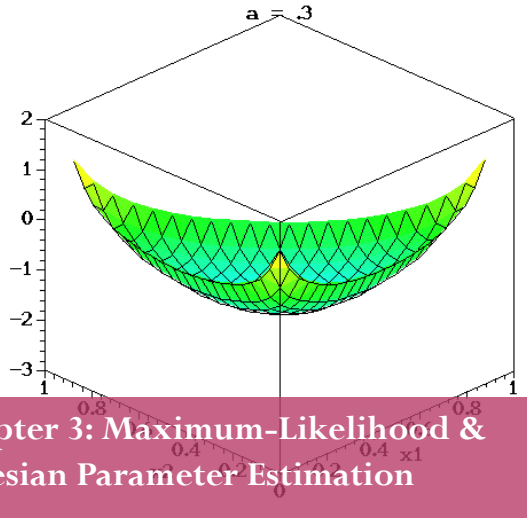
PATTERN RECOGNITION

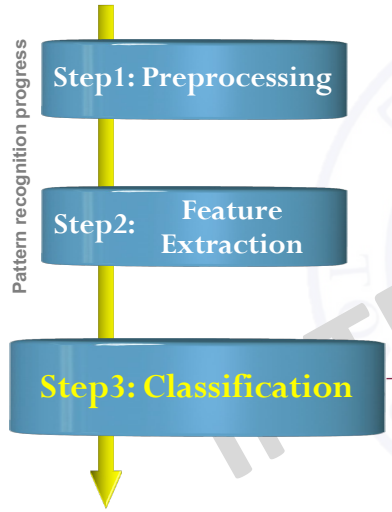


Chapter 3: Maximum-Likelihood & Bayesian Parameter Estimation

Ch.3 Content

1. Introduction

- Pattern Recognition System



$P(\omega_i) = ?$
$p(x|\omega_i) = ?$

Bayesian Classification

Minimize error rate decision

$$- P(\omega_i|x) = \frac{p(x|\omega_i)P(\omega_i)}{\sum_{j=1}^{c} p(x|\omega_j)P(\omega_j)}, j = 1, \dots, c$$

Minimum-Risk Decision decision rule

$$- R(\alpha_i|x) = \sum_{j=1}^{c} \lambda(\alpha_i|\omega_j)P(\omega_j|x), \quad i = 1,2, \dots a$$

1. Introduction

- **Parameter Estimation** —— Data availability in a Bayesian framework

$P(\omega_i) = ?$
$p(x|\omega_i) = ?$

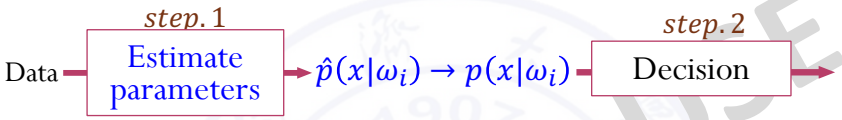
  - In pattern recognition applications we rarely have this kind of complete knowledge about the probabilistic structure of the problem.
    - We have a number of design samples or training data.
    - The problem is to find some way to use this information to design or train the classifier.
    - One approach is to use the samples to estimate the unknown probabilities and probability densities, and to use the resulting estimates as if they were the true values.

DATA→ $\hat{P}(\omega_i) \rightarrow P(\omega_i)$
DATA→ $\hat{p}(x|\omega_i) \rightarrow p(x|\omega_i)$

→ **Bayesian Classifier**

1. Introduction

- **Bayes Decision based on samples**



Example: $p(x|\omega_i) \sim N(\mu_i, \Sigma_i)$

Samples are independent and identically distributed (i.i.d.)

- **Class-conditional densities $p(x|\omega_i)$**
  - Nonparameter Estimation
  - Parameter Estimation: If we know the **number of parameters** in advance and our general knowledge about the problem permits us to **parameterize the conditional densities** then severity of the problem can be reduced significantly.

