

• Objectif

- estimer la densité $p(\mathbf{x})$
- étant donné $D_n = (X_1, X_2, \dots, X_n)$ tiré iid de $p(\mathbf{x})$

• Classification

- estimateurs a-posteriori: $\hat{p}(\mathbf{x}|C_i)$:
- estimateurs a-priori: $\hat{P}(C_i)$:
- on utilise $\hat{P}(C_i)\hat{p}(\mathbf{x}|C_i)$ comme des fonctions discriminantes

• Densité paramétrique: $p(\mathbf{x}) = p(\mathbf{x}|\theta)$

- θ : vecteur de paramètres

• Approche de maximum de vraisemblance

- les paramètres sont fixes mais inconnus
- maximiser la probabilité des données

• Approche bayésienne

- les paramètres sont aléatoires par nature
- les données sont utilisées pour raffiner la distribution a-priori des paramètres

• Principe de maximum de vraisemblance

- le vraisemblance de θ par rapport à D_n :

$$p(D_n|\theta) = \prod_{i=1}^n p(\mathbf{x}_i|\theta)$$

- fonction de log-vraisemblance:

$$l(\theta) = \ln p(D_n|\theta) = \sum_{i=1}^n \ln p(\mathbf{x}_i|\theta)$$

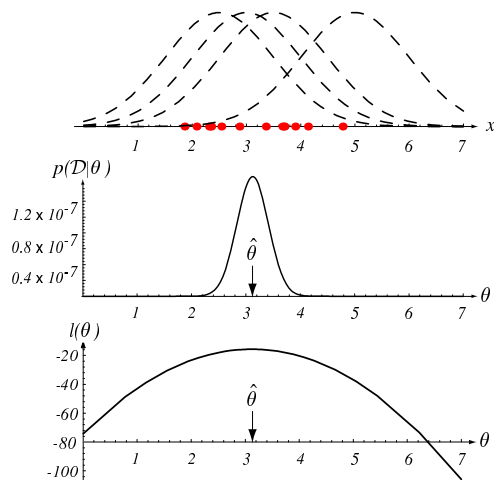
• Principe de maximum de vraisemblance

- estimation de maximum de vraisemblance

$$\begin{aligned} \hat{\theta} &= \arg \max_{\theta} p(D_n|\theta) = \arg \max_{\theta} \prod_{i=1}^n p(\mathbf{x}_i|\theta) \\ &= \arg \max_{\theta} \sum_{i=1}^n \ln p(\mathbf{x}_i|\theta) = \arg \max_{\theta} l(\theta) \end{aligned}$$

- conditions nécessaires:

$$\nabla_{\theta} l(\theta) = \sum_{i=1}^n \nabla_{\theta} \ln p(\mathbf{x}_i|\theta) = \mathbf{0}$$



• Le principe de minimisation du risque empirique

- perte: $L(\mathbf{x}, \hat{p}) = -\ln \hat{p}(\mathbf{x})$
- risque: $R(\hat{p}) = -\int p(\mathbf{x}) \ln \hat{p}(\mathbf{x}) d\mathbf{x}$
- pour une densité \hat{p} quelconque: $R(p) \leq R(\hat{p})$
- entropie:

$$H(p) = -\int p(\mathbf{x}) \ln p(\mathbf{x}) d\mathbf{x}$$

- "distance" de Kullback-Leibler:

$$d(\hat{p}, p) = -\int p(\mathbf{x}) \ln \frac{\hat{p}(\mathbf{x})}{p(\mathbf{x})} d\mathbf{x}$$

- Exemple: densité normale, μ inconnu

$$\hat{\mu} = \frac{1}{n} \sum_{i=1}^n x_i$$

- Exemple: densité normale, μ, Σ inconnus

$$\hat{\mu} = \frac{1}{n} \sum_{i=1}^n x_i$$

$$\hat{\Sigma} = \frac{1}{n} \sum_{i=1}^n (x_i - \hat{\mu})(x_i - \hat{\mu})^t$$

- Estimation bayésienne

- densité a-priori connue: $p(\theta)$

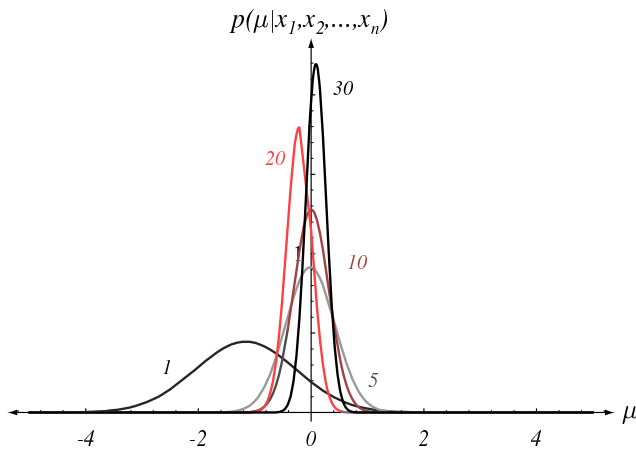
- densité a-posteriori $p(\theta|D_n) = "p(\theta) + D_n"$

- utilisation:

$$p(\mathbf{x}|D_n) = \int p(\mathbf{x}|\theta)p(\theta|D_n)d\theta \neq p(\mathbf{x}|\theta^*)$$

où

$$\theta^* = \arg \max_{\theta} p(\theta|D_n)$$



- Estimation bayésienne: cas normal

- densité a-posteriori $p(\mu|D_n) = ?, p(\sigma^2|D_n) = ?$

- cas univarié: $p(x|\mu) \sim N(\mu, \sigma^2)$

- densité a-priori $p(\mu) = N(\mu_0, \sigma_0^2)$

- Estimation bayésienne: cas normal

- théorème de Bayes:

$$p(\mu|D_n) = \frac{p(D_n|\mu)p(\mu)}{\int p(D_n|\mu)p(\mu)d\mu} = \alpha \prod_{i=1}^n p(x_i|\mu)p(\mu)$$

$$= \frac{1}{\sqrt{2\pi}\sigma_n} \exp\left[-\frac{1}{2}\left(\frac{\mu - \mu_n}{\sigma_n}\right)^2\right]$$

où

$$\mu_n = \left(\frac{n\sigma_0^2}{n\sigma_0^2 + \sigma^2}\right) \hat{\mu}_n + \frac{\sigma^2}{n\sigma_0^2 + \sigma^2} \mu_0$$

$$\sigma_n^2 = \frac{\sigma_0^2 \sigma^2}{n\sigma_0^2 + \sigma^2}$$

- Estimation bayésienne: cas normal

- densité conditionnelle de classe

$$p(x|D_n) = \int p(x, \mu|D_n)d\mu = \int p(x|\mu, D_n)p(\mu|D_n)d\mu$$

$$= \int p(x|\mu)p(\mu|D_n)d\mu$$

$$\sim N(\mu_n, \sigma^2 + \sigma_n^2)$$

- Avantages de l'approche bayésienne
 - connaissances a-priori intégrées doucement
 - tendance à mieux fonctionner pour les petites données
- Avantages de l'approche de maximum de vraisemblance
 - simplicité
 - interprétabilité
 - vitesse du calcul