Estimation de densité paramétrique par maximum de vraisemblance (ML) ou maximum a posteriori (MAP)

September 29, 2016

1 Cadre

Soit un ensemble de données $D_n = \{x^{(1)}, \ldots, x^{(n)}\}$. Typiquement avec $x^{(i)} \in \mathbb{R}^d$, qu'on suppose tirée i.i.d d'une distribution inconnue $p^*(x)$.

Soit une famille de lois de probabilité (fonction de densité ou fonction de masse de probabilité) paramétrée par θ.

On notera $p_\theta(x)$ ou $p(x;\theta)$ ou $p(x|\theta)$ la (densité de) probabilité associée à un point $x \in \mathbb{R}^d$ par cette loi pour la valeur de paramètre θ.

Par exemple, si on choisit comme famille les Gaussiennes isotropique, paramétrée par leur centre μ et leur variance σ^2, on aura:

$$p(x|\theta = \{\mu, \sigma^2\}) = \frac{1}{(2\pi)^{\frac{d}{2}}\sigma^d} \exp\left(-\frac{1}{2}\frac{\|x-\mu\|^2}{\sigma^2}\right)$$

On va **chercher** parmi cette famille, la **loi/fonction de densité de probabilité** qui **"explique le mieux les données"** de notre ensemble de données D_n. Chercher une loi/fonction de probabilité dans cette famille paramétrée, revient à trouver une bonne valeur des paramètres θ.

2 Maximum de vraisemblance

Vraisemblance (*likelihood*)

"probabilité que les données aient été générées par la loi de paramètre θ".

$$\begin{aligned}
\mathcal{L}(\theta) &= p(D_n|\theta) \\
&= p(x^{(1)}|\theta)p(x^{(2)}|\theta)\ldots p(x^{(n)}|\theta) \\
&= \prod_{i=1}^{n} p(x^{(i)}|\theta)
\end{aligned}$$

Log-vraisemblance (*log-likelihood*)

$$\begin{aligned}\ell(\theta) &= \log \mathcal{L}(\theta) \\ &= \log \left(\prod_{i=1}^{n} p(x^{(i)}|\theta) \right) \\ &= \sum_{i=1}^{n} \log(p(x^{(i)}|\theta))\end{aligned}$$

Principe de maximisation de la (log) vraisemblance (*maximum likelihood*)

$$\begin{aligned}\theta^* &= \arg\max_{\theta} \mathcal{L}(\theta) \\ &= \arg\max_{\theta} \ell(\theta)\end{aligned}$$

Lien avec le principe de minimisation du risque empirique

$$\begin{aligned}\theta^* &= \arg\max_{\theta} \ell(\theta) \\ &= \arg\max_{\theta} \sum_{i=1}^{n} \log(p(x^{(i)}|\theta)) \\ & \arg\min_{\theta} - \sum_{i=1}^{n} \log(p(x^{(i)}|\theta)) \\ &= \arg\min_{\theta} \sum_{i=1}^{n} \Big(-\log(p(x^{(i)}|\theta)) \Big) \\ &= \arg\min_{\theta} \frac{1}{n} \sum_{i=1}^{n} \Big(-\log(p(x^{(i)}|\theta)) \Big) \\ &= \arg\min_{\theta} \hat{R}(p_\theta, D_n)\end{aligned}$$

avec

$$\hat{R}(p_\theta, D_n) = \frac{1}{n} \sum_{i=1}^{n} L(p_\theta(x^{(i)}))$$

si on pose

$$L(p_\theta(x^{(i)})) = -\log p_\theta(x^{(i)})$$

Le principe de maximisation de la (log) vraissemblance es tun cas particulier de minimisation de risque empirique.

Comment résoudre le problème d'optimisation. Ex: trouver les paramètres de la densité Gaussienne

Il s'agit de résoudre un problème d'optimisation: trouver la valeur des paramètres qui maximisent la log-vraisemblance ou (ce qui est équivalent) qui minimisent le risque empirique défini précédemment:

$$
\begin{aligned}
\theta^* &= \arg\max_\theta \ell(\theta) \\
&= \arg\max_\theta \sum_{i=1}^n \log(p(x^{(i)}|\theta)) \\
&= \arg\max_{\theta=\{\mu,\sigma^2\}} \sum_{i=1}^n \left(\log\left(\frac{1}{(2\pi)^{\frac{d}{2}}\sigma^d}\exp\left(-\frac{1}{2}\frac{\|x^{(i)}-\mu\|^2}{\sigma^2}\right)\right)\right)
\end{aligned}
$$

Dans les cas les plus simples seulement (tel que la simple d'une Gaussienne ici) le problème d'optimisation peut se résoudre analytiquement en trouvant analytiquement la valeur des paramètres qui annule la dérivée:

