Apprentissage: cours 4b
Modélisation probabiliste, régression et classification linéaire

Guillaume Obozinski

8 mars 2012

Principe : proposer un modèle probabiliste des données. Déterminer les paramètres du modèle en utilisant le *principe du maximum de vraisemblance*, prédire grâce au modèle obtenu. Avant de considérer des modèles impliquant des entrées et des sorties, on considère un modèle de données simple.

Soit ν une mesure de référence (la mesure de comptage sur $\mathbb{N}$, la mesure de Lebesgue sur $\mathbb{R}$).

Définition 1. (Modèle) Soit $\Theta \subset \mathbb{R}^p$ un ensemble de paramètres. On appelle modèle $\mathcal{P}$ un ensemble de lois de probabilités à valeur dans $\mathcal{Y}$, possédant un densité par rapport à la mesure de référence sur $\mathcal{Y}$ et indexés par Θ : $\mathcal{P} = \{p_\theta d\mu \mid \theta \in \Theta\}$

Exemple 1. Modèles Binomial, Multiomial, Gaussien univarié et multivarié.

Définition 2. (Vraisemblance) Soit une donnée $y \in \mathcal{Y}$. On appelle *vraisemblance* la fonction $\theta \mapsto p_\theta(x)$

On considère un ensemble d'entraînement i.i.d. $y_1, \dots, y_n$ (dans ce contexte aussi *échantillon*). La vraisemblance de l'ensemble d'entraînement est

$$L(\theta) := \Pi_{i=1}^n p_\theta(y_i)$$

Principe du maximum de vraisemblance

Principe : un bon choix de paramètre est un choix de paramètre qui maximise la probabilité des données observées, i.e. qui maximise la vraisemblance.

- principe du à Sir Ronald Fisher
- validé a posteriori par les bonnes propriétés du maximum de vraisemblance

Reformulation en terme de risque

On définit comme fonction de perte la log-vraisemblance $\ell(\theta, y) = -\log(p_\theta(y))$. Le risque associé est

$$\mathcal{R}(\theta) = -\mathbb{E}[\log(p_\theta(Y))]$$

En particulier si $Y \sim p_{\theta_0} d\mu$ pour $\theta_0 \in \Theta$ alors la le paramètre cible est $\theta^* = \theta_0$. Le risque empirique est alors par définition

$$\widehat{\mathcal{R}}_n(\theta) = -\frac{1}{n}\sum_{i=1}^n \log(p_\theta(y_i))$$

Le principe de minimisation du risque empirique coïncide alors avec le principe du maximum de vraisemblance de Fisher.

On considère les modèles de Bernoulli, multinomial et gaussien univarié et multivarié.

Exercice 1. Calculer l'estimateur du maximum de vraisemblance pour ces modèles.

La formulation s'étend au cas de paires de données d'entrées et de sortie.
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 2. 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 3. (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*

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