Chapter 3
Maximum-Likelihood and Bayesian Parameter Estimation (3,4,5)

- Bayesian Estimation (BE)
- Bayesian Parameter Estimation: Gaussian Case
- Bayesian Parameter Estimation: General Estimation

3.3 Bayesian Estimation

- In MLE θ was supposed fix
- In BE θ is a random variable
- The computation of posterior probabilities $P(\omega_i \mid x)$ lies at the heart of Bayesian classification
- Goal: compute $P(\omega_i \mid x, \mathrm{D})$

  Given the sample D, Bayes formula can be written

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

- To demonstrate the preceding equation, use:

$$D = D_1 \cup D_2 \ldots \cup D_c \quad x \in D_i \rightarrow x \text{ is } \omega_i$$

$$D_i \text{ has no influlence on } \; p(x \mid \omega_j, D_j) \text{ if } i <> j$$

$$P(\omega_i) = P(\omega_i \mid \mathrm{D}) \;\; (\text{Training sample provides this!})$$

$$\text{Thus:}$$

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

Parameter Distribution

- $p(x)$ is unknown, we assume it has a known parametric for $p(x \mid \theta)$, and value of parameter θ is unknown
- Know prior density $p(\theta)$
- Training data convert $p(\theta)$ to a posterior $p(\theta \mid D)$ density
- Our path:

$$
\begin{aligned}
& p(x \mid \omega_i, D) = p(x) \cong p(x \mid D) \\
& = \int p(x, \theta \mid D) d\theta \\
& = \int p(x \mid \theta) p(\theta \mid D) d\theta
\end{aligned}
$$

- If $p(\theta \mid D)$ peaks very sharply about $\hat{\theta}$ parameter $p(x \mid \theta)$

  and is smooth, and the tails of the integral are not important, then

$$p(x \mid D) \cong p(x \mid \hat{\theta})$$

3.4 Bayesian Parameter Estimation: Gaussian Case

- Goal: Estimate θ using the a-posteriori density $P(\theta \mid D)$
- The univariate case: $P(\mu \mid D)$

  μ is the only unknown parameter

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

  (μ_0 and σ_0 are known!)

$$P(\mu \mid D) = \frac{P(D \mid \mu).P(\mu)}{\int P(D \mid \mu).P(\mu)d\mu} \quad (1)$$

$$= \alpha \prod_{k=1}^{k=n} P(x_k \mid \mu).P(\mu)$$

- Reproducing density

$$P(\mu \mid D) \sim N(\mu_n, \sigma_n^2) \quad (2)$$

Identifying (1) and (2) yields:

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

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

- Understanding
  - μ_n represents our best guess for μ after observing n samples
  - σ_n^2 measures our uncertainty about this guess
  - Add samples to decrease the uncertainty
  - Bayse Learning: as n increase, p(μ | D) becomes more and more sharply peaked, approaching a Dirac delta function as n approaches infinity