1. Introduction

- **Parameter Estimation techniques**
  - Maximum-Likelihood Estimations(MLE)
  - Bayesian Estimations(BE)

MLE	BE
• Parameters are **fixed** but unknown. • Best parameters are obtained by maximizing the probability of obtaining the samples observed.	• Parameters are **random variables** having some known distribution. • Best parameters are obtained by estimating them given the data.

- Results are nearly identical, but the approaches are different.

Ch.3 Content

2. Maximum-Likelihood Estimation

- **Maximum-likelihood Estimation (MLE)**
  - GOAL: **determine the most likely values** of the population parameter value (e.g, μ, σ, β, ρ, …) given an observed sample value (e.g., x-bar, s, b, r, ….).

  - **Simpler** than any other alternative techniques.

- General principle
- Gaussian Case: unknown μ
- Gaussian Case: unknown μ and σ

2. Maximum-Likelihood Estimation

- **General principle**
  - Preliminaries and Notations

$p(x|\omega_i) = p(x|\omega_i, \theta_i)$

$\theta = (\theta_1, \theta_2 \dots \theta_c)$: unknown parameters

$i = 1,2, \dots c$

$\theta_i = (\theta_1, \theta_2 \dots \theta_p)^t$: Each θ_i is associated with each category.

$\hat{\theta}$: estimated parameter

$D_i = \{x_1, \dots, x_n\}$: n samples (i.i.d.)

- **Task:** find $\hat{\theta}$ of parameters θ
- **Method:** ML estimation of θ is the value $\hat{\theta}$ that maximizes $p(D|\theta)$.

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

— $p(D|\theta)$ *is called the likelihood of* θ *with respect to the set of samples.*

"It is the value of $\hat{\theta}$ that best agrees with the actually observed training sample"

$$\text{Max}_{\theta}\, p(D|\theta) \Rightarrow \hat{\theta} = \underset{\theta}{\text{argmax}}\, l(\theta) = \underset{\theta}{\text{argmax}}\, p(D|\theta)$$
$$= \underset{\theta}{\text{argmax}}\, p(x_1, \dots, x_n|\theta) = \underset{\theta}{\text{argmax}} \prod_{k=1}^{n} p(x_k|\theta)$$

2. Maximum-Likelihood Estimation

- **General principle**

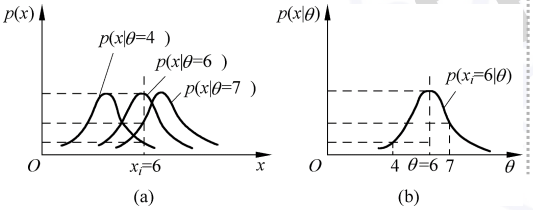
Example $p(x|\omega_i) \sim N(\mu_i, \Sigma_i)$

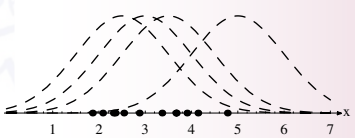
$p(x|\omega_i) = p(x|\omega_i, \theta_i)$

—$\theta_i = (\mu_i) = (\mu_i^1, \mu_i^2 \dots), i = 1,2, \dots c$, μ_i *is unknown.*

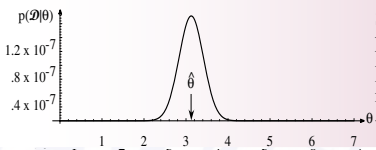
—*Use the information provided by the training samples to estimate* $\theta = (\theta_1, \theta_2 \dots \theta_c)$, *each* θ_i *is associated with each category.*

—*log-likelihood function.*

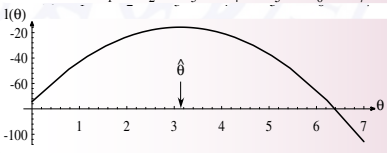




Shows several training points in one dimension, known variance, but unknown mean. Four of the infinite number of candidate source distributions are shown in dashed lines.



Shows the likelihood $p(D|\theta)$. If we had a very large number of training points, this likelihood would be very narrow. The value that maximizes the likelihood is marked $\hat{\theta}$.



