Lecture 7

*) LDA, QDA

→ $\Sigma_i = \sigma^2 I$, $\Sigma_i = \Sigma$, Σ_i's are not the same

*) Bernoulli pmf

*) Error bounds

*) MLE

Chernoff bound

$$\min(a, b) \leq a^{\beta} b^{1-\beta} \quad \text{for } a, b \geq 0 \text{ and } 0 \leq \beta \leq 1$$

Often it is difficult to find P(error) as it needs decision boundaries.
In high dimensions it is very hard to find decision boundaries.

Thus instead of exact P(error), we seek a bound on this error.

$$P(\text{error}) = \int \min\left[P(\omega_1|x), P(\omega_2|x)\right] p(x)\, dx$$

$$\leq \text{Upper bound}$$

W.L.O.G. $a > b$ | $b \leq \left(\frac{a}{b}\right)^{\beta} \cdot b$

Applications: https://en.wikipedia.org/wiki/Chernoff_bound

$$P(\text{error}) \le \int_{-\infty}^{\infty} P(\omega_1|x)^{\beta} \, P(\omega_2|x)^{1-\beta} \, p(x)\, dx$$

$$= \int P(\omega_1)^{\beta} \, P(\omega_2)^{1-\beta} \cdot p(x|\omega_1)^{\beta} \, p(x|\omega_2)^{1-\beta} \, dx$$

$$= P(\omega_1)^{\beta} \, P(\omega_2)^{1-\beta} \underbrace{\int_{-\infty}^{\infty} p(x|\omega_1)^{\beta} \, p(x|\omega_2)^{1-\beta} \, dx}_{e^{-K(\beta)}}$$

$$p(x|\omega_1) \sim N(\mu_1, \Sigma_1)$$

$$p(x|\omega_2) \sim N(\mu_2, \Sigma_2)$$

$$P(\text{error}) = P(\omega_1)^{\beta} \, P(\omega_2)^{1-\beta} \, e^{-K(\beta)}$$

$$\int p(x)\,dx = f(x)$$

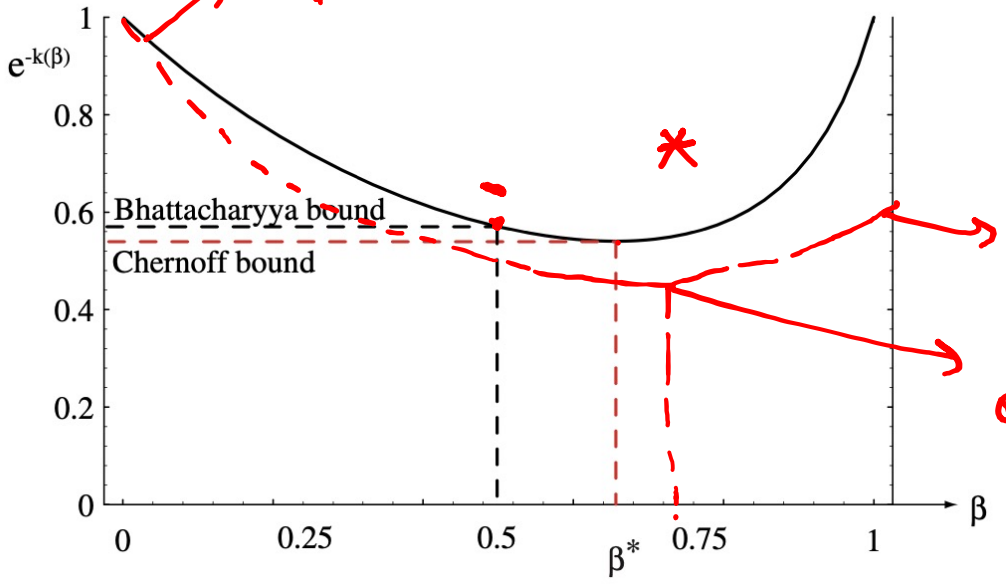
Bhattacharya bound, $\beta = 1/2$

$$P(\text{error}) \leq \int\int \sqrt{P(\omega_1)\, P(\omega_2)}\; e^{-k\left(\frac{1}{2}\right)}$$

$$\leq \sqrt{P(\omega_1) P(\omega_2)}\; e^{-\left\{\frac{1}{8} \cdot \delta\right\}}$$

