Apprentissage: cours 7
Modélisation probabiliste, régression et classification linéaire

Guillaume Obozinski

10 avril 2013

Maximum de vraisemblance : cas génératif et conditionnel

Modèle génératif :

$$\mathcal{R}(\theta) = -\mathbb{E}[\log(p_\theta(X,Y))] \qquad \widehat{\mathcal{R}}_n(\theta) = -\frac{1}{n}\sum_{i=1}^n \log(p_\theta(x_i,y_i))$$

Modèle conditionnel :

$$\mathcal{R}(\theta) = -\mathbb{E}[\log(p_\theta(Y|X)) \mid X] \qquad \widehat{\mathcal{R}}_n(\theta) = -\frac{1}{n}\sum_{i=1}^n \log(p_\theta(y_i|x_i))$$

Modèle probabiliste pour la régression linéaire

On considère la modélisation probabiliste d'un couple entrée sortie (X,Y) avec $\mathcal{X} = \mathbb{R}^p$ et $\mathcal{Y} = \mathbb{R}$ Précisément on ne modélise que la loi conditionnelle de Y sachant X comme étant $Y = \mathbf{w}^\top X + \varepsilon$ avec $\varepsilon \sim \mathcal{N}(0,\sigma^2)$ pour les paramètres $\theta = (\mathbf{w},\sigma^2) \in \mathbb{R}^p \times \mathbb{R}_+$. La log-vraisemblance du modèle est

$$-\log(p_\theta(y|\mathbf{x})) = \frac{1}{2\sigma^2}(y - \mathbf{w}^\top\mathbf{x})^2 + \frac{1}{2}\log(2\pi\sigma^2).$$

L'estimateur du maximum de vraisemblance en $\mathbf{w}$ est donc celui de la régression linéaire.

Exercice 1. Calculer l'EMV de σ^2

Complement sur la régression linéaire

En haute dimension, i.e., quand $p > n$, le prédicteur de la régression linéaire se calcule plus efficacement qu'avec la formule issue des équations normales

Exercice 2. **(Lemme d'inversion de matrice)** Soit $\mathbf{X} \in \mathbb{R}^{n\times p}$ tel que $\mathbf{I} + \mathbf{X}^\top\mathbf{X}$ est inversible. Quel est la complexité de l'inversion matricielle d'une matrice $p \times p$ en général ? Si $p > n$, comment calculer $(\mathbf{I}+\mathbf{X}^\top\mathbf{X})^{-1}$ plus efficacement ?

En se basant sur le résultat de l'exercice on

Proposition 1. *Soit $\hat{f}_\lambda$ le prédicteur de la régression linéaire régularisée pour une matrice de design $\mathbf{X}$ et un vecteur de variables de sortie $\mathbf{y}$. Dénotons $\mathbf{K} = \mathbf{X}\mathbf{X}^\top$ la matrice de Gram des données. On a*

$$f : \mathbf{x}' \mapsto \mathbf{y}^\top(n\lambda\mathbf{I} + \mathbf{K})^{-1}\mathbf{X}\,\mathbf{x}'.$$

Modèle de la régression logistique

On considère de le problème de la classification binaire, i.e. $\mathcal{X} = \mathbb{R}^p$ et $\mathcal{Y} = \{0, 1\}$. Il s'agit de modéliser $\mathbb{P}(Y|X = x)$. Cette distribution est entièrement caractérisée par le rapport de vraisemblance

$$\frac{p(X = x|Y = 1)}{p(X = x|Y = 0)} = \frac{p(Y = 1|X = x)}{p(Y = 0|X = x)} \frac{1 - \pi}{\pi} \quad \text{avec} \quad \pi = \mathbb{P}(Y = 1).$$

On modélise donc la fonction $f(x) = \log(\frac{p(Y=1|X=x)}{p(Y=0|X=x)})$ ce qui conduit au modèle

$$\mathbb{P}(Y = 1|X = x) = \sigma(f(x)) \qquad \text{avec} \qquad \sigma(z) = \frac{1}{1 + e^{-z}}.$$

La fonction σ appelée *fonction logistique* satisfait les propriétés :
- $\sigma(-z) = 1 - \sigma(z)$
- $\sigma'(z) = \sigma(z)(1 - \sigma(z)) = \sigma(z)\sigma(-z)$

On se restreint aux fonctions $f_{\mathbf{w}} : \mathbf{x} \mapsto \mathbf{w}^\top \mathbf{x}$ linéaires.

Comme $-\log(\mathbb{P}(Y = 1|X = x) = \log(1 + e^{-z})$ le problème de maximisation de la vraisemblance est équivalent à la minimisation du risque empirique pour la *perte logistique* définie par

$$-\ell(y, a) = y \log(\sigma(a)) + (1 - y) \log(\sigma(-a))$$

On a donc : $\widehat{\mathcal{R}}_n(\mathbf{w}) = -\frac{1}{n} \sum_{i=1}^{n} y_i \log(\sigma(\mathbf{w}^\top \mathbf{x}_i)) + (1 - y_i) \log(1 - \sigma(\mathbf{w}^\top \mathbf{x}_i))$.

Comme $\frac{\partial}{\partial a}\ell(y, a) = \sigma(a) - y$, on a $\nabla_{\mathbf{w}} \widehat{\mathcal{R}}_n(\mathbf{w}) = \sum_{i=1}^{n} \mathbf{x}_i(y_i - \sigma(\mathbf{w}^\top x_i))$ qui ne se résout pas sous forme analytique. On doit donc recourir à un algorithme itératif

Moindres carrés pondérés itérés

Si on peut se permettre un algorithme quadratique en p on privilégiera l'algorithme de Newton.

La dérivée seconde de la perte logistique est $\frac{\partial}{\partial a} - \ell(a, y) = \sigma(a)\sigma(-a)$ d'où le développement de Taylor du risque empirique à l'ordre 2. On note $\eta_i = \sigma(\mathbf{x}_i^\top \mathbf{w}^{(t)})$, $\boldsymbol{\eta} = (\eta_i)_{1 \leqslant i \leqslant p} \in \mathbb{R}^p$ et $\mathbf{D}(\boldsymbol{\eta}) = \mathrm{Diag}\big((\eta_i(1 - \eta - i))_{1 \leqslant i \leqslant p}\big)$.

$$\begin{aligned}
\widehat{\mathcal{R}}_n(\mathbf{w}) &\approx \widehat{\mathcal{R}}_n\left(\mathbf{w}^{(t)}\right) + \sum_{i=1}^{n} (y_i - \eta_i)\mathbf{x}_i^\top (\mathbf{w} - \mathbf{w}^{(t)}) + \frac{1}{2}(\mathbf{w} - \mathbf{w}^{(t)})^\top \Big[\sum_{i=1}^{n} \eta_i(1 - \eta_i)\mathbf{x}_i\mathbf{x}_i^\top\Big](\mathbf{w} - \mathbf{w}^{(t)}) \\
&\approx (\mathbf{y} - \boldsymbol{\eta})\mathbf{X}(\mathbf{w} - \mathbf{w}^{(t)}) + \frac{1}{2}(\mathbf{w} - \mathbf{w}^{(t)})\mathbf{X}^\top \mathbf{D}(\boldsymbol{\eta})\mathbf{X}(\mathbf{w} - \mathbf{w}^{(t)})
\end{aligned}$$

Exercice 3. (Implémentation) En déduire l'algorithme de Newton pour la régression logistique. Expliquer le terme de "moindres carrés pondérés itérés" (algorithme IRLS : Iterated Reweighted Least Squares) au vu de l'algorithme de Newton.