Chapter 3 Maximum-Likelihood and Bayesian Parameter Estimation

Bayes Theorem for Classification

$$P\left(\omega_j \mid \mathbf{x}\right)=\frac{p\left(\mathbf{x} \mid \omega_j\right) \cdot P\left(\omega_j\right)}{p(\mathbf{x})}(1 \leq j \leq c)$$

To compute posterior probability $P(w_j \mid \mathbf{x})$, we need to know

- Prior probability $P(w_j)$
- Likelihood $p(\mathbf{x} \mid w_j)$

Collection of Training Examples

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

- Composed of c data sets
- Each example in $\mathcal{D}_j$ is drawn according to the class-conditional pdf $p\left(\mathbf{x} \mid \omega_j\right)$
- Examples in $\mathcal{D}_j$ are i.i.d. random variables

To get prior probability

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

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

To get class-conditional pdf

Case 1: $p(\mathbf{x} \mid w_j)$ has certain parametric form

e.g.

$$p\left(\mathbf{x} \mid \omega_j\right) \sim N\left(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j\right)$$

the parameters are

$$\boldsymbol{\theta}_j=\left\{\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j\right\}$$

we know that $\mathbf{x} \in R^d$ so $\boldsymbol{\theta}_j$ contains $d+\frac{d(d+1)}{2}$ free parameters (corresponding to $\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j$)

To show the dependence, we also write $p(\mathbf{x} \mid w_j)$ as $p(\mathbf{x} \mid w_j, \boldsymbol{\theta})$

Case 2: $p(\mathbf{x} \mid w_j)$ has no parametric form

Note that it **doesn't mean no parameters**

Estimation Under Parametric Form

Maximum-Likelihood (ML) estimation

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

Bayesian estimation

- 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

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 $\mathcal{D}_j \rightarrow \mathcal{D}; \boldsymbol{\theta}_j \rightarrow \boldsymbol{\theta}$

Definitions

- $\mathbf{x}_k \sim p(\mathbf{x} \mid \boldsymbol{\theta}) \quad (k = 1, \ldots, n)$
- $\boldsymbol{\theta}$: 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

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

The likelihood of $\boldsymbol{\theta}$ w.r.t. the set of observed examples

The goal is

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

Define **log-likelihood function**

$$l(\boldsymbol{\theta}) = \ln p(\mathcal{D} \mid \boldsymbol{\theta})$$

so the goal could be rewrite as

$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} l(\boldsymbol{\theta})$$

the necessary conditions for ML estimate $\hat{\boldsymbol{\theta}}$

$$\nabla_{\boldsymbol{\theta}} l_{|_{\theta=\hat{\theta}}} = \mathbf{0}$$

The Gaussian Case

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

Case 1: Σ is known

For each component

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

$$\begin{aligned}
\ln p\left(\mathbf{x}_k \mid \boldsymbol{\mu}\right) & =-\tfrac{1}{2}\ln\left[(2\pi)^d|\boldsymbol{\Sigma}|\right]-\tfrac{1}{2}\left(\mathbf{x}_k-\boldsymbol{\mu}\right)^t \boldsymbol{\Sigma}^{-1}\left(\mathbf{x}_k-\boldsymbol{\mu}\right) \\
& =-\tfrac{1}{2}\ln\left[(2\pi)^d|\boldsymbol{\Sigma}|\right]-\tfrac{1}{2}\mathbf{x}_k^t \boldsymbol{\Sigma}^{-1}\mathbf{x}_k+\boldsymbol{\mu}^t \boldsymbol{\Sigma}^{-1}\mathbf{x}_k-\tfrac{1}{2}\boldsymbol{\mu}^t \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}
\end{aligned}$$

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

so we have

$$\nabla_\mu l=\sum_{k=1}\Sigma^{-1}\left(\mathbf{x}_k-\mu\right)=0$$

then we multiply Σ on both sizes

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

that is

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

which is a intuitive result

Case 2: Σ is unknown

Firstly we consider univariate case:

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

where

$$\boldsymbol{\theta}=\begin{bmatrix}\theta_1 \\ \theta_2\end{bmatrix}=\begin{bmatrix}\mu \\ \sigma^2\end{bmatrix}$$

use $\ln$ function

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

so we get

$$\boldsymbol{\nabla}_{\boldsymbol{\theta}} \ln l(\hat{\boldsymbol{\theta}})=\boldsymbol{\nabla}_{\boldsymbol{\theta}} \ln p\left(x_k \mid \boldsymbol{\theta}\right)=\begin{bmatrix}\frac{1}{\theta_2}\left(x_k-\theta_1\right) \\ -\frac{1}{2\theta_2}+\frac{\left(x_k-\theta_1\right)^2}{2\theta_2^2}\end{bmatrix}=\begin{bmatrix}0 \\ 0\end{bmatrix}$$

that is

$$\hat{\theta}_1=\frac{1}{n}\sum_{k=1}^n x_k$$

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

in multivariate case

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

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

Bayesian Estimation

- The **parametric form** of the likelihood function for each category is known $p\left(\mathbf{x} \mid \omega_j, \boldsymbol{\theta}_j\right)(1 \le j \le c)$
- Consider $\boldsymbol{\theta}$ as **random variables**
- Fully exploit training examples

$$P\left(\omega_j \mid \mathbf{x}, \mathcal{D}^*\right)$$
$$\left(\mathcal{D}^* = \mathcal{D}_1 \cup \mathcal{D}_2 \cup \cdots \cup \mathcal{D}_c\right)$$

Analyze

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

where

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

so

$$\begin{aligned} P\left(\omega_j \mid \mathbf{x}, \mathcal{D}^*\right) &= \frac{p(\mathcal{D}^*) \cdot P(\omega_j \mid \mathcal{D}^*) \cdot p(\mathbf{x} \mid \omega_j, \mathcal{D}^*)}{p(\mathcal{D}^*) \cdot \sum_{i=1}^{c} P(\omega_i \mid \mathcal{D}^*) \cdot p(\mathbf{x} \mid \omega_i, \mathcal{D}^*)} \\ &= \frac{P(\omega_j \mid \mathcal{D}^*) \cdot p(\mathbf{x} \mid \omega_j, \mathcal{D}^*)}{\sum_{i=1}^{c} P(\omega_i \mid \mathcal{D}^*) \cdot p(\mathbf{x} \mid \omega_i, \mathcal{D}^*)} \end{aligned}$$

with two assumptions

$$P\left(\omega_j \mid \mathcal{D}^*\right) = P\left(\omega_j\right)$$
$$p\left(\mathbf{x} \mid \omega_j, \mathcal{D}^*\right) = p\left(\mathbf{x} \mid \omega_j, \mathcal{D}_j\right)$$

(in first equation, Prior Probability is independent of Dataset)

we have

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

to calculate, the key problem is to determine $p\left(\mathbf{x} \mid \omega_j, \mathcal{D}_j\right)$

since we treat each class independently, we simplify the class-conditional pdf notation

$$p\left(\mathbf{x} \mid \omega_j, \mathcal{D}_j\right) \rightarrow p(\mathbf{x} \mid \mathcal{D})$$

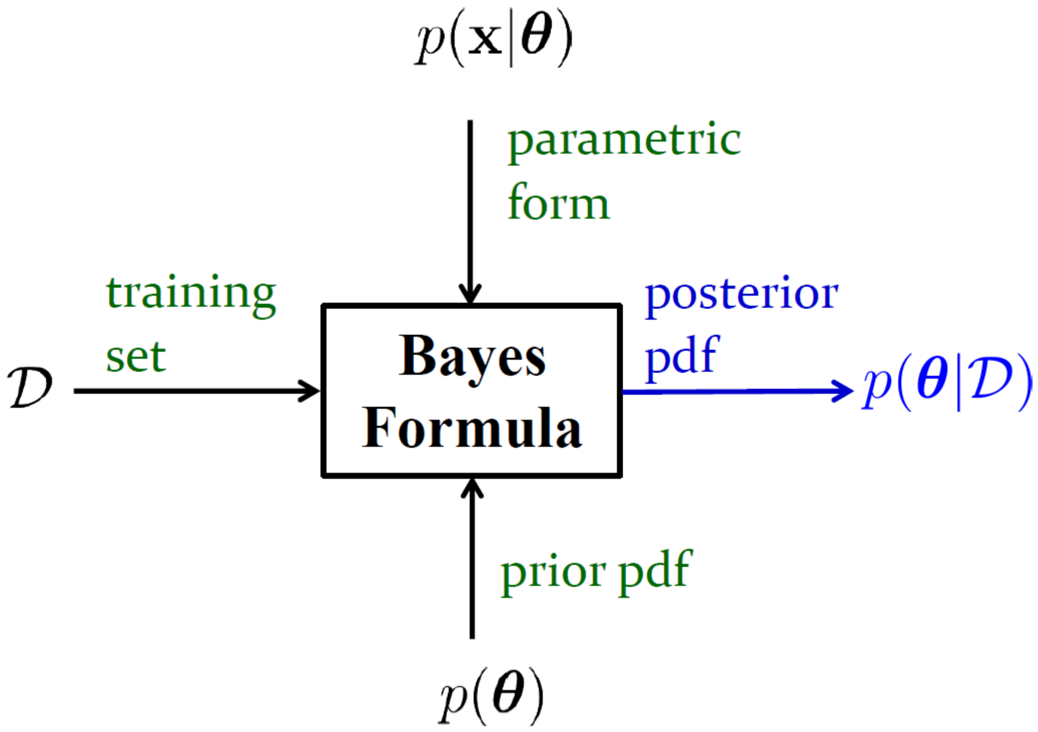
introducing $\boldsymbol{\theta}$ which is a random variable w.r.t. parametric form

