Machine Learning in Robotics
Lecture 3: Bayesian Decision Theory and Classifier

Prof. Dongheui Lee

Institute of Automatic Control Engineering
Technische Universität München

dhlee@tum.de

Summary of last lecture : Linear Regression

- Method of Least Mean Square (LMS)

$$E = \frac{1}{n}\sum_{i=1}^{n}(y^{(i)} - wx^{(i)})^2$$

- Normal equation

$$\boldsymbol{w}^{\star} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}$$

- Gradient descent

$$w_j := w_j - \alpha\frac{\partial}{\partial w_j}E(\boldsymbol{w})$$

- Probabilistic Approach

$$L(\boldsymbol{w}) = \prod_{i=1}^{n} p(y^{(i)}|\boldsymbol{x}^{(i)};\boldsymbol{w}) = \prod_{i=1}^{n}\frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{(y^{(i)} - \boldsymbol{x}^{(i)}\boldsymbol{w})^2}{2\sigma^2}\right)$$

- Polynomial Curve Fitting

$$f(x,\boldsymbol{w}) = \sum_{j=0}^{m} w_j x^j$$

Today Lecture Outline

- Linear classifiers
- Logistic regression
- Bayesian Decision
- The Likelihood Ratio Test
- The Probability of Error
- The Bayes Risk
- Discriminant Functions
  - Bayes, MAP, and ML Criterion
- Multi-class problems
- Bayes classifiers for Normally distributed classes

Bayesian Decision Theory

- Bayesian Decision Theory is a fundamental statistical approach to the problem of pattern classification.
- Quantifies the tradeoffs between various classifications using probability and the costs that accompany such decisions/classifications.
- Assumptions
  - Decision problem is posed in probabilistic terms.
  - All relevant probability values are known.

Feature, Class, Prior

- Feature is any distinctive aspect, quality or characteristics. Pattern is a set of features.

$$\boldsymbol{x} = [x_1, \cdots, x_m]^T$$

- Class (State of nature) ω_i
- Priors $p(\omega_i)$

Fish Sorting Example Revisited

- Features are the length and lightness of the fish.

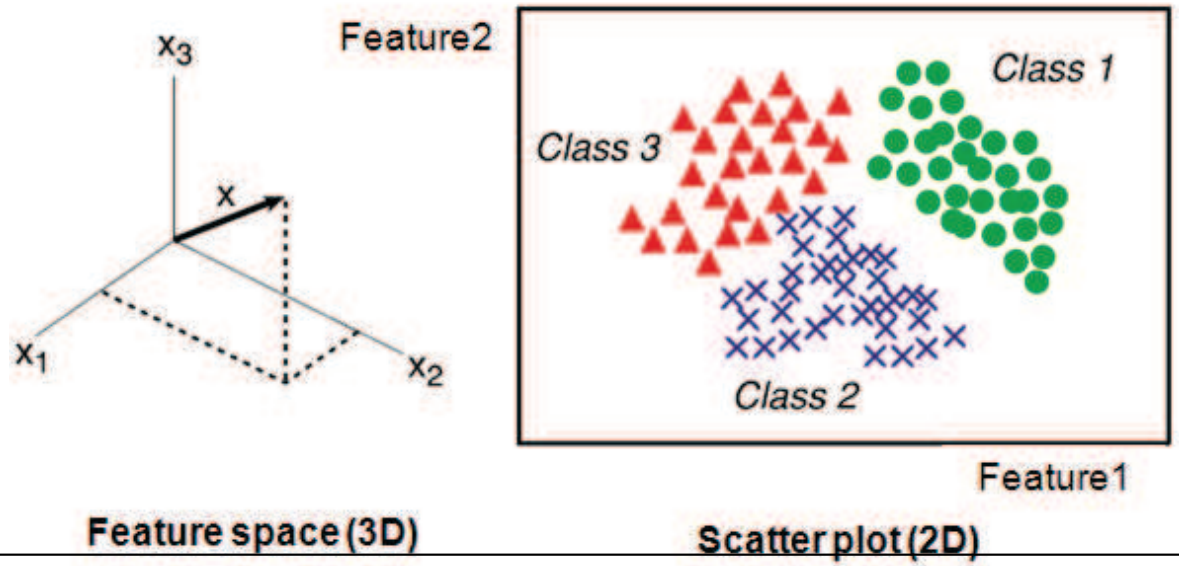
$$\boldsymbol{x} = [length, lightness]^T$$

- Class (State of nature): the type of fish we observed.
  ω_1 for salmon and ω_2 for sea bass.
- Priors $p(\omega_i)$
  $p(\omega_1)$ the a priori probability that the next fish is a salmon.
  $p(\omega_2)$ the a priori probability that the next fish is a sea bass.

Feature

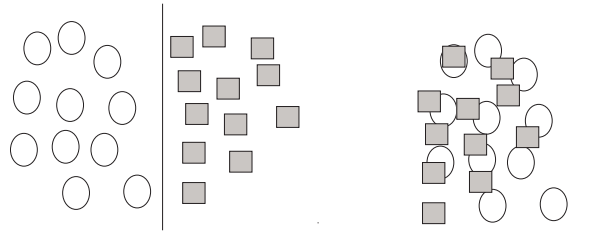
- A feature is an observable variable. Pattern is a set of features.
- A feature space is a set from which we can sample or observe values.
- Examples of features: Length, Width, Lightness, Location of Dorsal Fin

$$\boldsymbol{x} = [x_1, \cdots, x_m]^T$$

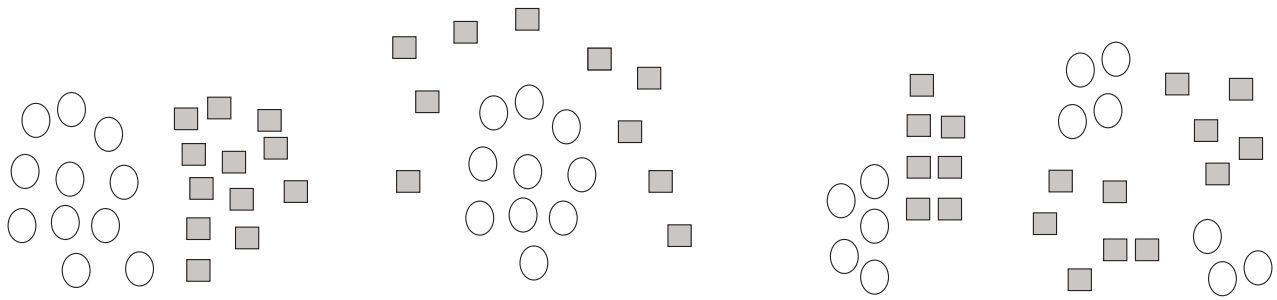


Features and Patterns

- Good Features and Bad Features

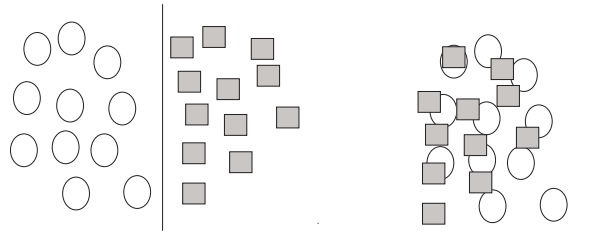


- More Feature Properties
  linear separability, nonlinear separability, highly correlated, and multi-modal features

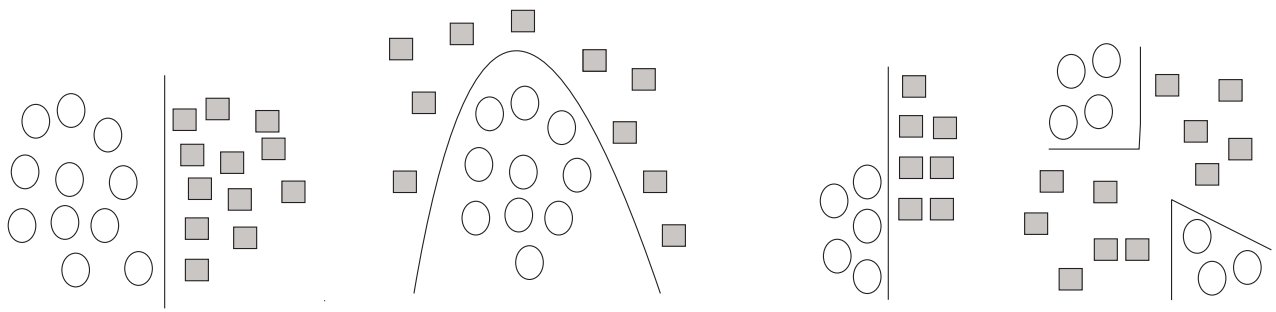


Features and Patterns

- Good Features and Bad Features



- More Feature Properties
  linear separability, nonlinear separability, highly correlated, and multi-modal features



Bayes Theorem

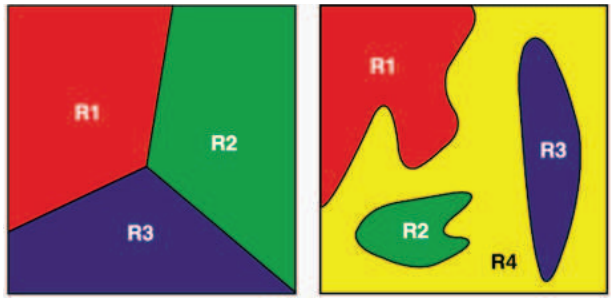
Thomas Bayer (1702-1761)

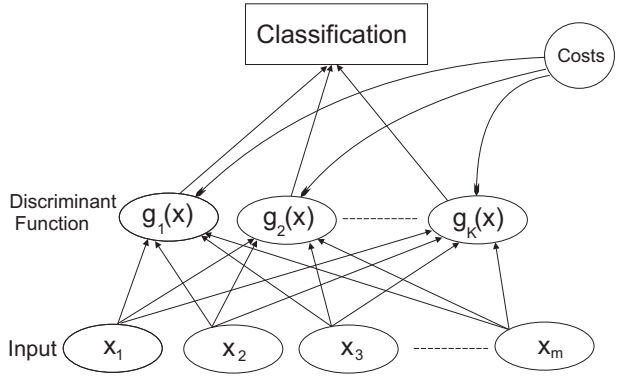
$$p(\omega_i|\boldsymbol{x}) = \frac{p(\boldsymbol{x}|\omega_i)p(\omega_i)}{p(\boldsymbol{x})}$$

- $p(\omega_i)$: prior probability (of class ω_i)
- $p(\omega_i|\boldsymbol{x})$: posterior probability (of class ω_i given the observation $\boldsymbol{x}$)
- $p(\boldsymbol{x}|\omega_i)$: likelihood (conditional probability of observation $\boldsymbol{x}$ given class ω_i)
- $p(\boldsymbol{x})$: a normalized constant that does not affect the decision

Classifier

- The task of a classifier is to partition feature space into class-labeled decision regions
- **decision boundary** : borders between **decision regions**
- The classification of feature vector $\boldsymbol{x}$ consists of determining which decision region it belongs to and assign $\boldsymbol{x}$ to this class
- A classifier can be represented as a set of discriminant functions
  - The classifier assigns a feature vector $\boldsymbol{x}$ to a class by using discriminant functions

