Chapter 3

Maximum-Likelihood and Bayesian Parameter Estimation

Bayes Theorem for Classification

$$P(\omega_j|\mathbf{x}) = \frac{p(\mathbf{x}|\omega_j) \cdot P(\omega_j)}{p(\mathbf{x})} \quad (1 \leq j \leq c)$$ (Bayes Formula)

To compute posterior probability $P(\omega_j|\mathbf{x})$, we need to know:

Prior probability: $P(\omega_j)$ Likelihood: $p(\mathbf{x}|\omega_j)$

The collection of training examples is composed of *c* data sets

$\mathcal{D}_j \ (1 \leq j \leq c)$

- Each example in $\mathcal{D}_j$ is drawn according to the class-conditional pdf, i.e. $p(\mathbf{x}|\omega_j)$
- Examples in $\mathcal{D}_j$ are *i.i.d.* random variables, i.e. **independent and identically distributed (独立同分布)**

Bayes Theorem for Classification (Cont.)

For prior probability: → **no difficulty**

$$P(\omega_j) = \frac{|\mathcal{D}_j|}{\sum_{i=1}^{c} |\mathcal{D}_i|}$$

(Here, $|\cdot|$ returns the **cardinality**, i.e. number of elements, of a set)

For class-conditional pdf:

Ch. 3 ← □ **Case I:** $p(\mathbf{x}|\omega_j)$ has certain **parametric form**

$p(\mathbf{x}|\omega_j)$

e.g.: $p(\mathbf{x}|\omega_j) \sim N(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)$ **(parameters:** $\boldsymbol{\theta}_j = \{\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j\}$**)**

$\mathbf{x} \in \mathbf{R}^d$ → $\boldsymbol{\theta}_j$ contains "$d + d(d+1)/2$" $free$ parameters

To show the dependence of $p(\mathbf{x}|\omega_j)$ on $\boldsymbol{\theta}_j$ **explicitly:**

$p(\mathbf{x}|\omega_j)$ → $p(\mathbf{x}|\omega_j, \boldsymbol{\theta}_j)$

Ch. 4 ← □ **Case II:** $p(\mathbf{x}|\omega_j)$ doesn't have **parametric form**

Estimation Under Parametric Form

Parametric class-conditional pdf: $p(\mathbf{x}|\omega_j, \boldsymbol{\theta}_j)\ (1 \leq j \leq c)$

- **Assumption I: Maximum-Likelihood (ML) estimation (极大似然估计)**

View parameters as quantities whose values are **fixed but unknown** → Estimate parameter values by **maximizing the likelihood** (probability) of observing the actual training examples

- **Assumption II: Bayesian estimation (贝叶斯估计)**

View parameters as **random variables** having some known prior distribution → Observation of the actual training examples transforms parameters' **prior distribution into posterior distribution** (via Bayes theorem)

Maximum-Likelihood Estimation

Settings

Likelihood function for each category is governed by some **fixed but unknown** parameters, i.e. $p(\mathbf{x}|\omega_j, \boldsymbol{\theta}_j)\ (1 \leq j \leq c)$

Task: Estimate $\{\boldsymbol{\theta}_j\}_{j=1}^{c}$ from $\{\mathcal{D}_j\}_{j=1}^{c}$

A simplified treatment

Examples in $\mathcal{D}_j$ gives no information about $\boldsymbol{\theta}_i$ if $i \neq j$

Work with each category **separately** and therefore simplify the notations by dropping subscripts w.r.t. categories

without loss of generality: $\mathcal{D}_j \Longrightarrow \mathcal{D}$; $\boldsymbol{\theta}_j \Longrightarrow \boldsymbol{\theta}$

Maximum-Likelihood Estimation (Cont.)

$\mathbf{x}_k \sim p(\mathbf{x}|\boldsymbol{\theta})$

$(k = 1, \ldots, n)$

$\boldsymbol{\theta}$: Parameters to be estimated

$\mathcal{D}$: A set of *i.i.d.* examples $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$

The objective function

$$p(\mathcal{D}|\boldsymbol{\theta}) = \prod_{k=1}^{n} p(\boldsymbol{x}_k|\boldsymbol{\theta})$$

The likelihood of θ w.r.t. the set of observed examples

The maximum-likelihood estimation

$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} p(\mathcal{D}|\boldsymbol{\theta})$$

Intuitively, $\hat{\theta}$ best agrees with the actually observed examples

Maximum-Likelihood Estimation (Cont.)

Gradient Operator (梯度算子)

- ✓ Let $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)^t \in \mathbf{R}^p$ be a *p*-dimensional vector
- ✓ Let $f : \mathbf{R}^p \to \mathbf{R}$ be *p*-variate real-valued function over $\boldsymbol{\theta}$

$$\boldsymbol{\nabla}_{\boldsymbol{\theta}} \equiv \begin{bmatrix} \frac{\partial}{\partial \theta_1} \\ \vdots \\ \frac{\partial}{\partial \theta_p} \end{bmatrix}$$

$$f(\boldsymbol{\theta}) = \theta_1^2 + 3\theta_1\theta_2$$

$$\boldsymbol{\nabla}_{\boldsymbol{\theta}} f = \begin{bmatrix} \frac{\partial f}{\partial \theta_1} \\ \frac{\partial f}{\partial \theta_2} \end{bmatrix} = \begin{bmatrix} 2\theta_1 + 3\theta_2 \\ 3\theta_1 \end{bmatrix}$$

$l(\boldsymbol{\theta}) = \ln p(\mathcal{D}|\boldsymbol{\theta})$ is named as the **log-likelihood function**

$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} p(\mathcal{D}|\boldsymbol{\theta}) \Longleftrightarrow \hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} l(\boldsymbol{\theta})$$

Maximum-Likelihood Estimation (Cont.)

