Lecture 5

- **Bayesian Decision Theory**
  - A fundamental statistical approach to quantifying the tradeoffs between various decisions using probabilities and costs that accompany such decisions. Reasoning is based on Bayes' Rule.
- **Discriminant Functions for the Gaussian Density / Quadratic Classifiers**
  - Apply the results of **Bayesian Decision Theory** to derive the discriminant functions for the case of Gaussian class-conditional probabilities.

Bayesian Decision Theory

- The Likelihood Ratio Test

Likelihood Ratio Test

- Want to classify an object based on the evidence provided by a measurement (a feature vector) $\mathbf{x}$.
- A reasonable decision rule would be - Choose the class that is most probable given $\mathbf{x}$. Or mathematically choose class i such that

$$P(\omega_i|\mathbf{x}) \geq P(\omega_j|\mathbf{x}) \quad \text{for } i = 1, \cdots, C$$

- Consider the decision rule for a 2-class problem:

$$\text{Class}\,(\mathbf{x}) = \begin{cases} \omega_1 & \text{if } P(\omega_1|\mathbf{x}) > P(\omega_2|\mathbf{x}) \\ \omega_2 & \text{if } P(\omega_1|\mathbf{x}) < P(\omega_2|\mathbf{x}) \end{cases}$$

Likelihood Ratio Test

Choose class ω_1 if $P(\omega_1|\mathbf{x}) > P(\omega_2|\mathbf{x})$,

$$\iff \frac{P(\mathbf{x}|\omega_1)P(\omega_1)}{P(\mathbf{x})} > \frac{P(\mathbf{x}|\omega_2)P(\omega_2)}{P(\mathbf{x})}, \quad \text{Bayes' Rule}$$

$$\iff P(\mathbf{x}|\omega_1)P(\omega_1) > P(\mathbf{x}|\omega_2)P(\omega_2), \quad \text{eliminate } P(\mathbf{x}) > 0$$

$$\iff \frac{P(\mathbf{x}|\omega_1)}{P(\mathbf{x}|\omega_2)} > \frac{P(\omega_2)}{P(\omega_1)}, \quad \text{as } P(\cdot) > 0$$

Let:

$$\Lambda(\mathbf{x}) = \underbrace{\frac{P(\mathbf{x}|\omega_1)}{P(\mathbf{x}|\omega_2)}}_{\text{likelihood ratio}}$$

then

Likelihood Ratio Test:

$$\text{Class}\,(\mathbf{x}) = \begin{cases} \omega_1 & \text{if } \Lambda(\mathbf{x}) > \frac{P(\omega_2)}{P(\omega_1)} \\ \omega_2 & \text{if } \Lambda(\mathbf{x}) < \frac{P(\omega_2)}{P(\omega_1)} \end{cases}$$

An example

Derive a decision rule for the 2-class problem based on the *Likelihood Ratio Test* assuming equal priors and class conditional densities:

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

Solution: Substitute the likelihoods and priors into the expressions in the LRT

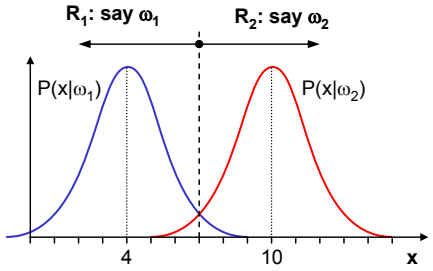
$$\Lambda(x) = \frac{(\sqrt{2\pi})^{-1} \exp\left(-.5(x-4)^2\right)}{(\sqrt{2\pi})^{-1} \exp\left(-.5(x-10)^2\right)}, \qquad \frac{P(\omega_2)}{P(\omega_1)} = \frac{.5}{.5} = 1$$

Choose class ω_1 if:

$$\begin{aligned} & \Lambda(x) > 1 \\ \iff & \exp\left(-.5(x-4)^2\right) > \exp\left(-.5(x-10)^2\right) \\ \iff & (x-4)^2 < (x-10)^2, \quad \text{by taking logs and changing signs} \\ \iff & x < 7 \end{aligned}$$

The LRT decision rule is:

$$\text{Class}\,(x) = \begin{cases} \omega_1 & \text{if } x < 7 \\ \omega_2 & \text{if } x > 7 \end{cases}$$



How would the LRT decision rule change if $P(w_1) = 2P(\omega_2)$?

Bayesian Decision Theory

- The Likelihood Ratio Test
- The Probability of Error

Probability of Error

- Performance of a decision rule is measured by its *probability of error*:

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

- The class conditional probability of error is:

$$P(\text{error}|\omega_i) = \sum_{j \neq i} P(\text{choose } \omega_j|\omega_i) = \sum_{j \neq i} \int_{\mathcal{R}_j} P(\mathbf{x}|\omega_i) d\mathbf{x}$$

  where $\mathcal{R}_j = \{\mathbf{x} : \text{Class}\,(\mathbf{x}) = \omega_j\}$.
- For the 2-class problem

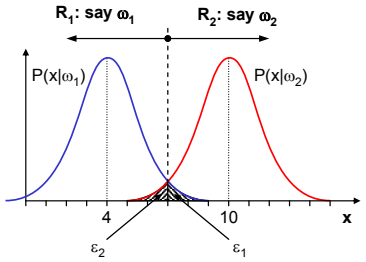
$$P(\text{error}) = P(\omega_1) \underbrace{\int_{\mathcal{R}_2} P(\mathbf{x}|\omega_1) d\mathbf{x}}_{\epsilon_1} + P(\omega_2) \underbrace{\int_{\mathcal{R}_1} P(\mathbf{x}|\omega_2) d\mathbf{x}}_{\epsilon_2}$$

  ϵ_1 is the integral of the likelihood $P(\mathbf{x}|\omega_1)$ over the region where ω_2 is chosen.

Back to the Example

For the decision rule of the previous example, the integrals ϵ_1 and ϵ_2 are depicted below.

Since we assumed equal priors, then $P(\text{error}) = .5(\epsilon_1 + \epsilon_2)$

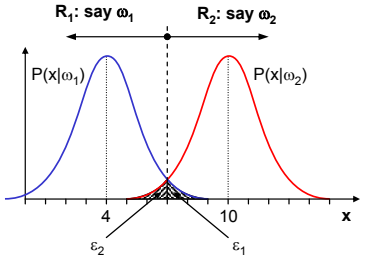


Write out the expression for $P(\text{error})$ for this example.

Back to the Example

For the decision rule of the previous example, the integrals ϵ_1 and ϵ_2 are depicted below.

