

# Apprentissage: cours 7

## Modélisation probabiliste, maximum de vraisemblance

Simon Lacoste-Julien

30 octobre 2015

### Résumé

Modèle probabiliste pour la régression et la classification. Maximum de vraisemblance. Divergence de Kullback-Leibler.

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  $\mu$  une mesure de référence sur  $\mathcal{Y}$  (la mesure de comptage sur  $\mathbb{N}$ , la mesure de Lebesgue sur  $\mathbb{R}$ ).

**Définition 1.** (Modèle paramétrique de distributions) 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 une densité par rapport à la mesure de référence sur  $\mathcal{Y}$  et indexés par  $\Theta : \mathcal{P} = \{p_\theta d\mu \mid \theta \in \Theta\}$

*Exemple 1.* Modèles Binomial, Multinomial, 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

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

## 1 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 dû à Sir Ronald Fisher
- validé a posteriori par les bonnes propriétés du maximum de vraisemblance
- À noter que maximiser la vraisemblance est équivalent à maximiser le *log* de la vraisemblance car  $\log$  est une fonction strictement monotone.

### 1.1 Reformulation en terme de risque

Ici, nous sommes dans le cadre de l'*estimation de densité* (étant données des observations, nous voulons identifier la distribution qui a généré les données). Les actions possibles sont donc  $\mathcal{A} = \Theta$ ; une action  $a$  est de choisir une distribution  $p_\theta$ . La perte standard utilisée dans ce cadre est le négatif de la log-vraisemblance  $\ell(\theta, y) = -\log(p_\theta(y))$  (appelée aussi *perte-log*). Le risque associé est alors

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

En particulier si  $Y \sim p_{\theta_0} d\mu$  pour  $\theta_0 \in \Theta$  alors 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.

## 1.2 Divergence de Kullback-Leibler

Considérons l'excès de risque (en supposant que les données sont générés à partir de  $p_{\theta_0}$ ) :

$$\begin{aligned} \mathcal{R}(\theta) - \mathcal{R}(\theta_0) &= -\mathbb{E}_{\theta_0}[\log(p_{\theta}(Y))] + \mathbb{E}_{\theta_0}[\log(p_{\theta_0}(Y))] \\ &= \mathbb{E}_{\theta_0}\left[\log\left(\frac{p_{\theta_0}(Y)}{p_{\theta}(Y)}\right)\right] =: KL(p_{\theta_0}||p_{\theta}) \end{aligned}$$

$KL(p||q)$  est la *divergence de Kullback-Leibler* :

$$KL(p||q) := \int_{\mathcal{Y}} p(y) \log \frac{p(y)}{q(y)} d\mu(y)$$

**Propriétés :**

- $KL(p||q) \geq 0$  (par Jensen).
- $KL(p||p) = 0$  (et donc on voit pourquoi  $\theta_0$  minimisait le risque).
- KL n'est pas une métrique (non-symétrique, pas d'inégalité triangulaire, etc.), mais est souvent interprétée comme une distance généralisée sur les distributions.
- Avec un abus de notation, on peut écrire la minimisation du risque empirique pour la perte-log avec la KL, où  $\hat{p}_n$  est la distribution empirique :

$$\min_{\theta \in \Theta} KL(\hat{p}_n || p_{\theta}).$$

## 1.3 Exemples de maximum de vraisemblance

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

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

### 1.3.1 Surapprentissage dans le modèle multinomial

Pour le modèle multinomial : chaque observation  $Y_i$  est une variable discrète qui peut prendre  $k$  valeurs. On l'encode avec un vecteur  $\mathbf{y}_i \in \{0, 1\}^k$  tel que  $y_{ij} = 1$  si  $Y_i$  prend la valeur  $j$  et 0 autrement (pour  $j$  dans  $\{1, \dots, k\}$ ). Le modèle multinomial  $\mathbf{y}_i \sim \text{Mult}(\pi, 1)$  donne la probabilité aux vecteurs  $\mathbf{y}_i$  valides (donc une seule entrée égale à 1) de :  $p(\mathbf{y}_i) = \prod_{j=1}^k \pi_j^{y_{ij}}$ .

Soit  $n_j := \sum_{i=1}^n y_{ij}$  le nombre de fois que l'option  $j$  a été observée dans les données. L'estimateur du maximum de vraisemblance est :

$$\hat{\pi}_j = \frac{n_j}{n}.$$

Si le nombre d'options  $k$  est élevé (par exemple, on pourrait modéliser la probabilité des mots dans un texte ; chaque mot est une option ; donc  $k$  peut facilement être dans les centaines de milliers), i.e.  $p = k > n$ , alors plusieurs options ne sont pas observées et donc on estime leur probabilité à zéro. La perte-log pour ces options est infinie ; le modèle est dans le surapprentissage ! Le maximum de vraisemblance surapprend quand  $p > n$ .

Pour le modèle multinomial, on peut régler ce problème en régularisant. Une approche est avec la *méthode bayésienne* avec un à priori sur les paramètres  $p(\theta)$  et en utilisant le maximum à-posteriori  $\text{argmax}_{\theta} p(\theta|D_n) = \text{argmax}_{\theta} p(D_n|\theta)p(\theta)$  (avec la règle de Bayes) plutôt que le maximum de vraisemblance  $\text{argmax}_{\theta} p(D_n|\theta)$ . Les méthodes bayésiennes donnent souvent des estimateurs avec de bonnes propriétés fréquentistes.

## 1.4 Modèles conditionnels

La formulation s'étend au cas de paires de données d'entrées et de sortie.

Modèle génératif :

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

Modèle conditionnel :

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

### Modèle probabiliste pour la régression linéaire (modèle conditionnel)

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 *conditionnelle* du modèle est

$$\widehat{\mathcal{R}}_n(\theta) = -\frac{1}{n} \sum_{i=1}^n \log(p_\theta(y_i|\mathbf{x}_i)) = \frac{1}{2n\sigma^2} \sum_{i=1}^n (y_i - \mathbf{w}^\top \mathbf{x}_i)^2 + \frac{1}{2} \log(2\pi\sigma^2).$$

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

*Exercice 2.* Calculer l'EMV de  $\sigma^2$