Review of Estimation Theory

Berlin 2003

References:

1. X. Huang et. al., Spoken Language Processing, Chapter 3

Introduction

- Estimation theory is the most important theory and method in *statistical inference*
- Statistical inference
  - Data generated in accordance with some unknown probability distribution must be analyzed
  - Some type of inference about the unknown distribution must be made like the characteristics (parameters) of the distribution generating the experimental data, the mean and variance etc.

The vector of random variables

$$\boldsymbol{X} = \{X_1, X_2, \ldots, X_n\} \xrightarrow[\text{estimator}]{\theta(\boldsymbol{X})} g(\boldsymbol{X} | \boldsymbol{\Phi})$$

The vector of sample values

$$\boldsymbol{x} = \{x_1, x_2, \ldots, x_n\} \xrightarrow[\text{estimate}]{\theta(\boldsymbol{x})} g(\boldsymbol{x} | \boldsymbol{\Phi})$$

$\boldsymbol{\Phi}$:the parameters of the distribution

Introduction

- Three common estimators (estimation methods)
  - Minimum mean square estimator
    - Estimate the random variable itself
    - Function approximation, curve fitting, …
  - Maximum likelihood estimator
    - Estimate the parameters of the distribution of the random variables
  - Bayes’ estimator
    - Estimate the parameters of the distribution of the random variables

Minimum Mean Square Error Estimation and Least Square Error Estimation

- There are two random variables X and Y. When observing the value of X, we want to find a transform $\hat{Y}=g(X,\boldsymbol{\Phi})$ ($\boldsymbol{\Phi}$ the parameter vectors of function g) to predict the value of Y
  - **Minimum Mean Square Error Estimation**

$$\boldsymbol{\Phi}_{MMSE} = \arg\min_{\boldsymbol{\Phi}} \left[E\left[\left(Y - g\left(X, \boldsymbol{\Phi}\right)\right)^2\right]\right]$$

If the joint distribution $f_{X,Y}(X,Y)$ Is known

  - **Least Square Error Estimation**

$$\boldsymbol{\Phi}_{LSE} = \arg\min_{\boldsymbol{\Phi}} \sum_{i=1}^{n} \left[y_i - g\left(x_i, \boldsymbol{\Phi}\right)\right]^2$$

When samples of (x_i, y_i) pairs are observed

- Base on the law of large numbers, when the joint probability $\hat{Y}=f_{X,Y}(X,Y)$ is uniform or the number of samples approaches to infinity, MMSE and LSE are equivalent

Minimum Mean Square Error Estimation and Least Square Error Estimation

- Constant functions $g(X) = c$
  - **MMSE**

$$\nabla_c E\left[(Y - c)^2\right] = 0$$

$$\therefore c_{MMSE} = E[Y]$$

mean

  -- **LSE**

$$\nabla_c \sum_{i=1}^{n} (y_i - c)^2 = 0$$

$$\therefore c_{LSE} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

sample mean

- Linear functions $g(X) = aX + b$
  - **MMSE**

$$\nabla_a E\left[(Y - (aX + b))^2\right] = 0 \quad aE\left[X^2\right] + bE[X] - E[XY] = 0$$

$$\nabla_b E\left[(Y - (aX + b))^2\right] = 0 \quad aE[X] + b - E[Y] = 0$$

$$a = \frac{\operatorname{cov}(X, Y)}{Var(X)}$$

$$b = E[Y] - \rho_{XY} \frac{\sigma_Y}{\sigma_X} E[X]$$

Minimum Mean Square Error Estimation and Least Square Error Estimation

- Linear functions
  - **LSE**
    - Suppose that ***x*** are d-dimensional vectors and y are scalars

$$\hat{\boldsymbol{Y}} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \boldsymbol{XA} = \begin{array}{c} \begin{matrix} \boldsymbol{c}_0 & \boldsymbol{c}_1 & & \boldsymbol{c}_d \end{matrix} \\ \begin{bmatrix} 1 & x_1^1 & \cdots & x_1^d \\ 1 & x_2^1 & \cdots & x_2^d \\ \vdots & \vdots & & \vdots \\ 1 & x_n^1 & \cdots & x_n^d \end{bmatrix} \end{array} \begin{bmatrix} a_0 \\ a_1 \\ \vdots \\ a_n \end{bmatrix}$$