Shows the log-likelihood $l(\theta)$ ($\ln l(\theta)$) which marks the maximum $\hat{\theta}$.

2. Maximum-Likelihood Estimation

- **General principle**
  - **Method: MLE log-likelihood function**
    - D contains n samples: $\{x_1, x_2 \dots x_n\}$
    - Unknown parameters: $\theta = (\theta_1, \theta_2 \dots \theta_p)^t$

STEP

① We define $l(\theta)$ as the log-likelihood function:

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

② determine $\hat{\theta}$ that maximizes the log-likelihood $\hat{\theta} = \underset{\theta}{\operatorname{argmax}}\, l(\theta)$

③ Set of necessary conditions for an optimum is:

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

④ Set the derivative of $l(\theta)$ equal to zero and solving for $\hat{\theta}$.

$$\nabla_\theta l = 0 \Longrightarrow \frac{d[l(\theta)]}{d\theta} = \frac{d[\ln p(x_k|\theta)]}{d\theta} = 0$$

∇_θ is the gradient operator

$$\nabla_\theta = \begin{bmatrix} \frac{\partial}{\partial\theta_1} \\ \frac{\partial}{\partial\theta_2} \\ \dots \\ \frac{\partial}{\partial\theta_p} \end{bmatrix}$$

2. Maximum-Likelihood Estimation

- **Gaussian Case: unknown μ**
  - Multivariate normal population: $p(x_k|\theta) \sim N(\mu, \Sigma)$
    (Samples are drawn from a multivariate normal population)
    - Unknown parameter: $\theta = \mu$
    - Sample set $D: \{x_1, x_2 \dots x_n\}$, a sample point: x_k

MLE

① log-likelihood function: $l(\theta) = \ln p(x_k|\mu) = -\frac{1}{2}\ln[(2\pi)^d|\Sigma|] - \frac{1}{2}(x_k - \mu)^t\Sigma^{-1}(x_k - \mu)$

② gradient operation: $\nabla_\theta l(\theta) = \nabla_\mu \ln p(x_k|\mu) = \Sigma^{-1}(x_k - \mu) = 0$

③ ML estimate for μ: $\sum_{k=1}^{n} \Sigma^{-1}(x_k - \hat{\mu}) = 0, \quad D: \{x_1, x_2 \dots x_n\}$

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

Just the arithmetic average of the samples of the training samples.

2. Maximum-Likelihood Estimation

- **Gaussian Case: unknown μ and σ**
  - Univariate case: $p(x_k|\theta) \sim N(\mu, \sigma^2)$

(Samples are drawn from a univariate normal population)

  - Unknown parameters: $\theta = (\theta_1, \theta_2) = (\mu, \sigma^2)$
  - Sample set D: $\{x_1, x_2 \dots x_n\}$, a sample point x_k

MLE

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

② $\nabla_\theta l = \nabla_\theta \ln P(x_k|\theta) = \begin{bmatrix} \frac{\partial}{\partial\theta_1}\ln p(x_k|\theta) \\ \frac{\sigma}{\partial\theta_2}\ln p(x_k|\theta) \end{bmatrix} = 0$

$$\begin{cases} \frac{1}{\theta_2}(x_k - \theta_1) = 0 \\ -\frac{1}{2\theta_2} + \frac{(x_k-\theta_1)^2}{2{\theta_2}^2} = 0 \end{cases}$$

③
$$\begin{cases} \sum_{k=1}^{n} \frac{1}{\hat{\theta}_2}(x_k - \hat{\theta}_1) = 0 & (1) \\ -\sum_{k=1}^{n} \frac{1}{\hat{\theta}_2} + \sum_{k=1}^{n} \frac{(x_k-\hat{\theta}_1)^2}{{\hat{\theta}_2}^2} = 0 & (2) \end{cases}$$

