Basics of Machine Learning

Likelihood Ratio Test

Professor: Cheolsoo Park

BCML
Bio Computing & Machine Learnnig Lab

What is Machine Learning (ML)?

↗ **Components of Machine Learning (ML)**

   ↗ Computer Program learning experience **(E)**

   ↗ Class of task (T) corresponding to E

   ↗ Performance Measure (P) of Task

➔ **Machine Learning Algorithm : Program (or algorithm) improving performance (P) of task (T) using the experience (E)**

↗ ML finds a target function f between data $X = (x_1, x_2, \ldots, x_3)$ and states $Y = (y_1, y_2, \ldots, y_n)$

↗ ML sets a hypothesis function, f', which best approximates target function f with some performance measure

↗ Problem Setting

   ↗ Set of possible instance(domain): X

   ↗ Output: Y

   ↗ Unknown target function $f : X \rightarrow Y$

   ↗ Set of hypothesis function space $H \subset \{g | g : X \rightarrow Y\}$

   ↗ **Input**: Training example $\{\langle x_i, y_i \rangle\}$

   ↗ **Output**: $g' \in H$ that best approximates target function f with some performance measure

It is impossible to check all functions when setting hypothesis function space. Thus limit it into a set to find

What is Machine Learning (ML)?

↗ Learning using input data (training data)

    ↗ Supervised learning : with label y $\{\langle x_i, y_i \rangle\}$

    ↗ Unsupervised learning : without label y

↗ Return hypothesis, g′ among g $\in H$, best approximating target function g'
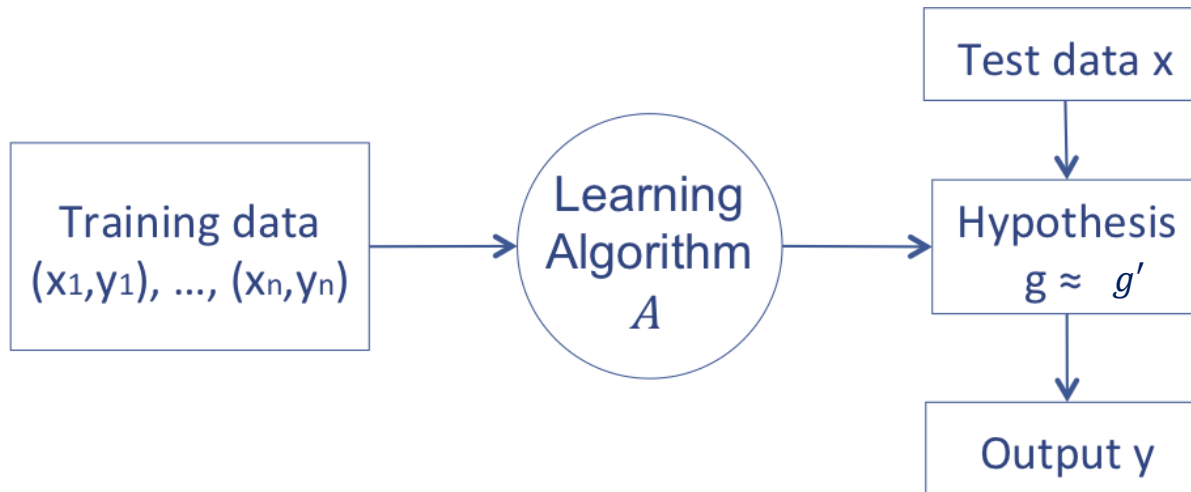
Probability Theory

↗ ML– Probabilistic Perspective

　　↗ ML sets a probability density function (PDF), rather than a deterministic function of hypothesis, and find the parameters of the PDF

　　Ex) Assume the data has a Gaussian PDF, and then mean and covariance need to be found

　　↗ function parameter → PDF parameter

Estimation

↗ In most case, the information of real probability density distribution can't be gotten

↗ This must be determined by the test data

↗ Parameter Estimation

  ↗ Estimate parameter about density by MLE (maximum likelihood estimation) method

↗ Non-parametric Density Estimation

  ↗ k-NNR

Maximum Likelihood Estimation

↗ Maximum Likelihood Estimation (MLE)

  ↗ Estimate random variable's parameter using observation or data

Ex) when flipping coin, we can get the p (probability of the front) by counting the number of the front among all trials

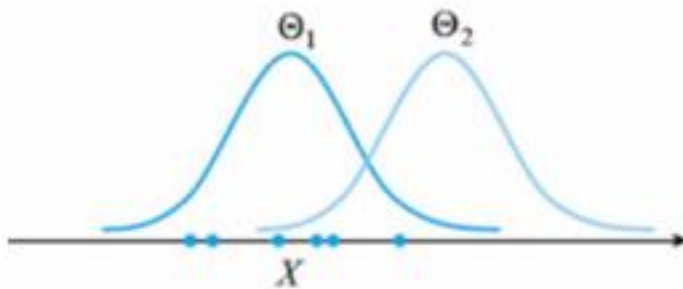 ↗ PDF $f = \{f(\cdot\,|\theta)\}$ and observation $X = (x_1, x_2, \dots, x_n)$

If f is a Gaussian function, θ includes mean μ and covariance Σ.

If f is a Bernoulli function, θ includes $0 \leq \text{p} \leq 1$

$$L(\theta; x_1, x_2, \dots, x_n) = L(\theta; X) = f(X|\theta) = f(x_1, x_2, \dots, x_n|\theta)$$
$$\hat{\theta} = \underset{\theta}{argmax}\, L(\theta; X) = \underset{\theta}{argmax}\, f(X|\theta)$$

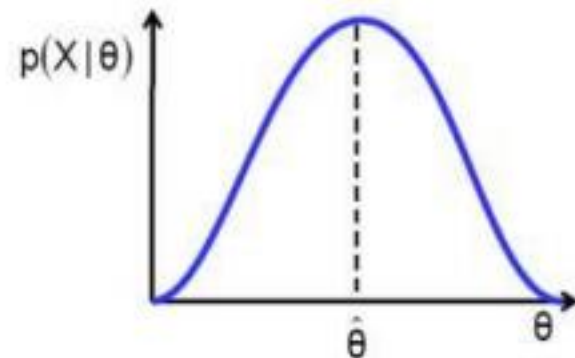# Maximum Likelihood Estimation

- Find the Θ that has maximum likelihood about given X

- $P(X| Θ1) > P(X| Θ2)$

$$\hat{\theta} = \arg\max[p(X \mid \theta)]$$

MAXIMUM LIKELIHOOD

Maximum Likelihood Estimation

- Estimate parameter θ=[θ1, θ2, …, θN] using sample data X={x1,x2,…,xN} observed at probability density function