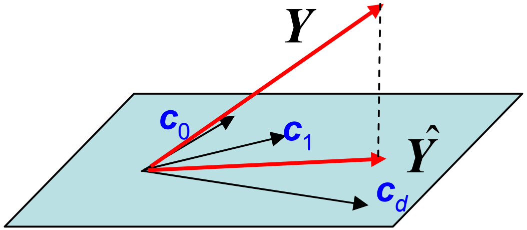
$$e(\boldsymbol{A}) = \left\| \hat{\boldsymbol{Y}} - \boldsymbol{Y} \right\| = \sum_{i=1}^{n} \left(\boldsymbol{A}^t \boldsymbol{x}_i - y_1 \right)^2$$

$$\nabla e(\boldsymbol{A}) = \sum_{i=1}^{n} 2\left(\boldsymbol{A}^t \boldsymbol{x}_i - y_i \right)\boldsymbol{x}_i = 2\boldsymbol{X}^t (\boldsymbol{XA} - \boldsymbol{Y}) = 0$$

$$\Rightarrow \boldsymbol{X}^t \boldsymbol{XA} = \boldsymbol{X}^t \boldsymbol{Y}$$

$$\Rightarrow \boldsymbol{A} = \left(\boldsymbol{X}^t \boldsymbol{X} \right)^{-1} \boldsymbol{X}^t \boldsymbol{Y}$$

.....



Maximum Likelihood Estimation (MLE/ML)

- ML is the most widely used parametric estimation method
- A set of random samples $\boldsymbol{X}=\{X_1,X_2,\dots,X_n\}$ is to be drawn independently according to a distribution with the pdf $p(x|\boldsymbol{\Phi})$
  - Given a sequence of random samples $\boldsymbol{x}=(x_1,x_2,\dots,x_n)$ the likelihood of it is defined as $p_n(\boldsymbol{x}|\boldsymbol{\Phi})$, a joint pdf of $(x_1,x_2,\dots,x_n)$

    $$p_n(\boldsymbol{x}|\boldsymbol{\Phi})=\prod_{k=1}^{n}p(x_k|\boldsymbol{\Phi}),\ \because X_1, X_2, \dots X_n \text{ are iid}$$

  - Maximum likelihood estimator of $\boldsymbol{\Phi}$ is denoted as

    $$\boldsymbol{\Phi}_{ML}=\arg\max_{\boldsymbol{\Phi}} p_n(\boldsymbol{x}|\boldsymbol{\Phi})=\arg\max_{\boldsymbol{\Phi}}\prod_{k=1}^{n}p(x_k|\boldsymbol{\Phi})$$

  - Since the logarithm function is *monotonically increasing function*, the parameter set $\boldsymbol{\Phi}_{ML}$ that maximizes the log-likelihood should also maximize the likelihood. The log-likelihood can be expressed as: $l(\boldsymbol{\Phi})=log\ \ p_n(\boldsymbol{x}|\boldsymbol{\Phi})=\sum_{k=1}^{n}\log p(x_k|\boldsymbol{\Phi})$

Maximum Likelihood Estimation (MLE/ML)

- If $p_n(\boldsymbol{x}|\boldsymbol{\Phi})$ is differentiable function of $\boldsymbol{\Phi}$, $\boldsymbol{\Phi}_{ML}$ can be attained by taking the partial derivative with respect to $\boldsymbol{\Phi}$ and setting it to zero
  - Let $\boldsymbol{\Phi}$ be a M-component parameter vector $\boldsymbol{\Phi} = (\Phi_1, \Phi_2, ..., \Phi_M)^t$

$$\nabla_{\boldsymbol{\Phi}} l(\boldsymbol{\Phi}) = \nabla_{\boldsymbol{\Phi}} \sum_{k=1}^{n} log\ p(x_k|\boldsymbol{\Phi}) = \begin{bmatrix} \frac{\partial l(\boldsymbol{\Phi})}{\partial \Phi_1} \\ . \\ . \\ \frac{\partial l(\boldsymbol{\Phi})}{\partial \Phi_M} \end{bmatrix} = 0$$