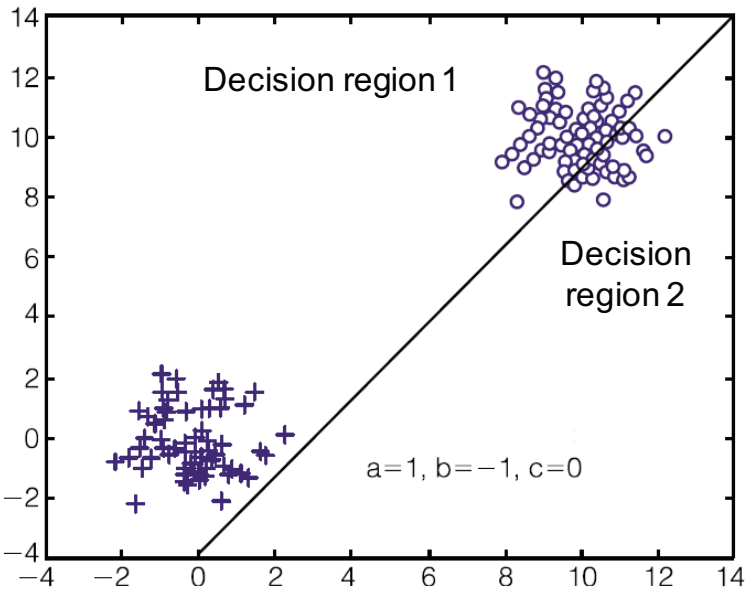




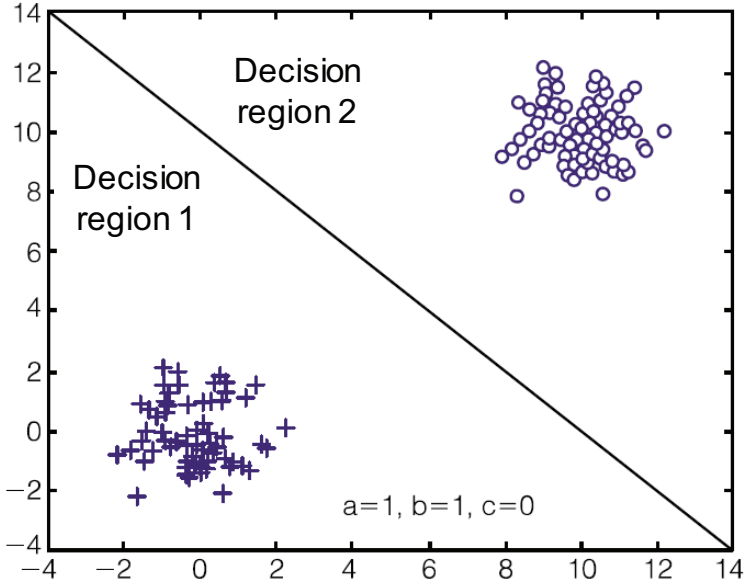
Linear Classifier

Linear models for classification

- Decision boundaries are linear functions of the input vectors $\boldsymbol{x}$ and defined by $m-1$ dimensional hyperplanes within the m dimensional input space. $g(x_1, x_2) = ax_1 + bx_2 + c$
- Data set whose classes can be separated by linear decision surfaces are said to be *linearly separable*.



Decision Boundary before Learning



Decision Boundary after Learning

Linear Classifier

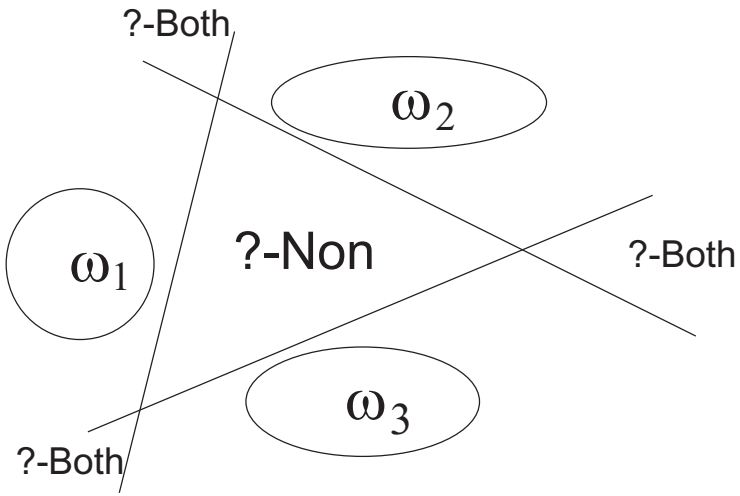
- Linear regression
  - Model prediction is given by a linear function of the parameter $\boldsymbol{w}$.
- Linear Models for classification
  - To predict discrete class label

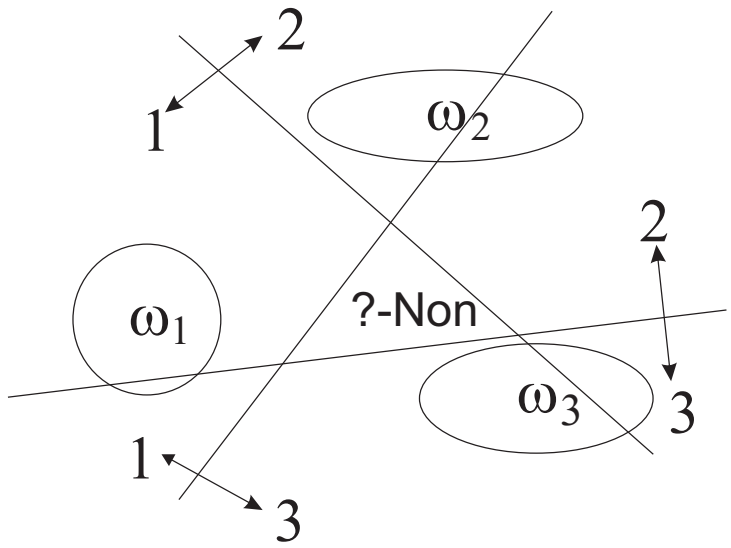
$$g(\boldsymbol{x}) = \boldsymbol{w}^T\boldsymbol{x} + w_0$$

  - $\boldsymbol{w}$: weight vector
  - w_0 : bias
- How to find the weight vector?
  - Normally minimization of classification errors

Decision Boundary

- The case of two classes
  - If $g(\boldsymbol{x}) \geq 0$, the input $\boldsymbol{x}$ is assigned to class ω_1
  - Else, assign to class ω_2
- The case of multiple classes
  - Approach 1: one versus the rest classifier
  - Approach 2: one versus one classifier





Decision Boundary - multiple classes

- Approach 3
  - Single K-class discriminant comprising K linear functions

$$g_k(\boldsymbol{x}) = \boldsymbol{w}_k^T \boldsymbol{x} + w_{k0}$$

  - Assign a point $\boldsymbol{x}$ into class ω_k if $g_k(\boldsymbol{x}) \geq g_j(\boldsymbol{x})$ for all $j \neq k$

Least Squares for Classification

Linear model

$$f(\boldsymbol{x}) = w_0 + w_1 x_1 + \cdots + w_m x_m$$

The parameters are estimated by minimizing the error

$$E = \frac{1}{n} \sum_{i=1}^{n} (y^{(i)} - f(x^{(i)}))^2$$

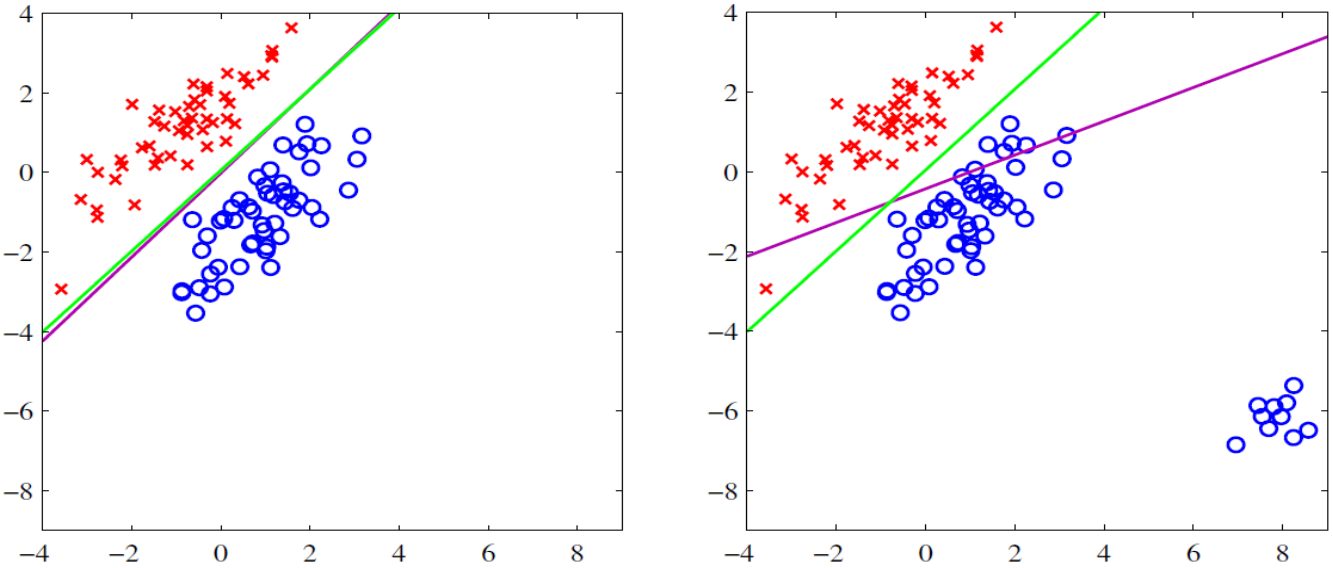
The result becomes

$$\boldsymbol{w}^\star = (\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{X}^T \boldsymbol{y}$$

In the case of binary classification, $y \in \{0, 1\}$

Least Squares for Classification

For 2D input vector



- Magenta : Linear regression
- Green : Logistic regression

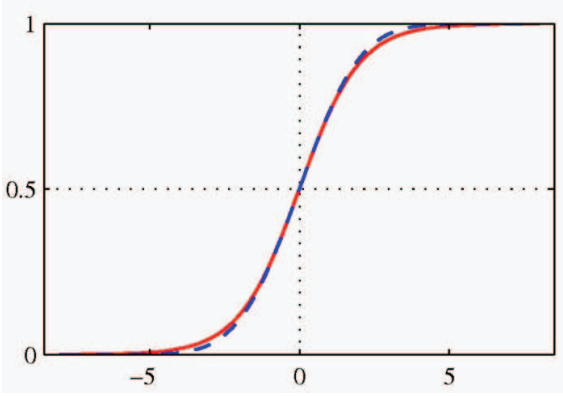
Logistic Regression

- for 2 class problem: classification label $y \in \{0, 1\}$
- Hypothesis function is based on the logistic function (sigmoid function)

$$f(\boldsymbol{x}) \in [0, 1]$$

$$f(\boldsymbol{x}) = \sigma(\boldsymbol{w}^T \boldsymbol{x}) = \frac{1}{1 + exp(-\boldsymbol{w}^T \boldsymbol{x})}$$