$$l(\boldsymbol{\theta}) = \ln p(\mathcal{D}|\boldsymbol{\theta}) = \sum_{k=1}^{n} \ln\, p(\mathbf{x}_k|\boldsymbol{\theta})$$

$$\underline{\underline{\boldsymbol{\nabla}_{\boldsymbol{\theta}}\, l}} = \boldsymbol{\nabla}_{\boldsymbol{\theta}} \left(\sum_{k=1}^{n} \ln\, p(\mathbf{x}_k|\boldsymbol{\theta}) \right) = \sum_{k=1}^{n} \underline{\underline{\boldsymbol{\nabla}_{\boldsymbol{\theta}} \ln\, p(\mathbf{x}_k|\boldsymbol{\theta})}}$$

p-dimensional vector with each component being a function over $\boldsymbol{\theta}$

p-variate real-valued function over $\boldsymbol{\theta}$ (not over $\mathbf{x}_k$)

Necessary conditions for ML estimate $\hat{\boldsymbol{\theta}}$

$$\boldsymbol{\nabla}_{\boldsymbol{\theta}}\, l\, |_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} = \mathbf{0}$$ **(a set of *p* equations)**

The Gaussian Case: Unknown $\boldsymbol{\mu}$

$\mathbf{x}_k \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$

$(k = 1, \ldots, n)$

suppose $\boldsymbol{\Sigma}$ is known $\boldsymbol{\theta} = \{\boldsymbol{\mu}\}$

$$p(\mathbf{x}_k|\boldsymbol{\mu}) = \frac{1}{(2\pi)^{d/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left[-\frac{1}{2}(\mathbf{x}_k - \boldsymbol{\mu})^t \boldsymbol{\Sigma}^{-1} (\mathbf{x}_k - \boldsymbol{\mu})\right]$$

$$\ln p(\mathbf{x}_k|\boldsymbol{\mu}) = -\frac{1}{2}\ln\left[(2\pi)^d|\boldsymbol{\Sigma}|\right] - \frac{1}{2}(\mathbf{x}_k - \boldsymbol{\mu})^t\boldsymbol{\Sigma}^{-1}(\mathbf{x}_k - \boldsymbol{\mu})$$

$$= -\frac{1}{2}\ln\left[(2\pi)^d|\boldsymbol{\Sigma}|\right] - \frac{1}{2}\mathbf{x}_k^t\boldsymbol{\Sigma}^{-1}\mathbf{x}_k + \boldsymbol{\mu}^t\boldsymbol{\Sigma}^{-1}\mathbf{x}_k - \frac{1}{2}\boldsymbol{\mu}^t\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}$$

$$\boldsymbol{\nabla}_{\boldsymbol{\mu}} \ln p(\mathbf{x}_k|\boldsymbol{\mu}) = \boldsymbol{\Sigma}^{-1}(\mathbf{x}_k - \boldsymbol{\mu})$$

The Gaussian Case: Unknown $\boldsymbol{\mu}$ (Cont.)

$$l(\boldsymbol{\mu}) = \sum_{k=1}^{n} \ln p(\mathbf{x}_k|\boldsymbol{\mu})$$

$$\nabla_{\boldsymbol{\mu}} \ln p(\mathbf{x}_k|\boldsymbol{\mu}) = \boldsymbol{\Sigma}^{-1}(\mathbf{x}_k - \boldsymbol{\mu})$$

$$\nabla_{\boldsymbol{\mu}}\, l = \sum_{k=1}^{n} \boldsymbol{\Sigma}^{-1}(\mathbf{x}_k - \boldsymbol{\mu})$$

$\nabla_{\boldsymbol{\mu}} l = \mathbf{0}$ (necessary condition for ML estimate $\hat{\boldsymbol{\mu}}$)

$$\sum_{k=1}^{n} \boldsymbol{\Sigma}^{-1}(\mathbf{x}_k - \hat{\boldsymbol{\mu}}) = \mathbf{0}$$

Multiply $\boldsymbol{\Sigma}$ on both sides

$$\sum_{k=1}^{n} (\mathbf{x}_k - \hat{\boldsymbol{\mu}}) = \mathbf{0}$$

$$\hat{\boldsymbol{\mu}} = \frac{1}{n}\sum_{k=1}^{n} \mathbf{x}_k$$

Intuitive result

ML estimate for the unknown $\boldsymbol{\mu}$ is just the arithmetic average of training samples – ***sample mean***

The Gaussian Case: Unknown $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$

$\mathbf{x}_k \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$

$(k = 1, \ldots, n)$

$\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ unknown $\Rightarrow$ $\boldsymbol{\theta} = \{\boldsymbol{\mu}, \boldsymbol{\Sigma}\}$

Consider *univariate* case

$$p(x_k|\boldsymbol{\theta}) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right] \qquad \left(\boldsymbol{\theta} = \begin{bmatrix}\theta_1 \\ \theta_2\end{bmatrix} = \begin{bmatrix}\mu \\ \sigma^2\end{bmatrix}\right)$$

$$\ln p(x_k|\boldsymbol{\theta}) = -\frac{1}{2}\ln 2\pi\theta_2 - \frac{1}{2\theta_2}(x_k - \theta_1)^2$$

$$\boldsymbol{\nabla}_{\boldsymbol{\theta}} \ln p(x_k|\boldsymbol{\theta}) = \begin{bmatrix} \frac{1}{\theta_2}(x_k - \theta_1) \\ -\frac{1}{2\theta_2} + \frac{(x_k-\theta_1)^2}{2\theta_2^2} \end{bmatrix}$$

The Gaussian Case: Unknown $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ (Cont.)