- Example: $p(x|\boldsymbol{\Phi})$ is a univariate Gaussian pdf with the parameter set (μ, σ^2)

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

$$log\ p_n(\boldsymbol{x}|\boldsymbol{\Phi}) = \sum_{k=1}^{n} log\ p(x_k|\boldsymbol{\Phi}) = \sum_{k=1}^{n} log\left(\frac{1}{\sqrt{2\pi}\,\sigma} exp\left[-\frac{(x_k-\mu)^2}{2\sigma^2}\right]\right) = -\frac{n}{2} log\left(2\pi\sigma^2\right) - \frac{1}{2\sigma^2}\sum_{k=1}^{n}(x_k-\mu)^2$$

Maximum Likelihood Estimation (MLE/ML)

- Example: univariate Gaussian pdf (cont.)
  - Take the partial derivatives of the above expression and set them to zero

$\boldsymbol{\Phi}$ itself is fixed but unkown

$$\frac{\partial}{\partial \mu} \log p_n(\boldsymbol{x}|\boldsymbol{\Phi}) = \frac{1}{\sigma^2} \sum_{k=1}^{n} (x_k - \mu) = 0$$

$$\frac{\partial}{\partial \sigma^2} \log p_n(\boldsymbol{x}|\boldsymbol{\Phi}) = -\frac{n}{\sigma^2} + \sum_{k=1}^{n} \frac{(x_k - \mu)^2}{\sigma^4} = 0$$

  - The maximum likelihood estimates for μ and σ^2 are

$$\mu_{ML} = \sum_{k=1}^{n} x_k = E(x)$$

$$\sigma^2_{ML} = \frac{1}{n}(x_k - \mu_{ML})^2 = E\left[(x_k - \mu_{ML})^2\right]$$

  - The maximum likelihood estimation for mean and variance is just **the sample mean and variance**

Maximum Likelihood Estimation (MLE/ML)

- Example: multivariate Gaussian pdf (cont.)

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

  - The maximum likelihood estimates for $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are

$$\hat{\boldsymbol{\mu}}_{MLE}=\frac{1}{n}\sum_{k=1}^{n}\boldsymbol{x}_{k}$$

$$\begin{aligned}\hat{\boldsymbol{\Sigma}}_{MLE}&=\frac{1}{n}\sum_{k=1}^{n}\left(\boldsymbol{x}_{k}-\hat{\boldsymbol{\mu}}_{MLE}\right)\left(\boldsymbol{x}_{k}-\hat{\boldsymbol{\mu}}_{MLE}\right)^{t}\\&=E\left[\left(\boldsymbol{x}_{k}-\hat{\boldsymbol{\mu}}_{MLE}\right)\left(\boldsymbol{x}_{k}-\hat{\boldsymbol{\mu}}_{MLE}\right)^{t}\right]\end{aligned}$$

    - The maximum likelihood estimation for mean vector and variance matrix is just **the sample mean vector and variance matrix**

- **In fact, $\boldsymbol{\Phi}_{MLE}$ itself is also a Gaussian distribution**

Bayesian Estimation

- Bayesian estimation has a different philosophy than maximum likelihood (ML) estimation
  - ML assumes the parameter set $\boldsymbol{\Phi}$ is fixed but unknown (non-informative, uniform prior)
  - Bayesian estimation assumes the parameter set $\boldsymbol{\Phi}$ itself is a random variable with a prior distribution $p(\boldsymbol{\Phi})$
  - Given a sequence of random samples $\boldsymbol{x}=(x_1,x_2,\ldots,x_n)$, which are i.i.d. with a joint pdf $p(\boldsymbol{x}|\boldsymbol{\Phi})$, the posterior distribution of $\boldsymbol{\Phi}$ can be the following according to the Bayes' rule

$$p(\boldsymbol{\Phi}|\boldsymbol{x})=\frac{p(\boldsymbol{x}|\boldsymbol{\Phi})p(\boldsymbol{\Phi})}{p(\boldsymbol{x})}\propto p(\boldsymbol{x}|\boldsymbol{\Phi})p(\boldsymbol{\Phi})$$