$$\begin{aligned}
p(\mathbf{x} \mid \mathcal{D}) & =\int p(\mathbf{x}, \boldsymbol{\theta} \mid \mathcal{D}) d \boldsymbol{\theta} \\
& =\int p(\mathbf{x} \mid \boldsymbol{\theta}, \mathcal{D}) p(\boldsymbol{\theta} \mid \mathcal{D}) d \boldsymbol{\theta} \\
& =\int p(\mathbf{x} \mid \boldsymbol{\theta}) p(\boldsymbol{\theta} \mid \mathcal{D}) d \boldsymbol{\theta}
\end{aligned}$$

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

General Procedure

Phase 1: prior pdf $\rightarrow$ posterior pdf (for θ)

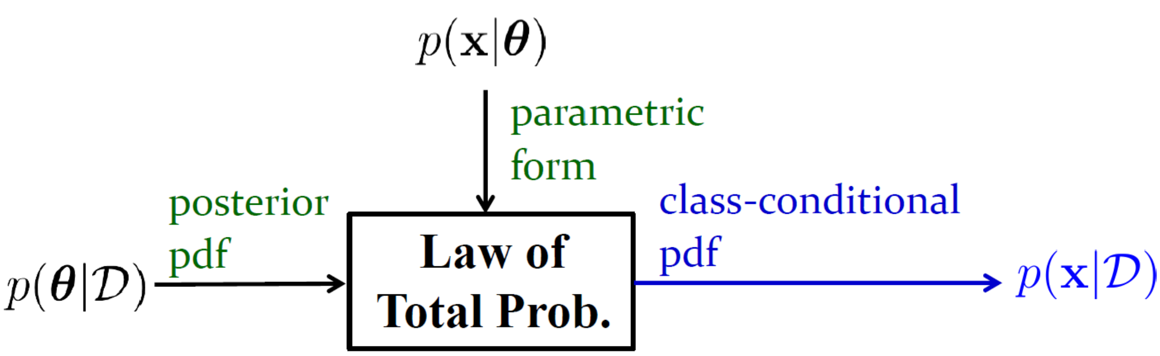


where

$$\begin{aligned}
p(\boldsymbol{\theta} \mid \mathcal{D}) & =\frac{p(\boldsymbol{\theta}, \mathcal{D})}{p(\mathcal{D})} \\
& =\frac{p(\boldsymbol{\theta}) p(\mathcal{D} \mid \boldsymbol{\theta})}{\int p(\boldsymbol{\theta}, \mathcal{D}) d \boldsymbol{\theta}} \\
& =\frac{p(\boldsymbol{\theta}) p(\mathcal{D} \mid \boldsymbol{\theta})}{\int p(\boldsymbol{\theta}) p(\mathcal{D} \mid \boldsymbol{\theta}) d \boldsymbol{\theta}}
\end{aligned}$$

$$p(\mathcal{D} \mid \boldsymbol{\theta})=\prod_{k=1}^{n} p\left(\mathbf{x}_{k} \mid \boldsymbol{\theta}\right)$$

Phase 2: posterior pdf (for θ) $\rightarrow$ class-conditional pdf (for $\mathbf{x}$)



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

Phase 3: posterior pdf (for $\mathbf{x}, D^*$)

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

Example 1: Gaussian Case & Unknown μ

we assume

$$\begin{aligned} & p(x \mid \mu) \sim N\left(\mu, \sigma^2\right) \\ & p(\mu) \sim N\left(\mu_0, \sigma_0^2\right)\end{aligned}$$