$$l(\boldsymbol{\theta}) = \sum_{k=1}^{n} \ln\, p(x_k|\boldsymbol{\theta})$$

$$\nabla_{\boldsymbol{\theta}} \ln p(x_k|\boldsymbol{\theta}) = \begin{bmatrix} \frac{1}{\theta_2}(x_k - \theta_1) \\ -\frac{1}{2\theta_2} + \frac{(x_k-\theta_1)^2}{2\theta_2^2} \end{bmatrix}$$

$$\nabla_{\boldsymbol{\theta}} l = \begin{bmatrix} \sum_{k=1}^{n} \frac{1}{\theta_2}(x_k - \theta_1) \\ \sum_{k=1}^{n} \left(-\frac{1}{2\theta_2} + \frac{(x_k-\theta_1)^2}{2\theta_2^2}\right) \end{bmatrix}$$

$\nabla_{\boldsymbol{\theta}} l = 0$ (necessary condition for ML estimate $\hat{\theta}_1$ and $\hat{\theta}_2$)

$$\sum_{k=1}^{n} \frac{1}{\hat{\theta}_2}(x_k - \hat{\theta}_1) = 0$$

$$-\sum_{k=1}^{n} \frac{1}{\hat{\theta}_2} + \sum_{k=1}^{n} \frac{(x_k - \hat{\theta}_1)^2}{\hat{\theta}_2^2} = 0$$

The Gaussian Case: Unknown $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ (Cont.)

$$\sum_{k=1}^{n} \frac{1}{\hat{\theta}_2}(x_k - \hat{\theta}_1) = 0 \Rightarrow \sum_{k=1}^{n}(x_k - \hat{\theta}_1) = 0 \Rightarrow \hat{\theta}_1 = \frac{1}{n}\sum_{k=1}^{n} x_k$$

$$-\sum_{k=1}^{n} \frac{1}{\hat{\theta}_2} + \sum_{k=1}^{n} \frac{(x_k - \hat{\theta}_1)^2}{\hat{\theta}_2^2} = 0 \Rightarrow \hat{\theta}_2 = \frac{1}{n}\sum_{k=1}^{n}(x_k - \hat{\theta}_1)^2$$

ML estimate in *univariate* case

$$\hat{\mu} = \frac{1}{n}\sum_{k=1}^{n} x_k \qquad \hat{\sigma}^2 = \frac{1}{n}\sum_{k=1}^{n}(x_k - \hat{\mu})^2$$

The Gaussian Case: Unknown $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ (Cont.)

Intuitive result as well!

ML estimate in *multivariate* case

$$\hat{\boldsymbol{\mu}} = \frac{1}{n}\sum_{k=1}^{n}\mathbf{x}_k$$

Arithmetic average of n vectors $\mathbf{x}_k$

$$\hat{\boldsymbol{\Sigma}} = \frac{1}{n}\sum_{k=1}^{n}(\mathbf{x}_k - \hat{\boldsymbol{\mu}})(\mathbf{x}_k - \hat{\boldsymbol{\mu}})^t$$

Arithmetic average of n matrices $(\mathbf{x}_k - \hat{\boldsymbol{\mu}})(\mathbf{x}_k - \hat{\boldsymbol{\mu}})^t$

Biased/Unbiased Estimator

$$\hat{\boldsymbol{\Sigma}} = \frac{1}{n}\sum_{k=1}^{n}(\mathbf{x}_k - \hat{\boldsymbol{\mu}})(\mathbf{x}_k - \hat{\boldsymbol{\mu}})^t$$

$$\mathcal{E}[\hat{\boldsymbol{\Sigma}}] = \mathcal{E}\left[\frac{1}{n}\sum_{k=1}^{n}(\mathbf{x}_k - \hat{\boldsymbol{\mu}})(\mathbf{x}_k - \hat{\boldsymbol{\mu}})^t\right] = \frac{n-1}{n}\boldsymbol{\Sigma}$$

Biased estimator (有偏估计) of $\boldsymbol{\Sigma}$

$$\mathbf{C} = \frac{n}{n-1}\hat{\boldsymbol{\Sigma}}$$

Unbiased estimator (无偏估计) of $\boldsymbol{\Sigma}$

$\lim_{n\to\infty}\mathcal{E}[\hat{\boldsymbol{\Sigma}}] = \boldsymbol{\Sigma}$

Asymptotically unbiased estimator (渐进无偏估计) of $\boldsymbol{\Sigma}$

Bayesian Estimation

Settings

- The **parametric form** of the likelihood function for each category is known $p(\mathbf{x}|\omega_j, \boldsymbol{\theta}_j)$ $(1 \leq j \leq c)$
- However, $\boldsymbol{\theta}_j$ is considered to be **random variables** instead of being fixed (but unknown) values

In this case, we can no longer make a single ML estimate $\hat{\boldsymbol{\theta}}_j$ and then infer $P(\omega_j|\mathbf{x})$ based on $P(\omega_j)$ and $p(\mathbf{x}|\omega_j, \hat{\boldsymbol{\theta}}_j)$

How can we proceed under this situation

Fully exploit training examples!

$$P(\omega_j|\mathbf{x}) \Rightarrow P(\omega_j|\mathbf{x}, \mathcal{D}^*)$$

$$(\mathcal{D}^* = \mathcal{D}_1 \cup \mathcal{D}_2 \cup \cdots \cup \mathcal{D}_c)$$

Bayesian Estimation (Cont.)

