Maximum-Likelihood and Bayesian Parameter Estimation (part 2)

Bayesian Estimation

Bayesian Parameter Estimation: Gaussian Case

Bayesian Parameter Estimation: General Estimation

Bayesian Estimation

- The parameter θ is a random variable
  - Computation of posterior probabilities $P(\omega_i \mid x)$ lies at the heart of Bayesian classification
  - Goal: compute $P(\omega_i \mid x, D)$
  - Given the sample D, Bayes formula is written

$$\mathbf{P}(\omega_i \mid \mathbf{x}, \mathrm{D}) = \frac{\mathbf{p}(\mathbf{x} \mid \omega_i, \mathrm{D}).\mathbf{P}(\omega_i \mid \mathrm{D})}{\sum_{j=1}^{c} \mathbf{p}(\mathbf{x} \mid \omega_j, \mathrm{D}).\mathbf{P}(\omega_j \mid \mathrm{D})}$$

  - c separate parameter estimation problems $p(x|D)$

Parameter Distribution

- Although desired pdf p(x) is unknown, we assume that it has a known parametric form.
- Only thing unknown is the value of parameter vector theta

$$p(x \mid D) = \int p(x, \theta \mid D)\, d\theta$$

or

$$p(x \mid D) = \int p(x \mid \theta)\, p(\theta \mid D)\, d\theta$$

Core Task remaining

Bayesian Parameter Estimation: Univariate Gaussian Case

Assume μ is only unknown, σ, μ_0 and σ_0 known, ie,

$$p(x \mid \mu) \sim N(\mu, \sigma^2)$$
$$p(\mu) \sim N(\mu_0, \sigma_0^2)$$

We need

$$p(\mu \mid D) = \frac{p(D \mid \mu).p(\mu)}{\int p(D \mid \mu).p(\mu)d\mu} \qquad (1)$$
$$= \alpha \prod_{k=1}^{k=n} p(x_k \mid \mu).p(\mu)$$

$$p(\mu \mid D) \sim N(\mu_n, \sigma_n^2) \qquad (2)$$

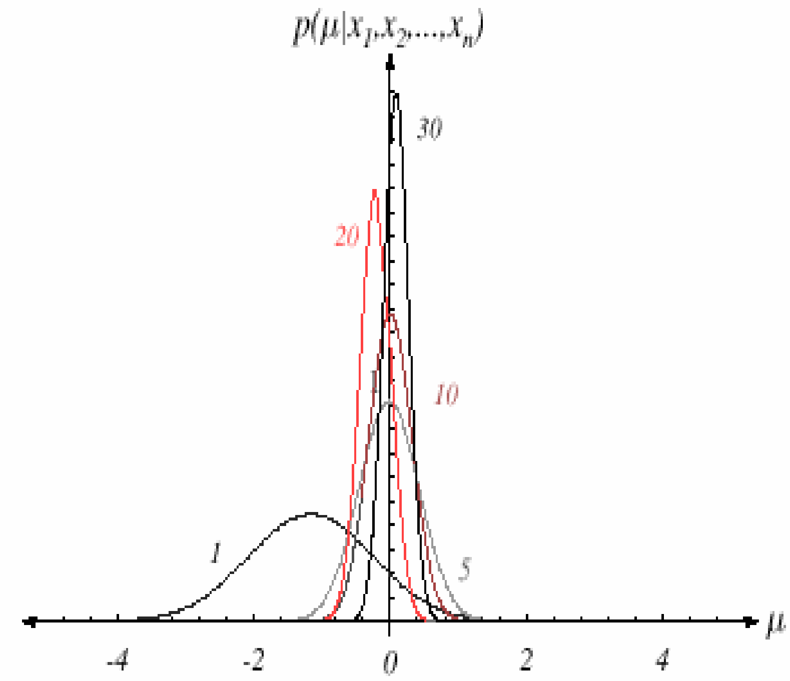
Reproducing density

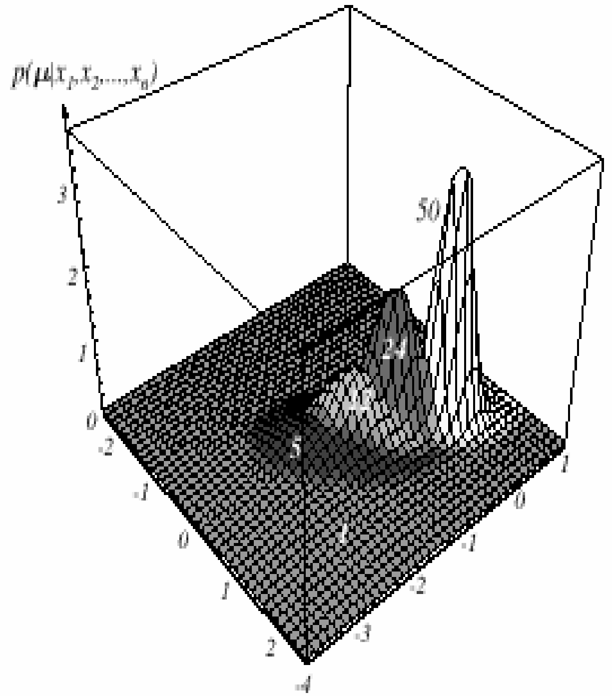
$$\mu_n = \left(\frac{n\sigma_0^2}{n_0\sigma_0^2 + \sigma^2} \right) \hat{\mu}_n + \frac{\sigma^2}{n\sigma_0^2 + \sigma^2} . \mu_0$$

