathbf{x}) \leq 1$ n'est pas prise en compte. Il peut sembler plus judicieux de choisir une famille de fonctions $g(\mathbf{x}; \mathbf{w})$ à valeurs dans l'intervalle $[0, 1]$. Pour cela, supposons que l'on a :

$$\ln \frac{\mathbb{P}(Y = 1|\mathbf{x})}{\mathbb{P}(Y = 0|\mathbf{x})} = \mathbf{w}^{*\prime}\mathbf{x}.$$

On en déduit

$$\frac{\mathbb{P}(Y = 0|\mathbf{x})}{\mathbb{P}(Y = 1|\mathbf{x})} = \exp(-\mathbf{w}^{*\prime}\mathbf{x})$$

et, en utilisant l'égalité $\mathbb{P}(Y = 0|\mathbf{x}) + \mathbb{P}(Y = 1|\mathbf{x}) = 1$, on obtient :

$$\mathbb{P}(Y = 1|\mathbf{x}) = \frac{1}{1 + \exp(-\mathbf{w}^{*\prime}\mathbf{x})} = g^*(\mathbf{x}).$$

On peut dans ce cas approcher $g^*(\mathbf{x})$ par une fonction dans l'ensemble

$$\mathcal{G} = \{g : \mathbf{x} \to g(\mathbf{x}; \mathbf{w}) = \frac{1}{1 + \exp(-\mathbf{w}'\mathbf{x})} | \mathbf{w} \in \mathcal{W}\}.$$

Le problème est que la fonction $\widehat{R}(\mathbf{w})$ n'est alors plus quadratique, et la solution du problème de minimisation n'a plus de solution analytique. Il faut donc avoir recours à une algorithme itératif.

L'algorithme est le plus simple est la méthode de *descente de gradient à pas constant*. Cet algorithme consiste à initaliser $\mathbf{w}$ à une valeur initiale $\mathbf{w}(0)$ (par exemple, par tirage aléatoire selon une loi $\mathcal{N}(0, \sigma^2 I)$), puis à appliquer itérativement la règle d'apprentissage suivante :

$$\mathbf{w}(k + 1) = \mathbf{w}(k) - \eta \frac{d\widehat{R}}{d\mathbf{w}}(\mathbf{w}(k)),$$

η étant un coefficient fixé appelé *pas d'apprentissage*, jusqu'à ce que l'un des critères d'arrêt suivants soit satisfait :

$$\| \frac{d\widehat{R}}{d\mathbf{w}}(\mathbf{w}(k)) \| \leq \epsilon,$$

ou

$$\|\mathbf{w}(k + 1) - \mathbf{w}(k)\| \leq \epsilon.$$

Soit $\varphi(u) = (1 + \exp(-u))^{-1}$. On a :

$$\begin{aligned}\frac{d\widehat{R}}{d\mathbf{w}}(\mathbf{w}) &= \frac{1}{n}\sum_{i=1}^{n}\frac{d}{d\mathbf{w}}\left[(\varphi(\mathbf{w}'\mathbf{x}_i) - y_i)^2\right] \\ &= \frac{1}{n}\sum_{i=1}^{n} 2(\varphi(\mathbf{w}'\mathbf{x}_i) - y_i)\varphi'(\mathbf{w}'\mathbf{x}_i)\mathbf{x}_i.\end{aligned}$$

Or $\varphi'(u) = \varphi(u)(1 - \varphi(u))$. On a donc finalement :

$$\frac{d\widehat{R}}{d\mathbf{w}}(\mathbf{w}) = \frac{1}{n}\sum_{i=1}^{n} 2(g(\mathbf{x}_i;\mathbf{w}) - y_i)g(\mathbf{x}_i;\mathbf{w})\left(1 - g(\mathbf{x}_i;\mathbf{w})\right)\mathbf{x}_i.$$

Cet algorithme converge vers le voisinage d'un minimimum local de $\widehat{R}$, à condition que η soit suffisamment petit. Etant donné le caractère local de la solution, il est nécessaire de relancer plusieurs fois l'algorithme avec des conditions initiales aléatoires différentes, en conservant la meilleure solution.

Un inconvénient de l'algorithme de descente de gradient à pas constant est la sensibilité au choix de η : s'il est trop petit, la convergence est très lente et, s'il est trop grand, l'algorithme peut diverger. Il existe des algorithmes d'optimisation plus rapides, tels que la *méthode de Newton*. Cette méthode consiste à faire à chaque itération k un développement limité au second ordre de la fonction à minimiser, ici $\widehat{R}(\mathbf{w})$ au voisinage de l'estimation courante $\mathbf{w}(k)$ de la solution. On a :

$$\begin{aligned}\widehat{R}(\mathbf{w}) = \widehat{R}(\mathbf{w}(k)) + (\mathbf{w} - \mathbf{w}(k))'\frac{d\widehat{R}}{d\mathbf{w}}(\mathbf{w}(k)) + \\ \frac{1}{2}(\mathbf{w} - \mathbf{w}(k))'\frac{d^2\widehat{R}}{d\mathbf{w}d\mathbf{w}'}(\mathbf{w}(k))(\mathbf{w} - \mathbf{w}(k)) + \epsilon,\end{aligned}$$

la notation $\frac{d^2 \ln L}{d\mathbf{w}d\mathbf{w}'}(\mathbf{w}(k))$ désigne la *matrice hessienne* en $\mathbf{w}(k)$. Notons H cette matrice. On obtient, en dérivant par rapport à $\mathbf{w}$:

$$\frac{d\widehat{R}}{d\mathbf{w}}(\mathbf{w}) \approx \frac{d\widehat{R}}{d\mathbf{w}}(\mathbf{w}(k)) + H(\mathbf{w} - \mathbf{w}(k)).$$

On a donc, en négligeant l'approximation :

$$\frac{d\widehat{R}}{d\mathbf{w}}(\mathbf{w}) = 0 \Leftrightarrow \mathbf{w} = \mathbf{w}(k) - H^{-1}\frac{d\widehat{R}}{d\mathbf{w}}(\mathbf{w}(k)).$$

La méthode de Newton consiste à appliquer itérativement cette formule. On part d'une valeur initiale aléatoire $\mathbf{w}(0)$ du vecteur $\mathbf{w}$, puis on calcule à chaque itération $k+1$ une nouvelle estimation en fonction de l'estimation précédente par

$$\mathbf{w}(k+1) = \mathbf{w}(k) - H^{-1}\frac{d\widehat{R}}{d\mathbf{w}}(\mathbf{w}(k)),$$

jusqu'à ce qu'un critère d'arrêt (par exemple, sur la norme du gradient) soit vérifié.