  - The entire sample set is expressed by the joint probability density

  $$p(X \mid \theta) = p(x^{(1} \mid \theta)\, p(x^{(2} \mid \theta)...p(x^{(N} \mid \theta) = \prod_{k=1}^{N} p(x^{(k} \mid \theta)$$
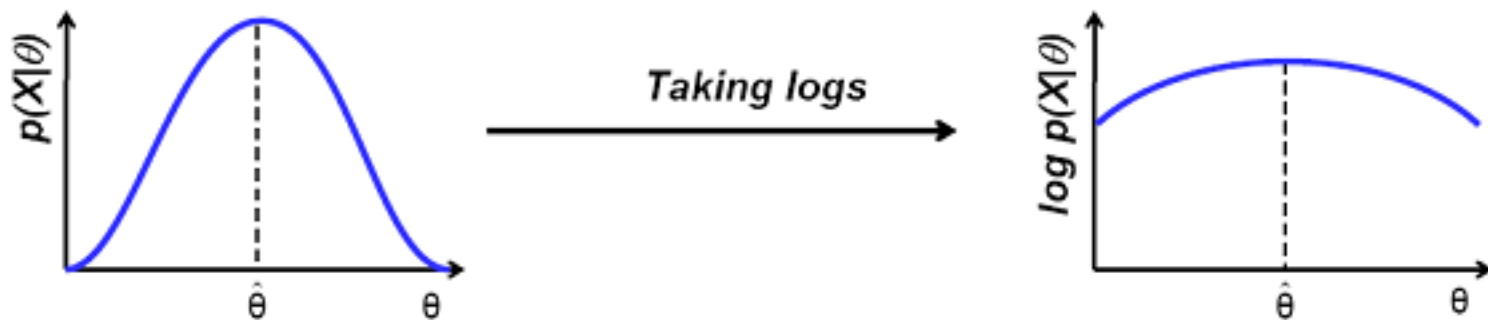
  - Θ , having the highest probability, becomes the estimated parameter

  - Take the log

  $$\hat{\theta} = \text{argmax}\left[\log \prod_{k=1}^{N} p\left(x^{(k} \mid \theta\right)\right] = \text{argmax}\left[\sum_{k=1}^{N} \log p\left(x^{(k} \mid \theta\right)\right]$$

Maximum Likelihood Estimation

- The log function is convenient to calculate
  - Maximize $\sum$(sum) is easier than $\prod$(product)
  - If distribution is similar to gaussian, increase the efficiency

$$\hat{\theta} = \text{argmax}\big[p(X|\theta)\big] = \text{argmax}\big[\log p(X|\theta)\big]$$

Maximum Likelihood Estimation

- Probability density function $p(x) = N(\mu, \sigma)$.
  If $X=\{x1,x2,...,xN\}$ and σ is fixed. Get the MLE of μ.

$$\theta = \mu \Rightarrow \hat{\theta} = \operatorname{argmax} \sum_{k=1}^{N} \log p\left(x^{(k} \mid \theta\right)$$

$$= \operatorname{argmax} \sum_{k=1}^{N} \log\left(\frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}\left(x^{(k} - \mu\right)^2\right)\right)$$

$$= \operatorname{argmax} \sum_{k=1}^{N} \left\{\log\left(\frac{1}{\sqrt{2\pi}\sigma}\right) - \frac{1}{2\sigma^2}\left(x^{(k} - \mu\right)^2\right\}$$

Maximum Likelihood Estimation

- Maximum point

$$\frac{\partial}{\partial \mu} \sum_{k=1}^{N} \left\{ \log\left(\frac{1}{\sqrt{2\pi}\sigma} \right) - \frac{1}{2\sigma^2} \left(x^{(k} - \mu \right)^2 \right\} = -\frac{1}{2\sigma^2} \frac{\partial}{\partial \mu} \sum_{k=1}^{N} \left(x^{2(k} - 2x^{(k}\mu + \mu^2 \right)$$

$$= \frac{1}{2\sigma^2} \frac{\partial}{\partial \mu} \sum_{k=1}^{N} \left(2x^{(k}\mu - \mu^2 \right) = \sum_{k=1}^{N} \left(x^{(k} - \mu \right) = \sum_{k=1}^{N} x^{(k} - N\mu = 0 \Rightarrow \mu = \frac{1}{N} \sum_{k=1}^{N} x^{(k}$$

- MLE of μ is average of X.

Maximum Likelihood Estimation

- Probability density function $p(x) = N(\mu, \sigma)$. If $X=\{x1,x2,\ldots,xN\}$ and σ is fixed. Get the MLE of μ and σ.

$$\hat{\theta} = \begin{bmatrix} \theta_1 = \mu \\ \theta_2 = \sigma^2 \end{bmatrix} \Rightarrow \nabla_\theta = \begin{bmatrix} \dfrac{\partial}{\partial \theta_1} \sum_{k=1}^{N} \log p(x^{(k} | \theta) \\ \dfrac{\partial}{\partial \theta_2} \sum_{k=1}^{N} \log p(x^{(k} | \theta) \end{bmatrix} = \sum_{k=1}^{N} \begin{bmatrix} \dfrac{1}{\theta_2}(x^{(k} - \theta_1) \\ -\dfrac{1}{2\theta_2} + \dfrac{(x^{(k} - \theta_1)^2}{2\theta_2^2} \end{bmatrix} = 0$$

- Simplify them with respect to θ_1 and θ_2

$$\hat{\theta}_1 = \frac{1}{N} \sum_{k=1}^{N} x^{(k}; \quad \hat{\theta}_2 = \frac{1}{N} \sum_{k=1}^{N} \left(x^{(k} - \hat{\theta}_1\right)^2$$

Maximum Likelihood Estimation

↗ Disadvantage of MLE

   ↗ Too sensitive to the observation

EX) when we only get the front of coin during flipping coin trials, then $p = 1$ by MLE

Solution ➡ Maximum a Posteriori Estimation (MAP)

Maximum a Posteriori Estimation (MAP)

↗ MLE : $\hat{\theta} = \underset{\theta}{argmax}\, f(X|Y,\theta)$

↗ MAP : $\hat{\theta} = \underset{\theta}{argmax}\, f(Y|X,\theta)$

↗ While MLE produces parameters based on the current observation, MAP uses general knowledge about the data as well as the observation (prior knowledge)

↗ EX) Flipping coin example

➜ MAP uses Bayes' Theorem

$$\hat{\theta} = \underset{\theta}{argmax}\, f(Y|X,\theta) = \underset{\theta}{argmax}\, \frac{f(X|Y,\theta)f(Y)}{f(X|\theta)}$$

$$\blacktriangleright\, \hat{\theta} = \underset{\theta}{argmax}\, f(X|Y,\theta)f(Y)$$

