Ch 03. Parameter Estimation

Maximum Likelihood Estimation & Bayesian Estimation

Part 1 Maximum Likelihood Estimation

Approaches to Pattern Classification

- **Approach 1**： Estimate class-conditional probability density $p(\mathbf{x} \mid \omega_i)$
  - Through $p(\mathbf{x} \mid \omega_i)$ and $P(\omega_i)$, calculate posterior probability $P(\omega_i \mid \mathbf{x})$ with Bayes' rule, then make decisions with maximum posterior probability
  - Two Methods
    - **Method 1a**： Parameter estimation of probability density

      Based on parametric description of $p(\mathbf{x} \mid \omega_i)$
    - **Method 1b**： Non-Parametric estimation of probability density

      Based on non-parametric description of $p(\mathbf{x} \mid \omega_i)$
- **Approach 2**： Estimate posterior probability $P(\omega_i \mid \mathbf{x})$
  - Don't have to estimate $p(\mathbf{x} \mid \omega_i)$ in advance
- **Approach 3**： Compute discrimination function
  - Don't have to estimate $p(\mathbf{x} \mid \omega_i)$ or $P(\omega_i \mid \mathbf{x})$

Probability Density Function Estimation & Parameter Estimation

- Parameter estimation is based on parameterized representation of $p(\mathbf{x} \mid \omega_i)$ by known function form
- The question of estimating unknown probability density function $p(\mathbf{x} \mid \omega_i)$ can be simplified to estimate unknown parameters in known function form
- All unknown parameters in $p(\mathbf{x} \mid \omega_i)$ can be written in vector form, which are called parameter vectors $\boldsymbol{\theta}_i$, the probability density function $p(\mathbf{x} \mid \omega_i)$ with unknown parameters can be expressed as $p(\mathbf{x} \mid \omega_i, \boldsymbol{\theta}_i)$
- Parameter vector in Gaussian density function

$$\boldsymbol{\theta}_i = (\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$$

Parameter Estimation in Bayes Decision

- Bayes decision is the optimal decision (minimum total risk, minimum error probability)
  - Precondition
    - Known prior probability $P(\omega_i)$
    - Known class-conditional probability density $p(\mathbf{x} \mid \omega_i)$
- Unfortunately……
  - In most cases, prior probability and class-conditional probability density are unknown
- What we can use……
  - Some vague and general knowledge about pattern recognition
  - Some design samples (training samples), which constitute a specific subset and representative of patterns to be classified

Parameter Estimation in Bayes Decision

- Solution
  - Suppose class-conditional probability density is a kind of probability density distribution function with parameters, and estimate the unknown parameters through training data
  - Take the probability density function after parameter estimation as the class-conditional probability density and utilize Bayes Decision to classify
  - Supervised Learning
    - The true category of each sample in training set is known

Parameter Estimation Method

- Maximum Likelihood Estimation（最大似然估计）
  - Hypothesis
    - Treat the parameters to be estimated as definite quantities, but the values are unknown
  - Estimation method
    - Treat the parameter values that maximize the probability of generating training data as the best estimation of these parameters
- Bayesian Estimation (Bayesian Learning)
  - Hypothesis
    - Treat the parameters to be estimated as random variables confirming to a certain prior distribution
  - Estimation method
    - By observing samples, transform the prior probability density into the posterior probability density through Bayes' rule

Parameter Estimation Methods

- The relationship between ML estimation and Bayesian estimation
  - ML estimation is usually simpler than Bayesian estimation
  - ML estimation can give the value of the parameter while Bayesian estimation can give the distribution of all possible parameter values
  - When there is so much available data that the effect of prior knowledge is reduced, Bayesian estimation can be reduced to ML estimation

Maximum Likelihood Estimation

- Given c classes： $\omega_1, \omega_2, \ldots \omega_c$
  - Suppose all class-conditional probability density functions $p(\mathbf{x} \mid \omega_i, \theta_i)$, $i=1,\ldots,c$ have known parameterized forms
  - Suppose each parameter vector θ_i has an independent effect on the category it belongs to
  - For example： $p(\mathbf{x} \mid \omega_i, \theta_i) \sim N(\mu_i, \Sigma_i)$ where $\theta_i = (\mu_i, \Sigma_i)$
- Given c datasets (each dataset corresponds to a category): $D_1, D_2, \ldots D_c$
  - Samples in each dataset D_i are independent and identically distributed (i.i.d) random variables, which are extracted independently from a certain probability density function $p(\mathbf{x} \mid \omega_i, \theta_i)$
  - It is impossible for D_i to provide any information to the estimation of θ_j $j \neq i$ due to parameters of different classes are independent to each other
  - Therefore, parameters can be estimated separately for each class, and the class subscripts can be omitted

$$p(\mathbf{x} \mid \omega_i, \theta_i) \Longrightarrow p(\mathbf{x} \mid \theta) \qquad D_i \Longrightarrow D \qquad \theta_i \Longrightarrow \theta$$

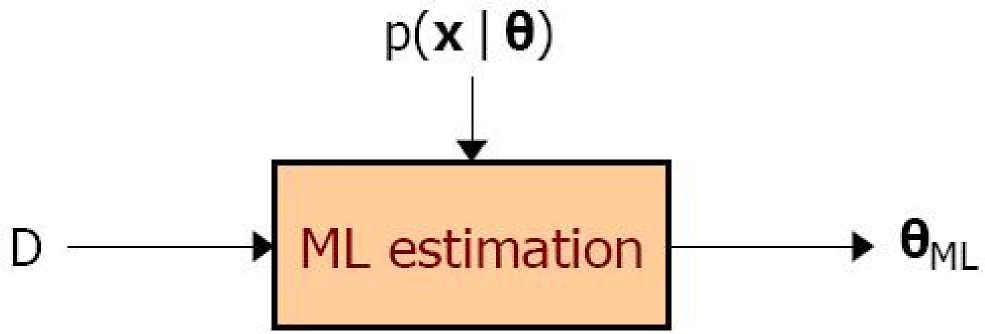
Maximum Likelihood Estimation

- The likelihood function $p(D \mid \boldsymbol{\theta}) = \prod_{k=1}^{n} p(\mathbf{x}_k \mid \boldsymbol{\theta})$ of θ

  relative to dataset $D = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$

- The ML estimation of θ is the value θ_{ML} that maximizes the likelihood function $p(D \mid \theta)$,where $\boldsymbol{\theta}_{ML} = \arg\max_{\boldsymbol{\theta}} p(D \mid \boldsymbol{\theta})$

  Intuitively speaking, θ_{ML} is the value that maximizes the possibility of observing samples in D



$p(\mathbf{x} \mid \boldsymbol{\theta})$
D
ML estimation
$\boldsymbol{\theta}_{ML}$

Maximum Likelihood Estimation

- After ML estimation is completed, the probability density function is fully known, that is, the form and value of its parameters are known $p(\mathbf{x} \mid \omega_i, \boldsymbol{\theta}_i)$
- The posterior probability of class ω_i can be computed by Bayes' formula

$$P(\omega_i \mid \mathbf{x}) = \frac{p(\mathbf{x} \mid \omega_i, \boldsymbol{\theta}_i)\, P(\omega_i)}{\sum_{i=1}^{c} p(\mathbf{x} \mid \omega_i, \boldsymbol{\theta}_i)\, P(\omega_i)}$$