Since we assumed equal priors, then $P(\text{error}) = .5(\epsilon_1 + \epsilon_2)$



Write out the expression for $P(\text{error})$ for this example.

$\epsilon_1 = (2\pi)^{-\frac{1}{2}} \int_{x=7}^{\infty} \exp(-.5(x-4)^2)dx,$

$\epsilon_2 = (2\pi)^{-\frac{1}{2}} \int_{x=-\infty}^{7} \exp(-.5(x-10)^2)dx$

Probability of Error

Thinking about the 2-class problem, not all decisions are equally good wrt minimizing $P(\text{error})$. For our example consider this (silly) rule:

$$\text{Class}\,(x) = \begin{cases} \omega_1 & \text{if } x < -100 \\ \omega_2 & \text{if } x > -100 \end{cases}$$

For this (silly) rule $\epsilon_1 \sim 1$ and $\epsilon_2 \sim 0$ which is much more than the error for the rule defined by the *likelihood ratio test*. In fact:

Bayes Error Rate: For any given problem, the minimum probability of error is achieved by the *Likelihood Ratio Test* decision rule. This probability of error is called the Bayes Error Rate and is the ***BEST*** any classifier can do.

Bayesian Decision Theory

- The Likelihood Ratio Test
- The Probability of Error
- Bayes' Risk

Bayes Risk

- So far have assumed that the penalty of misclassification of a class ω_1 example as class ω_2 is the same as that for the misclassification of a class ω_2 example as class ω_1. But consider misclassifying
  - a faulty airplane as a safe airplane (puts people's lives in danger)
  - a safe airplane as a faulty airplane (costs the airline company money)
- Can formalize this concept in terms of a cost function C_{ij}
  - Let C_{ij} denote the cost of choosing class ω_i when ω_j is the true class.
- **Bayes' Risk** is the expected value of the cost

$$\mathrm{E}\left[C\right] = \sum_{i=1}^{2}\sum_{j=1}^{2} C_{ij} P(\text{decide } \omega_i,\ \omega_j \text{ true class}) = \sum_{i=1}^{2}\sum_{j=1}^{2} C_{ij} P(\mathbf{x} \in \mathcal{R}_i|\omega_j) P(\omega_j)$$

Bayes Risk

What is the decision rule that minimizes the Bayes Risk ?

- First note: $P(\mathbf{x} \in \mathcal{R}_i|\omega_j) = \int_{\mathbf{x}\in\mathcal{R}_i} P(\mathbf{x}|\omega_j)d\mathbf{x}$
- Then the Bayes' Risk is:

$$\mathrm{E}\left[C\right] = \int_{\mathcal{R}_1} \left[C_{11}P(\omega_1)P(\mathbf{x}|\omega_1) + C_{12}P(\omega_2)P(\mathbf{x}|\omega_2)\right] d\mathbf{x} + \int_{\mathcal{R}_2} \left[C_{21}P(\omega_1)P(\mathbf{x}|\omega_1) + C_{22}P(\omega_2)P(\mathbf{x}|\omega_2)\right] d\mathbf{x}$$

- Now remember

$$\int_{\mathcal{R}_1} P(\mathbf{x}|\omega_j)d\mathbf{x} + \int_{\mathcal{R}_2} P(\mathbf{x}|\omega_j) = \int_{\mathcal{R}_1\cup\mathcal{R}_2} P(\mathbf{x}|\omega_j)d\mathbf{x} = 1.$$

$$\begin{aligned}
\mathrm{E}\left[C\right] = {} & C_{11}P(\omega_1)\int_{\mathcal{R}_1} P(\mathbf{x}|\omega_1)d\mathbf{x} + C_{12}P(\omega_2)\int_{\mathcal{R}_1} P(\mathbf{x}|\omega_2)d\mathbf{x} + \\
& C_{21}P(\omega_1)\int_{\mathcal{R}_2} P(\mathbf{x}|\omega_1)d\mathbf{x} + C_{22}P(\omega_2)\int_{\mathcal{R}_2} P(\mathbf{x}|\omega_2)d\mathbf{x} + \\
& \boxed{C_{21}P(\omega_1)\int_{\mathcal{R}_1} P(\mathbf{x}|\omega_1)d\mathbf{x} + C_{22}P(\omega_2)\int_{\mathcal{R}_1} P(\mathbf{x}|\omega_2)d\mathbf{x}} + \leftarrow +A \\
& \boxed{-C_{21}P(\omega_1)\int_{\mathcal{R}_1} P(\mathbf{x}|\omega_1)d\mathbf{x} - C_{22}P(\omega_2)\int_{\mathcal{R}_1} P(\mathbf{x}|\omega_2)d\mathbf{x}} \leftarrow -A \\
= {} & C_{21}P(\omega_1)\int_{\mathcal{R}_1\cup\mathcal{R}_2} P(\mathbf{x}|\omega_1)d\mathbf{x} + C_{22}P(\omega_2)\int_{\mathcal{R}_1\cup\mathcal{R}_2} P(\mathbf{x}|\omega_2)d\mathbf{x} + \\
& (C_{12}-C_{22})P(\omega_2)\int_{\mathcal{R}_1} P(\mathbf{x}|\omega_2)d\mathbf{x} - (C_{21}-C_{11})P(\omega_1)\int_{\mathcal{R}_1} P(x|\omega_1)d\mathbf{x} \\
= {} & C_{21}P(\omega_1) + C_{22}P(\omega_2) + \\
& (C_{12}-C_{22})P(\omega_2)\int_{\mathcal{R}_1} P(\mathbf{x}|\omega_2)d\mathbf{x} - (C_{21}-C_{11})P(\omega_1)\int_{\mathcal{R}_1} P(\mathbf{x}|\omega_1)d\mathbf{x}
\end{aligned}$$

We want to find the region $\mathcal{R}_1$ that minimizes the Bayes' Risk. From the previous slide we see the first two terms of $\mathrm{E}\left[C\right]$ are constant with respect to $\mathcal{R}_1$. Thus the optimal region is:

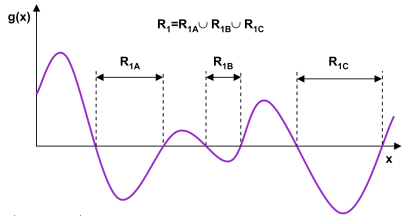
$$\begin{aligned}
\mathcal{R}_1^* &= \arg\min_{\mathcal{R}_1} \left\{ \int_{\mathcal{R}_1} \left[(C_{12}-C_{22})P(\omega_2)P(\mathbf{x}|\omega_2) - (C_{21}-C_{11})P(\omega_1)P(\mathbf{x}|\omega_1)\right] d\mathbf{x} \right\} \\
&= \arg\min_{\mathcal{R}_1} \left\{ \int_{\mathcal{R}_1} g(\mathbf{x})d\mathbf{x} \right\}
\end{aligned}$$