$$P(\omega_j|\mathbf{x},\mathcal{D}^*) = \frac{p(\omega_j,\mathbf{x},\mathcal{D}^*)}{p(\mathbf{x},\mathcal{D}^*)} = \frac{p(\omega_j,\mathbf{x},\mathcal{D}^*)}{\sum_{i=1}^{c} p(\omega_i,\mathbf{x},\mathcal{D}^*)}$$

$$p(\omega_j,\mathbf{x},\mathcal{D}^*) = p(\mathcal{D}^*)\cdot p(\omega_j,\mathbf{x}|\mathcal{D}^*) = p(\mathcal{D}^*)\cdot P(\omega_j|\mathcal{D}^*)\cdot p(\mathbf{x}|\omega_j,\mathcal{D}^*)$$

$$P(\omega_j|\mathbf{x},\mathcal{D}^*) = \frac{p(\mathcal{D}^*)\cdot P(\omega_j|\mathcal{D}^*)\cdot p(\mathbf{x}|\omega_j,\mathcal{D}^*)}{p(\mathcal{D}^*)\cdot \sum_{i=1}^{c} P(\omega_i|\mathcal{D}^*)\cdot p(\mathbf{x}|\omega_i,\mathcal{D}^*)}$$

$$= \frac{P(\omega_j|\mathcal{D}^*)\cdot p(\mathbf{x}|\omega_j,\mathcal{D}^*)}{\sum_{i=1}^{c} P(\omega_i|\mathcal{D}^*)\cdot p(\mathbf{x}|\omega_i,\mathcal{D}^*)}$$ **Eq.22** [pp.91]

Two assumptions

$P(\omega_j|\mathcal{D}^*) = P(\omega_j)$

$p(\mathbf{x}|\omega_j,\mathcal{D}^*) = p(\mathbf{x}|\omega_j,\mathcal{D}_j)$

$$= \frac{P(\omega_j)\cdot p(\mathbf{x}|\omega_j,\mathcal{D}_j)}{\sum_{i=1}^{c} P(\omega_i)\cdot p(\mathbf{x}|\omega_i,\mathcal{D}_i)}$$ **Eq.23** [pp.91]

Bayesian Estimation (Cont.)

$$P(\omega_j|\mathbf{x}, \mathcal{D}^*) = \frac{P(\omega_j) \cdot p(\mathbf{x}|\omega_j, \mathcal{D}_j)}{\sum_{i=1}^{c} P(\omega_i) \cdot p(\mathbf{x}|\omega_i, \mathcal{D}_i)}$$

Key problem

Determine $p(\mathbf{x}|\omega_j, \mathcal{D}_j)$

Treat each class independently → Simplify the *class-conditional pdf* notation $p(\mathbf{x}|\omega_j, \mathcal{D}_j)$ as $p(\mathbf{x}|\mathcal{D})$

$$p(\mathbf{x}|\mathcal{D}) = \int p(\mathbf{x}, \boldsymbol{\theta}|\mathcal{D})\, d\boldsymbol{\theta}$$

($\boldsymbol{\theta}$: random variables w.r.t. parametric form)

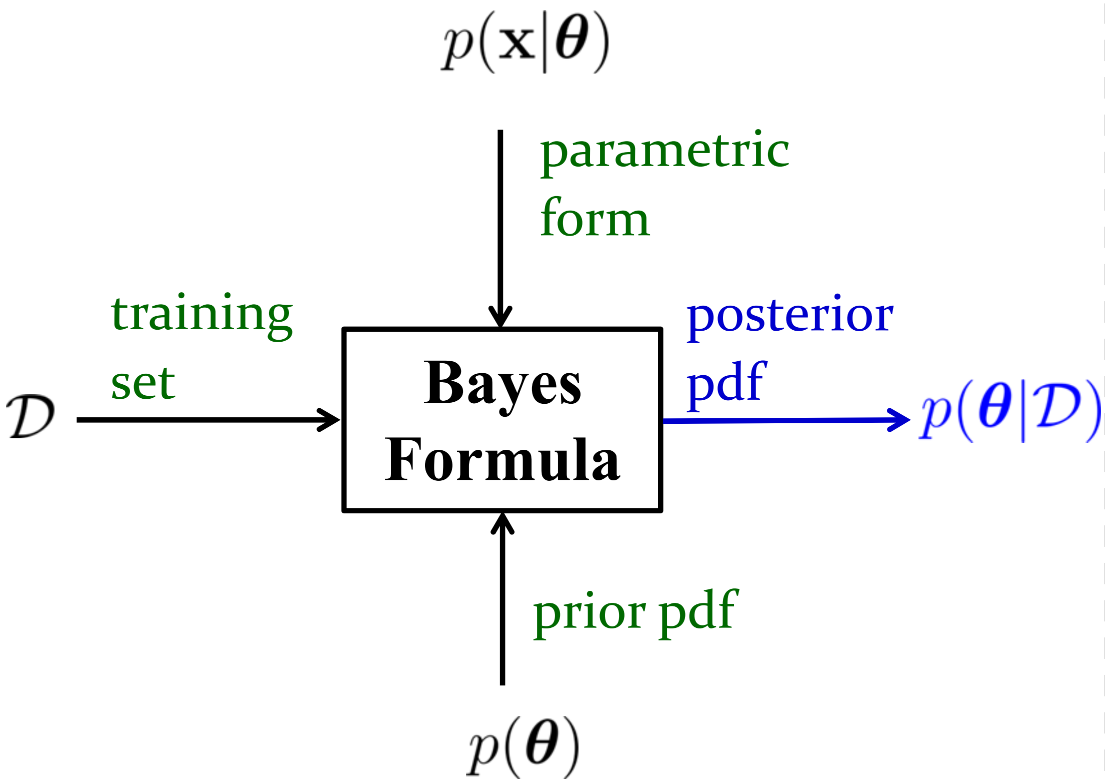
$$= \int p(\mathbf{x}|\boldsymbol{\theta}, \mathcal{D})\, p(\boldsymbol{\theta}|\mathcal{D})\, d\boldsymbol{\theta}$$

$$= \int p(\mathbf{x}|\boldsymbol{\theta})\, p(\boldsymbol{\theta}|\mathcal{D})\, d\boldsymbol{\theta}$$

($\mathbf{x}$ is independent of $\mathcal{D}$ given $\boldsymbol{\theta}$)