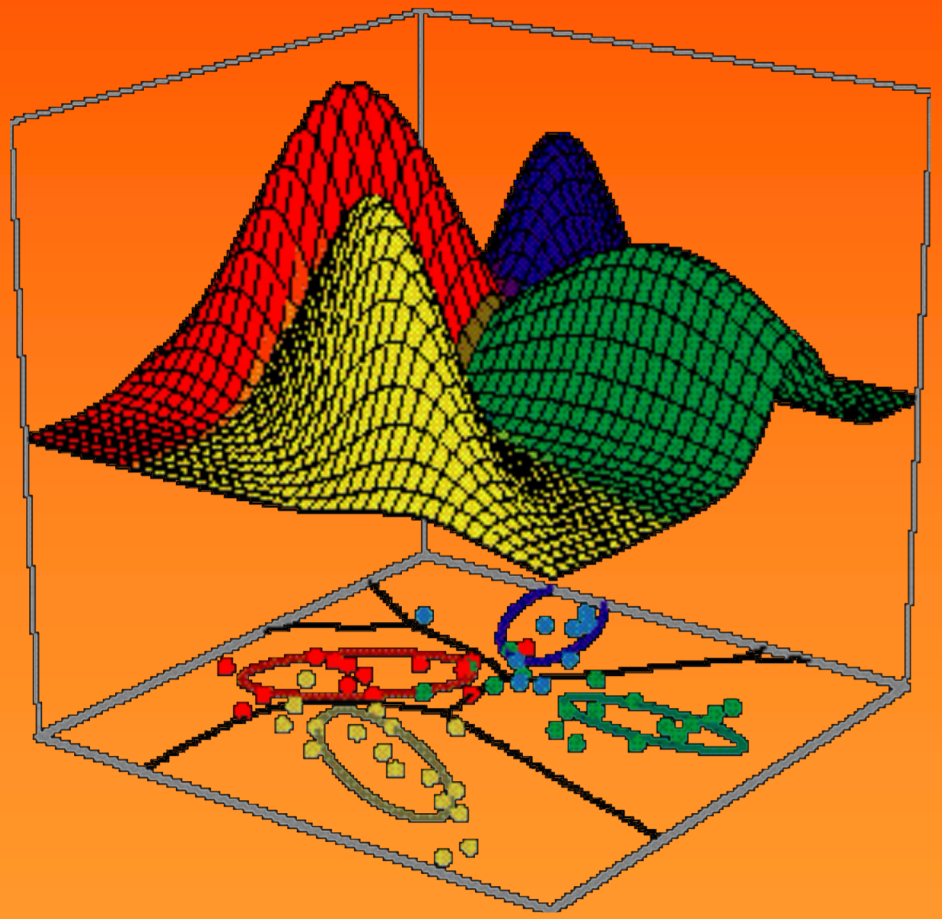
$$P(\text{error}) \leq \frac{1}{e} \sqrt{P(\omega_1)\, P(\omega_2)}$$

Chernoff bound



Bhattacharya bound named after Anil Kumar Bhattacharya

$$k(\beta) = \frac{\beta(1-\beta)}{2}(\boldsymbol{\mu}_2-\boldsymbol{\mu}_1)^t[\beta\boldsymbol{\Sigma}_1+(1-\beta)\boldsymbol{\Sigma}_2]^{-1}(\boldsymbol{\mu}_2-\boldsymbol{\mu}_1) + \frac{1}{2}\ln\frac{|\beta\boldsymbol{\Sigma}_1+(1-\beta)\boldsymbol{\Sigma}_2|}{|\boldsymbol{\Sigma}_1|^{\beta}|\boldsymbol{\Sigma}_2|^{1-\beta}}.$$



Pattern Classification

Chapter 3: Maximum-Likelihood & Bayesian Parameter Estimation (part 1)

- Introduction
- Maximum-Likelihood Estimation
  - Example of a Specific Case
  - The Gaussian Case: unknown μ and σ
  - Bias