Note we are assuming $C_{21} > C_{11}$ and $C_{12} > C_{22}$, that is the cost of a misclassification is higher than the cost of a correct classification. Thus:

$$(C_{12}-C_{22}) > 0 \quad \text{AND} \quad (C_{21}-C_{11}) > 0$$

Bayes Risk (2)

- Temporally forget about the specific expression of $g(\mathbf{x})$. Consider the type of decision region $\mathcal{R}_1^*$ we are looking for. Select the intervals that minimize the integral $\int_{\mathcal{R}_1} g(\mathbf{x})d\mathbf{x}$, that is the intervals where $g(\mathbf{x}) < 0$



- Thus we will choose $\mathcal{R}_1^*$ such that

$$(C_{21} - C_{11})P(\omega_1)P(\mathbf{x}|\omega_1) > (C_{12} - C_{22})P(\omega_2)P(\mathbf{x}|\omega_2)$$

Rearranging the terms yields:

$$\frac{P(\mathbf{x}|\omega_1)}{P(\mathbf{x}|\omega_2)} > \frac{(C_{12} - C_{22})P(\omega_2)}{(C_{21} - C_{11})P(\omega_1)}$$

Therefore we obtain the decision rule

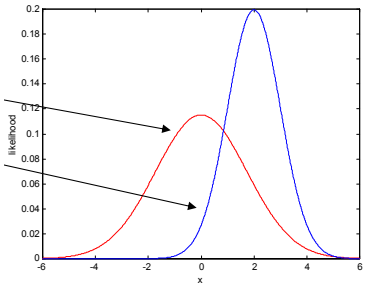
A Likelihood Ratio Test:

$$\text{Class}\,(\mathbf{x}) = \begin{cases} \omega_1 & \text{if } \frac{P(\mathbf{x}|\omega_1)}{P(\mathbf{x}|\omega_2)} > \frac{(C_{12}-C_{22})P(\omega_2)}{(C_{21}-C_{11})P(\omega_1)} \\ \omega_2 & \text{if } \frac{P(\mathbf{x}|\omega_1)}{P(\mathbf{x}|\omega_2)} < \frac{(C_{12}-C_{22})P(\omega_2)}{(C_{21}-C_{11})P(\omega_1)} \end{cases}$$

Bayes Risk: An Example

Consider the following 2 class classification problem. The likelihood functions for each class are:

$$P(x|\omega_1) = (2\pi 3)^{-\frac{1}{2}} \exp\left(-.5x^2/3\right), \quad P(x|\omega_2) = (2\pi)^{-\frac{1}{2}} \exp\left(-.5(x-2)^2\right)$$



The priors are: $P(\omega_1) = P(\omega_2) = .5$

Define the (mis)classification costs as: $C_{11} = C_{22} = 0, C_{12} = 1, C_{21} = \sqrt{3}$

Problem: Determine a decision rule minimizing the probability of error.

Bayes Risk: An Example (2)

Solution: $\Lambda(x) = \frac{(2\pi 3)^{-\frac{1}{2}} \exp\left(-.5x^2/3\right)}{(2\pi)^{-\frac{1}{2}} \exp\left(-.5(x-2)^2\right)} = \frac{(3)^{-\frac{1}{2}} \exp\left(-.5x^2/3\right)}{\exp\left(-.5(x-2)^2\right)}$.

Choose class ω_1 if $\Lambda(x) > \frac{.5(1-0)}{.5(\sqrt{3}-0)}$

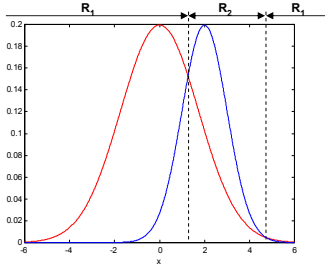
$$\iff \frac{(3)^{-\frac{1}{2}} \exp\left(-.5x^2/3\right)}{\exp\left(-.5(x-2)^2\right)} > 1$$

$$\iff \exp\left(-.5x^2/3\right) > \exp\left(-.5(x-2)^2\right)$$

$$\iff -\frac{1}{2}\frac{x^2}{3} > \frac{1}{2}(x-2)^2$$

$$\iff x^2 - 6x + 6 > 0$$

$$\iff x > 4.73 \text{ and } x < 1.27$$



Bayesian Decision Theory

- The Likelihood Ratio Test
- The Probability of Error
- Bayes' Risk
- Bayes, MAP and ML Criteria

- Finally, for the case of equal priors $P(\omega_i)$ and $C_{ij} = \delta_{ij}$ (a zero one cost function) the LRT decision rule is called the *Maximum Likelihood Criterion*, since it will minimize the likelihood $P(\mathbf{x}|\omega_i)$.

$$\textbf{ML Criterion} \rightarrow \text{Class}(\mathbf{x}) = \begin{cases} \omega_1 & \text{if } P(\mathbf{x}|\omega_1) > P(\mathbf{x}|\omega_2) \\ \omega_2 & \text{if } P(\mathbf{x}|\omega_1) < P(\mathbf{x}|\omega_2) \end{cases}$$

Variations of the LRT

- The LRT decision rule minimizing the Bayes Risk is also known as the Bayes Criterion

$$\textbf{Bayes Criterion} \rightarrow \text{Class}(\mathbf{x}) = \begin{cases} \omega_1 & \text{if } \Lambda(\mathbf{x}) > \frac{(C_{12}-C_{22})P(\omega_2)}{(C_{21}-C_{11})P(\omega_1)} \\ \omega_2 & \text{if } \Lambda(\mathbf{x}) < \frac{(C_{12}-C_{22})P(\omega_2)}{(C_{21}-C_{11})P(\omega_1)} \end{cases}$$

- Minimize the probability of error, that is the *Bayes Criterion* with $C_{ij} = \delta_{ij}$, this version of the LRT decision is referred to as the *Maximum A Posteriori Criterion*.

$$\textbf{MAP Criterion} \rightarrow \text{Class}(\mathbf{x}) = \begin{cases} \omega_1 & \text{if } P(\omega_1|\mathbf{x}) > P(\omega_2|\mathbf{x}) \\ \omega_2 & \text{if } P(\omega_1|\mathbf{x}) < P(\omega_2|\mathbf{x}) \end{cases}$$

Variations of the LRT (2)

Two more decision rules are commonly cited in the related literature.