Bayesian Estimation: The General Procedure

Phase I: *prior pdf* ➔ *posterior pdf* (for $\boldsymbol{\theta}$)



$$p(\boldsymbol{\theta}|\mathcal{D}) = \frac{p(\boldsymbol{\theta}, \mathcal{D})}{p(\mathcal{D})}$$

$$= \frac{p(\boldsymbol{\theta})p(\mathcal{D}|\boldsymbol{\theta})}{\int p(\boldsymbol{\theta}, \mathcal{D})d\boldsymbol{\theta}}$$

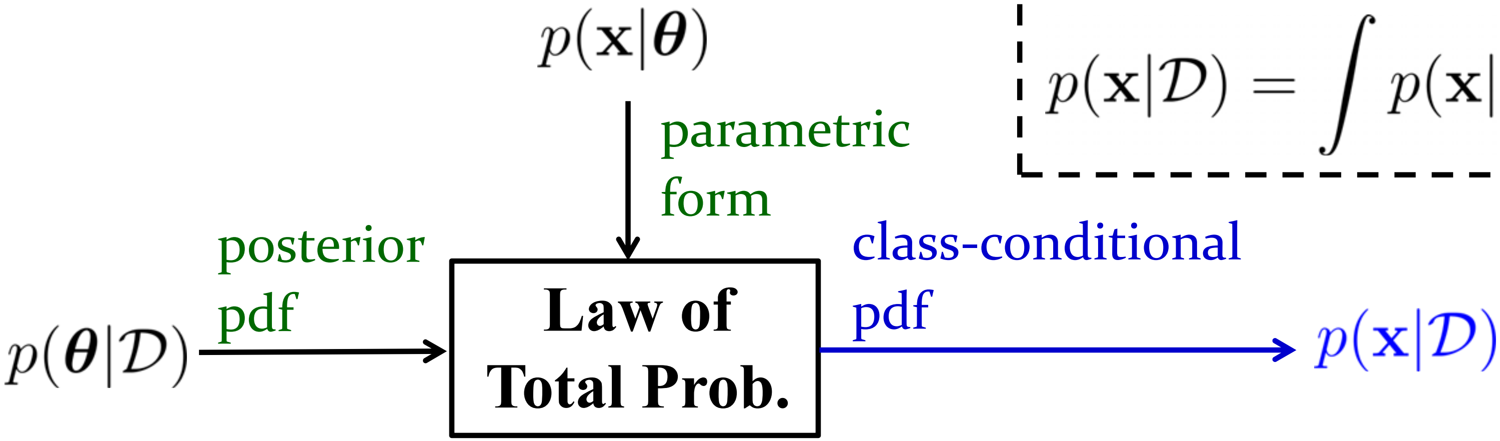
$$= \frac{p(\boldsymbol{\theta})p(\mathcal{D}|\boldsymbol{\theta})}{\int p(\boldsymbol{\theta})p(\mathcal{D}|\boldsymbol{\theta})d\boldsymbol{\theta}}$$

$$p(\mathcal{D}|\boldsymbol{\theta}) = \prod_{k=1}^{n} p(\mathbf{x}_k|\boldsymbol{\theta})$$

Bayesian Estimation: The General Procedure

Phase II: *posterior pdf* (for $\boldsymbol{\theta}$) ➔ *class-conditional pdf* (for $\mathbf{x}$)

$$\boxed{p(\mathbf{x}|\mathcal{D}) = \int p(\mathbf{x}|\boldsymbol{\theta})p(\boldsymbol{\theta}|\mathcal{D})d\boldsymbol{\theta}}$$



Phase III: $$P(\omega_j|\mathbf{x}, \mathcal{D}^*) = \frac{P(\omega_j) \cdot p(\mathbf{x}|\omega_j, \mathcal{D}_j)}{\sum_{i=1}^{c} P(\omega_i) \cdot p(\mathbf{x}|\omega_i, \mathcal{D}_i)}$$

The Gaussian Case: Unknown $\boldsymbol{\mu}$

Consider *univariate* case: $\boldsymbol{\theta} = \{\mu\}$ (σ^2 is known)

Phase I: *prior pdf* ➔ *posterior pdf* (for $\boldsymbol{\theta}$)

$$\underline{\underline{p(\mu)}} + \underline{\underline{p(x|\mu)}} + \mathcal{D} \Longrightarrow p(\mu|\mathcal{D})$$

$$p(x|\mu) \sim N(\mu, \sigma^2)$$

Gaussian parametric form

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

- Prior pdf still takes Gaussian form
- Other form of prior pdf could be assumed as well

How would $p(\mu|\mathcal{D})$ look like in this case?

The Gaussian Case: Unknown $\boldsymbol{\mu}$ (Cont.)

$$p(\mu|\mathcal{D}) = \frac{p(\mu,\mathcal{D})}{p(\mathcal{D})} = \frac{p(\mu)p(\mathcal{D}|\mu)}{\int p(\mu)p(\mathcal{D}|\mu)\,d\mu}$$

$$= \alpha\, p(\mu)\, p(\mathcal{D}|\mu)$$

($\int p(\mu)p(\mathcal{D}|\mu)\,d\mu$ is a constant not related to μ)

$$= \alpha\, p(\mu) \prod_{k=1}^{n} p(x_k|\mu)$$

(examples in $\mathcal{D}$ are *i.i.d.*)

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

$$p(\mu) = \frac{1}{\sqrt{2\pi}\sigma_0} \exp\left[-\frac{1}{2}\left(\frac{\mu-\mu_0}{\sigma_0}\right)^2\right]$$

$$p(x|\mu) \sim N(\mu, \sigma^2)$$

$$p(x_k|\mu) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{1}{2}\left(\frac{x_k-\mu}{\sigma}\right)^2\right]$$