↗ $f(\theta)$ is a prior knowledge or prior assumption
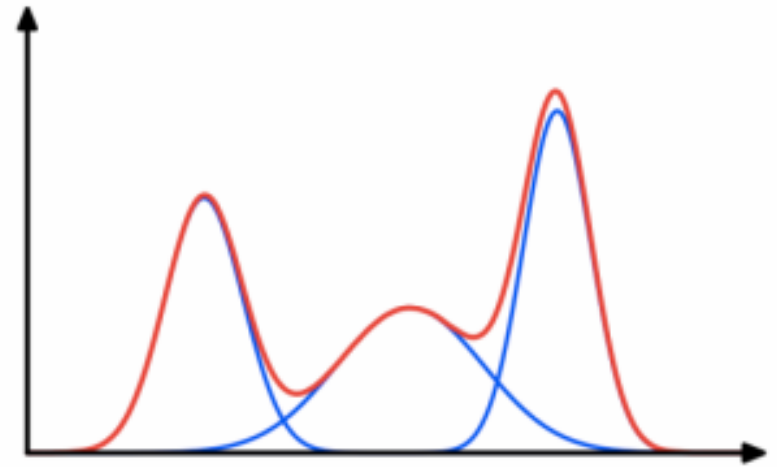
Maximum a Posteriori Estimation (MAP)

↗ Prior knowledge could be prejudice

↗ The results are depending on the prior, and thus it is very important to set the prior probability function carefully
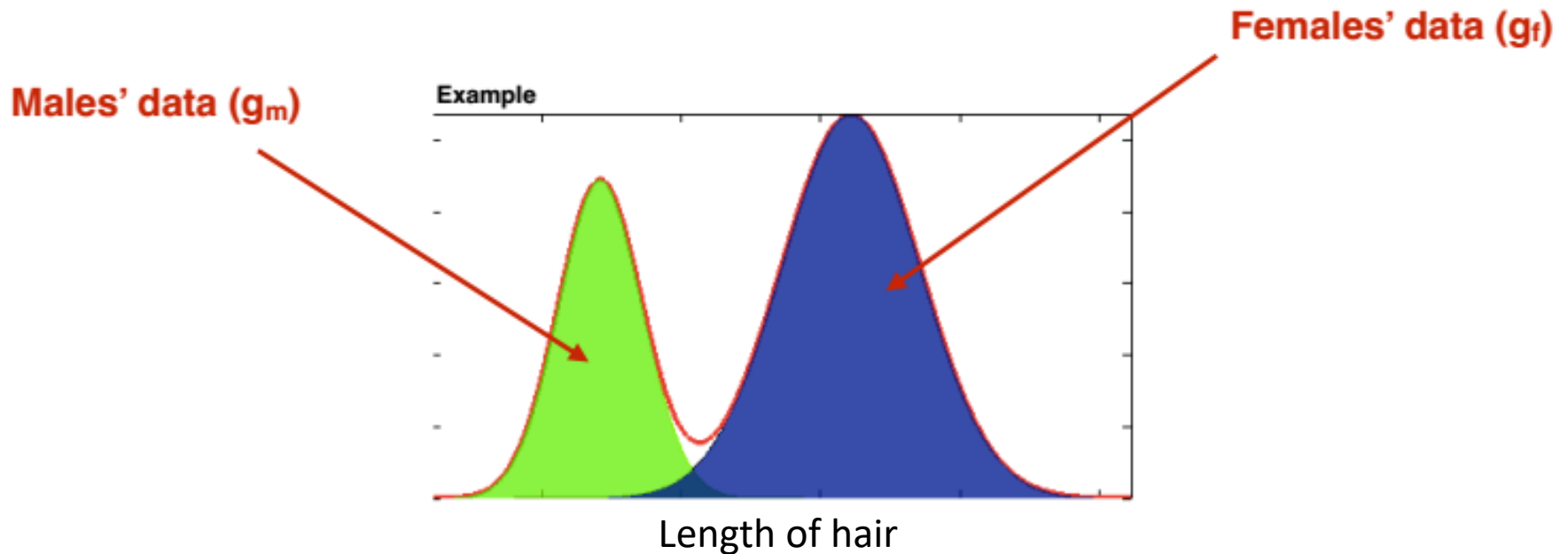
Likelihood Ratio Test

Gaussian Distribution

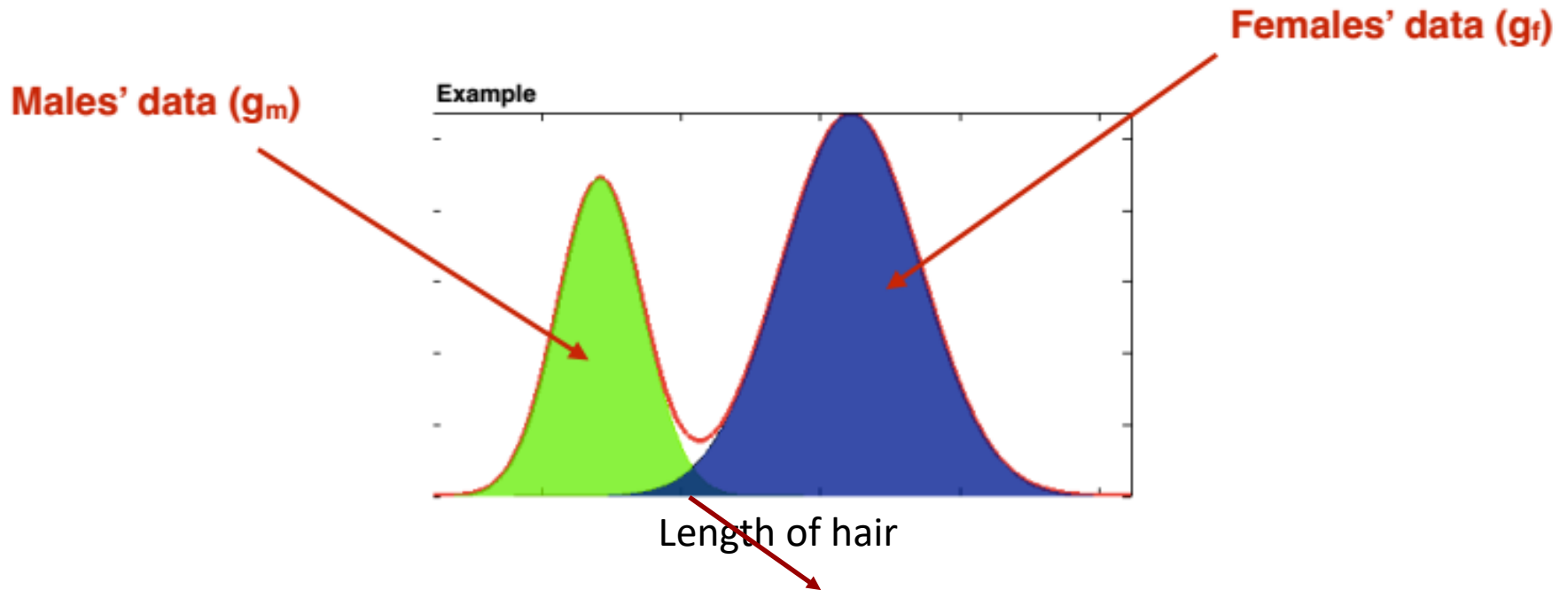**Gaussian Mixture Model**

Likelihood Ratio Test

* PDF of hair length of the students in our class

Males' data (g_m)

Females' data (g_f)

Example

Length of hair

https://www.youtube.com/watch?v=U8mjbpqsOTI

Likelihood Ratio Test

- PDF of hair length of the students in our class

Males' data (g_m)

Females' data (g_f)

Example

Length of hair

How can we decide whether this is for male or female?