- The *Neyman-Pearson Criterion* which also leads to a LRT decision rule. It fixes one class error probability, say $\epsilon < \alpha$ and seeks to minimize the other.
- The *Minimax Criterion*, derived from the *Bayes Criterion* and seeks to **mini**mize the **max**imum *Bayes Risk*.

Bayesian Decision Theory

- The Likelihood Ratio Test
- The Probability of Error
- Bayes' Risk
- Bayes, MAP and ML Criteria
- Multi-class functions

Decision rules for multi-class problems

The decision rule minimizing $P(\text{error})$ generalizes to multi-class problems.

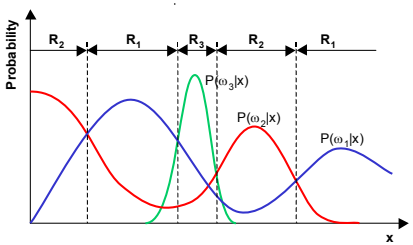
- The derivation is easier if we express $P(\text{error})$ in terms of making a correct assignment.

$$P(\text{error}) = 1 - P(\text{correct})$$

- Probability of making a correct assignment is

$$\begin{aligned} P(\text{correct}) &= \sum_{i=1}^{C} P(\omega_i) \int_{\mathcal{R}_i} P(\mathbf{x}|\omega_i) d\mathbf{x} \\ &= \sum_{i=1}^{C} \int_{\mathcal{R}_i} P(\mathbf{x}|\omega_i) P(\omega_i) d\mathbf{x} \\ &= \sum_{i=1}^{C} \underbrace{\int_{\mathcal{R}_i} P(\omega_i|\mathbf{x}) P(\mathbf{x}) d\mathbf{x}}_{\mathcal{I}_i} \end{aligned}$$

- The problem of minimizing $P(\text{error})$ is equivalent to that of maximizing $P(\text{correct})$.
- To maximize $P(\text{correct})$ we have to maximize each of the integrals $\mathcal{I}_i$. In turn, each integral $\mathcal{I}_i$ will be maximized by choosing the class ω_i that yields the maximum $P(\omega_i|\mathbf{x}) \implies$ we will define $\mathcal{R}_i$ to be the regions where $P(\omega_i|\mathbf{x})$ is maximum.



Therefore, the decision rule that minimizes $P(\text{error})$ is the *MAP Criterion*.

Minimum Bayes Risk

- Define the overall decision rule as a function

$$\boldsymbol{\alpha} : \mathbf{x} \rightarrow \{\alpha_1, \alpha_2, \cdots, \alpha_C\} \; s.t. \; \boldsymbol{\alpha}(\mathbf{x}) = \alpha_i \text{ if } \mathbf{x} \text{ is assigned to class } \omega_i.$$

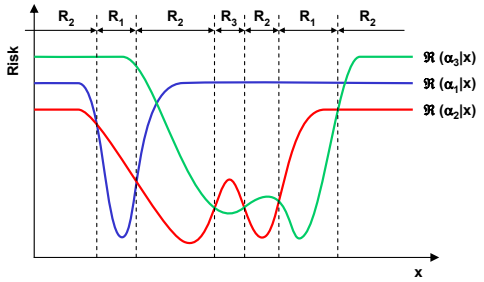
- The risk $R(\alpha(\mathbf{x})|\mathbf{x})$ of assigning $\mathbf{x}$ to class $\alpha(\mathbf{x}) = \alpha_i$ is

$$R(\alpha_i|\mathbf{x}) = \sum_{j=1}^{C} C_{ij} P(\omega_j|\mathbf{x})$$

- The Bayes Risk associate with the decision rule $\alpha(\mathbf{x})$ is

$$R(\alpha(\mathbf{x})) = \int R(\alpha(\mathbf{x})|\mathbf{x}) P(\mathbf{x}) d\mathbf{x}$$

- To minimize this expression we have to minimize the conditional risk $R(\alpha(\mathbf{x})|\mathbf{x})$ at each point $\mathbf{x}$ in the feature space.



Bayesian Decision Theory

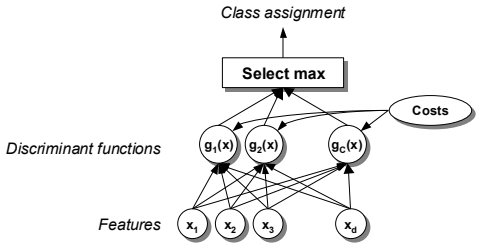
- The Likelihood Ratio Test
- The Probability of Error
- Bayes' Risk
- Bayes, MAP and ML Criteria
- Multi-class functions
- Discriminant Functions

Discriminant Functions

- All decision rules presented in this lecture have the same structure
  - At each $\mathbf{x}$ in feature space choose class ω_i which maximizes (or minimizes) some measure $g_i(\mathbf{x})$
- Formally, there is a set of discriminant functions $\{g_i(\mathbf{x})\}_{i=1}^{C}$ and the following decision rule

> assign $\mathbf{x}$ to class ω_i if $g_i(\mathbf{x}) > g_j(\mathbf{x})\ \forall j \neq i$

- Can visualize the decision rule as a network or machine



- The three basic decision rules *Bayes*, *MAP* and *ML* in terms of *Discriminant Functions*:

Criterion	Discriminant Function
Bayes	$g_i(\mathbf{x}) = R(\alpha_i\|\mathbf{x})$
MAP	$g_i(\mathbf{x}) = P(\omega_i\|\mathbf{x})$
ML	$g_i(\mathbf{x}) = P(\mathbf{x}\|\omega_i)$

Discriminant Functions for the class of Gaussian Distributions

Bayes classifiers for Normally distributed classes

- The (MAP) decision rule minimizing the probability of error can be formulated as a family of discriminant functions

  Choose class ω_i if $g_i(\mathbf{x}) > g_j(\mathbf{x})\ \forall i \neq j$ with $g_i(\mathbf{x}) = P(\omega_i|\mathbf{x})$

  For classes that are normally distributed, this family can be reduced to very simple expressions.