Bayes Decision can be made based on posterior probability

- Explicitly represents the role of D_i in parameter estimation

$$P(\omega_i \mid \mathbf{x}, \{D_i\}_{i=1}^{c}) = \frac{p(\mathbf{x} \mid \omega_i, D_i)\, P(\omega_i)}{\sum_{i=1}^{c} p(\mathbf{x} \mid \omega_i, D_i)\, P(\omega_i)}$$

Likelihood Function & Log-likelihood Function

- Given Dataset D, Define likelihood function $L(\boldsymbol{\theta})$ as

$$L(\boldsymbol{\theta}) \equiv p(D \mid \boldsymbol{\theta}) = \prod_{k=1}^{n} p(\mathbf{x}_k \mid \boldsymbol{\theta})$$

$L(\theta)$ can be written as $L(\theta; D)$

to emphasize that it depends on the dataset D

- Log-likelihood Function $l(\theta)$

$$l(\boldsymbol{\theta}) \equiv \log p(D \mid \boldsymbol{\theta}) = \log L(\boldsymbol{\theta}) = \sum_{k=1}^{n} \log p(\mathbf{x}_k \mid \boldsymbol{\theta})$$

The computation of log-likelihood function is usually simpler than likelihood function

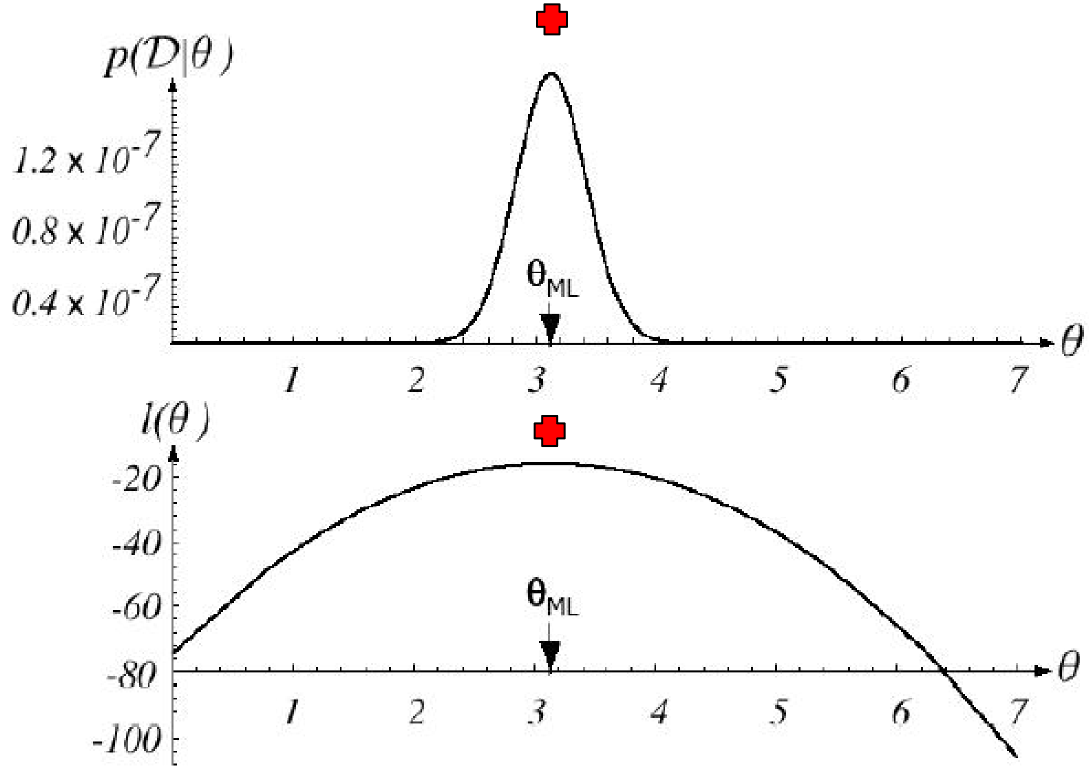
- Maximum Likelihood Estimation

$$\boldsymbol{\theta}_{ML} = \arg\max_{\boldsymbol{\theta}} p(D \mid \boldsymbol{\theta}) = \arg\max_{\boldsymbol{\theta}} L(\boldsymbol{\theta}) = \arg\max_{\boldsymbol{\theta}} l(\boldsymbol{\theta})$$

log(x) is a monotone increasing function

Maximization Problem

- The solution of ML estimation is realized by maximizing likelihood function or log-likelihood function



Maximization Problem

- Let $\boldsymbol{\theta}$ denote p-dimensional parameter vector $(\theta_1, \ldots, \theta_p)^t$, and $\nabla_{\boldsymbol{\theta}}$ denotes the gradient operator

$$\nabla_{\boldsymbol{\theta}} = \left(\frac{\partial}{\partial \theta_1}, \ \ldots, \ \frac{\partial}{\partial \theta_p} \right)^t$$

- Necessary conditions for global maximum (likelihood equation)

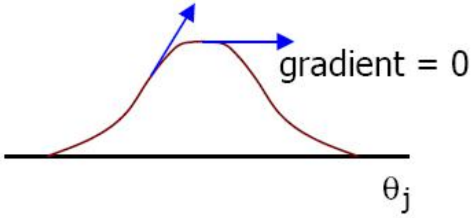
$$\frac{\partial L}{\partial \theta_j} = 0 \quad \forall j \quad \text{or} \quad \nabla_{\boldsymbol{\theta}} L = \mathbf{0} = (0, \ldots, 0)^t$$

Equally (likelihood equation)

$$\nabla_{\boldsymbol{\theta}} \log L = \mathbf{0} = (0, \ldots, 0)^t$$

- The solution of likelihood equation or log likelihood equation is not a sufficient condition to obtain global maximum
  - Might be

Glob$\underbrace{\text{al maximum / minimum, local maximum / minimu}}_{\text{Extremum}}$m, inflection point



ML estimation-Gaussian Case : μ is unknown

- $p(\mathbf{x}) \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$
- Log-likelihood of μ under $\mathbf{x}_k$

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

- Log-likelihood equation

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

- ML estimation of μ

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

→ The sample mean of dataset D

ML estimation-Gaussian Case: $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are unknown

- The case that x is the single variable
  - Parameter vector $\boldsymbol{\theta} = (\theta_1, \theta_2)^t = (\mu, \sigma^2)^t$
  - The log likelihood of $\boldsymbol{\theta}$ under x_k

$$\log p(x_k \mid \boldsymbol{\theta}) = -\frac{1}{2}\log\left(2\pi\sigma^2\right) - \frac{1}{2\sigma^2}(x_k - \mu)^2$$