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

Prior information is combined with the empirical information in the samples to obtain the *a posteriori* density

Bayesian Learning





$$\mu_n = \left(\frac{n\sigma_0^2}{n_0\sigma_0^2 + \sigma^2} \right) \hat{\mu}_n + \frac{\sigma^2}{n\sigma_0^2 + \sigma^2} \cdot \mu_0$$

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

$$p(\mu|\mathrm{D}) \sim N(\mu_n, \sigma_n^2)$$

A linear combination of assumed means

Computing p(x | D)

- p(μ | D) is computed
- p(x | D) remains

$$p(x|\mathrm{D}) = \int p(x|\mu).p(\mu|\mathrm{D})d\mu$$

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

$$= \frac{1}{2\pi\sigma\sigma_{\mathrm{n}}} \exp\left[-\frac{1}{2}\frac{(x-\mu_n)^2}{\sigma^2+\sigma_n^2}\right] f(\sigma,\sigma_n)$$

$$p(x|\mathrm{D}) \sim N(\mu_n, \sigma^2+\sigma_n^2)$$

(Desired class-conditional density $p(x \mid D_j, \omega_j)$)

Multivariate case is similar: $p(x|\mathrm{D}) \sim N(\mu_n, \Sigma+\Sigma_n)$

Bayesian Parameter Estimation: General Theory

$p(x \mid D)$ computation can be applied to any situation in which unknown density can be parameterized

Basic assumptions:

- Form of $p(x \mid \theta)$ known, value of θ not known exactly
- Initial knowledge of θ in known prior density $p(\theta)$
- Rest of knowledge about θ is contained in a set D of n random variables $x_1, x_2, \ldots, x_n$ that follows $p(x)$

General Bayesian Parameter Estimation

Compute posterior density $p(\theta \mid D)$ then $p(x \mid D)$ using

$$p(x \mid D) = \int p(x \mid \theta)\, p(\theta \mid D)\, d\theta$$

Using Bayes formula:

$$p(\theta \mid \mathrm{D}) = \frac{p(\mathrm{D} \mid \theta).p(\theta)}{\int p(\mathrm{D} \mid \theta).p(\theta)\,d\theta},$$

By independence assumption:

$$p(\mathrm{D} \mid \theta) = \prod_{k=1}^{k=n} p(x_k \mid \theta)$$

Recursive Bayes Incremental Learning

- Explicitly indicate number of samples in a set for a given category as $D^n = \{x_1,.., x_n\}$

- Then from $p(\mathrm{D} \mid \theta) = \prod_{k=1}^{k=n} p(x_k \mid \theta)$ we can write