Bayesian Estimation

- $p(\boldsymbol{\Phi}|\boldsymbol{x})$: the posterior probability, the distribution of $\boldsymbol{\Phi}$ after we observed the values of random variables
- $p(\boldsymbol{\Phi})$: a conjugate prior of the random variables (or vector) is defined as the prior distribution for the parameters of the density function (e.g. $\boldsymbol{\Phi}$) of the random variables (or vectors)
  - Before we observed the values of random variables

- The joint pdf/likelihood function

$$p(\boldsymbol{x}|\boldsymbol{\Phi}) = \frac{1}{(2\pi)^{n/2}\sigma^{n}} \exp\left[-\frac{1}{2}\sum_{i=1}^{n}\left(\frac{x_i-\Phi}{\sigma}\right)^2\right] \propto \exp\left[-\frac{1}{2}\sum_{i=1}^{n}\left(\frac{x_i-\Phi}{\sigma}\right)^2\right]$$

- The prior is also a Gaussian distribution

$$p(\boldsymbol{\Phi}) = \frac{1}{(2\pi)^{1/2}\nu} \exp\left[-\frac{1}{2}\left(\frac{\Phi-\mu}{\nu}\right)^2\right] \propto \exp\left[-\frac{1}{2}\left(\frac{\Phi-\mu}{\nu}\right)^2\right]$$

Maximum a Posterior Probability (MAP)

- The MAP chooses a estimate $\boldsymbol{\Phi}_{MAP}$ that maximizes the posterior probability $p(\boldsymbol{\Phi}|\boldsymbol{x})$ is the most common Bayesian estimator

$$\boldsymbol{\Phi}_{MAP} = \arg\max_{\boldsymbol{\Phi}} \; p(\boldsymbol{\Phi}|\boldsymbol{x}) = \arg\max_{\boldsymbol{\Phi}} \; p(\boldsymbol{x}|\boldsymbol{\Phi})p(\boldsymbol{\Phi})$$

$$\boldsymbol{\Phi}_{MAP} = \arg\max_{\boldsymbol{\Phi}} \left[log \;\; p(\boldsymbol{x}|\boldsymbol{\Phi}) + log \;\; p(\boldsymbol{\Phi})\right]$$

$$\frac{\partial \; log \;\; p(\boldsymbol{x}|\boldsymbol{\Phi})}{\partial \boldsymbol{\Phi}} + \frac{log \;\; p(\boldsymbol{\Phi})}{\partial \boldsymbol{\Phi}} = 0$$

  - For example, the conjugate prior for the mean of a Gaussian pdf is also a Gaussian pdf
    - Supposed in previous example, $\boldsymbol{X} = \{X_1, X_2, \ldots, X_n\}$ is drawn from a Gaussian which mean Φ is unknown and variance σ^2 is known, while the conjugate prior (is a Gaussian) with mean μ and variance ν^2
    - The MAP estimated Φ is:

$$\Phi_{MAP} = \frac{\sigma^2 \mu + n\nu^2 \bar{x}_n}{\sigma^2 + n\nu^2}, \; n \text{ is no. of training samples, } \bar{x}_n \text{ is the sample mean}$$

Bayes' Decision Theory

- A decision-making based on both the posterior knowledge obtained from specific observation data and prior knowledge of the categories
  - Prior class probabilities $P(\omega_i), \ \forall \text{ class } i$
  - Class-conditioned probabilities (likelihoods) $P(x|\omega_i), \ \forall \text{ class } i$

$$k = \arg\max_i P(\omega_i|x) = \arg\max_i \frac{P(x|\omega_i)P(\omega_i)}{P(x)} = \arg\max_i \frac{P(x|\omega_i)P(\omega_i)}{\sum_{j=1} P(x|\omega_j)P(\omega_j)}$$

$$\therefore k = \arg\max_i P(x|\omega_i)P(\omega_i)$$

Bayes' Decision Theory

- Bayes' decision rule designed to minimize the overall risk involved in making decision
  - The **expected loss** (conditional risk) when making decision δ_i

$$R(\delta_i|x) = \sum_j l(\delta_i|\omega_j, x)P(\omega_j|x), \quad \text{where} \quad l(\delta_i|\omega_j, x) = \begin{cases} 0, & i = j \\ 1, & i \neq j \end{cases}$$

a decision

The class x might belong to

loss function

  - The overall risk (Bayes' risk)

$$R = \int_{-\infty}^{\infty} R(\delta(x)|x)p(x)dx, \ \delta(x): \text{the selected decision for a sample } x$$

