Méthodes paramétriques

- Objectif
  - estimer la densité $p(\mathbf{x})$
  - étant donné $D_n = (X_1, X_2, \ldots, X_n)$ tiré iid de $p(\mathbf{x})$
- Classification
  - estimateurs a-posteriori: $\widehat{p}(\mathbf{x}|C_i)$:
  - estimateurs a-priori: $\widehat{P}(C_i)$:
  - on utilise $\widehat{P}(C_i)\widehat{p}(\mathbf{x}|C_i)$ comme des fonctions discriminantes

Méthodes paramétriques

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

Méthodes paramétriques

- 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:

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

Méthodes paramétriques

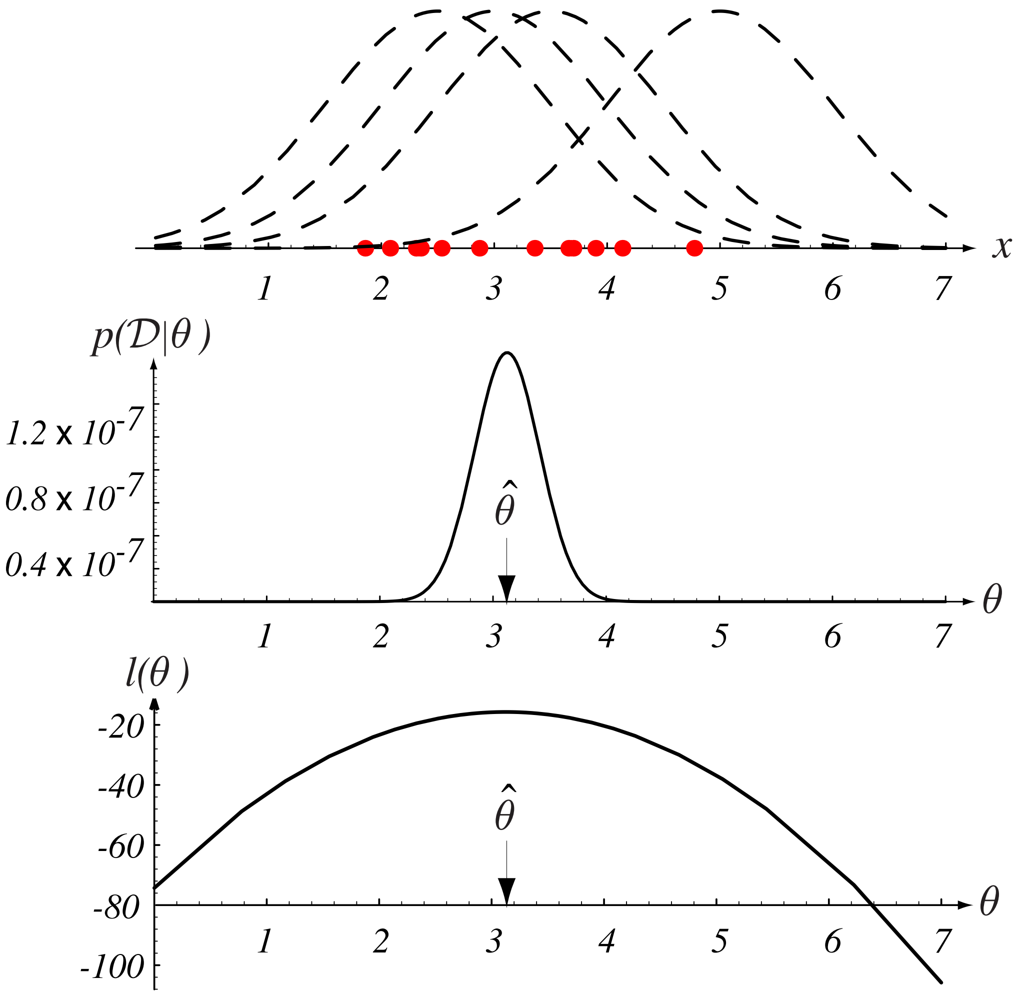
- Principe de maximum de vraisemblance
  - estimation de maximum de vraisemblance

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

  - conditions nécessaires:

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

Méthodes paramétriques



Méthodes paramétriques

- Le principe de minimisation du risque empirique
  - perte: $L(\mathbf{x}, \widehat{p}) = -\ln \widehat{p}(\mathbf{x})$
  - risque: $R(\widehat{p}) = -\int p(\mathbf{x}) \ln \widehat{p}(\mathbf{x}) d\mathbf{x}$
  - pour une densité $\widehat{p}$ quelconque: $R(p) \leq R(\widehat{p})$
  - entropie:
$$H(p) = -\int p(\mathbf{x}) \ln p(\mathbf{x}) d\mathbf{x}$$
  - “distance” de Kullback-Leibler:
$$d(\widehat{p}, p) = -\int p(\mathbf{x}) \ln \frac{\widehat{p}(\mathbf{x})}{p(\mathbf{x})} d\mathbf{x}$$

Méthodes paramétriques

- Exemple: densité normale, μ inconnu
  - $\widehat{\mu} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_i$
- Exemple: densité normale, μ, Σ inconnus
  - $\widehat{\mu} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_i$
  - $\widehat{\Sigma} = \frac{1}{n}\sum_{i=1}^{n} (\mathbf{x}_i - \widehat{\mu})(\mathbf{x}_i - \widehat{\mu})^t$