- Assume we are going to classify an object into classes based on the evidence provided by a measurement (or feature vector) **x**

Likelihood Ratio Test

↗ Choose a class most 'probable' given the observed feature vector **x**

↗ In formal

Evaluate the posterior probability of each class $P(w_i|x)$ and choose a class with the largest $P(w_i|x)$

Likelihood Ratio Test

↗ 2 class classification example

↗ The decision rule

$$if \ P(w_1|x) > P(w_2|x), choose \ w_1$$
$$else, \qquad\qquad\qquad choose \ w_2$$

↗ Or

$$P(w_1|x) \underset{w_2}{\overset{w_1}{\gtrless}} P(w_2|x)$$

↗ Using Bayes theorem

Likelihood ratio test

$$\frac{P(x|w_1)P(w_1)}{P(x)} \underset{w_2}{\overset{w_1}{\gtrless}} \frac{P(x|w_2)P(w_2)}{P(x)} \quad\Longrightarrow\quad \Lambda(x) = \frac{P(x|w_1)}{P(x|w_2)} \underset{w_2}{\overset{w_1}{\gtrless}} \frac{P(w_2)}{P(w_1)}$$
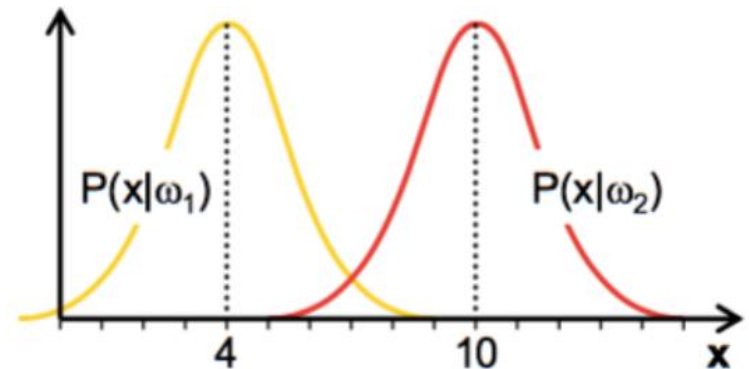
Likelihood ratio

Likelihood Ratio Test

↗ Example

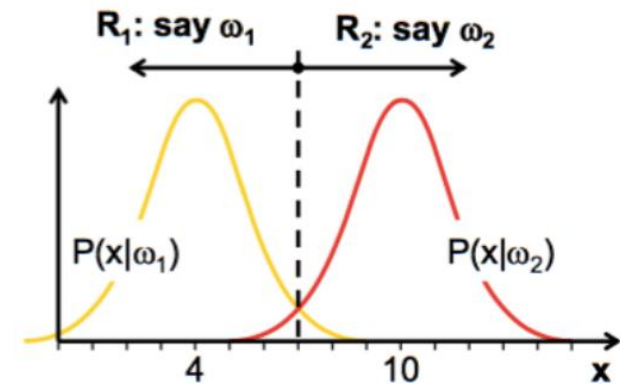Derive a classification rule using the likelihood ratio test with assuming the priors of two classes are equal

$$P(x|w_1) = \frac{1}{\sqrt{2\pi}} exp\{-\frac{1}{2}(x-4)^2\}$$

$$P(x|w_2) = \frac{1}{\sqrt{2\pi}} exp\{-\frac{1}{2}(x-10)^2\}$$

$$\Lambda(x) = \frac{\frac{1}{\sqrt{2\pi}} exp\{-\frac{1}{2}(x-4)^2\}}{\frac{1}{\sqrt{2\pi}} exp\{-\frac{1}{2}(x-10)^2\}} \underset{w_2}{\overset{w_1}{\gtrless}} \frac{1}{1}$$

$$(x-4)^2 - (x-10)^2 \underset{w_2}{\overset{w_1}{\gtrless}} 0 \implies x \underset{w_2}{\overset{w_1}{\lessgtr}} 7 \implies$$
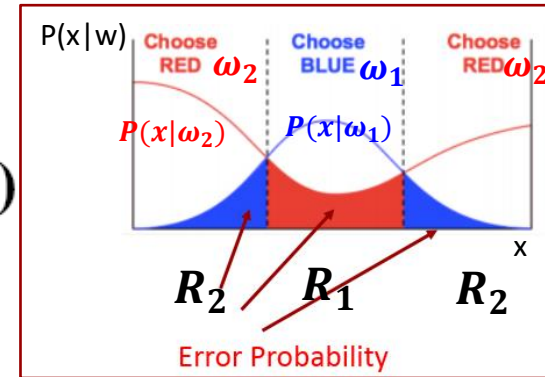
Comput

Error Probability

Decided by the borders

$$P(error) = P(x \in R_2, \omega_1) + P(x \in R_1, \omega_2)$$

$$= P(x \in R_2 \mid \omega_1)P(\omega_1) + P(x \in R_1 \mid \omega_2)P(\omega_2)$$

$$= \int_{R_2} p(x \mid \omega_1)P(\omega_1)dx + \int_{R_1} p(x \mid \omega_2)P(\omega_2)dx$$