④ $\hat{\mu} = \frac{1}{n}\sum_{k=1}^{n} x_k$

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

2. Maximum-Likelihood Estimation

- **Gaussian Case: unknown μ and Σ**
  - Multivariate case:: $p(x_k|\theta) \sim N(\mu, \Sigma)$

(Samples are drawn from a multivariate normal population)

  - Unknown parameters θ: $\theta_1 = \mu$, $\theta_2 = \Sigma$
  - Sample set D: $\{x_1, x_2 \dots x_n\}$, a sample point x_k

MLE

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

$$\hat{\Sigma} = \frac{1}{n}\sum_{k=1}^{n}(x_k - \hat{\mu})(x_k - \hat{\mu})^t$$

2. Maximum-Likelihood Estimation

- **Gaussian Case:** Bias

	Biased estimator	unbiased estimator
σ^2	$E\left[\frac{1}{n}\sum_{k=1}^{n}(x_k-\hat{\mu})^2\right]=\frac{n-1}{n}\sigma^2\neq\sigma^2$	$E\left[\frac{1}{n-1}\sum_{k=1}^{n}(x_k-\hat{\mu})^2\right]=\sigma^2$
covariance matrix	$\hat{\Sigma}=\frac{1}{n}\sum_{k=1}^{n}(x_k-\hat{\mu})(x_k-\hat{\mu})^t=\frac{n-1}{n}\Sigma$	$\Sigma=\frac{1}{n-1}\sum_{k=1}^{n}(x_k-\hat{\mu})(x_k-\hat{\mu})^t$

asymptotically unbiased when n is very large

2. Maximum-Likelihood Estimation

Exercise

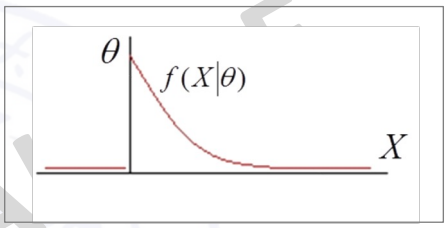
Consider an exponential distribution

$$f(x;\theta)=\begin{cases}\theta e^{-\theta x} & x\geq 0\\ 0 & otherwise\end{cases}$$

(single feature, single parameter),

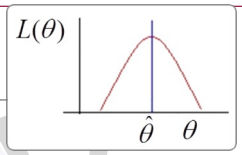
With a random sample set $\{x_1, x_2 \ldots x_n\}$, i. d. d.,

estimate $\boldsymbol{\theta}$?



2. Maximum-Likelihood Estimation

Exercise $f(x;\theta) = \begin{cases} \theta e^{-\theta x} & x \geq 0 \\ 0 & otherwise \end{cases}$



① $L(\theta) = f(x_1, x_2, \dots, x_n|\theta) = \prod_{k=1}^{n} \theta e^{-\theta x_k}$

② $l(\theta) = \ln L(\theta) = \sum_{k=1}^{n} \ln \theta - \theta \sum_{k=1}^{n} x_k$
$= n \ln \theta - \theta \sum_{k=1}^{n} x_k, \quad valid\ for\ x \geq 0$

③ $\frac{dl}{d\theta} = \frac{d \ln L(\theta)}{d\theta} = \frac{n}{\theta} - \sum_{k=1}^{n} x_k = 0$

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

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

2. Maximum-Likelihood Estimation

- **General principle: MLE log-likelihood function**
  - —**Note1**: the likelihood function must be differentiable and then this method can be used.
  - —**Note2**: $\hat{\theta}$ could represent a true global maximum, a local maximum or minimum, or an inflection point of $l(\theta)$. We should **check** each solution to identify which is the global optimum.