$\sigma(z) =$: logistic function



$$p(y = 1|\boldsymbol{x}) = f(\boldsymbol{x})$$

$$p(y = 0|\boldsymbol{x}) = 1 - f(\boldsymbol{x})$$

$$p(y|\boldsymbol{x}) =$$

Logistic Regression

- for 2 class problem: classification label $y \in \{0, 1\}$
- Hypothesis function is based on the logistic function (sigmoid function)

$$f(\boldsymbol{x}) \in [0, 1]$$

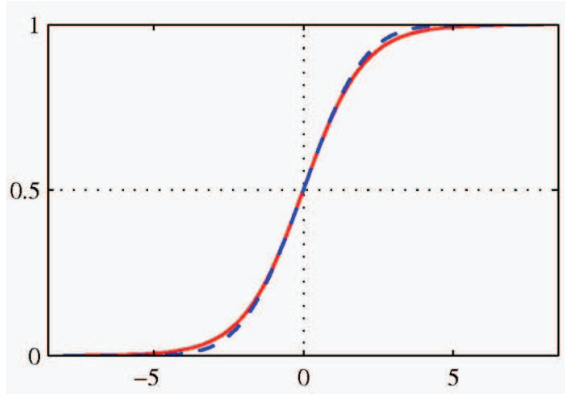
$$f(\boldsymbol{x}) = \sigma(\boldsymbol{w}^T\boldsymbol{x}) = \frac{1}{1 + exp(-\boldsymbol{w}^T\boldsymbol{x})}$$

$$\sigma(z) = \frac{1}{1 + exp(-z)}$$: logistic function

$$p(y = 1|\boldsymbol{x}) = f(\boldsymbol{x})$$

$$p(y = 0|\boldsymbol{x}) = 1 - f(\boldsymbol{x})$$

$$p(y|\boldsymbol{x}) = f(\boldsymbol{x})^y(1 - f(\boldsymbol{x}))^{1-y}$$



Logistic Regression

- How to estimate the parameters?
- To maximize the likelihood

$$L(\boldsymbol{w}) = p(\boldsymbol{y}|X;\boldsymbol{w}) = \prod_{i=1}^{n} f(\boldsymbol{x}^{(i)})^{y^{(i)}} \left\{1 - f(\boldsymbol{x}^{(i)})\right\}^{1-y^{(i)}}$$

$$l(\boldsymbol{w}) = lnL(\boldsymbol{w}) = \sum_{i=1}^{n} \left\{y^{(i)} \ln f(\boldsymbol{x}^{(i)}) + (1 - y^{(i)}) \ln(1 - f(\boldsymbol{x}^{(i)}))\right\}$$

- Batch gradient ascent

$$w_j := w_j - \frac{\alpha}{n} \sum_{i=1}^{n} (f(\boldsymbol{x}^{(i)}) - y^{(i)}) x_j^{(i)}$$

Likelihood Ratio Test (LRT)

- Given the problem of classifying a given measurement $\boldsymbol{x}$, a 'reasonable' heuristic decision rule would be:
  - Choose the class that is more 'likely' given the evidence provided by the measured feature $\boldsymbol{x}$
  - Evaluate the posterior probability of each class $p(\omega_i|\boldsymbol{x})$ and choose the class with the largest $p(\omega_i|\boldsymbol{x})$
- Let us examine the consequences of this decision rule for a 2-class problem
  - the decision rule becomes

$$if\ p(\omega_1|\boldsymbol{x}) > p(\omega_2|\boldsymbol{x})\ choose\ \omega_1$$
$$otherwise\ choose\ \omega_2$$

  - Likelihood ratio test

$$\boxed{\frac{p(\boldsymbol{x}|\omega_1)}{p(\boldsymbol{x}|\omega_2)} \underset{\omega_2}{\overset{\omega_1}{\gtrless}} \frac{p(\omega_2)}{p(\omega_1)}}$$

Likelihood Ratio Test : An Example

Given a classification problem with the following class conditional densities, derive a decision rule based on the Likelihood Ratio Test (assume equal priors)

$$p(x|\omega_1) = \frac{1}{\sqrt{2\pi}} \exp -\frac{(x-4)^2}{2} \quad , \quad p(x|\omega_2) = \frac{1}{\sqrt{2\pi}} \exp -\frac{(x-10)^2}{2}$$

Solution :

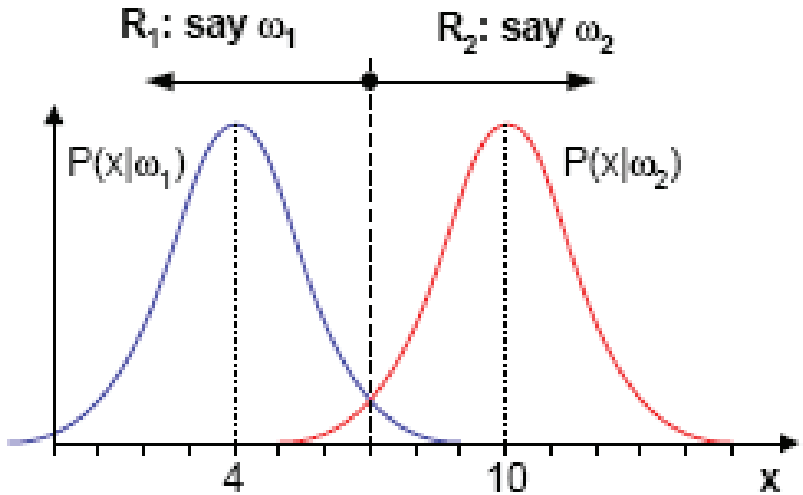
Likelihood Ratio Test : An Example

Given a classification problem with the following class conditional densities, derive a decision rule based on the Likelihood Ratio Test (assume equal priors)

$$p(x|\omega_1) = \frac{1}{\sqrt{2\pi}} \exp -\frac{(x-4)^2}{2} \quad , \quad p(x|\omega_2) = \frac{1}{\sqrt{2\pi}} \exp -\frac{(x-10)^2}{2}$$

Solution :

$$x \lessgtr_{\omega_2}^{\omega_1} 7$$



probability of error

The performance of any decision rule can be measured by its **probability of error** $p(error)$

$$p(error) = \sum_{i=1}^{K} p(error|\omega_i)p(\omega_i)$$

The class conditional probability of error $p(error|\omega_i)$ can be expressed as

$$p(error|\omega_i) = p(choose\ \omega_j|\omega_i) = \int_{R_j} p(\boldsymbol{x}|\omega_i)d\boldsymbol{x}$$

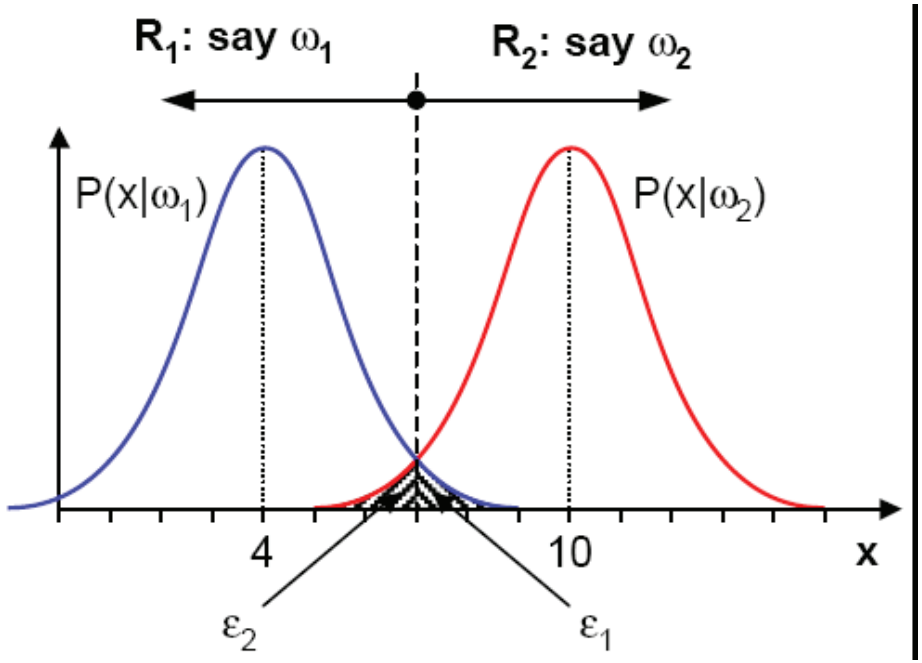
for our 2-class problem

$$\begin{aligned} p(error) &= p(\boldsymbol{x} \in R_2, \omega_1) + p(\boldsymbol{x} \in R_1, \omega_2) \\ &= p(\boldsymbol{x} \in R_2|\omega_1)p(\omega_1) + p(\boldsymbol{x} \in R_1|\omega_2)p(\omega_2) \\ &= \int_{R_2} p(\boldsymbol{x}|\omega_1)p(\omega_1)d\boldsymbol{x} + \int_{R_1} p(\boldsymbol{x}|\omega_2)p(\omega_2)d\boldsymbol{x} \end{aligned}$$

probability of error

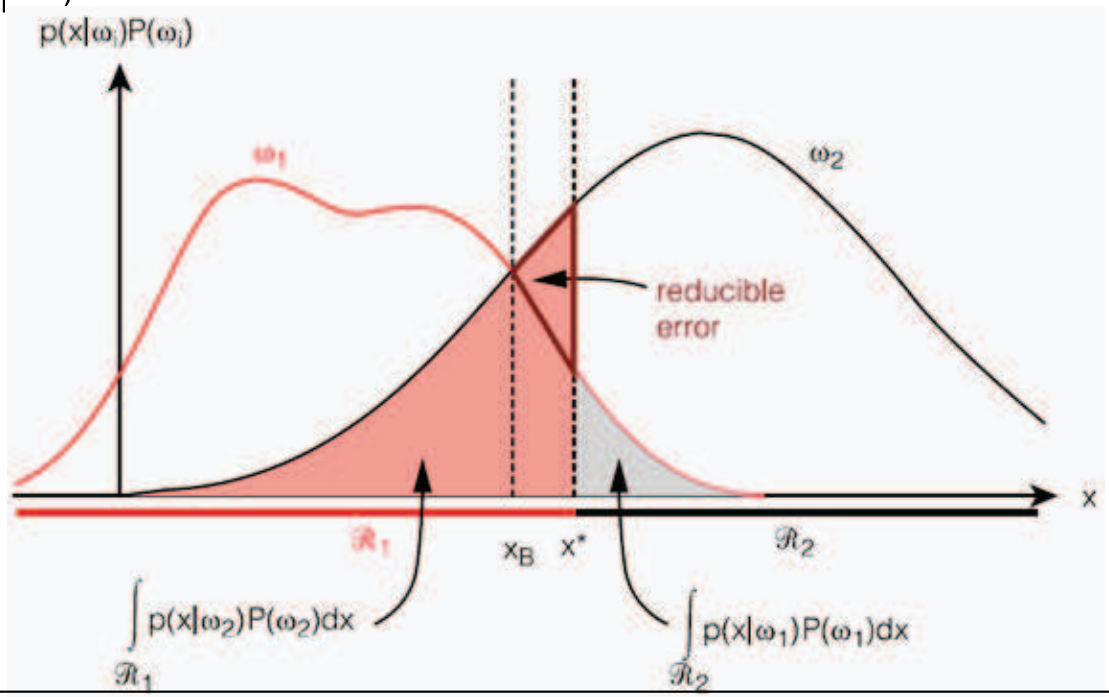
$$p(error) = p(\omega_1) \underbrace{\int_{R_2} p(\boldsymbol{x}|\omega_1)d\boldsymbol{x}}_{\epsilon_1} + p(\omega_2) \underbrace{\int_{R_1} p(\boldsymbol{x}|\omega_2)d\boldsymbol{x}}_{\epsilon_2}$$