  - Log likelihood equation

$$\sum_{k=1}^{n} \nabla_{\boldsymbol{\theta}} \log p(x_k \mid \boldsymbol{\theta}) = \begin{bmatrix} \displaystyle\sum_{k=1}^{n} \frac{1}{\sigma^2}(x_k - \mu) \\ \displaystyle\sum_{k=1}^{n} \left[-\frac{1}{2\sigma^2} + \frac{(x_k - \mu)^2}{2\sigma^4} \right] \end{bmatrix} = \begin{bmatrix} \displaystyle\sum_{k=1}^{n} \frac{1}{\sigma^2}(x_k - \mu) \\ \displaystyle -\sum_{k=1}^{n} \frac{1}{2\sigma^2} + \sum_{k=1}^{n} \frac{(x_k - \mu)^2}{2\sigma^4} \end{bmatrix} = \mathbf{0}$$

ML estimation-Gaussian Case: μ and Σ are unknown

- The case that x is the single variable
  - The ML estimation of θ

$$\mu_{ML} = \frac{1}{n}\sum_{k=1}^{n} x_k$$

$$\sigma_{ML}^2 = \frac{1}{n}\sum_{k=1}^{n} (x_k - \mu_{ML})^2$$

ML estimation-Gaussian Case: $\boldsymbol{\mu}$ and Σ are unknown

- The case that x is the multi-variable
  - Parameter vector $\boldsymbol{\theta} = (\theta_1, \theta_2) = (\boldsymbol{\mu}, \boldsymbol{\Sigma})$
  - Log-likelihood of $\boldsymbol{\theta}$ under $\mathbf{x}_k$

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

  - The ML estimation of $\boldsymbol{\theta}$

$$\boldsymbol{\mu}_{ML} = \frac{1}{n}\sum_{k=1}^{n}\mathbf{x}_k \longrightarrow \text{The sample mean of dataset D}$$

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

Bias of Estimation

- The ML estimation of Σ is biased estimation, that is, the mathematical expectation of ML estimation of covariance matrix for all possible sample sets of size n is not equal to the actual covariance matrix

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

- The unbiased estimation of Σ

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

the sample covariance matrix of dataset D

- Due to $\boldsymbol{\Sigma}_{ML}=\frac{n-1}{n}\mathbf{C}$
  - The ML estimation of Σ_{ML} is asymptotically unbiased estimation, that is, with the increase of sample number n, Σ_{ML} tends to C

Parameter Estimation Method

- Maximum Likelihood Estimation（最大似然估计）
  - Hypothesis
    - Treat the parameters to be estimated as definite quantities, but the values are unknown
  - Estimation method
    - Treat the parameter values that maximize the probability of generating training data as the best estimation of these parameters
- Bayesian Estimation (Bayesian Learning)
  - Hypothesis
    - Treat the parameters to be estimated as random variables confirming to a certain prior distribution
  - Estimation method
    - By observing samples, transform the prior probability density into the posterior probability density through Bayes' rule

Part 2 Bayes Estimation

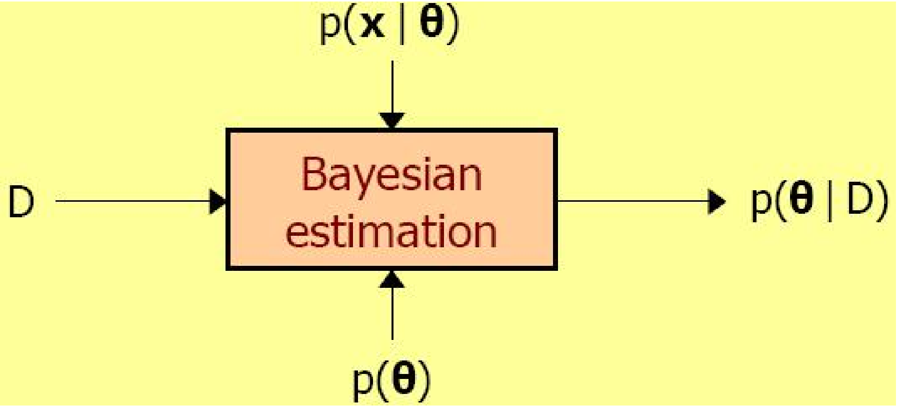
Bayes estimation

- Given
  - The probability density function $p(\mathbf{x} \mid \theta)$ in parametric form, where the unknown parameters are expressed as vectors $\boldsymbol{\theta}$
  - the prior probability density $p(\theta)$ about $\boldsymbol{\theta}$
  - Dataset $D = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$
- Solve
  - The posterior probability density $p(\theta \mid D)$ of parameter vector $\boldsymbol{\theta}$
  - The posterior probability density of x

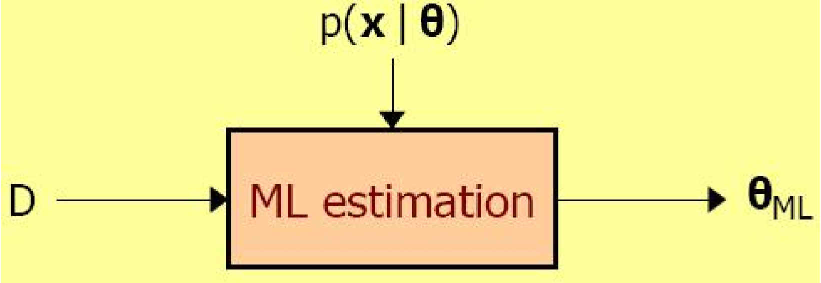
$$p(\mathbf{x} \mid D) = \int p(\mathbf{x}, \boldsymbol{\theta} \mid D)\, d\boldsymbol{\theta} = \int p(\mathbf{x} \mid \boldsymbol{\theta})\, p(\boldsymbol{\theta} \mid D)\, d\boldsymbol{\theta}$$

Bayes estimation

- Bayes estimation



- ML estimation



Bayes estimation

- To clarify the role of dataset D, which is similar to ML estimation, the posterior probability required by Bayes decision can be rewritten

$$P(\omega_i \mid \mathbf{x}, \{D_i\}_{i=1}^{c}) = \frac{\boxed{p(\mathbf{x} \mid \omega_i, D_i)}\, P(\omega_i)}{\sum_{i=1}^{c} p(\mathbf{x} \mid \omega_i, D_i)\, P(\omega_i)}$$

- Simplify

$$p(\mathbf{x} \mid \omega_i, D_i) \Longrightarrow \boxed{p(\mathbf{x} \mid D)}$$

Bayes estimation

- Core Problem
  - Given a set of training samples D, these samples are independently extracted from the fixed but unknown probability density function p(x), and $p(\mathbf{x} \mid D)$ are required to be estimated according to these samples
- Basic Idea