- General expression for Gaussian densities
  - The multi-variate Normal density is defined as

$$f_X(\mathbf{x}) = \frac{1}{(2\pi)^{\frac{n}{2}}|\Sigma|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu})\right)$$

Quadratic Classifiers

- Bayes classifiers for Normally distributed classes
  - Using Bayes' rule the MAP discriminant function becomes

$$\begin{aligned} g_i(\mathbf{x}) &= \frac{P(\mathbf{x}|\omega_i)P(\omega_i)}{P(\mathbf{x})} \\ &= \frac{1}{(2\pi)^{\frac{n}{2}}|\Sigma_i|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_i)^T\Sigma_i^{-1}(\mathbf{x}-\boldsymbol{\mu}_i)\right)\frac{P(\omega_i)}{P(\mathbf{x})} \end{aligned}$$

  - Eliminating constant terms

$$g_i(\mathbf{x}) = |\Sigma_i|^{-1/2} \exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\Sigma_i^{-1}(\mathbf{x}-\boldsymbol{\mu})\right)P(\omega_i)$$

  - Taking the log since it is a monotonically increasing function

$$\boxed{g_i(\mathbf{x}) = -\tfrac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_i)^T\Sigma_i^{-1}(\mathbf{x}-\boldsymbol{\mu}_i) - \tfrac{1}{2}\log\left(|\Sigma_i|\right) + \log\left(P(\omega_i)\right)}$$

Quadratic Discriminant Function

Quadratic Classifiers

- Bayes classifiers for Normally distributed classes
  - Case 1: $\Sigma_i = \sigma^2 I$

Case 1: $\Sigma_i = \sigma^2 I$

- The features are statistically independent with the same variance for all classes.
  - The quadratic function becomes

$$\begin{aligned} g_i(\mathbf{x}) &= -\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_i)^T(\sigma^2 I)^{-1}(\mathbf{x}-\boldsymbol{\mu}_i) - \frac{1}{2}\log\left(|\sigma^2 I|\right) + \log\left(P(\omega_i)\right) \\ &= -\frac{1}{2\sigma^2}(\mathbf{x}-\boldsymbol{\mu}_i)^T(\mathbf{x}-\boldsymbol{\mu}_i) - \frac{N}{2}\log\left(\sigma^2\right) + \log\left(P(\omega_i)\right) \\ &= -\frac{1}{2\sigma^2}(\mathbf{x}-\boldsymbol{\mu}_i)^T(\mathbf{x}-\boldsymbol{\mu}_i) + \log\left(P(\omega_i)\right), \quad \text{dropping the second term} \\ &= -\frac{1}{2\sigma^2}(\mathbf{x}^T\mathbf{x} - 2\boldsymbol{\mu}_i^T\mathbf{x} + \boldsymbol{\mu}_i^T\boldsymbol{\mu}_i) + \log\left(P(\omega_i)\right) \end{aligned}$$

  - Eliminate the term $\mathbf{x}^T\mathbf{x}$ as it is constant for all classes. Then

$$\begin{aligned} g_i(\mathbf{x}) &= -\frac{1}{2\sigma^2}(-2\boldsymbol{\mu}_i^T\mathbf{x} + \boldsymbol{\mu}_i^T\boldsymbol{\mu}_i) + \log\left(P(\omega_i)\right) \\ &= \mathbf{w}_i^T\mathbf{x} + w_{i0} \end{aligned}$$

    where

$$\mathbf{w}_i = \frac{\boldsymbol{\mu}_i}{\sigma^2} \quad \text{and} \quad w_{i0} = -\frac{1}{2\sigma^2}\boldsymbol{\mu}_i^T\boldsymbol{\mu}_i + \log\left(P(\omega_i)\right)$$

  - As the discriminant is linear, the decision boundaries $g_i(\mathbf{x}) = g_j(\mathbf{x})$ will be hyper-planes.

- If we assume equal priors (also know as the *nearest mean classifier*)

Minimum distance classifier: $g_i(\mathbf{x}) = -\frac{1}{2\sigma^2}(\mathbf{x}-\boldsymbol{\mu}_i)^T(\mathbf{x}-\boldsymbol{\mu}_i)$

- Properties of the class-conditional probabilities
  - the loci of constant probability for each class are hyper-spheres

Case 1: $\Sigma_i = \sigma^2 I$

- The decision boundaries are the hyperplanes $g_i(\mathbf{x}) = g_j(\mathbf{x})$, and can be written as

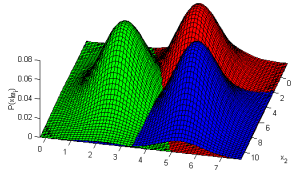
$$\tilde{\mathbf{w}}^T(\mathbf{x}-\mathbf{x}_0) = 0, \quad \text{after some algebra}$$

  where

$$\begin{aligned} \tilde{\mathbf{w}} &= \boldsymbol{\mu}_i - \boldsymbol{\mu}_j \\ \mathbf{x}_0 &= \frac{1}{2}(\boldsymbol{\mu}_i + \boldsymbol{\mu}_j) - \frac{\sigma^2}{\|\boldsymbol{\mu}_i - \boldsymbol{\mu}_j\|^2}\ln\frac{P(\omega_i)}{P(\omega_j)}(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j). \end{aligned}$$

- The hyperplane separating $\mathcal{R}_i$ and $\mathcal{R}_j$ passes through the point $\mathbf{x}_0$ and is orthogonal to the vector $\tilde{\mathbf{w}}$.

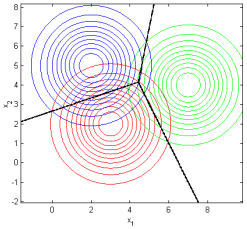
Case 1: $\Sigma_i = \sigma^2 I$, Example

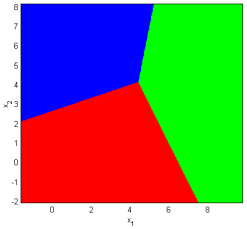


Compute decision boundaries for the above 3-class, 2D problem with class-conditional parameters and equal priors.

$$\boldsymbol{\mu}_1 = (3,2)^T,\ \boldsymbol{\mu}_2 = (7,4)^T,\ \boldsymbol{\mu}_3 = (2,5)^T$$

$$\Sigma_1 = \begin{pmatrix} 2 & 0 \\ 0 & 2 \end{pmatrix},\ \Sigma_2 = \begin{pmatrix} 2 & 0 \\ 0 & 2 \end{pmatrix},\ \Sigma_3 = \begin{pmatrix} 2 & 0 \\ 0 & 2 \end{pmatrix}$$





Case 2: $\Sigma_i = \Sigma$ (Σ diagonal)