Since we assumed equal priors, then $p(error) = (\epsilon_1 + \epsilon_2)/2$



probability of error

- How good is the Likelihood Ratio Test decision rule?
- The optimal decision rule will minimize $p(error|\boldsymbol{x})$ for every value of $\boldsymbol{x}$, so that the integral above is minimized.
- From the figure it becomes clear that, for any value of $\boldsymbol{x'}$, the Likelihood Ratio Test decision rule will always have a lower $p(error|\boldsymbol{x'})$



Bayes Risk

- Is the penalty of misclassifying a class ω_1 example as class ω_2 the same as the reciprocal?
  - misclassifying a cancer sufferer as a healthy patient
- Bayes Risk
  - the expected value of the overall risk for a 2-class problem

$$R = \int_{R_1} R(\alpha_1|\boldsymbol{x})p(\boldsymbol{x})d\boldsymbol{x} + \int_{R_2} R(\alpha_2|\boldsymbol{x})p(\boldsymbol{x})d\boldsymbol{x}$$

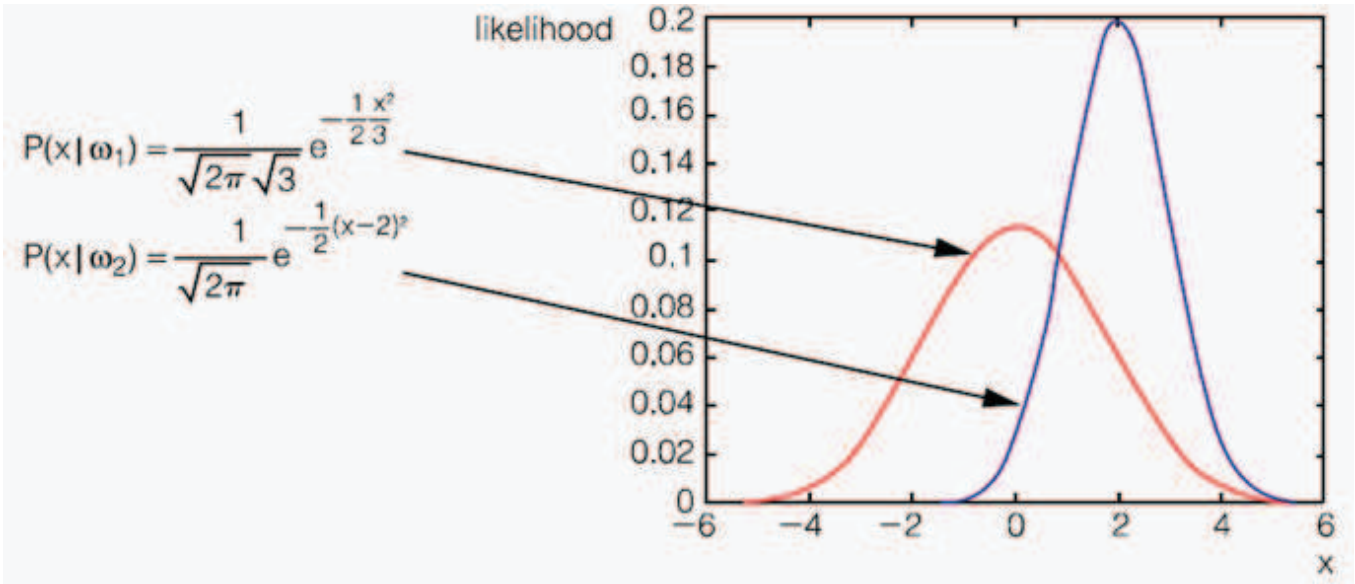
$$= \int_{R_1} (c_{11}p(\omega_1|\boldsymbol{x}) + c_{12}p(\omega_2|\boldsymbol{x}))p(\boldsymbol{x})d\boldsymbol{x} + \int_{R_2} (c_{21}p(\omega_1|\boldsymbol{x}) + c_{22}p(\omega_2|\boldsymbol{x}))p(\boldsymbol{x})d\boldsymbol{x}$$

Choose ω_1 if $R(\alpha_1|\boldsymbol{x}) < R(\alpha_2|\boldsymbol{x})$.

$$\Lambda(\boldsymbol{x}) = \frac{p(\boldsymbol{x}|\omega_1)}{p(\boldsymbol{x}|\omega_2)} \gtrless_{\omega_2}^{\omega_1} \frac{(c_{12} - c_{22})P(\omega_2)}{(c_{21} - c_{11})P(\omega_1)}$$

Bayes Risk : An Example

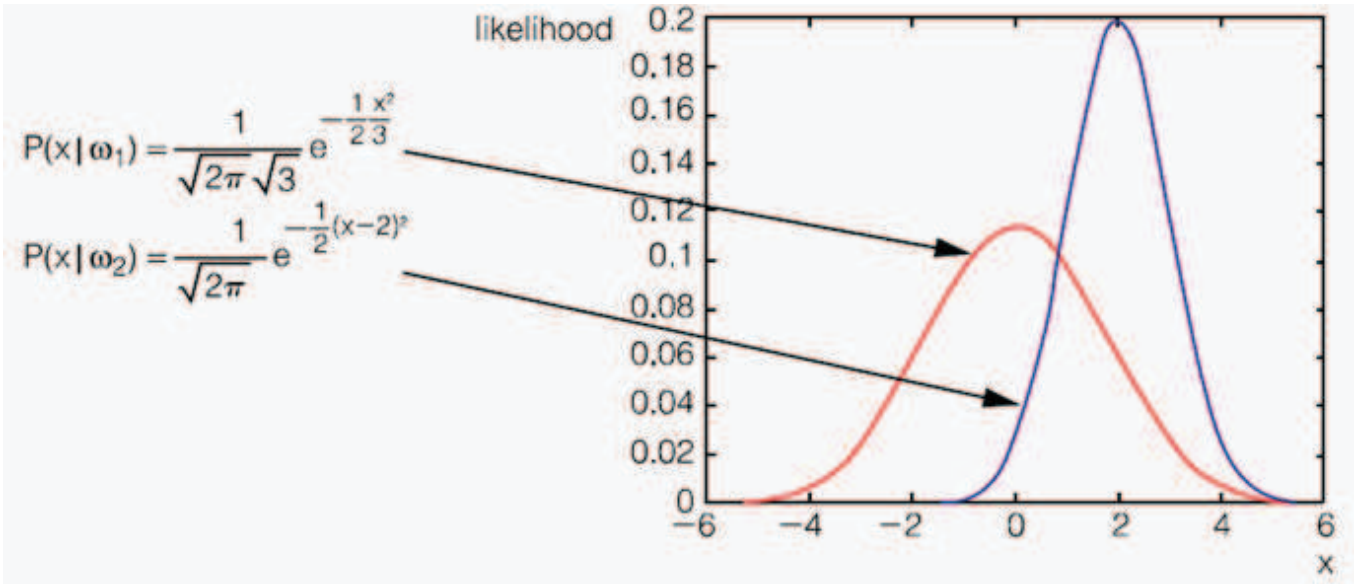
- Consider a classification problem with two classes defined by the following likelihood functions
  - What is the likelihood ratio?
  - When $p(\omega_1) = p(\omega_2) = 0.5$, $c_{11} = c_{22} = 0$, $c_{12} = 1$ and $c_{21} = 3^{1/2}$, determine a decision rule that minimizes the bayes risk.



Solution:

Bayes Risk : An Example

- Consider a classification problem with two classes defined by the following likelihood functions
  - What is the likelihood ratio?
  - When $p(\omega_1) = p(\omega_2) = 0.5$, $c_{11} = c_{22} = 0$, $c_{12} = 1$ and $c_{21} = 3^{1/2}$, determine a decision rule that minimizes the bayes risk.



Solution: If $1.27 < x < 4.73$, x belongs to ω_2.

Decision Criteria

- Bayes Criterion

$$\Lambda(\boldsymbol{x}) = \frac{p(\boldsymbol{x}|\omega_1)}{p(\boldsymbol{x}|\omega_2)} \gtrless_{\omega_2}^{\omega_1} \frac{(c_{12} - c_{22})P(\omega_2)}{(c_{21} - c_{11})P(\omega_1)}$$

- Maximum A Posteriori (MAP) Criterion

$$\Lambda(\boldsymbol{x}) = \frac{p(\boldsymbol{x}|\omega_1)}{p(\boldsymbol{x}|\omega_2)} \gtrless_{\omega_2}^{\omega_1} \frac{P(\omega_2)}{P(\omega_1)} \iff \frac{P(\omega_1|\boldsymbol{x})}{P(\omega_2|\boldsymbol{x})} \gtrless_{\omega_2}^{\omega_1} 1$$

- Maximum Likelihood (ML) Criterion

$$\Lambda(\boldsymbol{x}) = \frac{p(\boldsymbol{x}|\omega_1)}{p(\boldsymbol{x}|\omega_2)} \gtrless_{\omega_2}^{\omega_1} 1$$

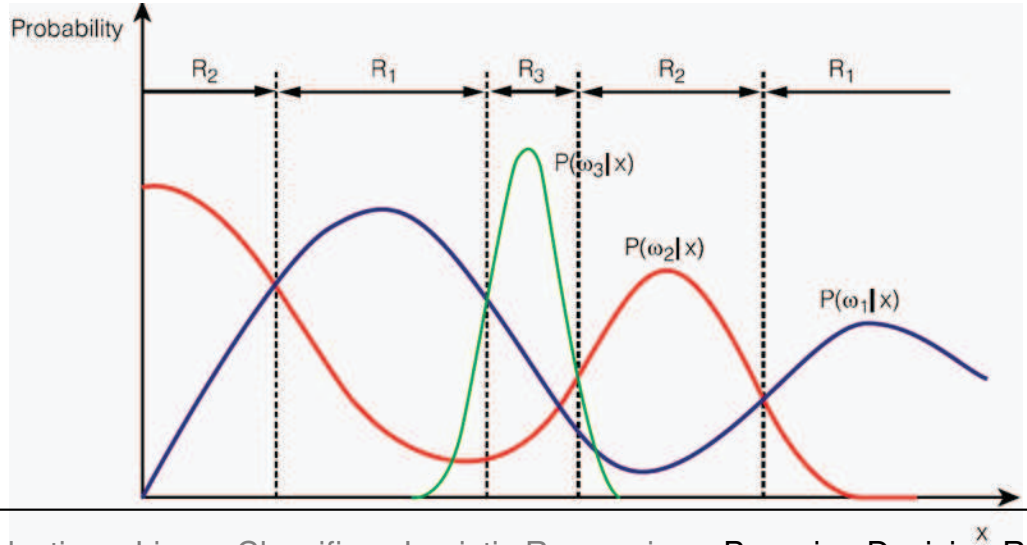
multi-class problems : probability of error