The Gaussian Case: Unknown $\boldsymbol{\mu}$ (Cont.)

$$p(\mu|\mathcal{D}) = \alpha\, p(\mu) \prod_{k=1}^{n} p(x_k|\mu)$$

$p(\mu|\mathcal{D})$ is an exponential function of a quadratic function of μ ⇒ $p(\mu|\mathcal{D})$ is a normal pdf as well

$$= \alpha \cdot \frac{1}{\sqrt{2\pi}\sigma_0} \exp\left[-\frac{1}{2}\left(\frac{\mu-\mu_0}{\sigma_0}\right)^2\right] \cdot \prod_{k=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{1}{2}\left(\frac{x_k-\mu}{\sigma}\right)^2\right]$$

$$= \alpha' \cdot \exp\left[-\frac{1}{2}\left(\left(\frac{\mu-\mu_0}{\sigma_0}\right)^2 + \sum_{k=1}^{n}\left(\frac{\mu-x_k}{\sigma}\right)^2\right)\right]$$

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

$$= \alpha'' \cdot \exp\left[-\frac{1}{2}\left[\left(\frac{n}{\sigma^2}+\frac{1}{\sigma_0^2}\right)\mu^2 - 2\left(\frac{1}{\sigma^2}\sum_{k=1}^{n} x_k + \frac{\mu_0}{\sigma_0^2}\right)\mu\right]\right]$$

The Gaussian Case: Unknown $\boldsymbol{\mu}$ (Cont.)

$$p(\mu|\mathcal{D}) = \alpha'' \cdot \exp\left[-\frac{1}{2}\left[\boxed{\left(\frac{n}{\sigma^2} + \frac{1}{\sigma_0^2}\right)}\mu^2 - 2\boxed{\left(\frac{1}{\sigma^2}\sum_{k=1}^{n} x_k + \frac{\mu_0}{\sigma_0^2}\right)}\mu\right]\right]$$

$$p(\mu|\mathcal{D}) = \frac{1}{\sqrt{2\pi}\sigma_n}\exp\left[-\frac{1}{2}\left(\frac{\mu-\mu_n}{\sigma_n}\right)^2\right] = \alpha'' \cdot \exp\left[-\frac{1}{2}\left[\boxed{\frac{1}{\sigma_n^2}}\mu^2 - 2\boxed{\frac{\mu_n}{\sigma_n^2}}\mu\right]\right]$$

Equating the coefficients in both form:

$$\frac{1}{\sigma_n^2} = \frac{n}{\sigma^2} + \frac{1}{\sigma_0^2}$$

$$\frac{\mu_n}{\sigma_n^2} = \frac{1}{\sigma^2}\sum_{k=1}^{n} x_k + \frac{\mu_0}{\sigma_0^2}$$

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

$$\mu_n = \frac{\sigma_n^2}{\sigma^2}\sum_{k=1}^{n} x_k + \frac{\sigma_n^2}{\sigma_0^2}\mu_0$$

The Gaussian Case: Unknown $\boldsymbol{\mu}$ (Cont.)

Phase II: *posterior pdf* (for $\boldsymbol{\theta}$) ➔ *class-conditional pdf* (for **x**)

$$\underline{\underline{p(\mu|\mathcal{D})}} + \underline{\underline{p(x|\mu)}} \Rightarrow p(x|\mathcal{D})$$

$$p(x|\mu) \sim N(\mu, \sigma^2)$$

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

How would $p(x|\mathcal{D})$ look like in this case?

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

$$\mu_n = \frac{\sigma_n^2}{\sigma^2}\sum_{k=1}^{n} x_k + \frac{\sigma_n^2}{\sigma_0^2}\mu_0$$

The Gaussian Case: Unknown $\boldsymbol{\mu}$ (Cont.)

Then, phase III follows naturally for prediction

$$p(x|\mathcal{D}) = \int p(x|\mu)p(\mu|\mathcal{D})d\mu$$ **Eq.25** [pp.92]

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

$$= \beta \cdot \exp\left[-\frac{1}{2}\frac{(x-\mu_n)^2}{\sigma^2+\sigma_n^2}\right]$$ **Eq.36** [pp.95]

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

$p(x|\mathcal{D})$ is an exponential function of a quadratic function of x → $p(x|\mathcal{D})$ is a normal pdf as well

The Gaussian Case: Unknown $\boldsymbol{\mu}$ (Multivariate)

$\boldsymbol{\theta} = \{\boldsymbol{\mu}\}$ ($\boldsymbol{\Sigma}$ is known)

$$p(\mathbf{x}|\boldsymbol{\mu}) \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

$$p(\boldsymbol{\mu}) \sim N(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)$$

$$p(\boldsymbol{\mu}|\mathcal{D}) \sim N(\boldsymbol{\mu}_n, \boldsymbol{\Sigma}_n) \qquad p(\mathbf{x}|\mathcal{D}) \sim N(\boldsymbol{\mu}_n, \boldsymbol{\Sigma} + \boldsymbol{\Sigma}_n)$$

$$\boldsymbol{\mu}_n = \boldsymbol{\Sigma}_0 \left(\boldsymbol{\Sigma}_0 + \frac{1}{n}\boldsymbol{\Sigma} \right)^{-1} \frac{1}{n} \sum_{k=1}^{n} \mathbf{x}_k + \frac{1}{n}\boldsymbol{\Sigma} \left(\boldsymbol{\Sigma}_0 + \frac{1}{n}\boldsymbol{\Sigma} \right)^{-1} \boldsymbol{\mu}_0$$

$$\boldsymbol{\Sigma}_n = \boldsymbol{\Sigma}_0 \left(\boldsymbol{\Sigma}_0 + \frac{1}{n}\boldsymbol{\Sigma} \right)^{-1} \frac{1}{n}\boldsymbol{\Sigma}$$

A Recursive View of Bayesian Estimation

Convergence property of Bayesian estimation

How would Bayesian estimation proceed as the number of training examples (i.e. n) increases?