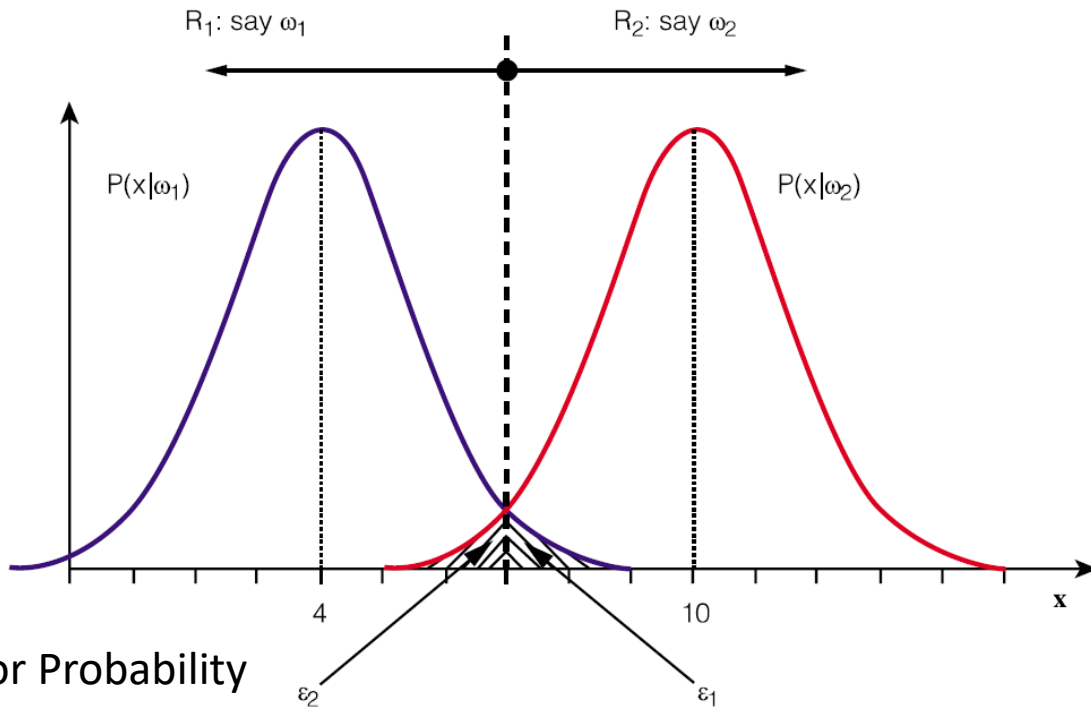$$P(error) = \sum_{i=1}^{C} P(error \mid \omega_i)P(\omega_i)$$

, where

$$P(error \mid \omega_i) = P(\text{choose } \omega_j \mid \omega_i) = \int_{R_j} P(x \mid \omega_i)dx$$
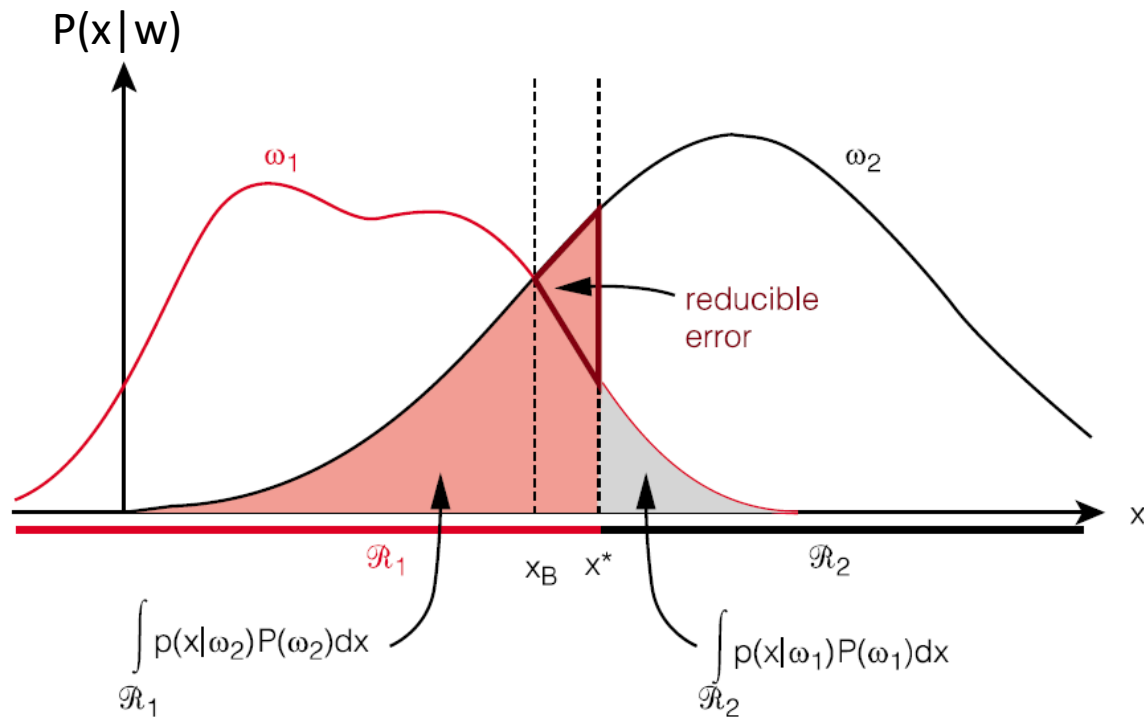
Error Probability

$$P[\text{error}] = P[\omega_1] \int_{R_2} P(x \mid \omega_1)dx + P[\omega_2] \int_{R_1} p(x \mid \omega_2)dx$$

$$\underbrace{}_{\varepsilon_1} \qquad \underbrace{}_{\varepsilon_2}$$

$$P(\text{error}) = (\varepsilon_1 + \varepsilon_2)/2, \text{ given } P(w_1) = P(w_2) = \frac{1}{2}$$

Error Probability

Error Probability

⬀ The optimal border = minimum error probability

[그림 5-3] 오류확률에 의한 결정 경계의 결정 [Duda, Hart, Stock, 2001]

Loss (or Cost) function

↗ In the previous slides, the error was 0 or 1 (misclassification number) when the classification was correct or wrong.

↗ However, the number of error can be a summation of squared difference between target data and estimated data. (This is one example)

    ➔ Cost function or Lost function

↗ For example, let's think about buying a wrong plane ticket to find a person in the world

↗ While true value is constant, the estimated value can be changed depending on the updated parameters

$$L(\theta, \hat{\theta}) : Loss\ function$$

$(\theta$: true parameter, $\hat{\theta}$: estimated parameter)

Bayes Risk

↗ So far we have assumed that the penalty of misclassifying a class w_1 as class w2 is the same as the opposite way $(w_2 \rightarrow w_1)$

↗ In general, this is not the case:

  ↗ For example, misclassifying a cancer sufferer as a healthy patient is a much more serious problem than the other way around

↗ This concept can be also formalised in terms of a cost (or lost) function C_{ij}

  ↗ C_{ij} represents the cost of choosing (wrong) class w_i when class w_j is the true class

↗ We define the Bayes Risk as the expected value of the cost

$$\Re = E[C] = \sum_{i=1}^{2} \sum_{j=1}^{2} C_{ij} \cdot P(choose \ w_i \ and \ x \in w_j) = \sum_{i=1}^{2} \sum_{j=1}^{2} C_{ij} \cdot P(x \in R_i | w_j) \cdot P(w_j)$$

Bayes Risk

$$\mathbf{R = E[C]} = \sum_{i=1}^{2}\sum_{j=1}^{2} \mathbf{C_{ij} \cdot P(choose \ \omega_i \ , \ x \in \omega_j)} = \sum_{i=1}^{2}\sum_{j=1}^{2} \mathbf{C_{ij} \cdot P(x \in R_i \mid \omega_j) \cdot P(\omega_j)}$$