Example

$D = \{x_1, x_2 \dots x_n\}$, *iid.*

$$p(x|\theta) = \begin{cases} \frac{1}{\theta_2 - \theta_1} & \theta_1 < x < \theta_2 \\ 0 & others \end{cases}$$

Estimate θ_1, θ_2?

$$l(\theta) = \ln p(D|\theta)$$

$$= \begin{cases} \ln p(x_1, x_2 \dots x_n|\theta) = -n \ln(\theta_2 - \theta_1) & \theta_1 < x < \theta_2 \\ 0 & others \end{cases}$$

$$= \begin{cases} \frac{\partial}{\partial \theta_1} \ln p(x_1, x_2 \dots x_n|\theta) = n \frac{1}{\theta_2 - \theta_1} \\ \frac{\sigma}{\partial \theta_2} \ln p(x_1, x_2 \dots x_n|\theta) = -n \frac{1}{\theta_2 - \theta_1} \end{cases}$$

It is uninteresting. Need to find a new method.

Ch.3 Content

3. Bayesian Estimation

- **MLE & BE**
  - In MLE, θ was supposed fix.
  - In BE, θ is a random variable.

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

3. Bayesian Estimation

- **Bayesian Estimation**
  - The computation of posterior probabilities $P(\omega_i|x)$ lies at the heart of Bayesian classification.

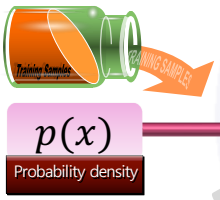
Prior information:
1. *The functional forms for unknown densities;*
2. *Ranges for the values of unknown parameters;*
3. *Training samples D.*

- **GOAL**: compute $P(\omega_i|x, D)$
- **GIVEN**:
  - the sample $\mathcal{D} = \mathcal{D}_1 \cup \mathcal{D}_2 \ldots \cup \mathcal{D}_c$,
  - $x \in \mathcal{D}_i$, $\mathcal{D}_i$ *has no influence on* $p\left(x \middle| \omega_j, \mathcal{D}_j\right)$ *if* $i \neq j$. $p(x|\omega_i, \mathcal{D}) = p\left(x|\omega_i, \mathcal{D}_i\right)$.
  - $P(\omega_i) = P(\omega_i|\mathcal{D})$
- Bayes formula can be written:

$$P(\omega_i|x, \mathcal{D}) = \frac{p(x|\omega_i, \mathcal{D}) \cdot P(\omega_i|\mathcal{D})}{\sum_{j=1}^{c} p\left(x|\omega_j, \mathcal{D}\right) \cdot P\left(\omega_j|\mathcal{D}\right)} \longrightarrow P(\omega_i|x, \mathcal{D}) = \frac{p\left(x|\omega_i, \mathcal{D}_i\right) \cdot P(\omega_i)}{\sum_{j=1}^{c} p\left(x \middle| \omega_j, \mathcal{D}_j\right) \cdot P\left(\omega_j\right)}$$

3. Bayesian Estimation

- **Bayesian Estimation**
  - $p(x)$ is unknown
    - We assume it has a known parametric form $p(x|\theta)$;
    - Value of parameter θ is unknown, $p(\theta) \rightarrow p(\theta|D)$.



Bayesian Estimation

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

– If $p(\theta|D)$ peaks very sharply about parameter $\hat{\theta}$ and $p(x|\theta)$ is smooth, then $p(x|D) \approx p(x|\hat{\theta})$.

3. Bayesian Estimation

- **Bayesian Parameter Estimation: Gaussian Case**

$$p(x|D) = \int p(x|\theta)p(\theta|D)d\theta$$

  - **GOAL:**
    - Calculate the a-posteriori density $p(\theta|D)$
    - Calculate the desired probability density $p(x|D)$.

  - Assume: $p(x|\mu) \sim N(\mu, \sigma^2)$, the only unknown parameter μ.
    - The univariate case: $p(\mu|D)$, $p(x|D)$
    - The multivariate case: $p(\mu|D)$, $p(x|D)$.

3. Bayesian Estimation

- **Bayesian Parameter Estimation: Gaussian Case**
  - The univariate case: $p(\mu|D)$
    - Known prior density $p(\mu) \sim N(\mu_0, {\sigma_0}^2)$
      - Assume μ_0 and σ_0 are known.
      - μ_0 represents our best a priori guess for μ.
      - ${\sigma_0}^2$ measures our uncertainty about this guess.