- The classes have the same covariance matrix, but the features are allowed to have different variances.
  - The quadratic function becomes

$$g_i(\mathbf{x}) = -\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_i)^T\Sigma_i^{-1}(\mathbf{x}-\boldsymbol{\mu}_i) - \frac{1}{2}\log\left(|\Sigma_i|\right) + \log\left(P(\omega_i)\right)$$

$$= -\tfrac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_i)^T \begin{pmatrix} \sigma_1^{-2} & & \\ & \ddots & \\ & & \sigma_N^{-2} \end{pmatrix}(\mathbf{x}-\boldsymbol{\mu}_i) - \frac{1}{2}\log\left(\left|\begin{pmatrix} \sigma_1^2 & & \\ & \ddots & \\ & & \sigma_N^2 \end{pmatrix}\right|\right) + \log\left(P(\omega_i)\right)$$

$$= -\frac{1}{2}\sum_{k=1}^{N}\frac{(x_k-\mu_{ik})^2}{\sigma_k^2} - \frac{1}{2}\log\left(\prod_{k=1}^{N}\sigma_k^2\right) + \log\left(P(\omega_i)\right)$$

$$= -\frac{1}{2}\sum_{k=1}^{N}\frac{x_k^2 - 2x_k\mu_{ik} + \mu_{ik}^2}{\sigma_k^2} - \frac{1}{2}\log\left(\prod_{k=1}^{N}\sigma_k^2\right) + \log\left(P(\omega_i)\right)$$

Quadratic Classifiers

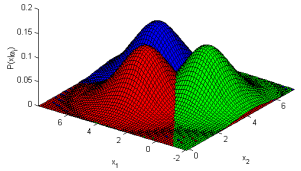
- Bayes classifiers for Normally distributed classes
  - Case 1: $\Sigma_i = \sigma^2 I$
  - Case 2: $\Sigma_i = \Sigma$ (Σ diagonal)
  - Eliminate the term x_k^2 as it is constant for all classes.

$$g_i(\mathbf{x}) = -\frac{1}{2}\sum_{k=1}^{N}\frac{-2x_k\mu_{ik} + \mu_{ik}^2}{\sigma_k^2} - \frac{1}{2}\log\left(\prod_{k=1}^{N}\sigma_k^2\right) + \log\left(P(\omega_i)\right)$$

Properties

- This discriminant is linear, so the decision boundaries $g_i(\mathbf{x}) = g_j(\mathbf{x})$ are hyper-planes.
- The loci of constant probability are hyper-ellipses aligned with the feature axes.
- The only difference to the previous classifier is that the distance of each axis is normalized by the variance of the axis.

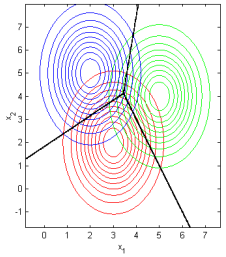
Case 2: $\Sigma_i = \Sigma$ (Σ diagonal), example

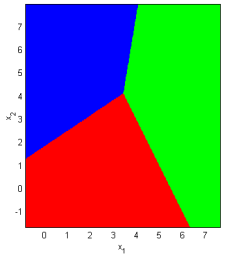


Compute decision boundaries for the above 3-class, 2D problem with class-conditional parameters and equal priors.

$$\boldsymbol{\mu}_1 = (3,2)^T,\ \boldsymbol{\mu}_2 = (5,4)^T,\ \boldsymbol{\mu}_3 = (2,5)^T$$

$$\Sigma_1 = \begin{pmatrix} 1 & 0 \\ 0 & 2 \end{pmatrix},\ \Sigma_2 = \begin{pmatrix} 1 & 0 \\ 0 & 2 \end{pmatrix},\ \Sigma_3 = \begin{pmatrix} 1 & 0 \\ 0 & 2 \end{pmatrix}$$





Quadratic Classifiers

- Bayes classifiers for Normally distributed classes
  - Case 1: $\Sigma_i = \sigma^2 I$
  - Case 2: $\Sigma_i = \Sigma$ (Σ diagonal)
  - Case 3: $\Sigma_i = \Sigma$ (Σ non-diagonal)

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

- All classes have the same covariance matrix, but it is not necessarily diagonal.
- The quadratic discriminant function becomes

$$\begin{aligned} g_i(\mathbf{x}) &= -\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_i)^T\Sigma_i^{-1}(\mathbf{x}-\boldsymbol{\mu}_i) - \frac{1}{2}\log(|\Sigma_i|) + \log(P(\omega_i)) \\ &= -\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_i)^T\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu}_i) - \frac{1}{2}\log(|\Sigma|) + \log(P(\omega_i)) \end{aligned}$$

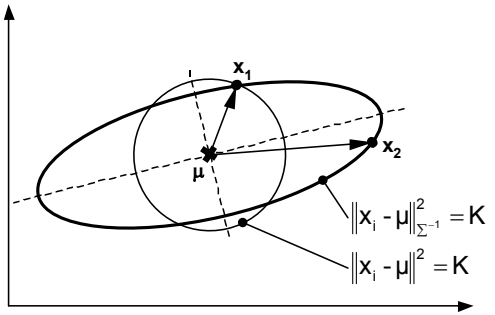
- Eliminate the term $\log(|\Sigma|)$, which is constant for all classes.

$$g_i(\mathbf{x}) = -\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_i)^T\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu}_i) + \log(P(\omega_i))$$

- The quadratic term is called the Mahalanobis distance, a very important term in Statistical PR.

Mahalanobis distance: $\|\mathbf{x}-\mathbf{y}\|^2_{\Sigma^{-1}} = (\mathbf{x}-\mathbf{y})^T\Sigma^{-1}(\mathbf{x}-\mathbf{y})$

- The Mahalanobis distance is a vector distance that uses a Σ^{-1} norm.
  - Σ^{-1} can be thought as a stretching factor on the space.
  - For $\Sigma = I$ the Mahalanobis distance becomes the Euclidean distance.