, where $\mathbf{P(x \in R_i \mid \omega_j) = \int_{R_i} P(x \mid \omega_j)dx}$
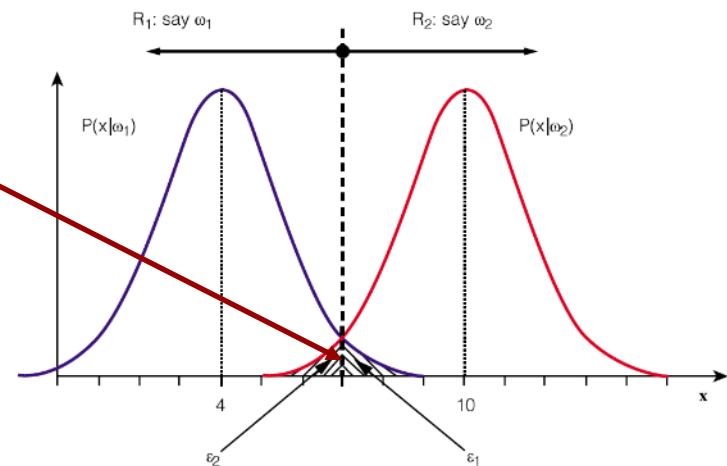
$$\mathbf{R = \int_{R_1} [C_{11}P(\omega_1)P(x \mid \omega_1) + C_{12}P(\omega_2)P(x \mid \omega_2)]dx +}$$

$$\mathbf{\int_{R_2} [C_{21}P(\omega_1)P(x \mid \omega_1) + C_{22}P(\omega_2)P(x \mid \omega_2)]dx}$$

Bayes Risk
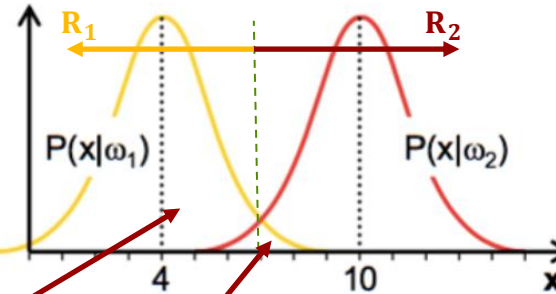
➚ Decision rule to reduce the Bayes risk

when i=1

$$\int_{R_1} P(x \mid \omega_i)dx + \int_{R_2} P(x \mid \omega_i)dx = \int_{R_1 \cup R_2} P(x \mid \omega_i)dx = 1$$
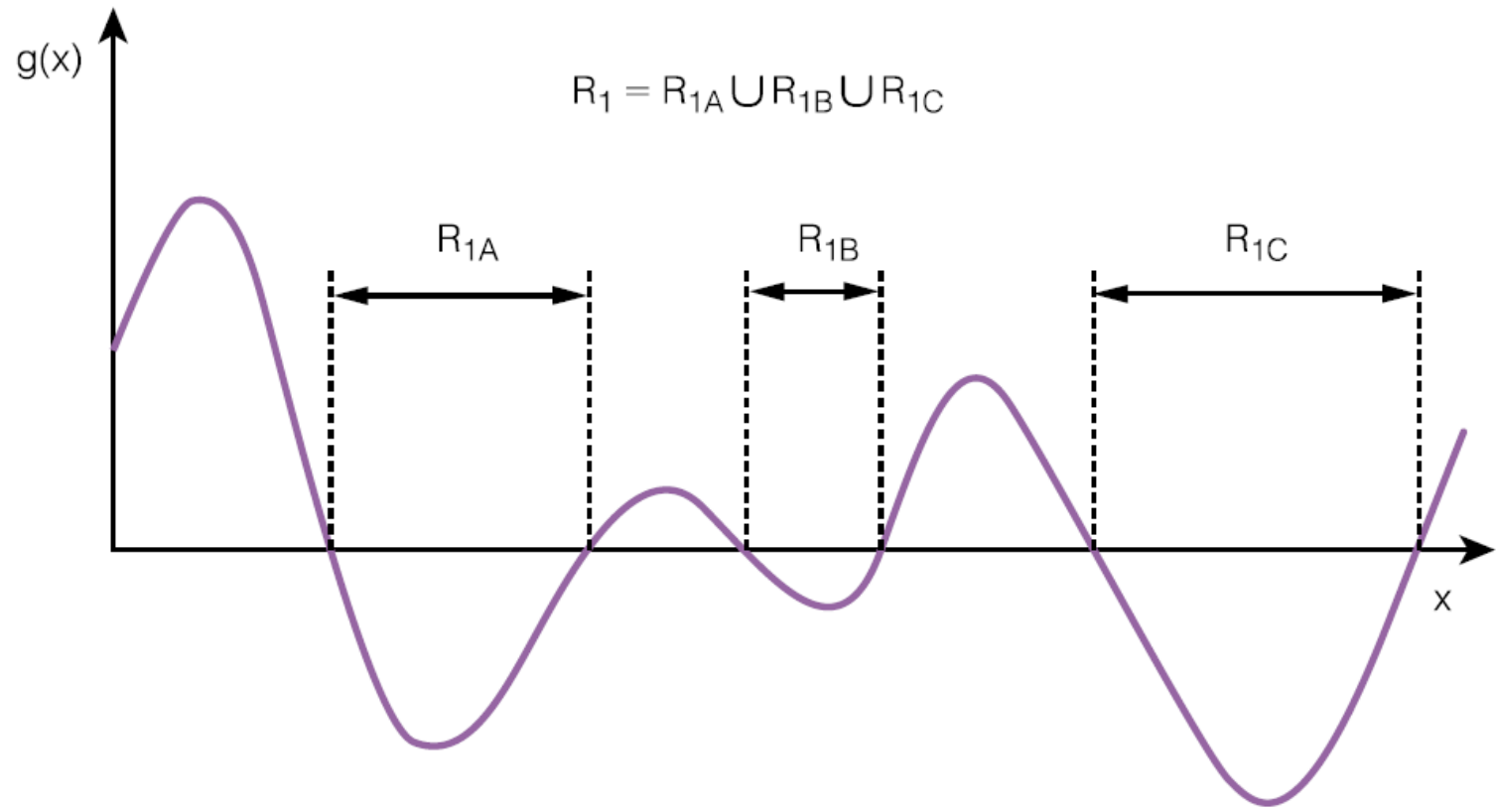
$$R = C_{11}P(\omega_1)\int_{R_1} P(x \mid \omega_1)dx + C_{12}P(\omega_2)\int_{R_1} P(x \mid \omega_2)dx +$$
$$+ C_{21}P(\omega_1)\int_{R_2} P(x \mid \omega_1)dx + C_{22}P(\omega_2)\int_{R_2} P(x \mid \omega_2)dx +$$
$$+ C_{21}P(\omega_1)\int_{R_1} P(x \mid \omega_1)dx + C_{22}P(\omega_2)\int_{R_1} P(x \mid \omega_2)dx +$$
$$- C_{21}P(\omega_1)\int_{R_1} P(x \mid \omega_1)dx - C_{22}P(\omega_2)\int_{R_1} P(x \mid \omega_2)dx$$