We denote the training set in terms of n explicitly as: $\mathcal{D}^n = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$

Eq.53

[pp.98]

$$p(\boldsymbol{\theta} \mid \mathcal{D}^n) = \frac{p(\boldsymbol{\theta})p(\mathcal{D}^n \mid \boldsymbol{\theta})}{p(\mathcal{D}^n)}$$

$$= \frac{p(\boldsymbol{\theta})p(\mathbf{x}_n \mid \boldsymbol{\theta})p(\mathcal{D}^{n-1} \mid \boldsymbol{\theta})}{p(\mathcal{D}^n)} \quad \left(p(\mathcal{D}^n \mid \boldsymbol{\theta}) = p(\mathbf{x}_n \mid \boldsymbol{\theta})p(\mathcal{D}^{n-1} \mid \boldsymbol{\theta})\right)$$

$$= \frac{p(\mathbf{x}_n \mid \boldsymbol{\theta})p(\boldsymbol{\theta} \mid \mathcal{D}^{n-1})}{p(\mathcal{D}^n)/p(\mathcal{D}^{n-1})}$$

$$= \frac{p(\mathbf{x}_n \mid \boldsymbol{\theta})p(\boldsymbol{\theta} \mid \mathcal{D}^{n-1})}{\int p(\mathbf{x}_n \mid \boldsymbol{\theta})p(\boldsymbol{\theta} \mid \mathcal{D}^{n-1})d\boldsymbol{\theta}}$$

($p(\boldsymbol{\theta} \mid \mathcal{D}^n)$ should be normalized w.r.t. $\boldsymbol{\theta}$)

A Recursive View of Bayesian Estimation (Cont.)

$$p(\boldsymbol{\theta} \mid \mathcal{D}^n) = \frac{p(\mathbf{x}_n \mid \boldsymbol{\theta})p(\boldsymbol{\theta} \mid \mathcal{D}^{n-1})}{\int p(\mathbf{x}_n \mid \boldsymbol{\theta})p(\boldsymbol{\theta} \mid \mathcal{D}^{n-1})d\boldsymbol{\theta}}$$

Recursive estimation:
(a.k.a. *incremental learning*)

$$p(\boldsymbol{\theta} \mid \mathcal{D}^0) = p(\boldsymbol{\theta})$$
$$p(\boldsymbol{\theta} \mid \mathcal{D}^k) \propto p(\mathbf{x}_k \mid \boldsymbol{\theta})p(\boldsymbol{\theta} \mid \mathcal{D}^{k-1})$$
$$k = 1, 2, \ldots, n$$

An illustrative example

$$p(x \mid \theta) \sim U(0, \theta) = \begin{cases} 1/\theta, & 0 \leq x \leq \theta \\ 0, & \text{otherwise} \end{cases}$$

$$p(\theta) \sim U(0, 10) = \begin{cases} 1/10, & 0 < \theta \leq 10 \\ 0, & \text{otherwise} \end{cases}$$

Given $\mathcal{D} = \{4, 7, 2, 8\}$, how would the recursive Bayes learning procedure proceed?

A Recursive View of Bayesian Estimation (Cont.)

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

$$p(\theta) \sim U(0,10) = \begin{cases} 1/10, & 0 < \theta \le 10 \\ 0, & \text{otherwise} \end{cases}$$

Given $\mathcal{D} = \{4,7,2,8\}$, *how would the recursive Bayes learning procedure proceed?*

$\mathcal{D}^1 = \{4\}:$

$$p(\theta \mid \mathcal{D}^1) \propto p(x=4 \mid \theta)p(\theta \mid \mathcal{D}^0) \propto \begin{cases} \frac{1}{\theta}, & 4 \le \theta \le 10 \\ 0, & \text{otherwise} \end{cases}$$

$\mathcal{D}^2 = \{4,7\}:$

$$p(\theta \mid \mathcal{D}^2) \propto p(x=7 \mid \theta)p(\theta \mid \mathcal{D}^1) \propto \begin{cases} \frac{1}{\theta^2}, & 7 \le \theta \le 10 \\ 0, & \text{otherwise} \end{cases}$$

A Recursive View of Bayesian Estimation (Cont.)

$$p(x \mid \theta) \sim U(0,\theta) = \begin{cases} 1/\theta, & 0 \leq x \leq \theta \\ 0, & \text{otherwise} \end{cases}$$

$$p(\theta) \sim U(0,10) = \begin{cases} 1/10, & 0 < \theta \leq 10 \\ 0, & \text{otherwise} \end{cases}$$

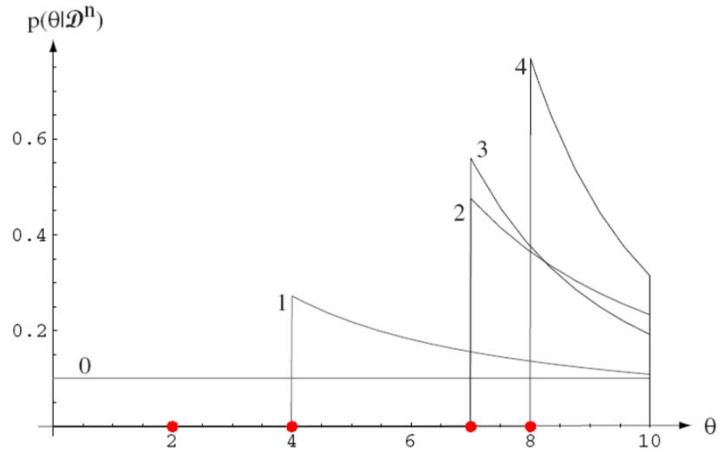
Given $\mathcal{D} = \{4, 7, 2, 8\}$, how would the recursive Bayes learning procedure proceed?