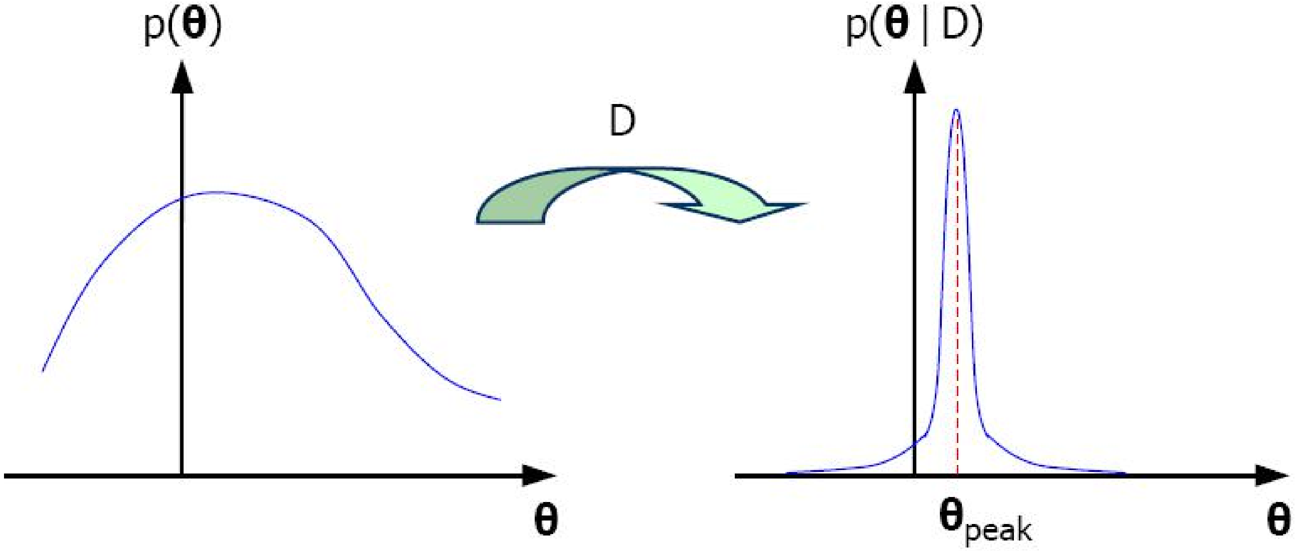
$$p(\mathbf{x} \mid D) = \int p(\mathbf{x}, \boldsymbol{\theta} \mid D)\, d\boldsymbol{\theta} = \int p(\mathbf{x} \mid \boldsymbol{\theta})\, p(\boldsymbol{\theta} \mid D)\, d\boldsymbol{\theta}$$

  - Suppose $p(\mathbf{x} \mid \theta)$ is the probability density of known parametric form
  - $p(\theta \mid D)$ is the posterior probability density of $\boldsymbol{\theta}$ under D -- by Bayes estimation
  - If $p(\theta \mid D)$ forms the most significant peak near a certain value θ_{peak} then

$$p(\mathbf{x} \mid D) \cong p(\mathbf{x} \mid \boldsymbol{\theta}_{peak})$$

Bayes estimation

- By observing dataset D, the prior probability density $p(\theta)$ is transformed into the posterior probability density $p(\theta|D)$, and it is expected to have a peak at the real value θ

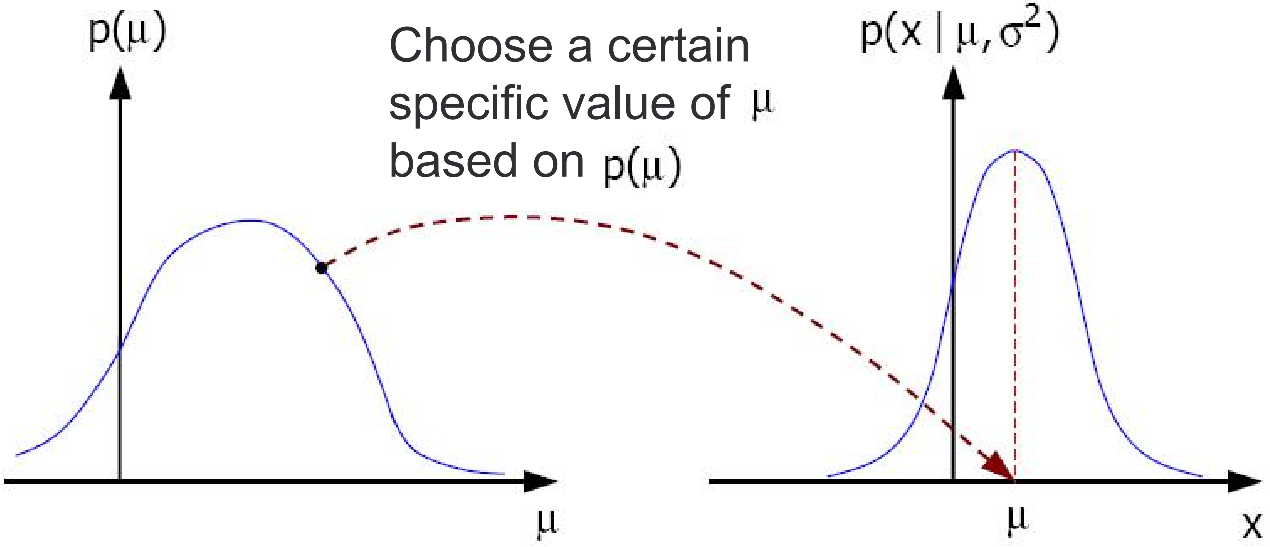


Gaussian Case: Single Variable, μ is unknown, σ^2 is known

- Target probability density function

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

  - μ is unknown, but the distribution $p(\mu)$ is known
  - σ^2 is known, $p(x \mid \mu, \sigma^2)$ can be simplified as $p(x \mid \mu)$



Choose a certain specific value of μ based on $p(\mu)$

Gaussian Case: Single Variable, μ is unknown, σ^2 is known

- Calculate posterior probability μ by Bayes rule

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

where α is a normalized coefficient dependent on sample set

$D = \{x_1, x_2, ..., x_n\}$, which is independent of μ

- Suppose $p(\mu) \sim N(\mu_0, \sigma_0^2)$, where μ_0 and σ_0^2 are known

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

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

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

Gaussian Case: Single Variable, μ is unknown, σ^2 is known

- $p(\mu \mid D)$ is also obey Gaussian distribution

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

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

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

  - $p(\mu)$ is called conjugate prior（共轭先验）, $p(\mu \mid D)$ is called reproducing density（复制密度）

  - Compute

$$p(\mu \mid D) = \frac{1}{\sqrt{2\pi}\,\sigma_n} \exp\left[-\frac{1}{2} \left(\frac{\mu - \mu_n}{\sigma_n} \right)^2 \right]$$

Sample mean $\hat{\mu}_n = \frac{1}{n} \sum_{k=1}^{n} x_k$

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

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

Gaussian Case: Single Variable, μ is unknown, σ^2 is known

- Observation

The sum is 1, indicating that μ_n is on the line between $\hat{\mu}_n$ and μ_0

$$\mu_n = \boxed{\frac{n\sigma_0^2}{n\sigma_0^2 + \sigma^2}}\hat{\mu}_n + \boxed{\frac{\sigma^2}{n\sigma_0^2 + \sigma^2}}\mu_0$$

Sample mean $\hat{\mu}_n = \frac{1}{n}\sum_{k=1}^{n} x_k$

- If $\sigma_0 \neq 0$

When $n \to +\infty$, $\mu_n \to \hat{\mu}_n$ → ML estimation

- If $\sigma_0 = 0$

Degradation situation $\mu_n = \mu_0$

The respective contributions of prior knowledge and empirical data are determined by the ratio of σ^2 and σ_0^2, which is called dogmatism（决断因子）