$$R = C_{21}P(\omega_1) + C_{22}P(\omega_2)$$
$$+ (C_{12} - C_{22})P(\omega_2)\int_{R_1} P(x \mid \omega_2)dx - (C_{21} - C_{11})P(\omega_1)\int_{R_1} P(x \mid \omega_1)dx$$

$$R_1 = \arg\min_{R_1}\left\{\int_{R_1} [(C_{12} - C_{22})P[\omega_2]P(x \mid \omega_2) - (C_{21} - C_{11})P[\omega_1]P(x \mid \omega_1)]dx\right\}$$

$$= \arg\min_{R_1}\left\{\int_{R_1} g(x)dx\right\}$$

Bayes Risk

↗ Find the decision area based on the Bayes risk

Bayes Risk

$$R_1 = \arg\min\left\{\int_{R_1}[(C_{12} - C_{22})P[\omega_2]P(x\,|\,\omega_2) - (C_{21} - C_{11})P[\omega_1]P(x\,|\,\omega_1)]dx\right\}$$

$$= \arg\min\left\{\int_{R_1} g(x)dx\right\}$$

$$(C_{12} - C_{22})P[\omega_2]P(x\,|\,\omega_2) \underset{\omega_1}{\overset{\omega_2}{\gtrless}} (C_{21} - C_{11})P[\omega_1]P(x\,|\,\omega_1)$$

$$\frac{P(x\,|\,\omega_1)}{P(x\,|\,\omega_2)} \underset{\omega_2}{\overset{\omega_1}{\gtrless}} \frac{(C_{12} - C_{22})P[\omega_2]}{(C_{21} - C_{11})P[\omega_1]}$$

Bayes Risk

Ex) Given the likelihood functions below for two classes, their prior probabilities, $P(w_1) = P(w_2) = 0.5$, and costs $C_{11} = C_{22} = 0, C_{12} = 1, C_{21} = 3^{\frac{1}{2}}$, find the likelihood ratio based on the Bayes risk and decision boundary to minimize the error probability

$$P(x|\omega_1) = \frac{1}{\sqrt{2\pi}\sqrt{3}} e^{-\frac{1}{2}\frac{x^2}{3}}$$

$$P(x|\omega_2) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(x-2)^2}$$
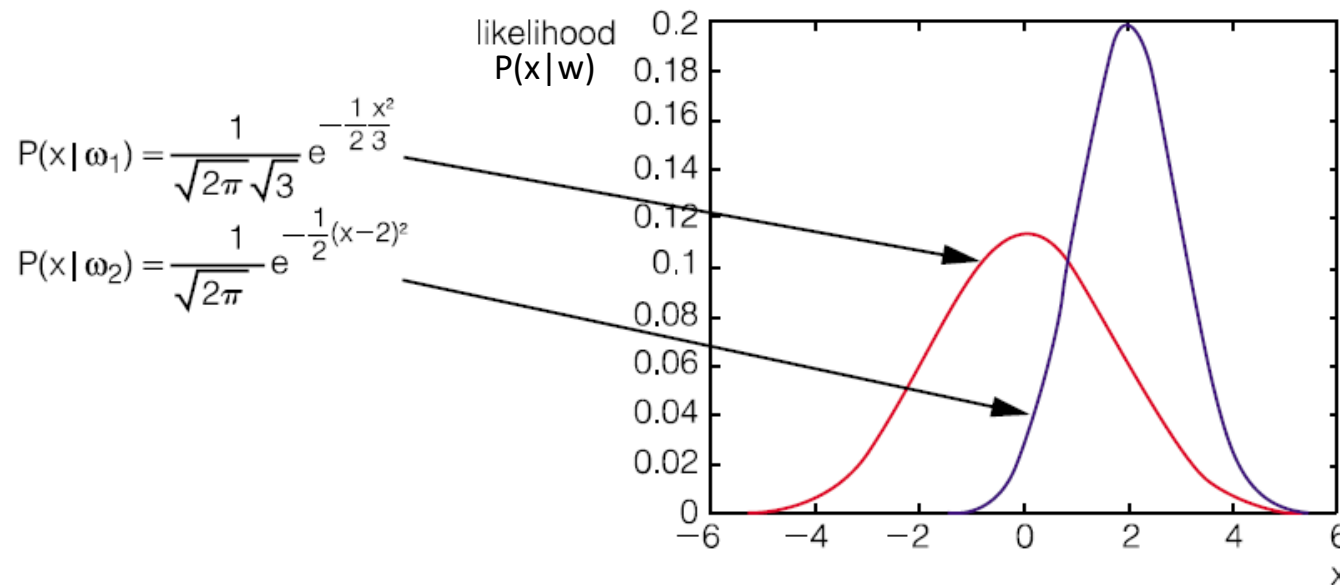
likelihood
P(x|w)

Minimise Bayes Risk

Likelihood Ratio

$$\Lambda(\mathbf{x}) = \frac{\dfrac{1}{\sqrt{2\pi}\sqrt{3}}\mathbf{e}^{-\frac{1}{2}\frac{x^2}{3}}}{\dfrac{1}{\sqrt{2\pi}}\mathbf{e}^{-\frac{1}{2}(x-2)^2}} \overset{\omega_1}{\underset{\omega_2}{\gtrless}} \frac{1}{\sqrt{3}}$$

$$\frac{\mathbf{e}^{-\frac{1}{2}\frac{x^2}{3}}}{\mathbf{e}^{-\frac{1}{2}(x-2)^2}} \overset{\omega_1}{\underset{\omega_2}{\gtrless}} 1$$

$$-\frac{1}{2}\frac{x^2}{3} + \frac{1}{2}(x-2)^2 \overset{\omega_1}{\underset{\omega_2}{\gtrless}} 0$$

$$2x^2 - 12x + 12 \overset{\omega_1}{\underset{\omega_2}{\gtrless}} 0 \Rightarrow x = 4.73, 1.27$$

$$\frac{\mathbf{P}(\mathbf{x}|\omega_1)}{\mathbf{P}(\mathbf{x}|\omega_2)} \overset{\omega_1}{\underset{\omega_2}{\gtrless}} \frac{(\mathbf{C}_{12} - \mathbf{C}_{22})\mathbf{P}[\omega_2]}{(\mathbf{C}_{21} - \mathbf{C}_{11})\mathbf{P}[\omega_1]}$$

