Démo 4
Apprentissage par maximum de vraisemblance dans un modèle gaussien isotropique

Pierre-Antoine Manzagol

September 23, 2009

Le modèle gaussien isotropique est le cas particulier du modèle gaussien multivarié pour lequel on a $\Sigma = \sigma^2\mathbf{I}$ (la matrice de covariance vaut un scalaire fois la matrice identité). Avec cette forme simplifiée le déterminant est très facile à calculer: le déterminant d'une matrice diagonale est tout simplement le produit des valeurs sur la diagonale. Dans notre cas $|\Sigma| = (\sigma^2)^d = \sigma^{2d}$ (où d est le nombre de dimensions). On a donc $|\Sigma|^{\frac{1}{2}} = \sigma^d$. On a aussi que $\Sigma^{-1} = \frac{1}{\sigma^2}\mathbf{I}$.

Grâce à ces simplifications, le modèle définit la fonction de densité $p(\mathbf{x})$ suivante:

$$p(\mathbf{x}) = \frac{1}{(2\pi)^{\frac{d}{2}}\sigma^d} e^{-\frac{(\mathbf{x}-\mu)^T(\mathbf{x}-\mu)}{2\sigma^2}}$$

Rappel: μ est la moyenne, un vecteur de dimension d (la dimension de x aussi), et σ est l'écart-type donc σ^2 est la variance. Essayons maintenant d'apprendre les valeurs des paramètres μ et σ de ce modèle paramétrique. Pour ce faire on va suivra l'approche du maximum de vraisemblance. Considérons la probabilité des données sous notre modèle. Notre objectif va être d'ajuster les paramètres de façon à ce que cette probabilité soit la plus forte possible.

Prenons un exemple d'entrainement au hasard: $\mathbf{x}_i$. (Ici, on fait de l'estimation de densité, donc il n'y a pas de y_i, ou alors on l'ignore!) Le modèle lui donne une probabilité de $p(\mathbf{x}_i)$. Maintenant, si on considère la probabilité de l'ensemble d'entrainement on doit combiner ces probabilités en les multipliant: $\prod_i p(\mathbf{x}_i)$ (parce que les exemples sont indépendemment et identiquement tirés) où i est un index sur les n exemples d'entraînement. C'est cette quantité que l'on veut maximiser.

Allons-y:

$$\prod_i p(\mathbf{x_i}) = \prod_i \frac{1}{(2\pi)^{\frac{d}{2}}\sigma^d} e^{-\frac{(\mathbf{x}_i-\mu)^T(\mathbf{x}_i-\mu)}{2\sigma^2}}$$

On va maintenant prendre le log de cette quantité. On parle ici du log en base e. Ça ne change pas grand chose (log est une fonction monotone) et ça va nous simplifier la vie. Ça ne change pas grand chose parce que bien que ce n'est pas la même quantité, les deux ont leur maximum au même endroit (valeurs de μ et σ).

$$\log\left(\prod_i p(\mathbf{x_i})\right) = \log\left(\prod_i \frac{1}{(2\pi)^{\frac{d}{2}}\sigma^d} e^{-\frac{(\mathbf{x}_i-\mu)^T(\mathbf{x}_i-\mu)}{2\sigma^2}}\right)$$

Le log d'un produit est égal à la somme des log:

$$\log\left(\prod_i p(\mathbf{x_i})\right) = \sum_i \log\left(\frac{1}{(2\pi)^{\frac{d}{2}}\sigma^d} e^{-\frac{(\mathbf{x}_i-\mu)^T(\mathbf{x}_i-\mu)}{2\sigma^2}}\right)$$

On peut encore simplifier! Quelques rappels:

$$\log ab = \log a + \log b$$

$$\log a^b = b \log a$$

$$\log e = 1$$

$$\log \frac{1}{a} = \log a^{-1} = -\log a$$

Simplifions encore cette quantité que l'on désire maximiser:

$$log\left(\prod_i p(\mathbf{x_i})\right) = \sum_i \left(-\log\left((2\pi)^{\frac{d}{2}}\sigma^d\right) - \frac{(\mathbf{x}_i - \mu)^T(\mathbf{x}_i - \mu)}{2\sigma^2}\right)$$

$$log\left(\prod_i p(\mathbf{x_i})\right) = \sum_i \left(-\frac{d}{2}\log(2\pi) - d\log\sigma - \frac{(\mathbf{x}_i - \mu)^T(\mathbf{x}_i - \mu)}{2\sigma^2}\right)$$

Maintenant on va sortir de la somme tout ce qu'on est capable (ce qui ne dépend pas de i):

$$log\left(\prod_i p(\mathbf{x_i})\right) = -\frac{nd}{2}\log(2\pi) - nd\log\sigma - \frac{\sum_i(\mathbf{x}_i - \mu)^T(\mathbf{x}_i - \mu)}{2\sigma^2}$$

En résumé, on vient simplement de mettre la quantité que l'on veut maximiser sous une autre forme. On va maintenant chercher le maximum. Un maximum est un point où deux choses se produisent: la dérivée s'annule (c'est vrai pour un minimum et un maximum) et la dérivée seconde est négative (positive avec un minimum). Calculons d'abord la dérivée par rapport à μ.

$$\frac{\partial}{\partial\mu} log\left(\prod_i p(\mathbf{x_i})\right) = 0 + 0 - \frac{\sum_i 2(\mu - \mathbf{x}_i)}{2\sigma^2}$$

(Notez la dérivée du produit scalaire des vecteurs qui est plus complexe.) Si maintenant on annule cette quantité, on obtient:

$$\frac{\partial}{\partial\mu} log\left(\prod_i p(\mathbf{x}_i)\right) = 0$$

$$\sum_i(\mu - \mathbf{x}_i) = \sum_i \mu - \sum_i \mathbf{x}_i = n\mu - \sum_i \mathbf{x}_i = 0$$

$$\mu_{MaxVraiss} = \frac{\sum_i \mathbf{x}_i}{n}$$

Notez qu'on est chanceux ici, la valeur de μ que l'on obtient ne dépend pas de σ. On fait de même en dérivant par rapport à σ (ou $sigma^2$ si on voulait).

$$\frac{\partial}{\partial\sigma} log\left(\prod_i p(\mathbf{x}_i)\right) = 0$$

$$-\frac{nd}{\sigma} + \frac{4\sigma\sum_i(\mathbf{x}_i - \mu)^T(\mathbf{x}_i - \mu)}{4\sigma^4} = 0$$

$$\sigma^2_{MaxVraiss} = \frac{\sum_i(\mathbf{x}_i - \mu)^T(\mathbf{x}_i - \mu)}{nd}$$

Normalement il faudrait vérifier la dérivée seconde dans les deux cas pour s'assurer d'avoir des maxima et non des minima ou des points de selle.