- If $\sigma_0 \quad \sigma$

$\mu_n \approx \hat{\mu}_n$

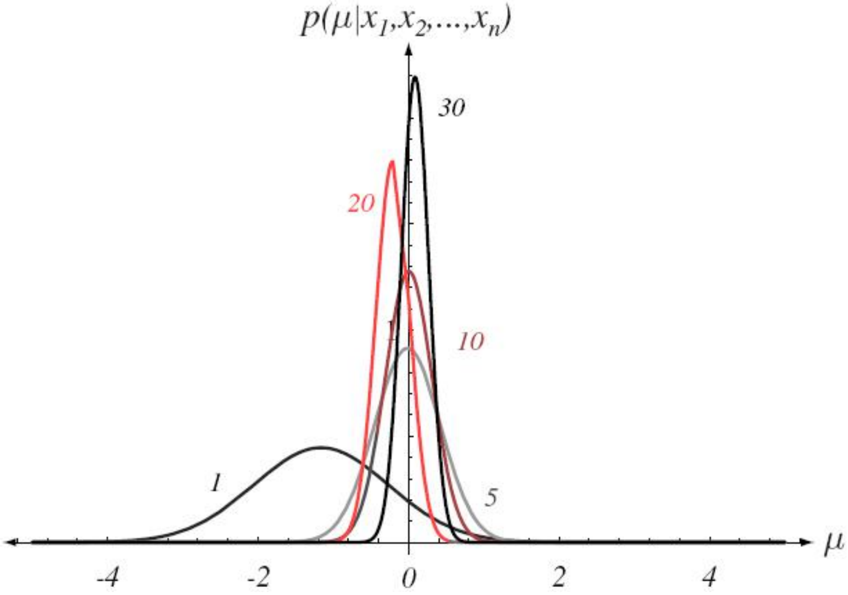
When enough samples are obtained, the exact assumption of the exact value of μ_0 and σ_0^2 becomes irrelevant, and μ_n will converge to the sample mean $\hat{\mu}_n$

Gaussian Case: Single Variable, μ is unknown, σ^2 is known

- Observation

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

- As the number of samples n increases, the σ_n^2 monotonically decreases, that is, the additional samples can reduce the uncertainty of the estimation of μ , as n increases, the waveform of $p(\mu \mid D)$ gets sharper and sharper



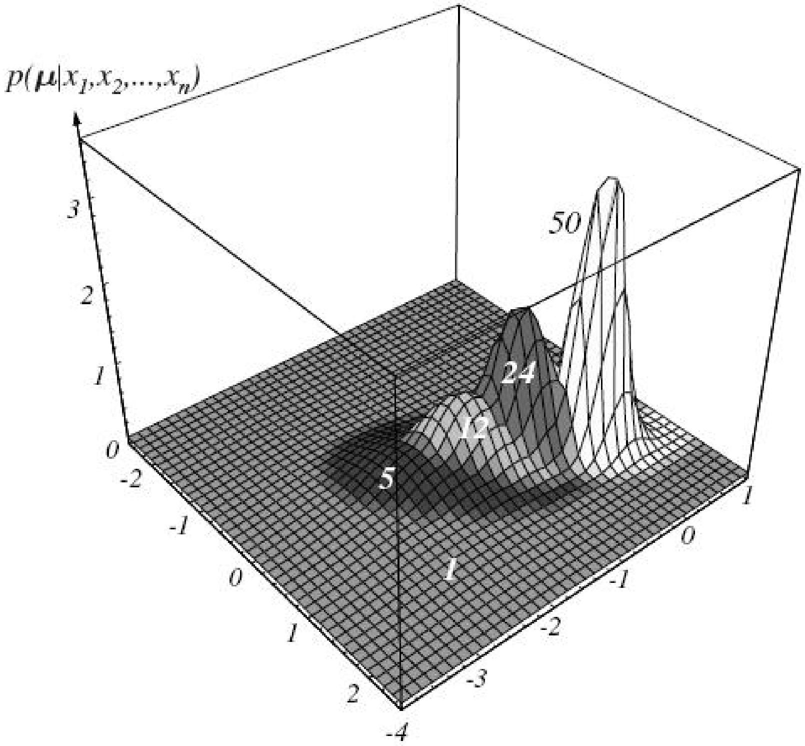
The Process of Bayes Learning

Gaussian Case: Single Variable, μ is unknown, σ^2 is known

- Observation

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

- As the number of samples n increases, the σ_n^2 monotonically decreases, that is, the additional samples can reduce the uncertainty of the estimation of μ , as n increases, the waveform of $p(\mu \mid D)$ gets sharper and sharper



The Process of Bayes Learning

ML estimation-Gaussian Case : μ is unknown

- $p(\mathbf{x}) \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$
- Log-likelihood of μ under $\mathbf{x}_k$

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

- Log-likelihood equation

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

- ML estimation of μ

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

The sample mean of dataset D

ML estimation-Gaussian Case: $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are unknown

- The case that x is the single variable
  - Parameter vector $\boldsymbol{\theta} = (\theta_1, \theta_2)^t = (\mu, \sigma^2)^t$
  - The log likelihood of $\boldsymbol{\theta}$ under x_k

$$\log p(x_k \mid \boldsymbol{\theta}) = -\frac{1}{2}\log\left(2\pi\sigma^2\right) - \frac{1}{2\sigma^2}(x_k - \mu)^2$$

  - Log likelihood equation

$$\sum_{k=1}^{n} \nabla_{\boldsymbol{\theta}} \log p(x_k \mid \boldsymbol{\theta}) = \begin{bmatrix} \displaystyle\sum_{k=1}^{n} \frac{1}{\sigma^2}(x_k - \mu) \\ \displaystyle\sum_{k=1}^{n}\left[-\frac{1}{2\sigma^2} + \frac{(x_k - \mu)^2}{2\sigma^4}\right] \end{bmatrix} = \begin{bmatrix} \displaystyle\sum_{k=1}^{n} \frac{1}{\sigma^2}(x_k - \mu) \\ \displaystyle -\sum_{k=1}^{n}\frac{1}{2\sigma^2} + \sum_{k=1}^{n}\frac{(x_k - \mu)^2}{2\sigma^4} \end{bmatrix} = \mathbf{0}$$

ML estimation-Gaussian Case: μ and Σ are unknown

- The case that x is the single variable
  - The ML estimation of θ

$$\mu_{ML} = \frac{1}{n}\sum_{k=1}^{n} x_k$$

$$\sigma_{ML}^2 = \frac{1}{n}\sum_{k=1}^{n} (x_k - \mu_{ML})^2$$

ML estimation-Gaussian Case: $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are unknown

- The case that x is the multi-variable
  - Parameter vector $\boldsymbol{\theta} = (\theta_1, \theta_2) = (\boldsymbol{\mu}, \boldsymbol{\Sigma})$
  - Log-likelihood of $\boldsymbol{\theta}$ under $\mathbf{x}_k$

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

  - The ML estimation of $\boldsymbol{\theta}$

$\boldsymbol{\mu}_{ML} = \boxed{\frac{1}{n}\sum_{k=1}^{n} \mathbf{x}_k} \longrightarrow$ The sample mean of dataset D