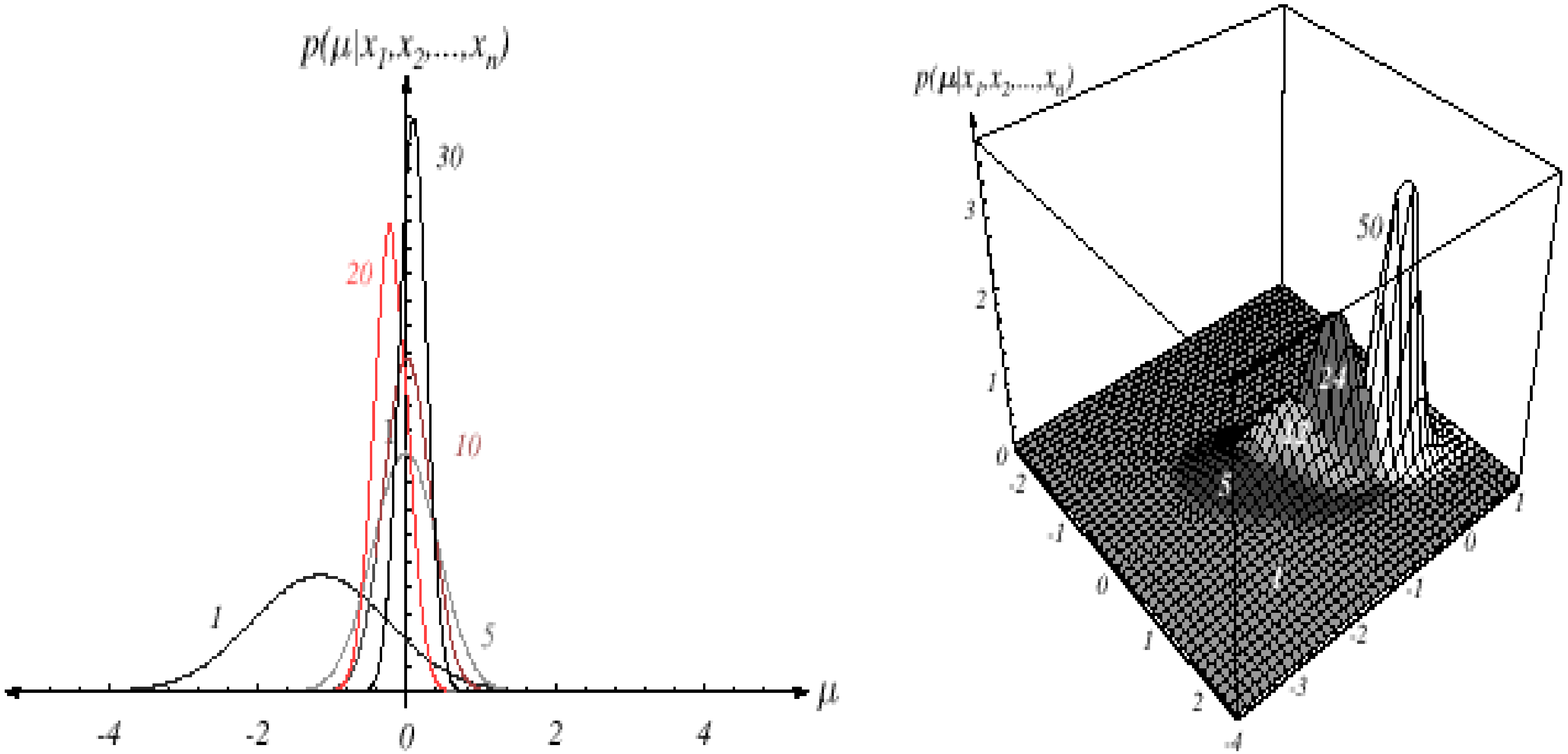

FIGURE 3.2. Bayesian learning of the mean of normal distributions in one and two dimensions. The posterior distribution estimates are labeled by the number of training samples used in the estimation. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

- The univariate case $P(x \mid \mathcal{D})$
  - $P(\mu \mid \mathcal{D})$ computed as above
  - $P(x \mid \mathcal{D})$ remains to be computed!

$$\mathbf{P(x \mid \mathrm{D}) = \int P(x \mid \mu).P(\mu \mid \mathrm{D})d\mu} \quad \textbf{is Gaussian}$$

  - It provides:

$$\mathbf{P(x \mid \mathrm{D}) \sim N(\mu_n, \sigma^2 + \sigma_n^2)}$$

(Desired class-conditional density $P(x \mid \mathcal{D}_j, \omega_j)$)

Therefore: $P(x \mid \mathcal{D}_j, \omega_j)$ together with $P(\omega_j)$

And using Bayes formula, we obtain the

Bayesian classification rule:

$$\underset{\omega_j}{Max}\left[P(\omega_j \mid x, \mathrm{D})\right] \equiv \underset{\omega_j}{Max}\left[P(x \mid \omega_j, \mathrm{D}_j).P(\omega_j)\right]$$

3.5 Bayesian Parameter Estimation: General Theory

- $P(x \mid \mathcal{D})$ computation can be applied to any situation in which the unknown density can be parametrized. the basic assumptions are:
  - The form of $P(x \mid \theta)$ is assumed known, but the value of θ is not known exactly
  - Our knowledge about θ is assumed to be contained in a known prior density $P(\theta)$
  - The rest of our knowledge θ is contained in a set $\mathcal{D}$ of n random variables $x_1, x_2, \ldots, x_n$ that follows unknown $P(x)$

- The basic problem is:

  "Compute the posterior density $P(\theta \mid D)$"

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

  Using Bayes formula, we have:

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

  And by independence assumption:

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

Bayse incremental learning

$$D^n = \{x_1, \ldots x_n\}$$

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

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

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

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

$$p(\theta \mid D^0) = p(\theta)$$

- Maximum Likelihood vs Bayse Estimation
  - Computational complexity
  - Interpretability
  - Confidence in prior information
- Source of classification error
  - Bayes Error
  - Model Error
  - Estimation Error

- Noninformative Priors and Invariance
  - If there is known or assumed invariance, there will be constraints on the form of the prior. If we can find a prior that satisfies such constraints, the resulting prior is noninformative with respect to that invariance