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

$D = \{x_1, x_2, \dots, x_n\}$

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

$p(x_k|\mu) \sim N(\mu, \sigma^2)$
$p(\mu) \sim N(\mu_0, {\sigma_0}^2)$

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

3. Bayesian Estimation

- **Bayesian Parameter Estimation: Gaussian Case**
  - The univariate case: $p(\mu|D)$

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

$$\frac{\mu_n}{\sigma_n^2} = \frac{n}{\sigma^2} x_n + \frac{\mu_0}{\sigma_0^2}$$

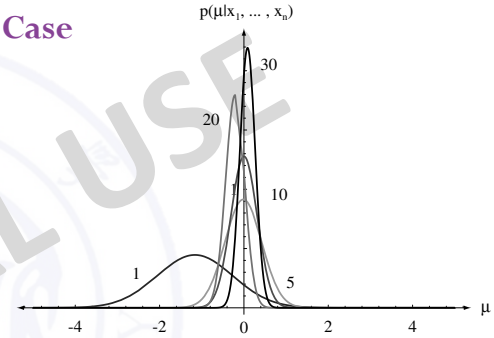
$$p(\mu|D) = \frac{1}{\sqrt{2\pi}\sigma_n} exp\left[-\frac{1}{2}\left(\frac{\mu-\mu_n}{\sigma_n}\right)^2\right]$$

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

$$\mu_n = \left(\frac{n\sigma_0^2}{n\sigma_0^2+\sigma^2}\right) x_n + \frac{\sigma^2}{n\sigma_0^2+\sigma^2}\mu_0$$

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

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



— μ_n represents our best guess for μ after observing n samples.

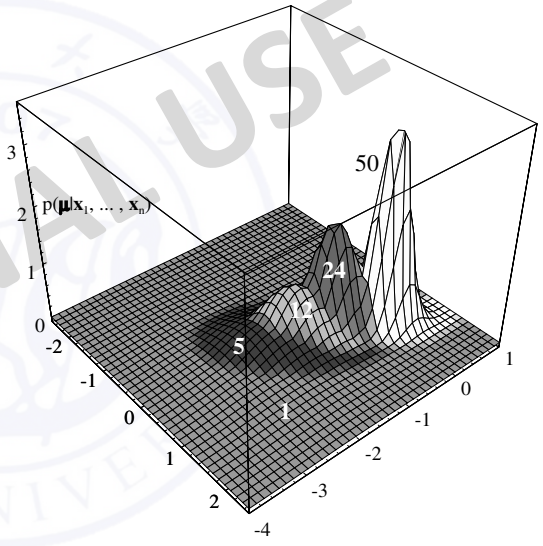
— ${\sigma_n}^2$ measures our uncertainty about this guess. ${\sigma_n}^2$ decreases monotonically as n increases. As the same time, $p(\mu|D)$ becomes more and more sharply peaked.

— This behavior is commonly known as ***Bayesian learning***.

3. Bayesian Estimation

- **Bayesian Parameter Estimation**: Gaussian Case
  - The 2D case: $p(\mu|D)$

— Bayesian learning of the mean of normal distributions in two dimensions. The posterior distribution estimates are labeled by the number of training samples used in the estimation.



3. Bayesian Estimation

- **Bayesian Parameter Estimation**: Gaussian Case
  - The univariate case: $p(x|D)$
    - $p(\mu|D)$ has been computed $p(\mu|D) \sim N(\mu_n, \sigma_n^2)$
    - $p(\mu|D) \rightarrow p(x|D)$

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

$$\mu_n = \left(\frac{n\sigma_0^2}{n\sigma_0^2+\sigma^2}\right)\hat{\mu}_n + \frac{\sigma^2}{n\sigma_0^2+\sigma^2}\mu_0$$
$$\sigma_n^2 = \frac{\sigma_0^2\sigma^2}{n\sigma_0^2+\sigma^2}$$
$$f(\sigma, \sigma_n) = \int \exp\left[-\frac{1}{2}\frac{\sigma^2+\sigma_n^2}{\sigma^2\sigma_n^2}\left(\mu - \frac{\sigma_n^2 x + \sigma^2\mu_n}{\sigma^2+\sigma_n^2}\right)^2\right] d\mu$$

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

— The density $p(x|D)$ is the desired class-conditional density $p(x|D_j, \omega_j)$, and together with the prior probabilities $P(\omega_j)$ it gives us the probabilistic information needed to design the classifier.