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

- Expansion of the quadratic term in the discriminant yields

$$g_i(\mathbf{x}) = -\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_i)^T\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu}_i) + \log\left(P(\omega_i)\right)$$
$$= -\frac{1}{2}\left(\mathbf{x}^T\Sigma^{-1}\mathbf{x} - 2\boldsymbol{\mu}_i^T\Sigma^{-1}\mathbf{x} + \boldsymbol{\mu}_i^T\Sigma^{-1}\boldsymbol{\mu}_i\right) + \log\left(P(\omega_i)\right)$$

- Removing the term $\mathbf{x}^T\Sigma^{-1}\mathbf{x}$ which is constant for all classes

$$g_i(\mathbf{x}) = -\frac{1}{2}\left(-2\boldsymbol{\mu}_i^T\Sigma^{-1}\mathbf{x} + \boldsymbol{\mu}_i^T\Sigma^{-1}\boldsymbol{\mu}_i\right) + \log\left(P(\omega_i)\right)$$

- Reorganizing terms get:

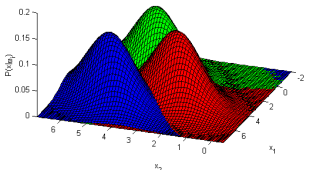
$$\boxed{g_i(\mathbf{x}) = \mathbf{w}_i^T\mathbf{x} + w_{i0}} \text{ with } \mathbf{w}_i = \Sigma^{-1}\boldsymbol{\mu}_i \text{ and } w_{i0} = -\frac{1}{2}\boldsymbol{\mu}_i^T\Sigma^{-1}\boldsymbol{\mu}_i + \log\left(P(\omega_i)\right)$$

Properties

- The discriminant is linear, so the decision boundaries are hyper-planes.
- The constant probability loci are hyper-ellipses aligned with the eigenvectors of Σ.
- If we can assume equal priors the classifier becomes a minimum (Mahalanobis) distance classifier.

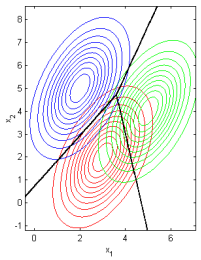
Equal Priors: $g_i(\mathbf{x}) = -\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_i)^T\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu}_i)$

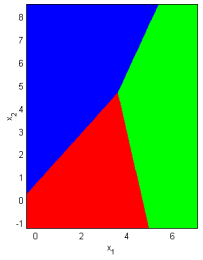
Case 3: Example



Compute decision boundaries for the above 3-class, 2D problem with class-conditional parameters and equal priors.

$$\boldsymbol{\mu}_1 = (3,2)^T,\ \boldsymbol{\mu}_2 = (5,4)^T,\ \boldsymbol{\mu}_3 = (2,5)^T$$
$$\Sigma_1 = \begin{pmatrix} .5 & .7 \\ .7 & 2 \end{pmatrix},\ \Sigma_2 = \begin{pmatrix} 1 & .7 \\ .7 & 2 \end{pmatrix},\ \Sigma_3 = \begin{pmatrix} 1 & .7 \\ .7 & 2 \end{pmatrix}$$





Quadratic Classifiers

- Bayes classifiers for Normally distributed classes
  - Case 1: $\Sigma_i = \sigma^2 I$
  - Case 2: $\Sigma_i = \Sigma$ (Σ diagonal)
  - Case 3: $\Sigma_i = \Sigma$ (Σ non-diagonal)
  - Case 4: $\Sigma_i = \sigma_i^2 I$

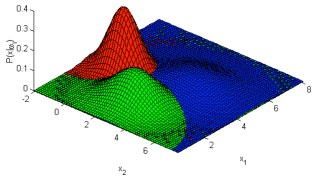
Case 4: $\Sigma_i = \sigma_i^2 I$

- Each class has a different covariance matrix, which is proportional to the identity matrix.
  - The quadratic discriminant becomes

$$g_i(\mathbf{x}) = -\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_i)^T\Sigma_i^{-1}(\mathbf{x}-\boldsymbol{\mu}_i) - \frac{1}{2}\log\left(|\Sigma_i|\right) + \log\left(P(\omega_i)\right)$$
$$= -\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_i)^T\sigma_i^{-2}(\mathbf{x}-\boldsymbol{\mu}_i) - \frac{N}{2}\log\left(\sigma_i^2\right) + \log\left(P(\omega_i)\right)$$

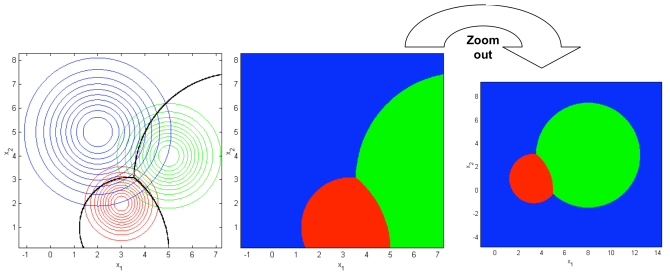
- The expression cannot be reduce further so
  - The decision boundaries are quadratic: hyper-ellipses
  - The loci of constant probability are hyper-spheres aligned with the feature axis.

Case 4: $\Sigma_i = \sigma_i^2 I$, example



Compute decision boundaries for the above 3-class, 2D problem with class-conditional parameters.

$$\boldsymbol{\mu}_1 = (3,2)^T,\ \boldsymbol{\mu}_2 = (5,4)^T,\ \boldsymbol{\mu}_3 = (2,5)^T$$
$$\Sigma_1 = \begin{pmatrix} .5 & 0 \\ 0 & .5 \end{pmatrix},\ \Sigma_2 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix},\ \Sigma_3 = \begin{pmatrix} 2 & 0 \\ 0 & 2 \end{pmatrix}$$



Quadratic Classifiers

- Bayes classifiers for Normally distributed classes
  - Case 1: $\Sigma_i = \sigma^2 I$
  - Case 2: $\Sigma_i = \Sigma$ (Σ diagonal)
  - Case 3: $\Sigma_i = \Sigma$ (Σ non-diagonal)
  - Case 4: $\Sigma_i = \sigma_i^2 I$
  - Case 5: $\Sigma_i \neq \Sigma_j$, General Case

Case 5: $\Sigma_i \neq \Sigma_j$, General Case

- Have already derived the expression for the general case it is:

$$g_i(\mathbf{x}) = -\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_i)^T\Sigma_i^{-1}(\mathbf{x}-\boldsymbol{\mu}_i) - \frac{1}{2}\log\left(|\Sigma_i|\right) + \log\left(P(\omega_i)\right)$$

  - Reorganizing terms in a quadratic form yields

$$\boxed{g_i(\mathbf{x}) = \mathbf{x}^T W_i \mathbf{x} + \mathbf{w}_i^T\mathbf{x} + w_{i0}}$$

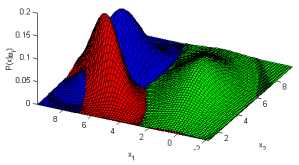
where

$$W_i = -\frac{1}{2}\Sigma_i^{-1},$$
$$\mathbf{w}_i = \Sigma_i^{-1}\boldsymbol{\mu}_i,$$
$$w_{i0} = -\frac{1}{2}\boldsymbol{\mu}_i^T\Sigma_i^{-1}\boldsymbol{\mu}_i - \frac{1}{2}\log\left(|\Sigma_i|\right) + \log\left(P(\omega_i)\right)$$

Properties

- The loci of constant probability for each class are hyper-ellipses, oriented with the eigenvectors of Σ_i for that class.
- The decision boundaries are quadratic: hyper-ellipses or hyper-parabolloids
- The quadratic expression in the discriminant is proportional to the Mahalanobis distance using the class-conditional variance Σ_i.