- The decision rule that minimizes p(error) generalizes very easily to multi-class problems

$$p(error) = 1 - p(correct)$$

- The problem of minimizing p(error) is equivalent to that of maximizing p(correct).

$$p(correct) = \sum_{i=1}^{K} p(\omega_i) \int_{R_i} p(\boldsymbol{x}|\omega_i) d\boldsymbol{x}$$



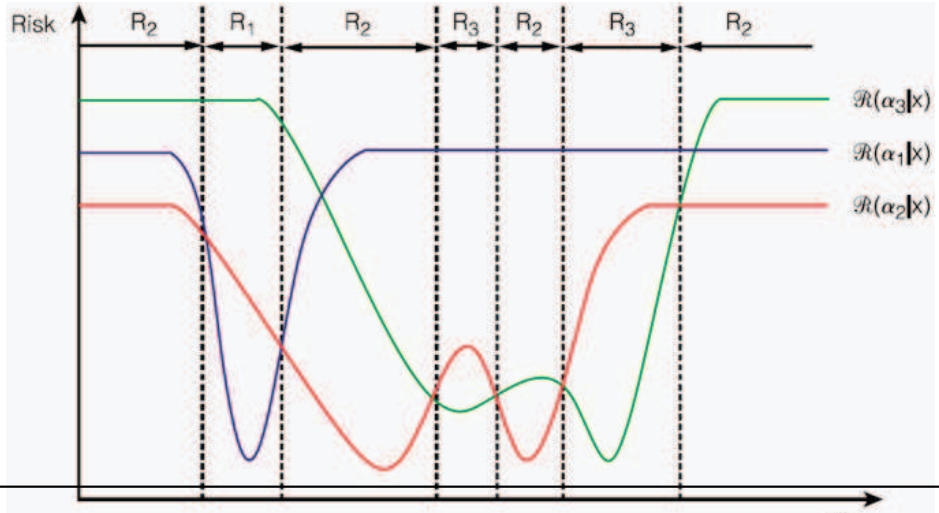
multi-class problems : Bayes Risk

To determine which decision rule yields the minimum Bayes Risk for the multi-class problem

$$R = \int R(\alpha(\boldsymbol{x})|\boldsymbol{x})p(\boldsymbol{x})d\boldsymbol{x}$$

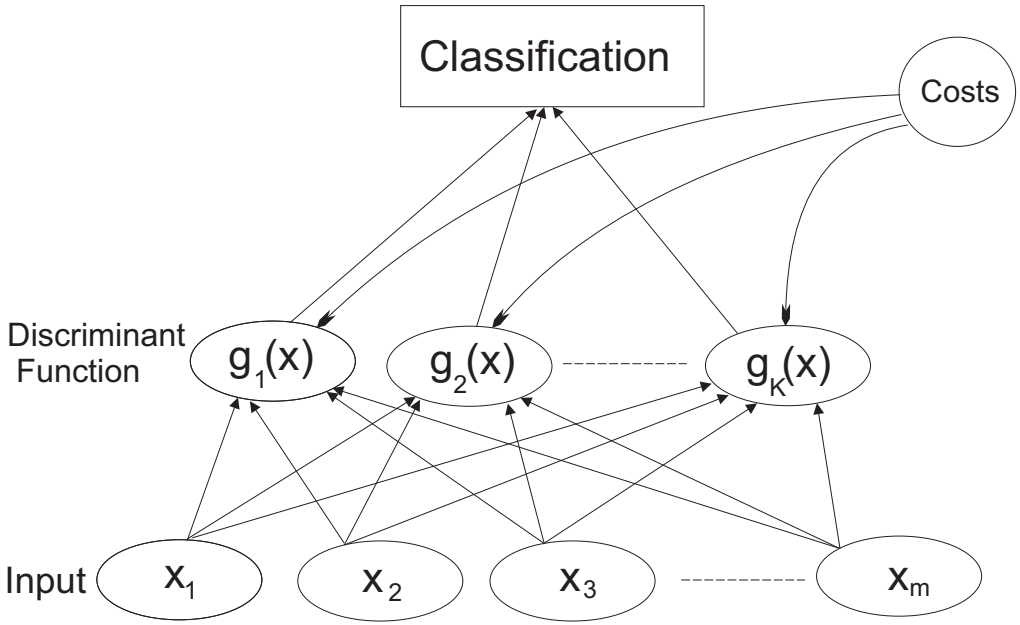
$$R(\alpha_i|\boldsymbol{x}) = \sum_{j=1}^{K} c_{ij}p(\omega_j|\boldsymbol{x})$$

Choose ω_i which has the minimum $R(\alpha_i|\boldsymbol{x})$



Discriminant functions

Assign $\boldsymbol{x}$ to class ω_i if $g_i(\boldsymbol{x}) > g_j(\boldsymbol{x}), \forall j \neq i$.

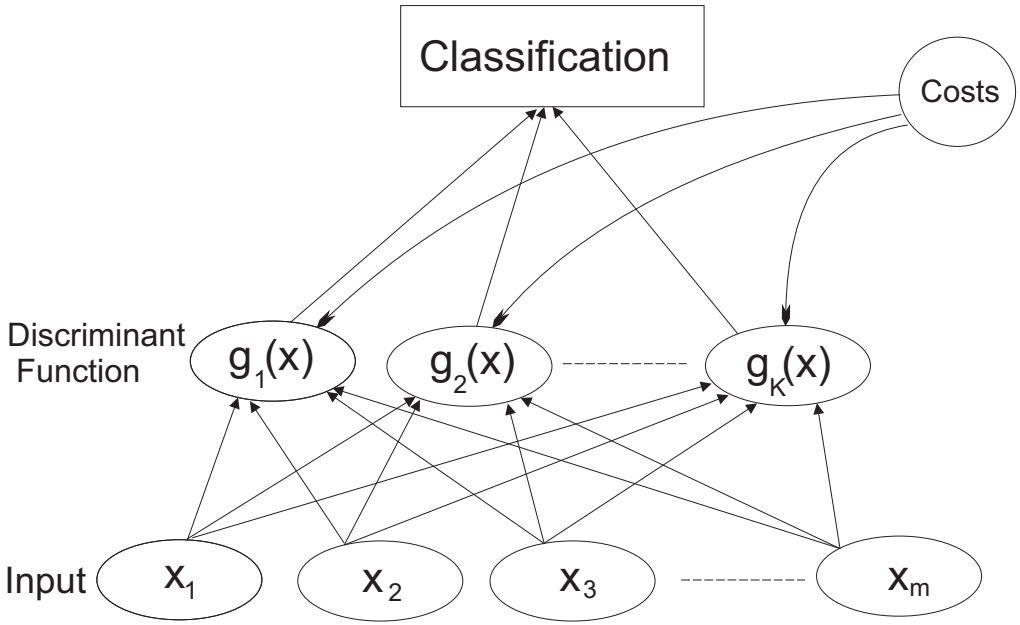


Three basic decision rules

Criterion	Discriminant function
Bayes	
MAP	
ML	

Discriminant functions

Assign $\boldsymbol{x}$ to class ω_i if $g_i(\boldsymbol{x}) > g_j(\boldsymbol{x}), \forall j \neq i$.



Three basic decision rules

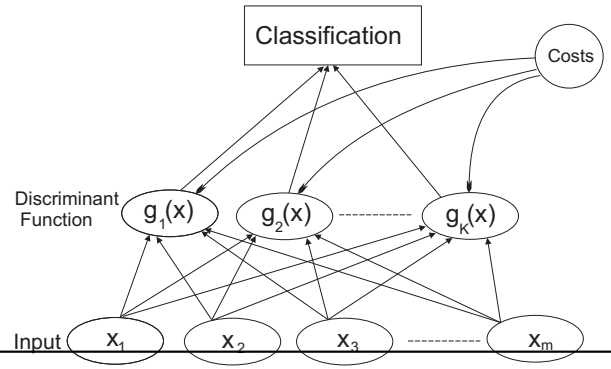
Criterion	Discriminant function
Bayes	$g_i(x) = -R(\alpha_i \vert \boldsymbol{x})$
MAP	$g_i(x) = p(\omega_i \vert \boldsymbol{x})$
ML	$g_i(x) = p(\boldsymbol{x} \vert \omega_i)$

Discriminant Functions for the Normal Density

- Decision rule based on the MAP discriminant function to choose ω_i if $g_i(\boldsymbol{x}) > g_j(\boldsymbol{x}), \forall j \neq i$ where $g_i(\boldsymbol{x}) = p(\omega_i|\boldsymbol{x})$
- Gaussian Discriminant Analysis
- Multivariate Gaussian distribution

$$p(\boldsymbol{x}) = \frac{1}{(2\pi)^{m/2}\,|\boldsymbol{\Sigma}|^{1/2}} \exp\left[-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right]$$

$$\boxed{g_k(\boldsymbol{x}) = -\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_k)^T\boldsymbol{\Sigma}_k^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_k) - \frac{1}{2}\ln|\boldsymbol{\Sigma}_k| + \ln p(\omega_k)} \quad (1)$$



Case 1 : $\Sigma_k = \sigma^2 \boldsymbol{I}$

- This situation occurs when the features are statistically independent with the same variance for all classes

$$g_k(\boldsymbol{x}) = -\frac{1}{2\sigma^2}(-2\boldsymbol{\mu}_k^T\boldsymbol{x} + \boldsymbol{\mu}_k^T\boldsymbol{\mu}_k) + \ln p(\omega_k) = \boldsymbol{w}_k^T\boldsymbol{x} + w_{k0}$$

$$\text{where} \begin{cases} \boldsymbol{w}_k = \\ \\ w_{k0} = \end{cases}$$

- If equal priors, a minimum-distance or nearest mean classifier

$$g_k(\boldsymbol{x}) = -\frac{1}{2\sigma^2}(\boldsymbol{x} - \boldsymbol{\mu}_i)^T(\boldsymbol{x} - \boldsymbol{\mu}_i)$$

Case 1 : $\Sigma_k = \sigma^2 \boldsymbol{I}$

- This situation occurs when the features are statistically independent with the same variance for all classes

$$g_k(\boldsymbol{x}) = -\frac{1}{2\sigma^2}(-2\boldsymbol{\mu}_k^T\boldsymbol{x} + \boldsymbol{\mu}_k^T\boldsymbol{\mu}_k) + \ln p(\omega_k) = \boldsymbol{w}_k^T\boldsymbol{x} + w_{k0}$$

$$\text{where} \begin{cases} \boldsymbol{w}_k = \frac{1}{\sigma^2}\boldsymbol{\mu}_k \\ \\ w_{k0} = -\frac{1}{2\sigma^2}\boldsymbol{\mu}_k^T\boldsymbol{\mu}_k + \ln p(\omega_k) \end{cases}$$

- If equal priors, a minimum-distance or nearest mean classifier

$$g_k(\boldsymbol{x}) = -\frac{1}{2\sigma^2}(\boldsymbol{x} - \boldsymbol{\mu}_k)^T(\boldsymbol{x} - \boldsymbol{\mu}_k)$$