Méthodes paramétriques

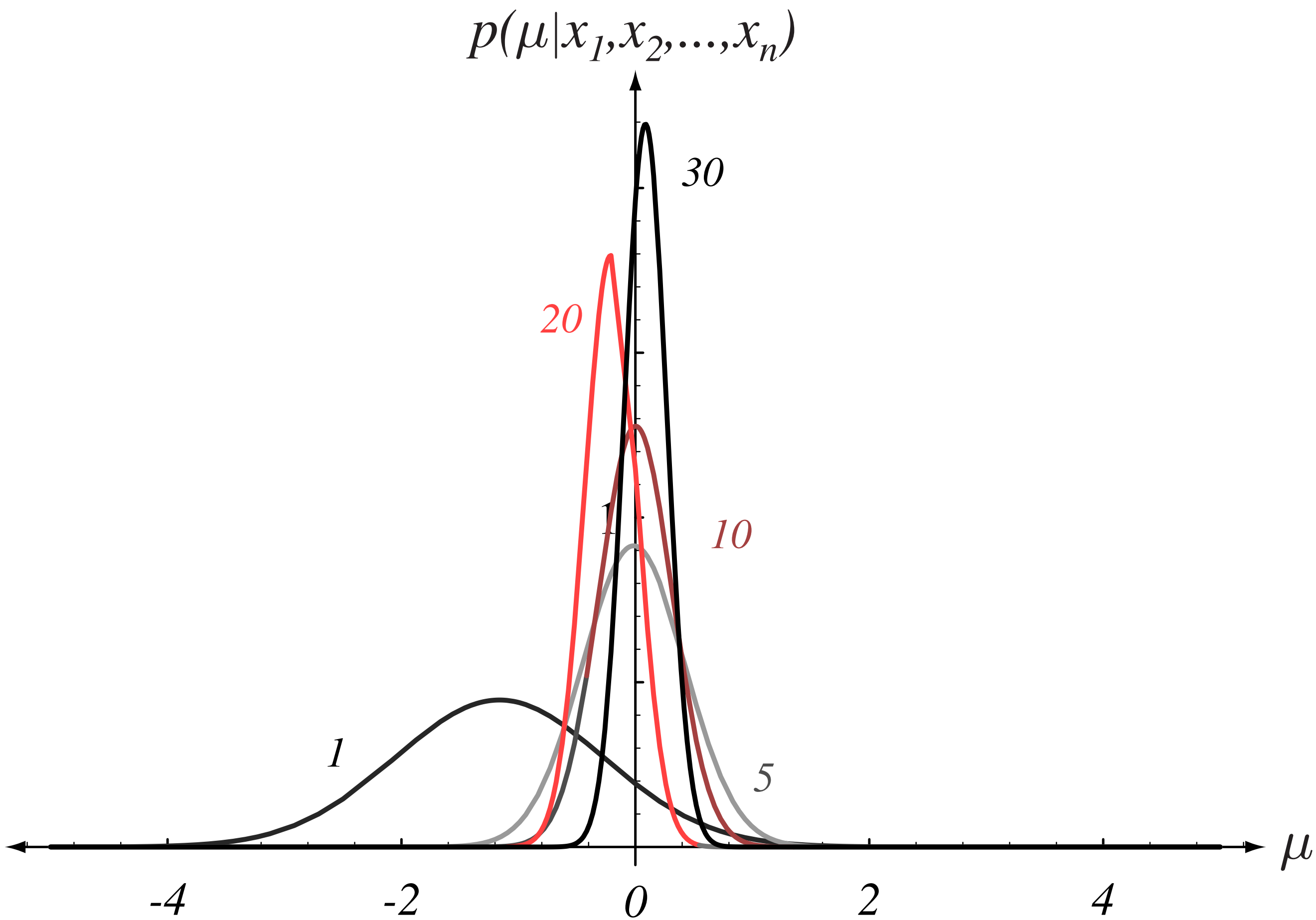
- Estimation bayesienne
  - densité a-priori connue: $p(\theta)$
  - densité a-posteriori $p(\theta|D_n) = \text{``}p(\theta) + D_n\text{''}$
  - utilisation:

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

où

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

Méthodes paramétriques



Méthodes paramétriques

- Estimation bayesienne: 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 $p(\mu) = N(\mu_0, \sigma_0^2)$

Méthodes paramétriques

- Estimation bayesienne: cas normal
  - théorème de Bayes:

$$
\begin{aligned}
p(\mu|D_n) &= \frac{p(D_n|\mu)p(\mu)}{\int p(D_n|\mu)p(\mu)d\mu} = \alpha \prod_{i=1}^{n} p(x_k|\mu)p(\mu) \\
&= \frac{1}{\sqrt{2\pi}\sigma_n} \exp\left[-\frac{1}{2}\left(\frac{\mu-\mu_n}{\sigma_n}\right)^2\right]
\end{aligned}
$$

où

$$
\begin{aligned}
\mu_n &= \left(\frac{n\sigma_0^2}{n\sigma_0^2+\sigma^2}\right)^2 \widehat{\mu}_n + \frac{\sigma^2}{n\sigma_0^2+\sigma^2}\mu_0 \\
\sigma_n^2 &= \frac{\sigma_0^2\sigma^2}{n\sigma_0^2+\sigma^2}
\end{aligned}
$$

Méthodes paramétriques

- Estimation bayesienne: cas normal
  - densité conditionnelle de classe

$$
\begin{aligned}
p(x|D_n) &= \int p(x,\mu|D_n)d\mu = \int p(x|\mu,D_n)p(\mu|D_n)d\mu \\
&= \int p(x|\mu)p(\mu|D_n)d\mu \\
&\sim N(\mu_n, \sigma^2+\sigma_n^2)
\end{aligned}
$$

Méthodes paramétriques

- Avantages de l’approche bayesienne
    - 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