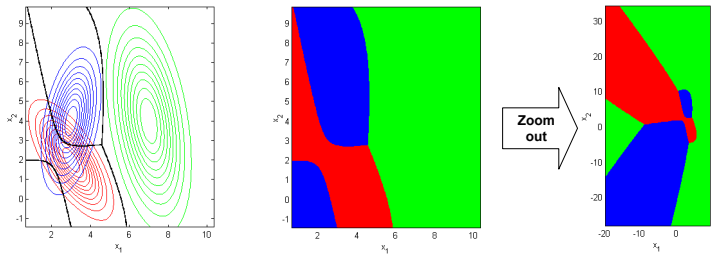
Case 5: $\Sigma_i \neq \Sigma_j$, Example



Compute the decision boundaries for the above 3-class, 2-dimensional problem with class-conditional parameters.

$$\boldsymbol{\mu}_1 = (3,2)^T \; \boldsymbol{\mu}_2 = (5,4)^T \; \boldsymbol{\mu}_3 = (2,5)^T$$

$$\Sigma_1 = \begin{pmatrix} 1 & -1 \\ -1 & 2 \end{pmatrix}, \; \Sigma_2 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \; \Sigma_3 = \begin{pmatrix} 2 & 0 \\ 0 & 2 \end{pmatrix}$$



Quadratic Classifiers

- Bayes classifiers for Normally distributed classes
  - Case 1: $\Sigma_i = \sigma^2 I$
  - Case 2: $\Sigma_i = \Sigma$ (Σ diagonal)
  - Case 3: $\Sigma_i = \Sigma$ (Σ non-diagonal)
  - Case 4: $\Sigma_i = \sigma_i^2 I$
  - Case 5: $\Sigma_i \neq \Sigma_j$, General Case
- **Numerical Example**

Numerical Example

Derive the discriminant function for the 2-class 3D classification problem defined by the following Gaussian Likelihoods

$$\boldsymbol{\mu}_1 = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix} \; ; \quad \boldsymbol{\mu}_2 = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} \; ; \quad \Sigma_1 = \Sigma_2 = \begin{pmatrix} \frac{1}{4} & 0 & 0 \\ 0 & \frac{1}{4} & 0 \\ 0 & 0 & \frac{1}{4} \end{pmatrix} \; ; \quad P(\omega_2) = 2P(\omega_1)$$

Solution:

$$g_1(\mathbf{x}) = \frac{1}{2\sigma^2}(\mathbf{x} - \boldsymbol{\mu}_1)^T(\mathbf{x} - \boldsymbol{\mu}_1) + \log\left(P(\omega_1)\right)$$

$$= -\frac{1}{2(\frac{1}{4})}(x_1 - 0, x_2 - 0, x_3 - 0) \begin{pmatrix} x_1 - 0 \\ x_2 - 0 \\ x_3 - 0 \end{pmatrix} + \log\left(\frac{1}{3}\right)$$

$$g_2(\mathbf{x}) = -\frac{1}{2(\frac{1}{4})}(x_1 - 1, x_2 - 1, x_3 - 1) \begin{pmatrix} x_1 - 1 \\ x_2 - 1 \\ x_3 - 1 \end{pmatrix} + \log\left(\frac{2}{3}\right)$$

Classify $\mathbf{x}$ as ω_1 if $g_1(\mathbf{x}) > g_2(\mathbf{x})$.

$$g_1(\mathbf{x}) > g_2(\mathbf{x})$$

$$\iff -2(\mathbf{x}^T\mathbf{x}) + \log\left(\frac{1}{3}\right) > -2((x_1-1)^2 + (x_2-1)^2 + (x_3-1)^2) + \log\left(\frac{2}{3}\right)$$

$$\iff x_1 + x_2 + x_3 < \frac{6 - \log 2}{4} = 1.32$$

Therefore the decision rule is:

$$\text{Class}\,(\mathbf{x}) = \begin{cases} \omega_1 & \text{if } x_1 + x_2 + x_3 < 1.32 \\ \omega_2 & \text{if } x_1 + x_2 + x_3 > 1.32 \end{cases}$$

Classify the test example $\mathbf{x}_u = (.1, .7, .8)^T$

$$.1 + .7 + .8 = 1.6 > 1.32 \implies \mathbf{x}_u \in \omega_2$$

Quadratic Classifiers

- Bayes classifiers for Normally distributed classes
  - Case 1: $\Sigma_i = \sigma^2 I$
  - Case 2: $\Sigma_i = \Sigma$ (Σ diagonal)
  - Case 3: $\Sigma_i = \Sigma$ (Σ non-diagonal)
  - Case 4: $\Sigma_i = \sigma_i^2 I$
  - Case 5: $\Sigma_i \neq \Sigma_j$, General Case
- **Numerical Example**
- **Conclusions**

Conclusions

- We can draw the following conclusions
  - The Bayes classifier for normally distributed classes (general case) is a quadratic classifier.
  - The Bayes classifier for normally distributed classes with equal covariance matrices is a linear classifier.
  - The minimum Mahalanobis distance classifier is Bayes-optimal for
    - normally distributed classes ***and***
    - equal covariance matrices ***and***
    - equal priors
  - The minimum Euclidean distance classifier is Bayes-optimal for
    - normally distributed classes ***and***
    - equal covariance matrices proportional to the identity matrix ***and***
    - equal priors
  - Both Euclidean and Mahalanobis distance classifiers are linear classifiers.
- Some of the most popular classifiers can be derived from decision-theoretic principles and some simplifying assumptions.
  - Using a specific (Euclidean or Mahalanobis) minimum distance classifier implicitly corresponds to certain statistical assumptions.
  - Can rarely answer the question if these assumptions hold for real problems. In most cases limited to answering *"Does this classifier solve our problem or not? "*