3. Bayesian Estimation

- Bayesian Parameter Estimation: Gaussian Case
  - The multivariate case: $p(x|D) \sim N(\mu, \Sigma)$
    - μ is the only unknown parameter
    - Σ is known
    - Known prior density $p(\mu) \sim N(\mu_0, {\Sigma_0}^2)$
      - μ_0 and Σ_0 are known!

  - $p(\mu|D)$
  - $p(x|D)$

3. Bayesian Estimation

- Bayesian Parameter Estimation: Gaussian Case
  - The multivariate case: $p(\mu|D)$

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

$D = \{x_1, x_2, \dots, x_n\}$

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

$$\Sigma_n^{-1} = n\Sigma^{-1} + \Sigma_0^{-1}$$

$$\Sigma_n^{-1}\mu_n = n\Sigma^{-1}\hat{\mu}_n + \Sigma_0^{-1}\mu_0$$

$$p(\mu|\mathcal{D}) = \alpha'' \exp\left[-\frac{1}{2}(\mu-\mu_n)^t \Sigma_n^{-1}(\mu-\mu_n)\right]$$

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

$$(A^{-1}+B^{-1})^{-1} = A(A+B)^{-1}B = B(A+B)^{-1}A$$

$$\mu_n = \Sigma_0\left(\Sigma_0 + \frac{1}{n}\Sigma\right)^{-1}\hat{\mu}_n + \frac{1}{n}\Sigma\left(\Sigma_0 + \frac{1}{n}\Sigma\right)^{-1}\mu_0$$

$$\Sigma_n = \Sigma_0\left(\Sigma_0 + \frac{1}{n}\Sigma\right)^{-1}\frac{1}{n}\Sigma$$

$$p(\mu|D) \sim N(\mu_n, \Sigma_n)$$

$p(x|D)$

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

$$\mu_n = \Sigma_0\left(\Sigma_0 + \frac{1}{n}\Sigma\right)^{-1}\hat{\mu}_n + \frac{1}{n}\Sigma\left(\Sigma_0 + \frac{1}{n}\Sigma\right)^{-1}\mu_0$$

$$\Sigma_n = \Sigma_0\left(\Sigma_0 + \frac{1}{n}\Sigma\right)^{-1}\frac{1}{n}\Sigma$$

$$p(x|D) \sim N(\mu_n, \Sigma + \Sigma_n)$$

3. Bayesian Estimation

- **Bayesian Parameter Estimation**:

General Theory

- This approach can be generalized to apply to any situation in which the unknown density can be parameterized. The basic **assumptions** are summarized as follows:
  - The form of the density $p(x|\theta)$ is assumed to be known, but the value of the parameter vector θ is not known exactly.
  - Our initial knowledge about θ is assumed to be contained in a known priori density $p(\theta)$.
  - The rest of our knowledge about θ is contained in a set D of n samples $x_1, x_2 \dots x_n$ drawn independently according to the unknown probability density $p(x)$.

$$p(x|D) = \int p(x|\theta)\, p(\theta|D)d\theta \approx p(x|\hat{\theta})$$

—If $p(\theta|D)$ peaks very sharply about parameter $\hat{\theta}$ and $p(x|\theta)$ is smooth, then $p(x|D) \approx p(x|\hat{\theta})$.