, $P(w_1) = P(w_2) = 0.5$, and
costs $C_{11} = C_{22} = 0, C_{12} = 1, C_{21} = 3^{\frac{1}{2}}$

$$P(x|\omega_1) = \frac{1}{\sqrt{2\pi}\sqrt{3}}\mathbf{e}^{-\frac{1}{2}\frac{x^2}{3}}$$

$$P(x|\omega_2) = \frac{1}{\sqrt{2\pi}}\mathbf{e}^{-\frac{1}{2}(x-2)^2}$$
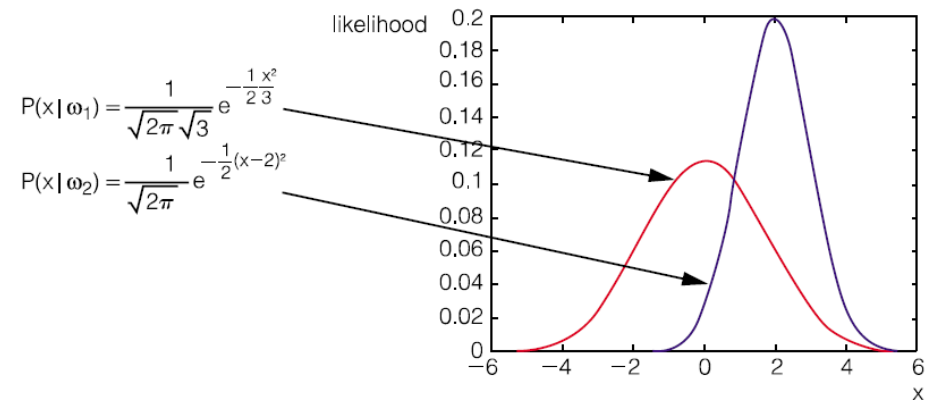
$$\therefore \omega_1 < 1.27, \ \omega_1 > 4.73$$
$$1.27 < \omega_2 < 4.73$$

Minimise Bayes Risk
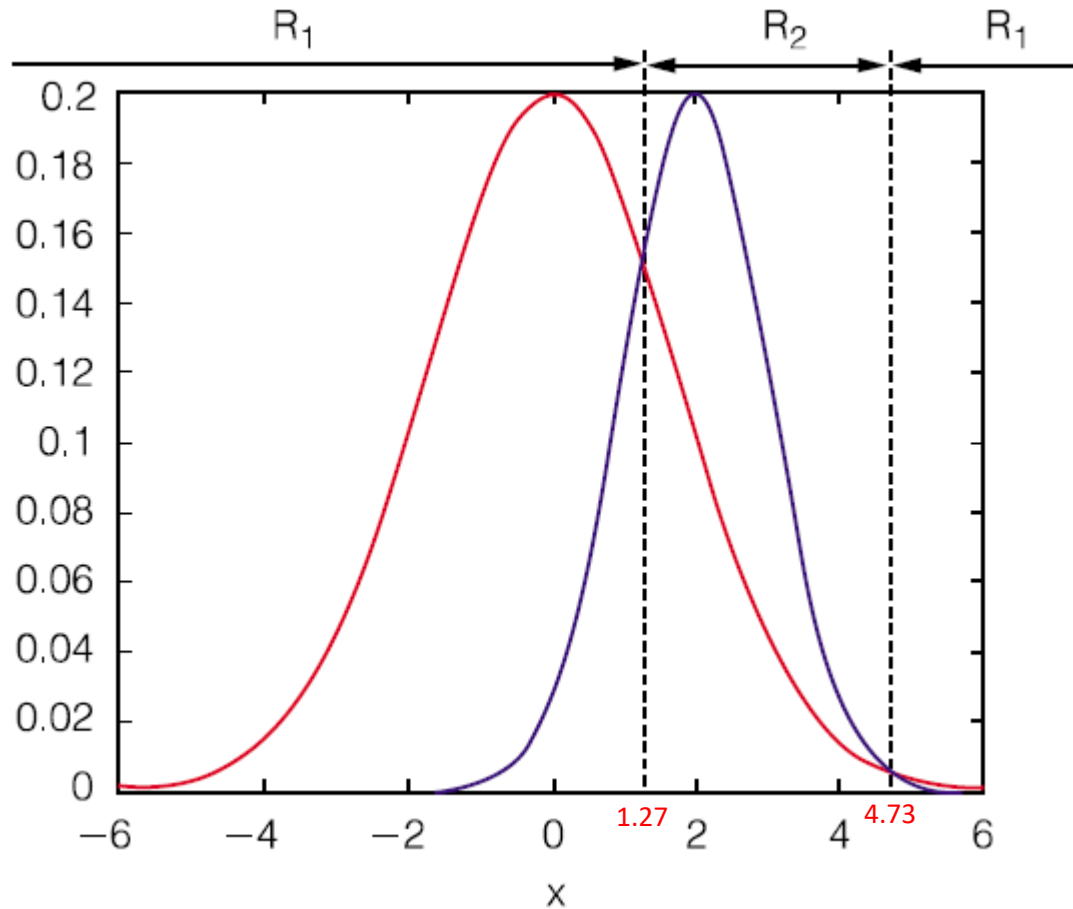
Variations of LRT Decision Rule

↗ Bayes' Criterion

   ↗ LRT decision rule with reducing the Bayes' risk

$$\Lambda(\mathbf{x}) = \frac{P(\mathbf{x}|\omega_1)}{P(\mathbf{x}|\omega_2)} \overset{\omega_1}{\underset{\omega_2}{\gtrless}} \frac{(C_{12} - C_{22})P[\omega_2]}{(C_{21} - C_{11})P[\omega_1]}$$

↗ MAP Criterion

   ↗ <u>Symmetric or zero-one cost functions</u> make Bayes' criterion a ratio of posteriori functions➜ called maximum a posteriori (MAP) criterion

$$C_{ij} = \begin{cases} 0 & i = j \\ 1 & i \neq j \end{cases} \Rightarrow \Lambda(\mathbf{x}) = \frac{P(\mathbf{x}|\omega_1)}{P(\mathbf{x}|\omega_2)} \overset{\omega_1}{\underset{\omega_2}{\gtrless}} \frac{P[\omega_2]}{P[\omega_1]} \Leftrightarrow \frac{P(\omega_1|\mathbf{x})}{P(\omega_2|\mathbf{x})} \overset{\omega_1}{\underset{\omega_2}{\gtrless}} 1$$

Variations of LRT Decision Rule

↗ ML Criterion

  ↗ <u>Equal prior probabilities and zero-one cost functions</u> make Bayes' criterion a ratio of likelihood functions ➜ called maximum likelihood (ML) criterion

$$C_{ij} = \begin{cases} 0 & i = j \\ 1 & i \neq j \end{cases}$$

$$P(\omega_i) = \frac{1}{C} \quad \forall i$$

$$\Rightarrow \Lambda(x) = \frac{P(x \mid \omega_1)}{P(x \mid \omega_2)} \underset{\omega_2}{\overset{\omega_1}{\underset{<}{>}}} 1$$