On essaye de résoudre l'équation

$$\frac{\partial \ell(\theta)}{\partial \theta} = 0$$

ex: cherchons pour commencer uniquement μ

$$
\begin{aligned}
\frac{\partial \ell}{\partial \mu} = \frac{\partial \ell}{\partial(\mu_1,\ldots,\mu_d)} &= 0 \\
\frac{\partial}{\partial \mu}\sum_{i=1}^n \left(\log\left(\frac{1}{(2\pi)^{\frac{d}{2}}\sigma^d}\exp\left(-\frac{1}{2}\frac{\|x^{(i)}-\mu\|^2}{\sigma^2}\right)\right)\right) &= 0 \\
\sum_{i=1}^n \frac{\partial}{\partial \mu}\left(\log\left(\frac{1}{(2\pi)^{\frac{d}{2}}\sigma^d}\right) + \log\left(\exp\left(-\frac{1}{2}\frac{\|x^{(i)}-\mu\|^2}{\sigma^2}\right)\right)\right) &= 0 \\
\sum_{i=1}^n \left(\frac{\partial}{\partial \mu}\left(\log\left(\frac{1}{(2\pi)^{\frac{d}{2}}\sigma^d}\right)\right) + \frac{\partial}{\partial \mu}\log\left(\exp\left(-\frac{1}{2}\frac{\|x^{(i)}-\mu\|^2}{\sigma^2}\right)\right)\right) &= 0 \\
\sum_{i=1}^n \left(\frac{\partial}{\partial \mu}\left(-\frac{1}{2}\frac{\|x^{(i)}-\mu\|^2}{\sigma^2}\right)\right) &= 0 \\
-\frac{1}{2\sigma^2}\sum_{i=1}^n \frac{\partial}{\partial \mu}\|x^{(i)}-\mu\|^2 &= 0 \\
\sum_{i=1}^n \frac{\partial}{\partial \mu}\|x^{(i)}-\mu\|^2 &= 0 \\
\sum_{i=1}^n \frac{\partial}{\partial \mu}\sum_{j=1}^d (x_j^{(i)}-\mu_j)^2 &= 0
\end{aligned}
$$

$$\sum_{i=1}^{n} \frac{\partial}{\partial(\mu_1,\ldots,\mu_d)} \sum_{j=1}^{d} (x_j^{(i)} - \mu_j)^2 \quad = \quad 0$$

Donc, concenrant la dérivée par rapport à la $k^{ième}$ composante μ_k du vecteur μ:

$$\begin{aligned}
\sum_{i=1}^{n} \frac{\partial}{\partial \mu_k} \sum_{j=1}^{d} (x_j^{(i)} - \mu_j)^2 &= 0 \\
\sum_{i=1}^{n} \frac{\partial}{\partial \mu_k} (x_k^{(i)} - \mu_k)^2 &= 0 \\
\sum_{i=1}^{n} -2(x_k^{(i)} - \mu_k) &= 0 \\
\sum_{i=1}^{n} (x_k^{(i)} - \mu_k) &= 0 \\
\left(\sum_{i=1}^{n} x_k^{(i)}\right) - \left(\sum_{i=1}^{n} \mu_k\right) &= 0 \\
\left(\sum_{i=1}^{n} x_k^{(i)}\right) - n\mu_k &= 0 \\
\sum_{i=1}^{n} x_k^{(i)} &= n\mu_k \\
\mu_k &= \frac{1}{n} \sum_{i=1}^{n} x_k^{(i)}
\end{aligned}$$

ce qui, pour toutes les composantes, $\mu_1, \ldots, \mu_d = \mu$ se résume, sous forme vectorielle:$\mu = \frac{1}{n} \sum_{i=1}^{n} x^{(i)}$.

En **conclusion**, la valeur de μ , paramètre d'une densité Gaussienne isotropique, qui permet de maximiser la vraisemblance $p(D_n|\mu, \sigma^2)$ est la moyenne empirique des données $\mu = \frac{1}{n} \sum_{i=1}^{n} x^{(i)}$.

Évaluation hors-échantillon

On a ainsi *entraîné* les paramètres θ de notre distribution paramétrée p_θ sur un *ensemble de données d'entainement*. $D_{train} = D_n$.

Il nous intéresse de savoir si la distribution ainsi apprise modélise bien la distribution inconnue p^* c.a.d. "généralise bien" sur d'autres données provenant de cette distribution inconnue p^*.

Pour cela on se sera au préalable réservé un ensemble de validation D_{valid} (et/ou de test D_{test}) séparé des données d'entrainement $D_{train} = D_n$

On a trouvé, par maximum de vraisemblance, la valeur des paramètres θ^* qui maximisent la (log) vraissemblance sur D_{train}.

On peut évaluer/estimer la performance de généralisation sur D_{valid} c.a.d. qu'on peut calculer la log-vraisemblance ou log-vraisemblance moyenne sur D_{valid} qui indiquera **"à quel point les données de D_{valid} sont "probables" selon notre modèle de distribution p_θ appris sur D_{train}"**).

Log-vraisemblance moyenne sur l'ensemble de validation est:

$$\frac{1}{|D_{valid}|} \sum_{x \in D_{valid}} \log(p(x|\theta^*))$$

Régularisation: maximisation de log-vraisemblance régularisée

3 Alternative: l'approche Bayesienne

Cadre Bayesien: ce qui le différencie de l'approche de maximum de vraisemblance

Probabilité a priori et a postériori Bayesien sur les paramètres

Pure approche Bayesienne pour la prédiction

Un compromis: le maximum a posteriori (MAP)

4 Maximum de vraisemblance conditionnelle (ou MAP conditionnel)