Ronald Aylmer Fisher (1890~1962)

- **Known for :**
  - 1912 : Maximum likelihood
  - 1922 : F-test
  - 1925 : Analysis of variance (Statistical Method for Research Workers)
- **Notable Prizes :**
  - Royal Medal (1938)
  - Copley Medal (1955)

- Introduction
  - Data availability in a Bayesian framework
    - We could design an optimal classifier if we knew:
      - $P(\omega_i)$ (priors)
      - $P(x \mid \omega_i)$ (class-conditional densities)

      Unfortunately, we rarely have this complete information!

  - Design a classifier from a training sample
    - No problem with prior estimation
    - Samples are often too small for class-conditional estimation (large dimension of feature space!)

- A priori information about the problem
- Normality of $P(x \mid \omega_i)$

$$P(x \mid \omega_i) \sim N(\mu_i, \Sigma_i)$$

  - Characterized by mean and cov parameters

- Estimation techniques

  - Maximum-Likelihood (ML) and the Bayesian estimations
  - Results are nearly identical, but the approaches are different

- Parameters in ML estimation are fixed but unknown!
- Best parameters are obtained by maximizing the probability of obtaining the samples observed
- Bayesian methods view the parameters as random variables having some known distribution
- In either approach, we use $P(\omega_i \mid x)$ for our classification rule!

Maximum-Likelihood Estimation

- Has good convergence properties as the sample size increases
- Simpler than any other alternative techniques

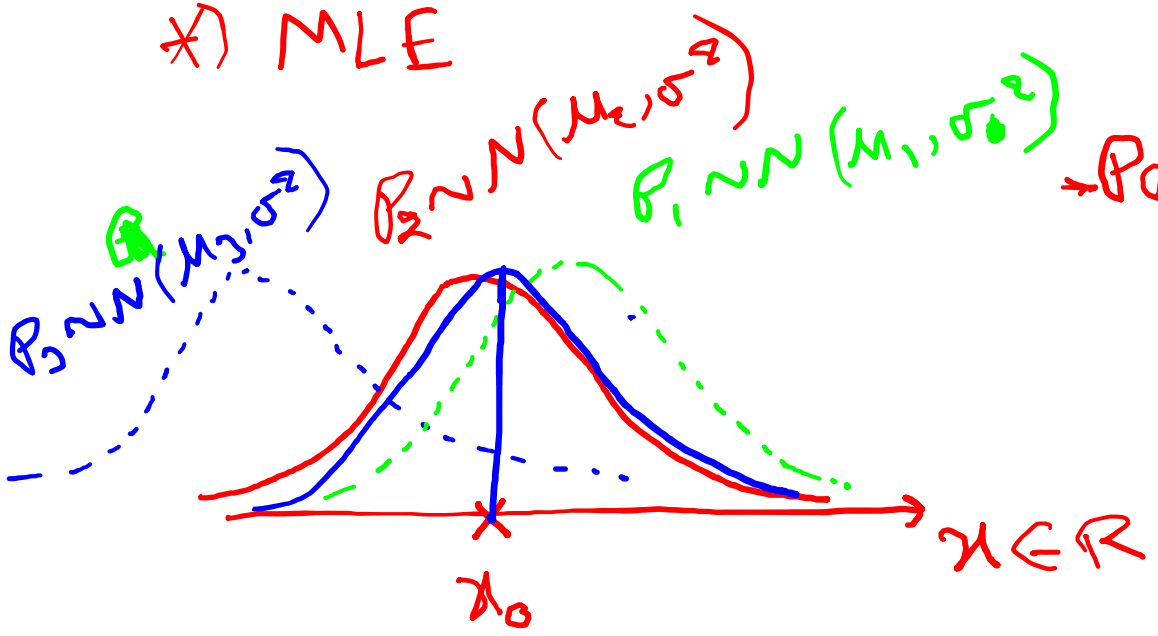
- General principle

- Assume we have c classes and

$$\theta = (\mu_j, \Sigma_j) = (\mu_j^1, \mu_j^2, \ldots, \sigma_j^{11}, \sigma_j^{22}, \mathbf{cov}(x_j^m, x_j^n)\ldots)$$

Total no of parameters to be computed?