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

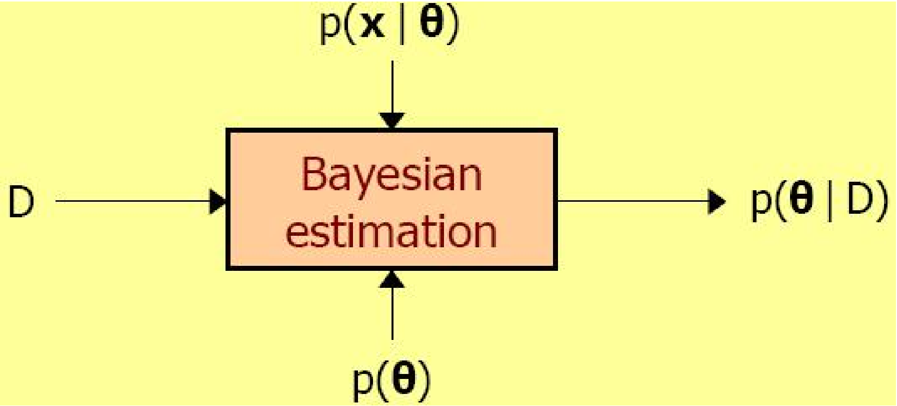
Summary - Bayes estimation

- Given
  - The probability density function $p(\mathbf{x} \mid \boldsymbol{\theta})$ in parametric form, where the unknown parameters are expressed as vectors $\boldsymbol{\theta}$
  - the prior probability density $p(\theta)$ about $\boldsymbol{\theta}$
  - Dataset $D = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$
- Solve
  - The posterior probability density $p(\boldsymbol{\theta} | D)$ of parameter vector $\boldsymbol{\theta}$
  - The posterior probability density of x

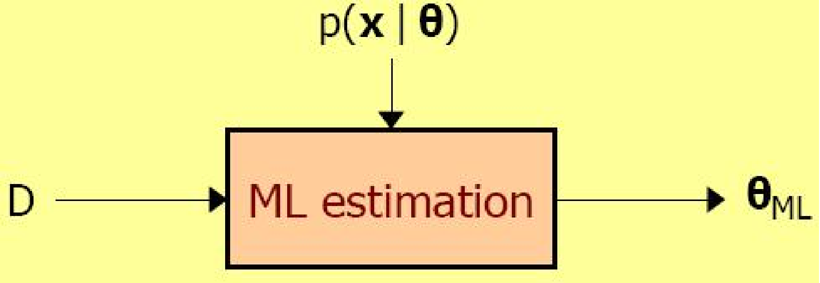
$$p(\mathbf{x} \mid D) = \int p(\mathbf{x}, \boldsymbol{\theta} \mid D)\, d\boldsymbol{\theta} = \int p(\mathbf{x} \mid \boldsymbol{\theta})\, p(\boldsymbol{\theta} \mid D)\, d\boldsymbol{\theta}$$

Bayes estimation

- Bayes estimation


$p(\mathbf{x} \mid \boldsymbol{\theta})$
D
Bayesian estimation
$p(\boldsymbol{\theta} \mid D)$
$p(\boldsymbol{\theta})$

- ML estimation


$p(\mathbf{x} \mid \boldsymbol{\theta})$
D
ML estimation
$\boldsymbol{\theta}_{ML}$

Bayes estimation

- Core Problem
  - Given a set of training samples D, these samples are independently extracted from the fixed but unknown probability density function p(x), and $p(\mathbf{x} \mid D)$ are required to be estimated according to these samples
- Basic Idea

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

  - Suppose $p(\mathbf{x} \mid \theta)$ is the probability density of known parametric form
  - $p(\theta \mid D)$ is the posterior probability density of $\boldsymbol{\theta}$ under D -- by Bayes estimation
  - If $p(\theta \mid D)$ forms the most significant peak near a certain value θ_{peak} then

$$p(\mathbf{x} \mid D) \cong p(\mathbf{x} \mid \boldsymbol{\theta}_{peak})$$

Gaussian Case: Single Variable, μ is unknown, σ^2 is known

- Calculate posterior probability μ by Bayes rule

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

where α is a normalized coefficient dependent on sample set

$D = \{x_1, x_2, \ldots, x_n\}$, which is independent of μ

- Suppose $p(\mu) \sim N(\mu_0, \sigma_0^2)$, where μ_0 and σ_0^2 are known

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

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

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

Gaussian Case: Single Variable, μ is unknown, σ^2 is known

- p(μ | D) is also obey Gaussian distribution

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

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

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

- p(μ) is called conjugate prior（共轭先验）, p(μ | D) is called reproducing density（复制密度）

- Compute

$$p(\mu \mid D) = \frac{1}{\sqrt{2\pi}\,\sigma_n} \exp\left[-\frac{1}{2} \left(\frac{\mu - \mu_n}{\sigma_n} \right)^2 \right]$$

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

Sample mean $\hat{\mu}_n = \frac{1}{n}\sum_{k=1}^{n} x_k$

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

Gaussian Case: Single Variable, μ is unknown, σ^2 is known

- Class-conditional probability density

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

$$p(x \mid D) = \int p(x \mid \mu) p(\mu \mid D) d\mu$$

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

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

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

Gaussian Case: Single Variable, μ is unknown, σ^2 is known

- Class-conditional probability density

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

  - The parametric form of $p(x \mid D)$ is $p(x \mid \mu) \sim N(\mu, \sigma^2)$
  - The result of Bayes Estimation

$$\mu_n \longrightarrow \mu$$

$$\sigma^2 + \sigma_n^2 \longrightarrow \sigma^2$$

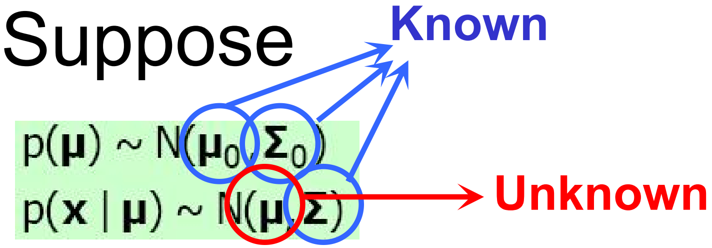
The uncertainty of the estimation of μ increases the uncertainty of ($\sigma^2 + \sigma_n^2 \longrightarrow \sigma^2$)

- Bayes decision rule

$$P(\omega_i \mid x, \{D_i\}_{i=1}^{c}) = \frac{p(x \mid \omega_i, D_i)\, P(\omega_i)}{\sum_{i=1}^{c} p(x \mid \omega_i, D_i)\, P(\omega_i)}$$

Gaussian Case: Multi Variable, $\boldsymbol{\mu}$ is unknown, $\boldsymbol{\Sigma}$ is known

- Suppose



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

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

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

$$= \alpha' \exp\left[-\frac{1}{2}\left(\boldsymbol{\mu}^t (n\boldsymbol{\Sigma}^{-1} + \boldsymbol{\Sigma}_0^{-1})\boldsymbol{\mu} - 2\boldsymbol{\mu}^t \left(\boldsymbol{\Sigma}^{-1} \sum_{k=1}^{n} \mathbf{x}_k + \boldsymbol{\Sigma}_0^{-1} \boldsymbol{\mu}_0 \right) \right) \right]$$

$$= \alpha'' \exp\left[-\frac{1}{2} (\boldsymbol{\mu} - \boldsymbol{\mu}_n)^t \boldsymbol{\Sigma}_n^{-1} (\boldsymbol{\mu} - \boldsymbol{\mu}_n) \right]$$