$$p(\theta|D) \rightarrow p(x|D)$$

① $$p(\theta|\mathcal{D}) = \frac{p(\mathcal{D}|\theta)p(\theta)}{\int p(\mathcal{D}|\theta)p(\theta)d\theta}$$

$$\mathcal{D}^n = \{x_1, \dots, x_n\}$$

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

② $$p(x|D) = \int p(x|\theta)\, p(\theta|D)d\theta$$

3. Bayesian Estimation

- **Recursive Bayes approach**

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

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

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

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

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

Incremental learning:

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

$$p(\theta|x_1) \xrightarrow{D^1 = \{x_1\}} p(\theta|D^1) \propto p(x_1|\theta)p(\theta|D^0)$$

$$\vdots \qquad\qquad \vdots$$

$$p(\theta|x_1, x_1 \dots x_n) \xrightarrow[D^n = \{x_1, x_2, \dots x_n\}]{} p(\theta|D^n) \propto p(x_n|\theta)p(\theta|D^{n-1})$$

3. Bayesian Estimation

- **Recursive Bayes Learning**

Example

Suppose we believe our one dimensional samples come from a uniform distribution

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

But initially we know only that our parameter is bounded $0 < \theta \le 10$ (uniform distribution). Use recursive Bayes methods to estimate θ and the underlying densities from the data $D = \{4,7,2,8\}$.

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

$x_1 = 4$ ..... $p(\theta|D^1) \propto p(x_1|\theta)p(\theta|D^0) = \begin{cases} 1/\theta & 4 \le \theta \le 10 \\ 0 & otherwise \end{cases}$

$x_2 = 7$ ..... $p(\theta|D^2) \propto p(x_2|\theta)p(\theta|D^1) = \begin{cases} 1/\theta^2 & 7 \le \theta \le 10 \\ 0 & otherwise \end{cases}$

... ...

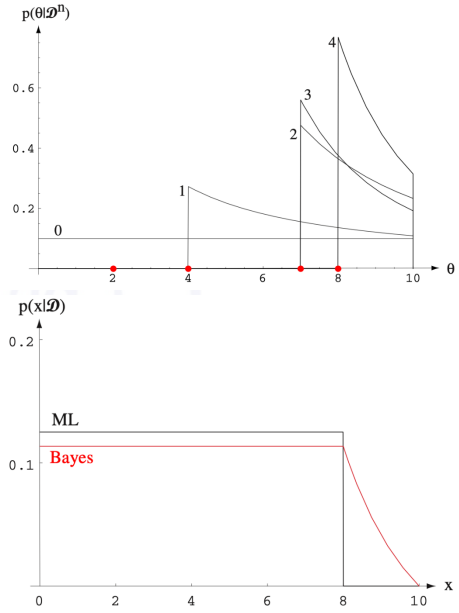
$$p(\theta|D^n) \propto \frac{1}{\theta^n}, \qquad \max_x[D^n] \le \theta \le 10$$

3. Bayesian Estimation

- **Recursive Bayes Learning**

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

The posterior $p(\theta|\mathcal{D}^n)$ for the model and n points in the data set in this example. For $n = 0$, the posterior starts our as a flat, uniform density from 0 to 10, denote $p(\theta) \sim U(1,10)$. As more points are incorporated, it becomes increasingly peaked at the value of the highest data point. $D = \{4,7,2,8\}$



— Given full data set, the maximum-likelihood solution here is clearly $\hat{\theta} = 8$, and this implies a uniform $p(x|D) \sim U(0,8)$.

— According to Bayesian methodology, which requires the integration, the density is uniform up to $x = 8$, but has a tail at higher values—an indication that the influence of our prior $p(\theta)$ has not yet been swamped by the information in the training data. $p(x|D) = \int p(x|\theta)\, p(\theta|D) d\theta$

3. Bayesian Estimation

Maximum-Likelihood Estimation and Bayesian Estimation		
	MLE	**BE**
Method	Maximum likelihood approach estimates a point in θ space	Bayesian approach estimates a distribution
Computational complexity	Simpler	Bayesian method has strong theoretical and methodological arguments supporting it.
results		Better
	When used for classifiers, they mostly give same result.	

Chapter 3

PART I