*) MLE



→ Pdf from which x_0 was drawn

Parameters of Pdf

$p_2(x_0) > p_1(x_0)$

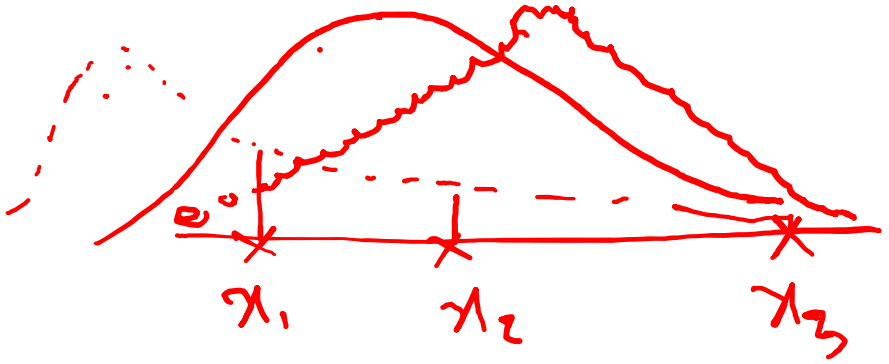
$p_2(x_0) > p_3(x_0)$

When you do not anything ?

→ 1st step.

→ Assumption. Gaussian.

→ $\mu = x_0$, $\sigma^2 \rightarrow$ X



σ is known.

$$p(x_1 | \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e\left\{-\frac{1}{2}\frac{(x_1-\mu)^2}{\sigma^2}\right\}$$

$\mu = ?$

$$p(x_2 | \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \left\{e^{-\frac{1}{2}\frac{(x_i-\mu)^2}{\sigma^2}}\right\}$$

① Joint prob. of obs. x_1, x_2, x_3 from parameters μ, σ^2

MLE Demo

x_1, x_2, x_3 are iids. → assumptions.

$$p(x_1, x_2, x_3 \mid \mu, \sigma^2) = \prod_{i=1}^{3} p(x_i \mid \mu, \sigma^2)$$

$$p(x_1, x_2 \cdots x_n \mid \mu, \sigma^2) = \prod_{i=1}^{n} p(x_i \mid \mu, \sigma^2)$$

likelihood w.r.t. samples.

Which Gaussian is most likely to have generated $x_1, x_2, \cdots x_n$

$$p(X|\theta) = \prod_{i=1}^{n} p(x_i|\mu, \sigma^2)$$

$$X = \{x_1, x_2, \cdots x_n\}$$

$\theta \rightarrow$ parameters, $\theta = \mu$,

$$\ln p(X|\theta) = \sum_{i=1}^{n} \ln p(x_i|\mu, \sigma^2)$$

$$= \sum_{i=1}^{n} \ln \frac{1}{\sqrt{2\pi\sigma^2}} e\left\{-\frac{1}{2}(x_i-\mu)^2/\sigma^2\right\}$$

$$= \sum_{i=1}^{n} \ln\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right) - \frac{1}{2}\left\{\frac{(x_i-\mu)^2}{\sigma^2}\right\}$$

We want to maximize log-likelihood to obtain θ, as here $\theta = \mu$

$$\frac{\partial}{\partial \theta} \ln p(X|\theta) = -\frac{1}{2\sigma^2} \sum_i -2(x_i - \mu) = 0$$

$$\mu = \frac{1}{n} \sum_{i=1}^{n} x_i$$

$$f(\theta) = \ln p(X|\theta)$$

$$\theta = \{\theta_1, \theta_2, \theta_3 \cdots \theta_m\}$$

$$\nabla_\theta f(\theta) = 0$$

We want

$\theta = \mu, \sigma^2$

$$\frac{\partial}{\partial \mu} F(\theta) = 0 \Rightarrow \mu = \frac{1}{n} \sum_{i=1}^{n} x_i$$

$$\frac{\partial}{\partial \sigma^2} F(\theta) = 0 \Rightarrow \frac{1}{n} \sum_{i=1}^{n} (x_i - \mu)^2$$

$$\mu_{ML} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

$$\sigma^2_{ML} = \frac{1}{n} \sum_{i=1}^{n} (x_i - \mu_{ML})^2$$

ML Estimation

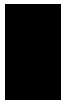
- Use the information
  provided by the training samples to estimate
  $\theta = (\theta_1, \theta_2, \ldots, \theta_c)$, each θ_i (i = 1, 2, …, c) is associated with each category