So $p(\boldsymbol{\mu} \mid D) \sim N(\boldsymbol{\mu}_n, \boldsymbol{\Sigma}_n)$

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

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

Gaussian Case: Multi Variable, $\boldsymbol{\mu}$ is unknown, $\boldsymbol{\Sigma}$ is known

- Class-conditional probability density

$$p(\mathbf{x} \mid D) \quad N(\boldsymbol{\mu}_n, \boldsymbol{\Sigma} + \boldsymbol{\Sigma}_n)$$

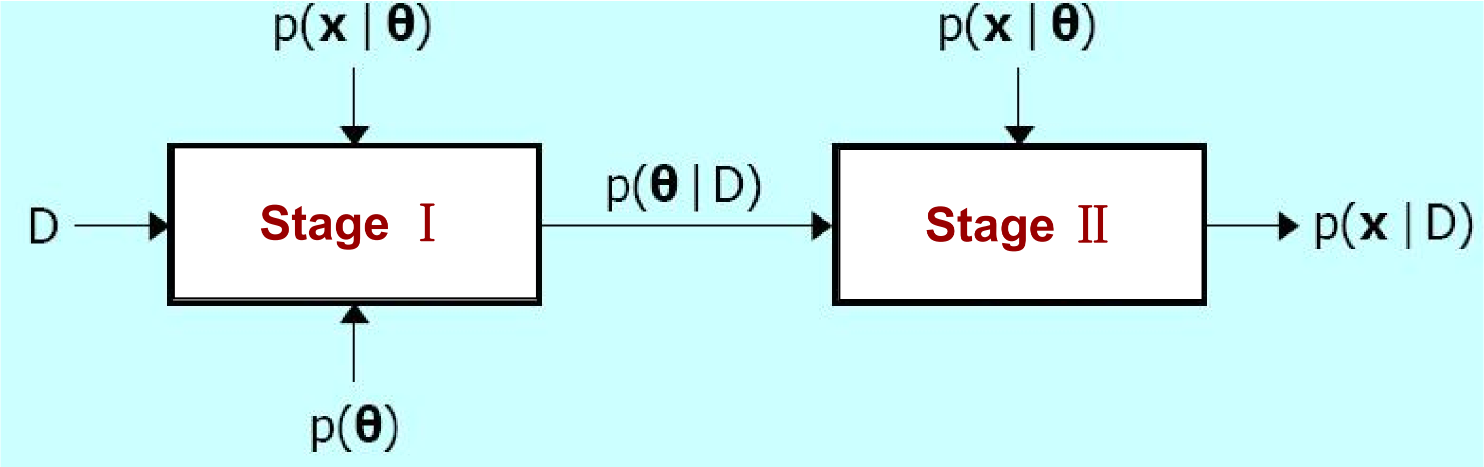
  - A simpler perspective

$$\mathbf{x} = \boldsymbol{\mu} + \mathbf{y}$$

$$p(\boldsymbol{\mu} \mid D) \quad N(\boldsymbol{\mu}_n, \boldsymbol{\Sigma}_n)$$

$$p(\mathbf{y}) \quad N(0, \boldsymbol{\Sigma})$$

General Process of Bayes Estimation



$$p(D\,|\,\boldsymbol{\theta}) = \prod_{k=1}^{n} p(x_k\ |\boldsymbol{\theta}\,)$$

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

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

Recursive Bayesian Learning

- Determine the number of samples in the sample set $D^n=\{\mathbf{x}_1,\mathbf{x}_2,\ldots,\mathbf{x}_n\}$, $D^{n-1}=\{\mathbf{x}_1,\mathbf{x}_2,\ldots,\mathbf{x}_{n-1}\},\ldots,D^1=\{\mathbf{x}_1\}$
- Bayesian learning

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

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

The posterior probability density of $\boldsymbol{\theta}$ under n samples

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

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

•The posterior probability density of $\boldsymbol{\theta}$ under n-1 samples

Recursive Bayesian Learning

- Recursive Learning Process

1. Before observing samples

$$p(\boldsymbol{\theta} | D^0) = p(\boldsymbol{\theta})$$

2. Observe sample $\mathbf{x}_1$

$$p(\boldsymbol{\theta} | D^1) = \frac{p(\mathbf{x}_1 | \boldsymbol{\theta}) p(\boldsymbol{\theta} | D^0)}{\int p(\mathbf{x}_1 | \boldsymbol{\theta}) p(\boldsymbol{\theta} | D^0) d\boldsymbol{\theta}}$$

3. Observe sample $\mathbf{x}_2$

$$p(\boldsymbol{\theta} | D^2) = \frac{p(\mathbf{x}_2 | \boldsymbol{\theta}) p(\boldsymbol{\theta} | D^1)}{\int p(\mathbf{x}_2 | \boldsymbol{\theta}) p(\boldsymbol{\theta} | D^1) d\boldsymbol{\theta}}$$

......

n. Observe sample $\mathbf{x}_n$

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

For each step, we only need to know the current sample $\mathbf{x}_i$ and the result of the previous step $p(\boldsymbol{\theta} | D^{i-1})$

incremental learning
（增量学习）

Example

- **Question**
  - One dimensional sample obeys uniform distribution

$$p(x|\theta) \quad U(0,\theta) = \begin{cases} 1/\theta & 0 \le x \le \theta \\ 0 & \text{其他} \end{cases}$$

  - Known: Parameter θ is bounded ,suppose $p(\theta) \quad U(0,10)$
  - Existing sample set $D^4 = \{4,7,2,8\}$
  - To solve $p(x|D^4)$ by recursive Bayesian learning

Example

- **Solution**
  - Before observing samples

$$p(\theta | D^0) = p(\theta) = U(0,10)$$

  - Observe sample $x_1 = 4$

$$p(4 | \theta) = \begin{cases} 1/\theta & \theta \geq 4 \\ 0 & \text{其他} \end{cases}$$

$$p(\theta | D^1) \propto p(4 | \theta) p(\theta | D^0) \propto \begin{cases} 1/\theta & 4 \leq \theta \leq 10 \\ 0 & \text{其他} \end{cases}$$

  - Observe sample $x_2 = 7$

$$p(7 | \theta) = \begin{cases} 1/\theta & \theta \geq 7 \\ 0 & \text{其他} \end{cases}$$

$$p(\theta | D^2) \propto p(7 | \theta) p(\theta | D^1) = \begin{cases} 1/\theta^2 & 7 \leq \theta \leq 10 \\ 0 & \text{其他} \end{cases}$$

Example

- **Solution**
  - Observe sample $x_3 = 2$

$$p(2 \mid \theta) = \begin{cases} 1/\theta & \theta \geq 2 \\ 0 & 其他 \end{cases}$$

$$p(\theta \mid D^3) \propto p(x \mid \theta) p(\theta \mid D^2) = \begin{cases} 1/\theta^3 & 7 \leq \theta \leq 10 \\ 0 & 其他 \end{cases}$$