Note that for $p(\mu)$ here, we could also assume other form of prior pdf;

$$\begin{aligned} p(\mu \mid \mathcal{D})=\frac{p(\mu, \mathcal{D})}{p(\mathcal{D})} & =\frac{p(\mu) p(\mathcal{D} \mid \mu)}{\int p(\mu) p(\mathcal{D} \mid \mu) d \mu} \\ & =\alpha p(\mu) p(\mathcal{D} \mid \mu) \\ & =\alpha p(\mu) \prod_{k=1}^n p\left(x_k \mid \mu\right)\end{aligned}$$

where $\int p(\mu) p(\mathcal{D} \mid \mu) d \mu$ is a constant not related to μ, rewrite as α; examples in $\mathcal{D}$ are i.i.d

continue

$$\begin{aligned} p(\mu \mid \mathcal{D}) \quad & =\alpha p(\mu) \prod_{k=1}^n p\left(x_k \mid \mu\right) \\ & =\alpha \cdot \frac{1}{\sqrt{2 \pi} \sigma_0} \exp \left[-\frac{1}{2}\left(\frac{\mu-\mu_0}{\sigma_0}\right)^2\right] \cdot \prod_{k=1}^n \frac{1}{\sqrt{2 \pi} \sigma} \exp \left[-\frac{1}{2}\left(\frac{x_k-\mu}{\sigma}\right)^2\right] \\ & =\alpha^{\prime} \cdot \exp \left[-\frac{1}{2}\left(\left(\frac{\mu-\mu_0}{\sigma_0}\right)^2+\sum_{k=1}^n\left(\frac{\mu-x_k}{\sigma}\right)^2\right)\right] \\ & =\alpha^{\prime \prime} \cdot \exp \left[-\frac{1}{2}\left[\left(\frac{n}{\sigma^2}+\frac{1}{\sigma_0^2}\right) \mu^2-2\left(\frac{1}{\sigma^2} \sum_{k=1}^n x_k+\frac{\mu_0}{\sigma_0^2}\right) \mu\right]\right]\end{aligned}$$

Note that $p(\mu \mid \mathcal{D})$ is an **exponential function** of a **quadratic function** of μ

So that $p(\mu \mid \mathcal{D})$ is a normal pdf as well

that is

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

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

with the result, considering

$$\begin{aligned} p(x \mid \mathcal{D}) \quad & =\int p(x \mid \mu) p(\mu \mid \mathcal{D}) d \mu \\ & =\int \frac{1}{\sqrt{2 \pi} \sigma} \exp \left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right] \frac{1}{\sqrt{2 \pi} \sigma_n} \exp \left[-\frac{1}{2}\left(\frac{\mu-\mu_n}{\sigma_n}\right)^2\right] d \mu \\ & =\beta \cdot \exp \left[-\frac{1}{2} \frac{\left(x-\mu_n\right)^2}{\sigma^2+\sigma_n^2}\right] \cdot \int[\text{pdf function of a Norm}] \\ & =\beta \cdot \exp \left[-\frac{1}{2} \frac{\left(x-\mu_n\right)^2}{\sigma^2+\sigma_n^2}\right]\end{aligned}$$

Note that $p(x \mid \mathcal{D})$ is an **exponential function** of a **quadratic function** of x

So that $p(x \mid \mathcal{D})$ is a normal pdf as well again

that is

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

Example 2: Gaussian Case & Multivariate & Unknown $\boldsymbol{\mu}$

$$\begin{gathered}
p(\mathbf{x} \mid \boldsymbol{\mu}) \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \\
p(\boldsymbol{\mu}) \sim N\left(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0\right)
\end{gathered}$$

so that

$$\begin{gathered}
p(\boldsymbol{\mu} \mid \mathcal{D}) \sim N\left(\boldsymbol{\mu}_n, \boldsymbol{\Sigma}_n\right) \\
p(\mathbf{x} \mid \mathcal{D}) \sim N\left(\boldsymbol{\mu}_n, \boldsymbol{\Sigma} + \boldsymbol{\Sigma}_n\right)
\end{gathered}$$

where

$$\begin{gathered}
\mu_n = \Sigma_0\left(\Sigma_0 + \frac{1}{n}\Sigma\right)^{-1} \frac{1}{n} \sum_{k=1}^{n} \mathbf{x}_k + \frac{1}{n}\Sigma\left(\Sigma_0 + \frac{1}{n}\Sigma\right)^{-1} \mu_0 \\
\Sigma_n = \Sigma_0\left(\Sigma_0 + \frac{1}{n}\Sigma\right)^{-1} \frac{1}{n}\Sigma
\end{gathered}$$

Comparing ML estimation & Bayes estimation

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

The source of classification error

$$\text{Bayes error } + \text{ Model error } + \text{ Estimation error}$$