- Suppose that D contains n samples, $x_1, x_2, \ldots, x_n$

$$P(D \mid \theta) = \prod_{k=1}^{k=n} P(x_k \mid \theta) = F(\theta)$$

$P(D \mid \theta)$ is called the likelihood of θ w.r.t. the set of samples)

- ML estimate of θ is, by definition the value that maximizes $P(D \mid \theta)$
  "It is the value of θ that best agrees with the actually observed training sample"



- Optimal estimation

$\mu, \sigma^2 \quad \theta = \{\theta_1, \theta_2\}$

  - Let $\theta = (\theta_1, \theta_2, \ldots, \theta_p)^t$ and let ∇_θ be the gradient operator

$$\nabla_\theta = \left[\frac{\partial}{\partial \theta_1}, \frac{\partial}{\partial \theta_2}, \ldots, \frac{\partial}{\partial \theta_p} \right]^t$$

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

$$l(\theta) = \ln P(D \mid \theta)$$

  - New problem statement:
    determine θ that maximizes the log-likelihood

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

Set of necessary conditions for an optimum is:

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

$$\nabla_\theta l = 0$$

- ML Estimation:
  - Gaussian Case: *unknown* μ *and* σ

    $\theta = (\theta_1, \theta_2) = (\mu, \sigma^2)$

$$l = \ln P(x_k \mid \theta) = -\frac{1}{2}\ln 2\pi\theta_2 - \frac{1}{2\theta_2}(x_k - \theta_1)^2$$

$$\nabla_\theta l = \begin{pmatrix} \frac{\sigma}{\sigma\theta_1}(\ln P(x_k \mid \theta)) \\ \frac{\sigma}{\sigma\theta_2}(\ln P(x_k \mid \theta)) \end{pmatrix} = 0$$

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

Summation:

$$\begin{cases} \sum_{k=1}^{k=n} \frac{1}{\hat{\theta}_2}(x_k - \theta_1) = 0 & (1) \\ -\sum_{k=1}^{k=n} \frac{1}{\hat{\theta}_2} + \sum_{k=1}^{k=n} \frac{(x_k - \hat{\theta}_1)^2}{\hat{\theta}_2^2} = 0 & (2) \end{cases}$$

Combining (1) and (2), one obtains:

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

ML estimate for multivariate case

$p(x|\theta) = \theta e^{-\theta x}, \quad x \geq 0$ pdf.

θ_{ML}

Likelihood $F(\theta) = \ln p(X|\theta)$

$X = \{x_1, x_2, \ldots, x_n\}$

$\theta_{ML} = \frac{1}{\ln \theta}$ ~~$E(x) = \frac{1}{\theta}$~~

$= \frac{1}{x \ln e}$

$\theta_{ML} = n / \sum_i x_i$

$F(\theta) = \ln \prod_{i=1}^{n} \theta e^{-\theta x_i}$

$= \sum_{i=1}^{n} \ln\theta - \theta x_i$

$\frac{\partial F(\theta)}{\partial \theta} = 0$

When $n \to \infty$ $\mu_{ML} = \mu$

Where 'μ' is the true value.

*) ~~Bias~~ Estimate → Biased
→ Unbiased.

x_i are i.i.d.s.
$\sim N(\mu, \sigma^2)$

$$\mu_{ML} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

and est. is called unbiased,

~~E =~~ ✓ $E(\mu_{ML}) = E\left(\frac{1}{n} \sum_{i=1}^{n} x_i\right) = \mu$

$= \frac{1}{n} \sum_{i=1}^{n} E(x_i) = \frac{1}{n} \cdot n\mu$