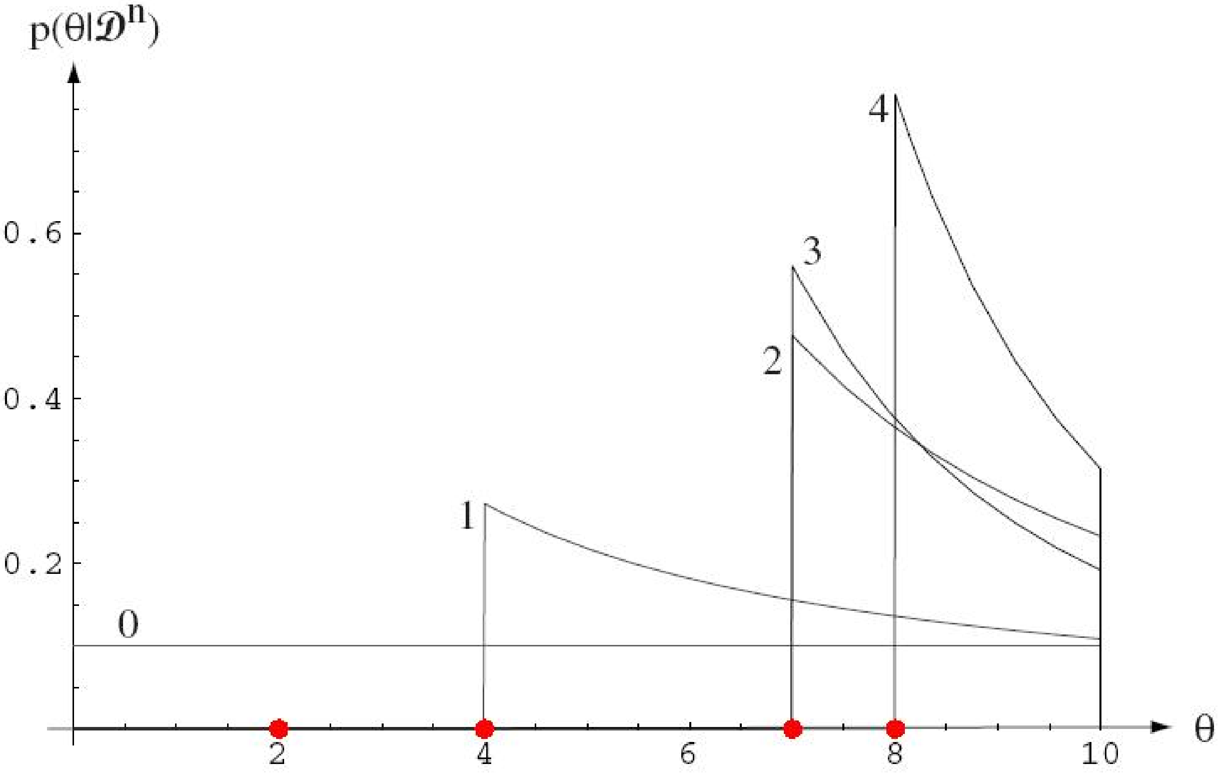
  - Observe sample $x_4 = 8$

$$p(8 \mid \theta) = \begin{cases} 1/\theta & \theta \geq 8 \\ 0 & 其他 \end{cases}$$

$$p(\theta \mid D^4) \propto p(x \mid \theta) p(\theta \mid D^3) = \begin{cases} 1/\theta^4 & 8 \leq \theta \leq 10 \\ 0 & 其他 \end{cases}$$

Example

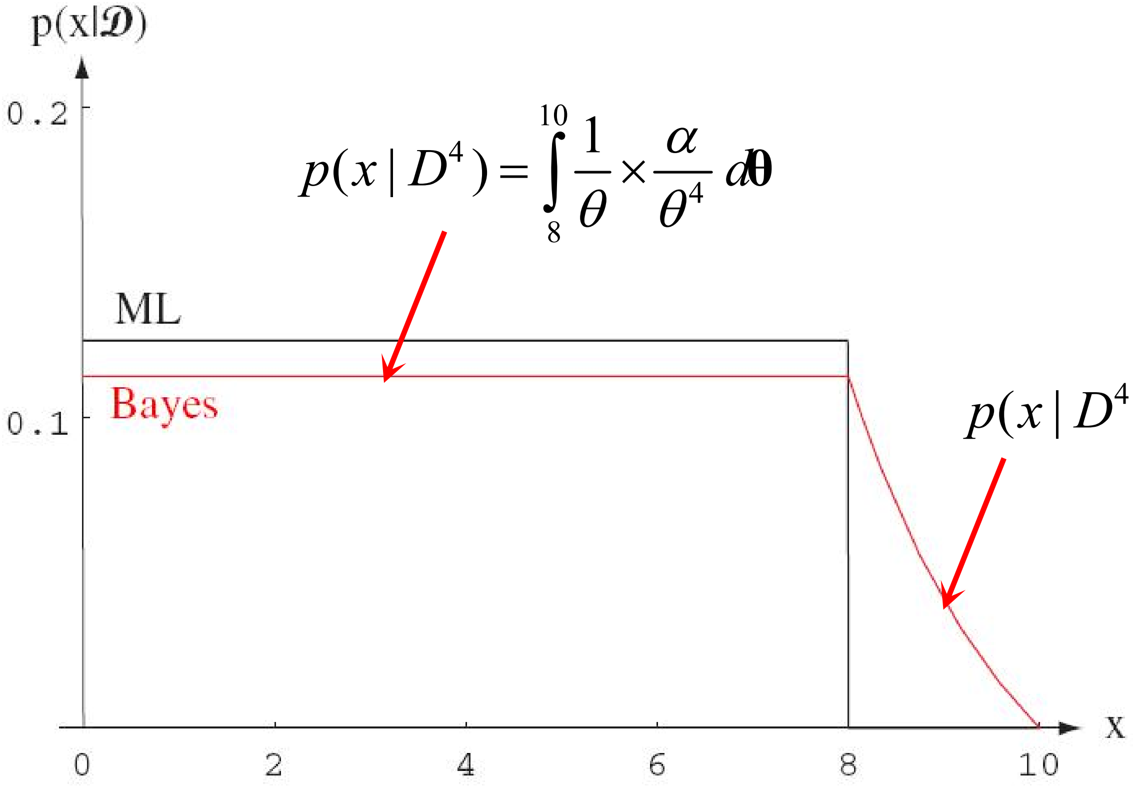
- **Solution**



Example

- **Solution**

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



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

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

Bayes Estimation vs. MLE

- When the number of samples tends to infinity

Bayes estimation = ML estimation

- Computation complexity

Bayes estimation > ML estimation

- Intelligibility

Bayes estimation < ML estimation

- Flexible application of prior knowledge

Bayes estimation > ML estimation

- Theoretical basis

Bayes estimation > ML estimation

Bayes Decision Based on Parameter Estimation

1. Suppose the parametric form of the class-conditional probability density
2. Use ML estimation or Bayesian estimation to estimate the conditional probability density

- Calculate the posterior probability by using Bayes formula
- Classify the test samples according to the maximum posterior probability

determined by the problem and cannot be eliminated

1. Sources of Classification Error

- Bayesian error (inseparability error)
- Model error
- Estimation error