Case 1 : $\Sigma_k = \sigma^2 I$, An example

Compute the decision boundaries for a 3-class, 2-dimensional problem with the following class mean vectors and covariance matrices and equal priors

$$\mu_1 = \begin{bmatrix} 3 & 2 \end{bmatrix}^T \quad \mu_2 = \begin{bmatrix} 7 & 4 \end{bmatrix}^T \quad \mu_3 = \begin{bmatrix} 2 & 5 \end{bmatrix}^T$$

$$\Sigma_1 = \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix} \quad \Sigma_2 = \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix} \quad \Sigma_3 = \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix}$$

Case 1 : $\Sigma_k = \sigma^2 \boldsymbol{I}$, An example

Compute the decision boundaries for a 3-class, 2-dimensional problem with the following class mean vectors and covariance matrices and equal priors

$$\mu_1 = \begin{bmatrix} 3 & 2 \end{bmatrix}^T \quad \mu_2 = \begin{bmatrix} 7 & 4 \end{bmatrix}^T \quad \mu_3 = \begin{bmatrix} 2 & 5 \end{bmatrix}^T$$

$$\Sigma_1 = \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix} \quad \Sigma_2 = \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix} \quad \Sigma_3 = \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix}$$

Case 2 : $\Sigma_k = \Sigma$ (diagonal)

- The classes still have the same covariance matrix, but the features are allowed to have different variances

$$g_k(\boldsymbol{x}) = -\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{x} - \boldsymbol{\mu}_k) + \ln p(\omega_k)$$

Case 2 : $\Sigma_k = \Sigma$ (diagonal), An example

Compute the decision boundaries for a 3-class, 2-dimensional problem with the following class mean vectors and covariance matrices and equal priors

$$\mu_1 = \begin{bmatrix} 3 & 2 \end{bmatrix}^T \quad \mu_2 = \begin{bmatrix} 5 & 4 \end{bmatrix}^T \quad \mu_3 = \begin{bmatrix} 2 & 5 \end{bmatrix}^T$$

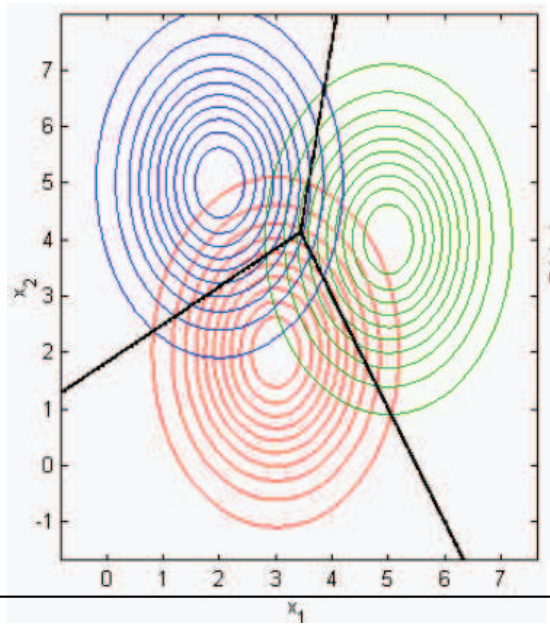
$$\Sigma_1 = \begin{bmatrix} 1 & 0 \\ 0 & 2 \end{bmatrix} \quad \Sigma_2 = \begin{bmatrix} 1 & 0 \\ 0 & 2 \end{bmatrix} \quad \Sigma_3 = \begin{bmatrix} 1 & 0 \\ 0 & 2 \end{bmatrix}$$

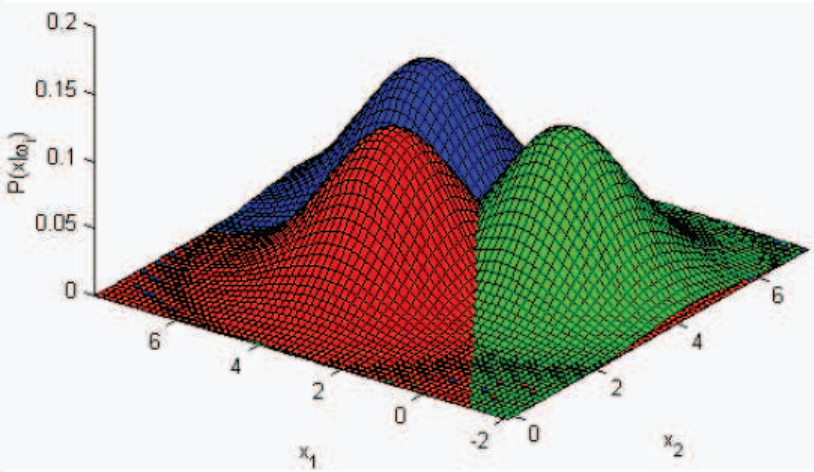
Case 2 : $\Sigma_k = \Sigma$ (diagonal), An example

Compute the decision boundaries for a 3-class, 2-dimensional problem with the following class mean vectors and covariance matrices and equal priors

$$\mu_1 = \begin{bmatrix} 3 & 2 \end{bmatrix}^T \quad \mu_2 = \begin{bmatrix} 5 & 4 \end{bmatrix}^T \quad \mu_3 = \begin{bmatrix} 2 & 5 \end{bmatrix}^T$$

$$\Sigma_1 = \begin{bmatrix} 1 & 0 \\ 0 & 2 \end{bmatrix} \quad \Sigma_2 = \begin{bmatrix} 1 & 0 \\ 0 & 2 \end{bmatrix} \quad \Sigma_3 = \begin{bmatrix} 1 & 0 \\ 0 & 2 \end{bmatrix}$$





Case 3 : $\Sigma_k = \Sigma$ (non-diagonal)

- The case that all the classes have the same covariance matrix, but this is no longer diagonal
- The quadratic discriminant becomes

$$g_k(\boldsymbol{x}) = -\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{x} - \boldsymbol{\mu}_k) + \ln p(\omega_k)$$

$$g_k(\boldsymbol{x}) = \boldsymbol{w}_k^T \boldsymbol{x} + w_{k0}$$

$$\text{where} \begin{cases} \boldsymbol{w}_k = \\ \\ w_{k0} = \end{cases}$$

- If equal priors, a Mahalanobis distance classifier

$$g_k(\boldsymbol{x}) = -\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{x} - \boldsymbol{\mu}_k)$$

Case 3 : $\Sigma_k = \Sigma$ (non-diagonal)

- The case that all the classes have the same covariance matrix, but this is no longer diagonal
- The quadratic discriminant becomes

$$g_k(\boldsymbol{x}) = -\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{x} - \boldsymbol{\mu}_k) + \ln p(\omega_k)$$

$$g_k(\boldsymbol{x}) = \boldsymbol{w}_k^T \boldsymbol{x} + w_{k0}$$

$$\text{where} \begin{cases} \boldsymbol{w}_k = \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_k \\ \\ w_{k0} = -\frac{1}{2}\boldsymbol{\mu}_k^T \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_k + \ln p(\omega_k) \end{cases}$$

- If equal priors, a Mahalanobis distance classifier

$$g_k(\boldsymbol{x}) = -\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{x} - \boldsymbol{\mu}_k)$$

Case 3 : $\Sigma_k = \Sigma$ (non-diagonal), An example

Compute the decision boundaries for a 3-class, 2-dimensional problem with the following class mean vectors and covariance matrices and equal priors

$$\mu_1 = \begin{bmatrix} 3 & 2 \end{bmatrix}^T \quad \mu_2 = \begin{bmatrix} 5 & 4 \end{bmatrix}^T \quad \mu_3 = \begin{bmatrix} 2 & 5 \end{bmatrix}^T$$

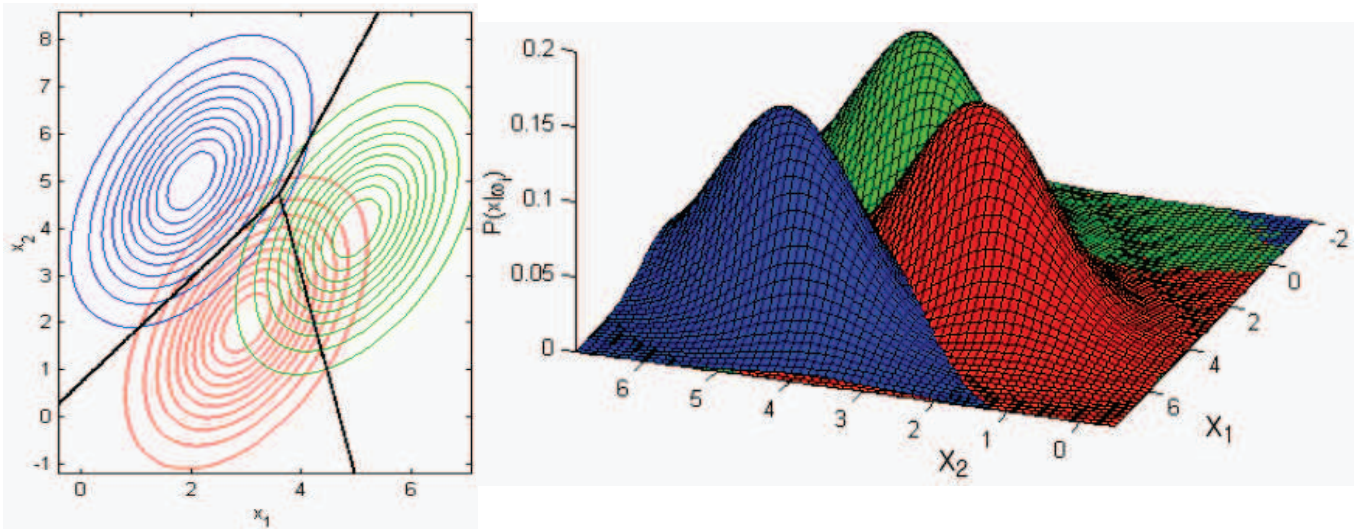
$$\Sigma_1 = \begin{bmatrix} 1 & 0.7 \\ 0.7 & 2 \end{bmatrix} \quad \Sigma_2 = \begin{bmatrix} 1 & 0.7 \\ 0.7 & 2 \end{bmatrix} \quad \Sigma_3 = \begin{bmatrix} 1 & 0.7 \\ 0.7 & 2 \end{bmatrix}$$

Case 3 : $\Sigma_k = \Sigma$ (non-diagonal), An example

Compute the decision boundaries for a 3-class, 2-dimensional problem with the following class mean vectors and covariance matrices and equal priors

$$\mu_1 = \begin{bmatrix} 3 & 2 \end{bmatrix}^T \quad \mu_2 = \begin{bmatrix} 5 & 4 \end{bmatrix}^T \quad \mu_3 = \begin{bmatrix} 2 & 5 \end{bmatrix}^T$$

$$\Sigma_1 = \begin{bmatrix} 1 & 0.7 \\ 0.7 & 2 \end{bmatrix} \quad \Sigma_2 = \begin{bmatrix} 1 & 0.7 \\ 0.7 & 2 \end{bmatrix} \quad \Sigma_3 = \begin{bmatrix} 1 & 0.7 \\ 0.7 & 2 \end{bmatrix}$$