Bayes' Decision Theory

- Minimize the overall risk (classification error) by computing the conditional risks and select the decision δ_i for which the conditional risk $R(\delta_i|x)$ is minimum, i.e., $P(\omega_i|x)$ is maximum

$$R(\delta_i|x) = \sum_j l(\delta_i|\omega_j, x)P(\omega_j|x) = \sum_{j \neq i} P(\omega_j|x)$$

$$= \sum_j P(\omega_j|x) - P(\omega_i|x)$$

$$= 1 - P(\omega_i|x)$$

the decision should be made

$$\delta(x) = \arg\max_i P(\omega_i|x) = \arg\max_i P(x|\omega_i)P(\omega_j)$$

  - Called the **minimum-error-rate decision rule** which minimizes the classification error rate

Bayes' Decision Theory

- Two-class pattern classification

$$d_1(x) = P(\omega_1|x) \cong P(x|\omega_1)P(\omega_1), \; d_2(x) = P(\omega_2|x) \cong P(x|\omega_2)P(\omega_2)$$

Bayes' Classifier

$$\underbrace{P(x|\omega_1)P(\omega_1)}_{P(\omega_2|x)} \underset{\omega_2}{\overset{\omega_1}{\gtrless}} \underbrace{P(x|\omega_2)P(\omega_2)}_{P(\omega_2|x)} \quad \Longleftrightarrow$$

Likelihood ratio or log-likelihood ratio:

$$l(x) = \frac{P(x|\omega_1)}{P(x|\omega_2)} \underset{\omega_2}{\overset{\omega_1}{\gtrless}} \frac{P(\omega_2)}{P(\omega_1)}$$

$$\log l(x) = \log P(x|\omega_1) - \log P(x|\omega_2) \underset{\omega_2}{\overset{\omega_1}{\gtrless}} \log P(\omega_2) - \log P(\omega_1)$$

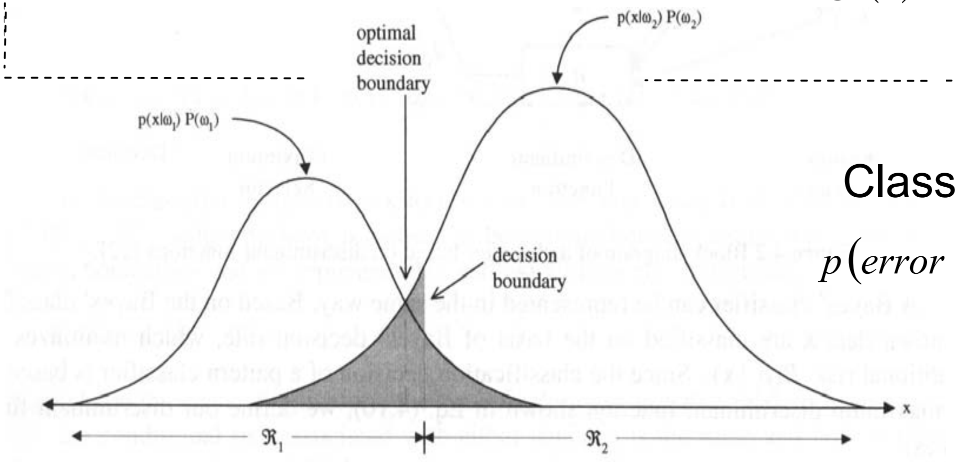


Figure 4.1 Calculation of the likelihood of classification error [22]. The shaded area represents the integral value in Eq. (4.9).

Classification error:

X falls in R_2, but the true class is ω_1

$$p(error) = P(x \in R_1, \omega_2) + P(x \in R_2, \omega_1)$$
$$= P(x \in R_1|\omega_2)P(\omega_2) + P(x \in R_2|\omega_1)P(\omega_1)$$
$$= \int_{R_1} P(x|\omega_2)P(\omega_2)dx + \int_{R_2} P(x|\omega_1)P(\omega_1)dx$$