$\mathcal{D}^3 = \{4,7,2\}:$ $$p(\theta \mid \mathcal{D}^3) \propto p(x=2 \mid \theta)p(\theta \mid \mathcal{D}^2) \propto \begin{cases} \frac{1}{\theta^3}, & 7 \leq \theta \leq 10 \\ 0, & \text{otherwise} \end{cases}$$

$\mathcal{D}^4 = \{4,7,2,8\}:$ $$p(\theta \mid \mathcal{D}^4) \propto p(x=8 \mid \theta)p(\theta \mid \mathcal{D}^3) \propto \begin{cases} \frac{1}{\theta^4}, & 8 \leq \theta \leq 10 \\ 0, & \text{otherwise} \end{cases}$$

general solution: $$p(\theta \mid \mathcal{D}^n) \propto \begin{cases} \frac{1}{\theta^n}, & \max_x [\mathcal{D}^n] \leq \theta \leq 10 \\ 0, & \text{otherwise} \end{cases}$$

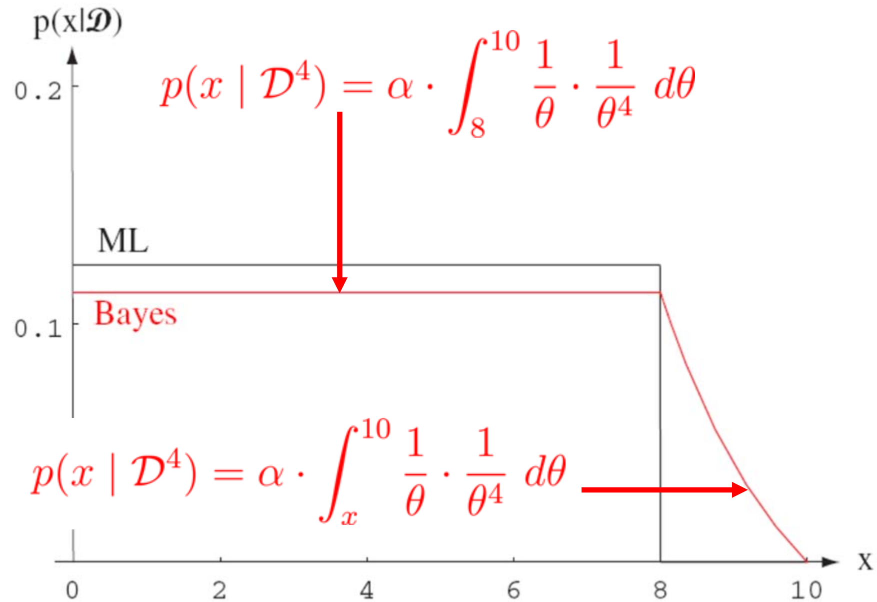
A Recursive View of Bayesian Estimation (Cont.)



$$p(x \mid \theta) \sim U(0,\theta) = \begin{cases} 1/\theta, & 0 \leq x \leq \theta \\ 0, & \text{otherwise} \end{cases}$$

$$p(\theta) \sim U(0,10) = \begin{cases} 1/10, & 0 < \theta \leq 10 \\ 0, & \text{otherwise} \end{cases}$$

$$\mathcal{D} = \{4, 2, 7, 8\}$$



$$p(x \mid \mathcal{D}^4) = \alpha \cdot \int_8^{10} \frac{1}{\theta} \cdot \frac{1}{\theta^4}\, d\theta$$

$$p(x \mid \mathcal{D}^4) = \alpha \cdot \int_x^{10} \frac{1}{\theta} \cdot \frac{1}{\theta^4}\, d\theta$$

$$p(x \mid \mathcal{D}^4) = \int p(x \mid \theta) p(\theta \mid \mathcal{D}^4) d\theta$$

A Few Notes on Parametric Techniques

ML estimation vs. Bayes estimation

- *Infinite examples*: ML estimation $=$ Bayes estimation
- *Complexity*: ML estimation $<$ Bayes estimation
- *Interpretability*: ML estimation $>$ Bayes estimation
- *Prior knowledge*: ML estimation $<$ Bayes estimation

Source of classification error

Bayes error + Model error + Estimation error

Summary

- Key issue for PR
  - Estimate prior and class-conditional pdf from training set
  - Basic assumption on training examples: *i.i.d.*
- Two strategies to the key issue
  - Parametric form for class-conditional pdf
    - Maximum likelihood (ML) estimation
    - Bayesian estimation
  - No parametric form for class-conditional pdf

Summary (Cont.)

- Maximum likelihood estimation
  - Settings: parameters as fixed but unknown values
  - The objective function: Log-likelihood function
  - Necessary conditions for ML estimation: gradient for the objective function should be zero vector
  - The Gaussian case
    - Unknown $\boldsymbol{\mu}$
    - Unknown $\boldsymbol{\mu}$ and Σ

Summary (Cont.)

- Bayesian estimation
  - Settings: parameters as random variables
  - The general procedure
    - Phase I: *prior pdf* ➔ *posterior pdf* (for $\boldsymbol{\theta}$)
    - Phase II: *posterior pdf* (for $\boldsymbol{\theta}$) ➔ *class-conditional pdf* (for $\mathbf{x}$)
    - Phase III: *prediction* (Eq.22 [pp.91])
  - The Gaussian case
    - Unknown $\boldsymbol{\mu}$: univariate and multivariate
  - A recursive view of Bayesian estimation
    - $p(\boldsymbol{\theta} \mid \mathcal{D}^n) \propto p(\mathbf{x}_k \mid \boldsymbol{\theta})p(\boldsymbol{\theta} \mid \mathcal{D}^{n-1})$