Case 4 : $\Sigma_k = \sigma_k^2 \boldsymbol{I}$

- In this case, each class has a different covariance matrix, which is proportional to the identity matrix
- The quadratic discriminant becomes

$$
\begin{aligned}
g_k(\boldsymbol{x}) &= -\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}_{\boldsymbol{k}}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_k) - \frac{1}{2}\ln|\boldsymbol{\Sigma}_k| + \ln p(\omega_k) \\
&= -\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_k)^T \boldsymbol{\sigma}_{\boldsymbol{k}}^{-2}(\boldsymbol{x}-\boldsymbol{\mu}_k) - \frac{m}{2}\ln(\boldsymbol{\sigma}_k^2) + \ln p(\omega_k)
\end{aligned}
$$

Case 4 : $\Sigma_k = \sigma_k^2 \boldsymbol{I}$, An example

Compute the decision boundaries for a 3-class, 2-dimensional problem with the following class mean vectors and covariance matrices and equal priors

$$\mu_1 = \begin{bmatrix} 3 & 2 \end{bmatrix}^T \quad \mu_2 = \begin{bmatrix} 5 & 4 \end{bmatrix}^T \quad \mu_3 = \begin{bmatrix} 2 & 5 \end{bmatrix}^T$$

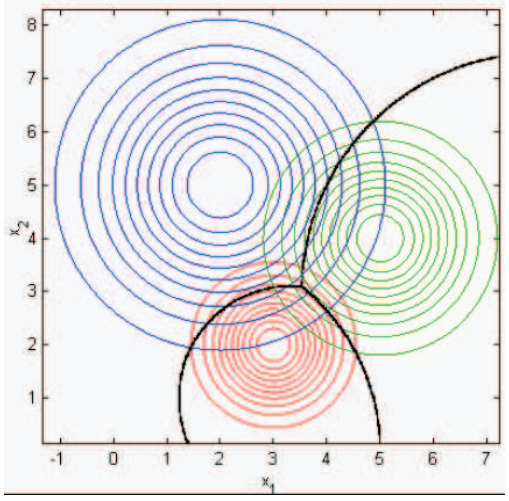
$$\Sigma_1 = \begin{bmatrix} 0.5 & 0 \\ 0 & 0.5 \end{bmatrix} \quad \Sigma_2 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \quad \Sigma_3 = \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix}$$

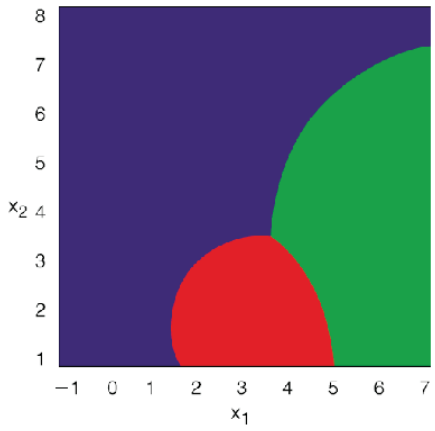
Case 4 : $\Sigma_k = \sigma_k^2 I$, An example

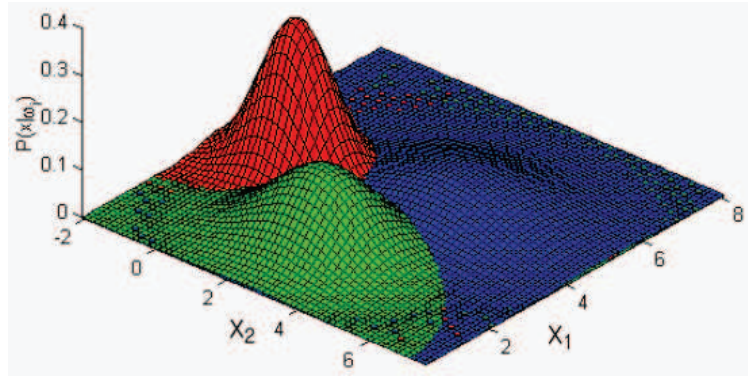
Compute the decision boundaries for a 3-class, 2-dimensional problem with the following class mean vectors and covariance matrices and equal priors

$$\mu_1 = \begin{bmatrix} 3 & 2 \end{bmatrix}^T \quad \mu_2 = \begin{bmatrix} 5 & 4 \end{bmatrix}^T \quad \mu_3 = \begin{bmatrix} 2 & 5 \end{bmatrix}^T$$

$$\Sigma_1 = \begin{bmatrix} 0.5 & 0 \\ 0 & 0.5 \end{bmatrix} \quad \Sigma_2 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \quad \Sigma_3 = \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix}$$







Case 5 : Σ_k = arbitrary

$$g_k(\boldsymbol{x}) = -\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_k)^T\boldsymbol{\Sigma}_k^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_k) - \frac{1}{2}\ln|\boldsymbol{\Sigma}_k| + \ln p(\omega_k)$$

Reorganizing terms in a quadratic form yields

$$g_k(\boldsymbol{x}) = \boldsymbol{x}^T\boldsymbol{W}_k\boldsymbol{x} + \boldsymbol{w}_k^T\boldsymbol{x} + w_{k0}$$

$$\text{where} \begin{cases} \boldsymbol{W}_k = \\ \boldsymbol{w}_k = \\ w_{k0} = \end{cases}$$

Case 5 : $\Sigma_k = \text{arbitrary}$

$$g_k(\boldsymbol{x}) - \frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}_k^{-1}(\boldsymbol{x} - \boldsymbol{\mu}_k) - \frac{1}{2}\ln|\boldsymbol{\Sigma}_k| + \ln p(\omega_k)$$

Reorganizing terms in a quadratic form yields

$$g_k(\boldsymbol{x}) = \boldsymbol{x}^T \boldsymbol{W}_k \boldsymbol{x} + \boldsymbol{w}_k^T \boldsymbol{x} + w_{k0}$$

$$\text{where} \begin{cases} \boldsymbol{W}_k = -\frac{1}{2}\boldsymbol{\Sigma}_k^{-1} \\ \boldsymbol{w}_k = \boldsymbol{\Sigma}_k^{-1}\boldsymbol{\mu}_k \\ w_{k0} = -\frac{1}{2}\boldsymbol{\mu}_k^T \boldsymbol{\Sigma}_k^{-1}\boldsymbol{\mu}_k - \frac{1}{2}\ln|\boldsymbol{\Sigma}_k| + \ln p(\omega_k) \end{cases}$$

Case 5 : $\Sigma_k = \text{arbitrary}$, An example

Compute the decision boundaries for a 3-class, 2-dimensional problem with the following class mean vectors and covariance matrices and equal priors

$$\mu_1 = \begin{bmatrix} 3 & 2 \end{bmatrix}^T \quad \mu_2 = \begin{bmatrix} 5 & 4 \end{bmatrix}^T \quad \mu_3 = \begin{bmatrix} 2 & 5 \end{bmatrix}^T$$

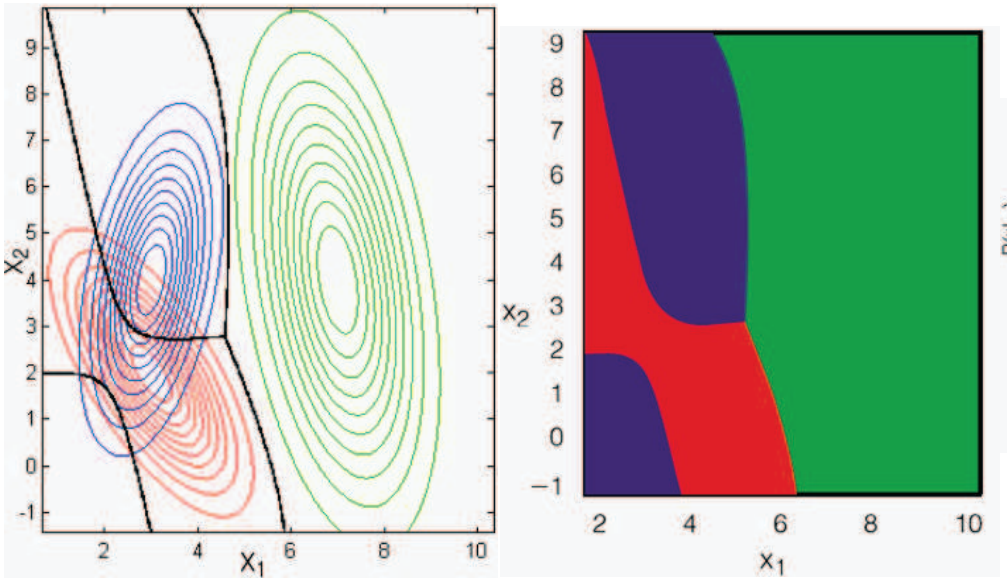
$$\Sigma_1 = \begin{bmatrix} 1 & -1 \\ -1 & 2 \end{bmatrix} \quad \Sigma_2 = \begin{bmatrix} 1 & -1 \\ -1 & 7 \end{bmatrix} \quad \Sigma_3 = \begin{bmatrix} 0.5 & 0.5 \\ 0.5 & 3 \end{bmatrix}$$

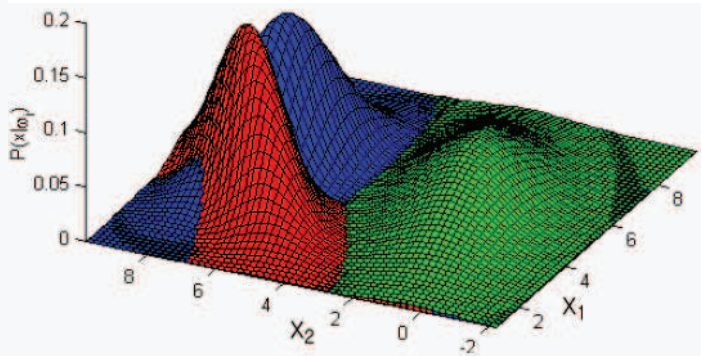
Case 5 : $\Sigma_k = \text{arbitrary}$, An example

Compute the decision boundaries for a 3-class, 2-dimensional problem with the following class mean vectors and covariance matrices and equal priors

$$\mu_1 = \begin{bmatrix} 3 & 2 \end{bmatrix}^T \quad \mu_2 = \begin{bmatrix} 5 & 4 \end{bmatrix}^T \quad \mu_3 = \begin{bmatrix} 2 & 5 \end{bmatrix}^T$$

$$\Sigma_1 = \begin{bmatrix} 1 & -1 \\ -1 & 2 \end{bmatrix} \quad \Sigma_2 = \begin{bmatrix} 1 & -1 \\ -1 & 7 \end{bmatrix} \quad \Sigma_3 = \begin{bmatrix} 0.5 & 0.5 \\ 0.5 & 3 \end{bmatrix}$$





An example

- Given 3-dimensional 2-classes with following mean vectors, covariance matrices, and priors

$$\mu_1 = \begin{bmatrix} 0 & 0 & 0 \end{bmatrix}^T \quad \mu_2 = \begin{bmatrix} 1 & 1 & 1 \end{bmatrix}^T$$

$$\Sigma_1 = \Sigma_2 = \begin{bmatrix} 1/4 & 0 & 0 \\ 0 & 1/4 & 0 \\ 0 & 0 & 1/4 \end{bmatrix} \quad p(\omega_2) = 2p(\omega_1)$$

- Which class does this vector $\boldsymbol{x} = \begin{bmatrix} 0.1 & 0.7 & 0.8 \end{bmatrix}^T$ belong to?