$$p(D^n \mid \theta) = p(x_n \mid \theta)\; p(D^{n-1} \mid \theta)$$

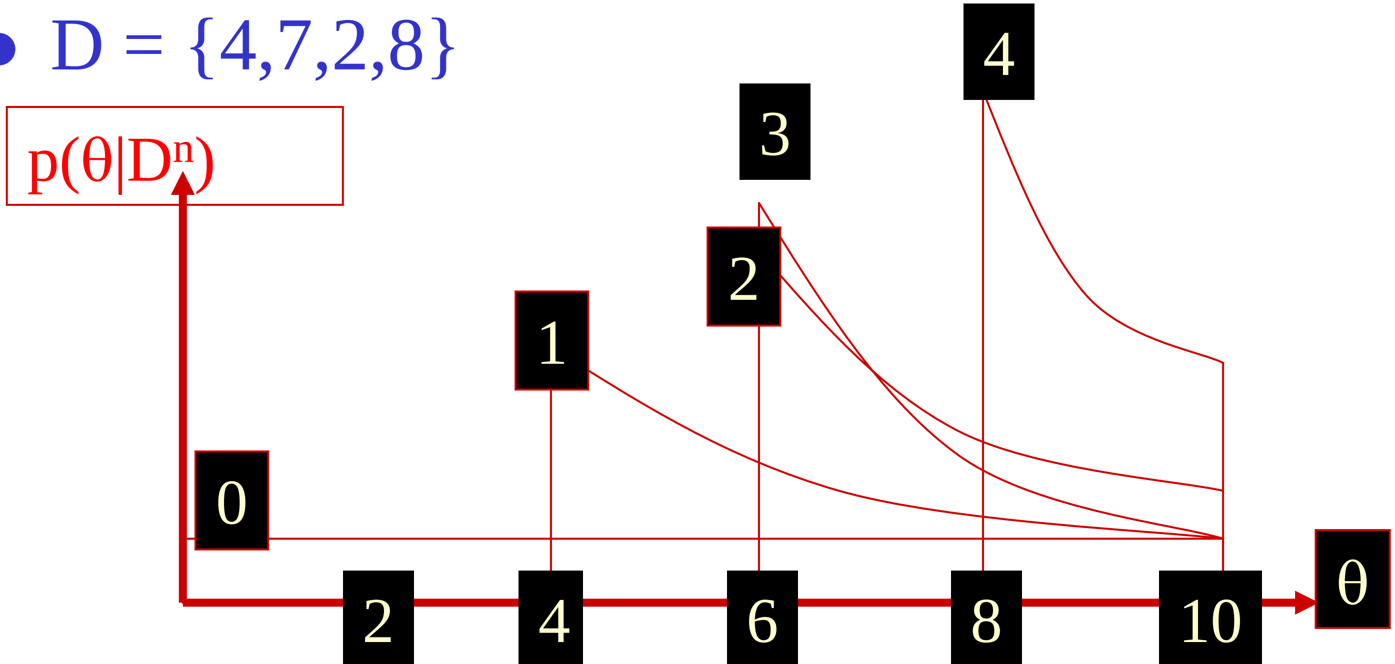
$$p(\theta \mid \mathrm{D}^{\mathrm{n}}) = \frac{p(\mathrm{x_n} \mid \theta).p(\theta \mid D^{n-1})}{\int p(\mathrm{x_n} \mid \theta).p(\theta \mid D^{n-1})d\theta},$$

Example of Recursive Bayes Learning

- One-dim samples from uniform distribution

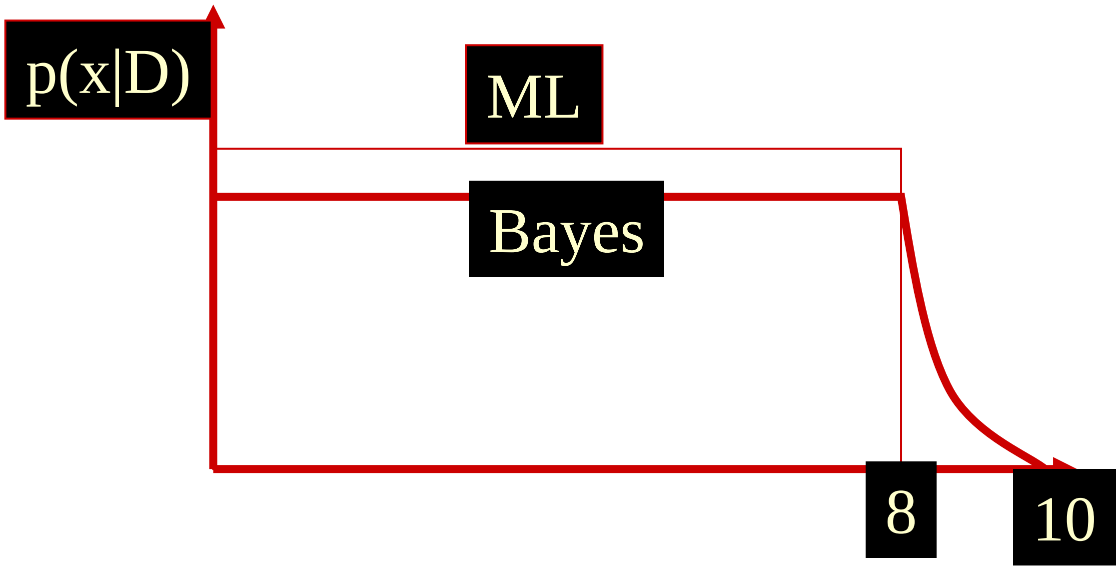
$$p(x\,|\,\theta) \sim U(0,\theta) = \begin{cases} \frac{1}{\theta} & 0 \le x \le \theta \\ 0 & \text{otherwise} \end{cases}$$

- Parameter distribution: Uniform over $0 \le \theta \le 10$
- $D = \{4,7,2,8\}$



Learning the parameter of Uniform Distribution using Recursive Bayes

- As more points are incorporated
- Bayes has a tail above 8 reflecting prior information



Three Sources of Error in Classification

- Bayes or Indistinguishability Error
  - Due to overlapping densities. Inherent property of given feature set
- Model Error
  - Due to an incorrect model. Can only be eliminated if designer specifies true model that generated the data.
- Estimation Error
  - Parameters are estimated from a finite sample.