Solution :

An example

- Given 3-dimensional 2-classes with following mean vectors, covariance matrices, and priors

$$\mu_1 = \begin{bmatrix} 0 & 0 & 0 \end{bmatrix}^T \quad \mu_2 = \begin{bmatrix} 1 & 1 & 1 \end{bmatrix}^T$$

$$\Sigma_1 = \Sigma_2 = \begin{bmatrix} 1/4 & 0 & 0 \\ 0 & 1/4 & 0 \\ 0 & 0 & 1/4 \end{bmatrix} \quad p(\omega_2) = 2p(\omega_1)$$

- Which class does this vector $\boldsymbol{x} = \begin{bmatrix} 0.1 & 0.7 & 0.8 \end{bmatrix}^T$ belong to?

Solution :

$$\boldsymbol{x} \in \omega_2$$

Sign Language Recognition

- American Sign Language recognition (10 signs)
- Features: 22 hand joint angles, 3D hand position, 3D hand orientation, 5 finger forces
- Comparison of naive Bayesian Classifier, Decision Tree, Neural Network

C. Shahabi, L. Kaghazian, S. Mehta, A. Ghoting, G. Shanbhag and M. McLaughlin, *Analysis of Haptic Data for Sign Language Recognition*, International conference on Universal Access in Human-Computer Interaction, 2001.

Sign Language Recognition

- The naive Bayesian Classifier has the highest average accuracy with 50 training examples.

C. Shahabi, L. Kaghazian, S. Mehta, A. Ghoting, G. Shanbhag and M. McLaughlin, *Analysis of Haptic Data for Sign Language Recognition*, International conference on Universal Access in Human-Computer Interaction, 2001.

In Door Place Classification

- Determining whether a cleaning robot (with bumper and vision sensors) is in a hall, a room, or a corridor
- From vision sensing an orientation histogram I is created, by grouping straight lines with a similar orientation.

C. Yi, Y. C. Oh, I. H. Suh and B.-U. Choi, *In door Place Classification Using Robot Behavior and Vision Data*, International Journal of Advanced Robotic Systems, 2011.

In Door Place Classification

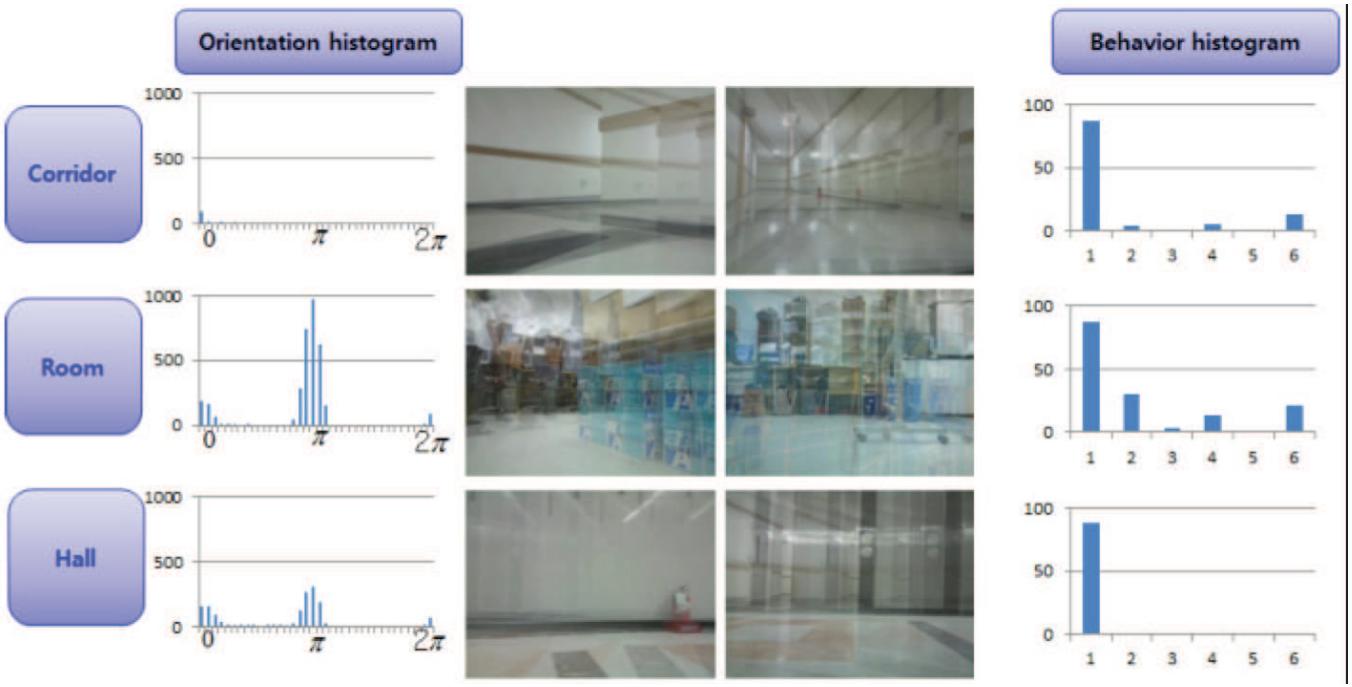
- From robot behavioral data a behavioral histogram B are calculated.
- From visual and behavioral histograms, $p(\omega_i|I, B)$, the probability that a place belongs to specific class, is calculated.

C. Yi, Y. C. Oh, I. H. Suh and B.-U. Choi, *In door Place Classification Using Robot Behavior and Vision Data*, International Journal of Advanced Robotic Systems, 2011.

Mobile Robot Navigation

$$p(\omega_i|I,B) = \frac{p(\omega_i)p(I,B|\omega_i)}{p(I,B)}$$

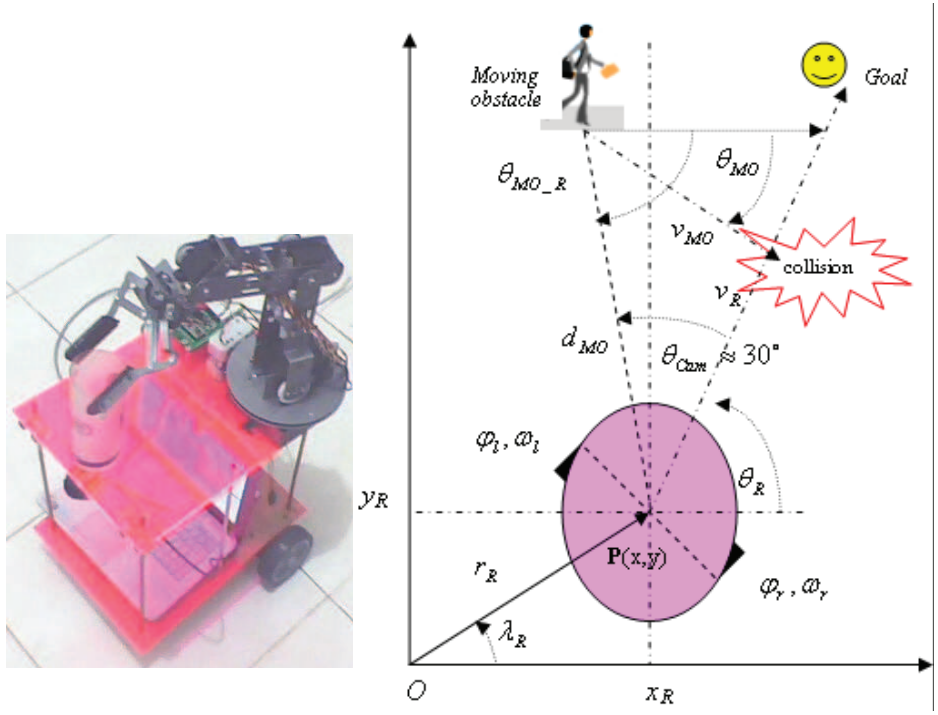
$$p(\omega_i)p(I,B|\omega_i) = p(\omega_i)p(I|\omega_i)p(B|\omega_i)$$



Application : Moving obstacle avoidance

Reference: W. Budiharto, D. Purwanto and A. Jazidie, *A Robust Obstacle Avoidance for Service Robot Using Bayesian Approach* , 2011.

The obstacle avoidance problem is formulated using decision theory, prior and posterior distribution and loss function to determine an optimal response based on inaccurate sensor data.



Moving obstacle avoidance using Bayesian approach

Reference: W. Budiharto, D. Purwanto and A. Jazidie, *A Robust Obstacle Avoidance for Service Robot Using Bayesian Approach* , 2011.

- $\Theta = (\theta_1,\ \theta_2) = (Obstacle,\ noObstacle)$ are two possible path states.
- The probability of path state θ given observation z is $p(\theta|z) = \frac{p(z|\theta)p(\theta)}{\sum p(z|\theta)p(\theta)}$.
- Define an action space $A = (a_1,\ a_2) = (maneuver, stop)$.
- Compute bayes risk for each action a : $\sum_\theta C(\theta, a)\ p(\theta|z)$.
- Take an action which leads to smaller risk

Summary

- The Bayes classifiers for normally distributed classes is a quadratic classifiers
- The Bayes classifiers for normally distributed classes with equal covariance matrices is a linear classifier

Summary

- The Bayes classifiers for normally distributed classes is a quadratic classifiers
- The Bayes classifiers for normally distributed classes with equal covariance matrices is a linear classifier
- The minimum Mahalanobis distance classifier is optimum for normal distributed classes, equal covariance matrices and equal priors.

Summary

- The Bayes classifiers for normally distributed classes is a quadratic classifiers
- The Bayes classifiers for normally distributed classes with equal covariance matrices is a linear classifier
- The minimum Mahalanobis distance classifier is optimum for normal distributed classes, equal covariance matrices and equal priors.
- The minimum Euclidean distance classifier is optimum for normal distributed classes, equal covariance matrices proportional to the identity matrix and equal priors.

Summary

- The Bayes classifiers for normally distributed classes is a quadratic classifiers
- The Bayes classifiers for normally distributed classes with equal covariance matrices is a linear classifier
- The minimum Mahalanobis distance classifier is optimum for normal distributed classes, equal covariance matrices and equal priors.
- The minimum Euclidean distance classifier is optimum for normal distributed classes, equal covariance matrices proportional to the identity matrix and equal priors.
- Both Euclidean and Mahalanobis minimum distance classifiers are linear classifiers

Summary and Next Lecture

- Summary
  - Linear classifiers
  - Logistic regression
  - The Likelihood Ratio Test
  - The Probability of Error
  - The Bayes Risk
  - Bayes, MAP, and ML Criterion
  - Multi-class problems
  - Quadratic classifiers
  - Bayes classifiers for Normally distributed classes
- Reading : Duda, Chap. 2.1-7, Bishop Chap. 1.5, Bishop Chap. 4
- Topics for next lecture : Unsupervised Clustering