Pattern Recognition and Image Analysis

Dr. Manal Helal – Fall 2014

Lecture 2

BAYES DECISION THEORY

In Action 1

Bayesian Decision Theory

- **The Basic Idea**
  - To minimize errors, choose the least risky class, i.e. the class for which the *expected loss* is smallest
- **Assumptions**
  - Problem posed in probabilistic terms, and all relevant probabilities are known

Probability Mass vs. Probability Density Functions

- Probability Mass Function, P(x)
  - Probability for values of discrete random variable *x*.
  - Each value has its own associated probability

$$\chi = \{v_1, \ldots, v_m\}$$
$$P(x) \geq 0, \text{ and } \sum_{x \in \chi} P(x) = 1$$

- Probability Density, p(x)
  - Probability for values of continuous random variable *x.*
  - Probability returned is for an *interval* within which the value lies (intervals defined by some unit distance)

$$Pr[x \in (a,b)] = \int_a^b p(x)\, dx$$
$$p(x) \geq 0 \text{ and } \int_{-\infty}^{\infty} p(x)\, dx = 1$$

Prior Probability

- Definition (P(w))
  - The likelihood of a value for a random variable representing the *state of nature (true class w for the current input)*, in the absence of other information
  - Informally, "what percentage of the time state X occurs"
- Example
  - The prior probability that an instance taken from two classes is provided as input, in the absence of any features (e.g. P(cat) = 0.3, P(dog) = 0.7)

Class-Conditional Probability Density Function (for Continuous Features)

- Definition (p(x|w))
  - The probability of a value for continuous random variable x, given a state of nature w
  - For each value of x, we have a different class-conditional pdf for each class in w (example next slide)

Example: Class-Conditional Probability Densities

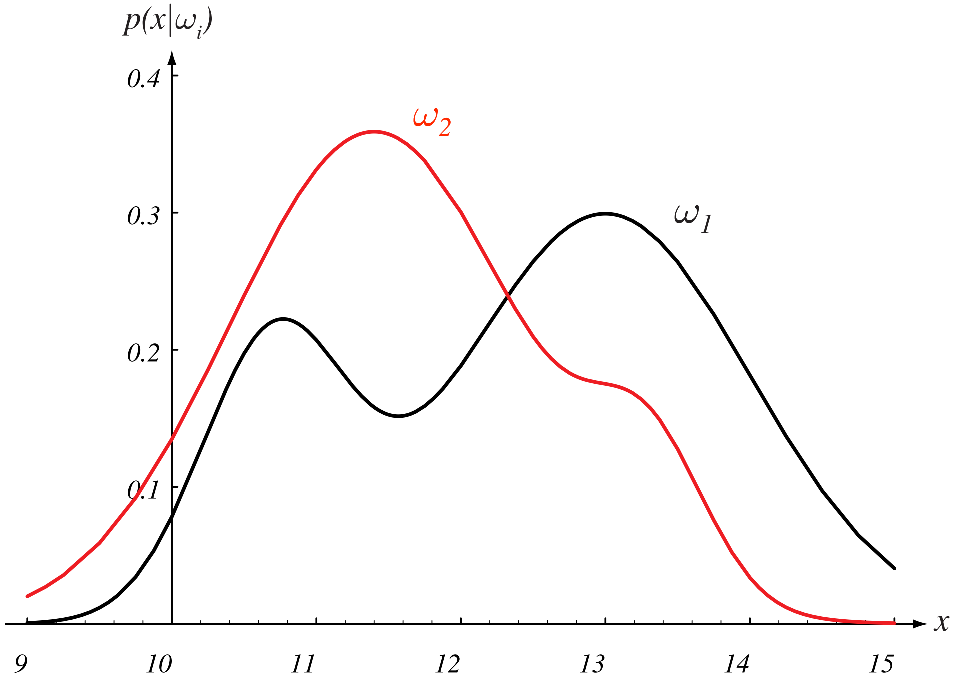


FIGURE 2.1. Hypothetical class-conditional probability density functions show the probability density of measuring a particular feature value x given the pattern is in category ω_i. If x represents the lightness of a fish, the two curves might describe the difference in lightness of populations of two types of fish. Density functions are normalized, and thus the area under each curve is 1.0. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

Bayes Formula

$$P(\omega_j|x) = \frac{p(x|\omega_j)P(w_j)}{p(x)}$$

posterior = $\frac{\text{likelihood x prior}}{\text{evidence}}$

$$\text{where} \quad p(x) = \sum_{j=1}^{c} p(x|\omega_j)P(\omega_j)$$

- Purpose
  - Convert class prior and class-conditional densities to a *posterior probability* for a class: the probability of a class given the input features ('post-observation')

Example: Posterior Probabilities

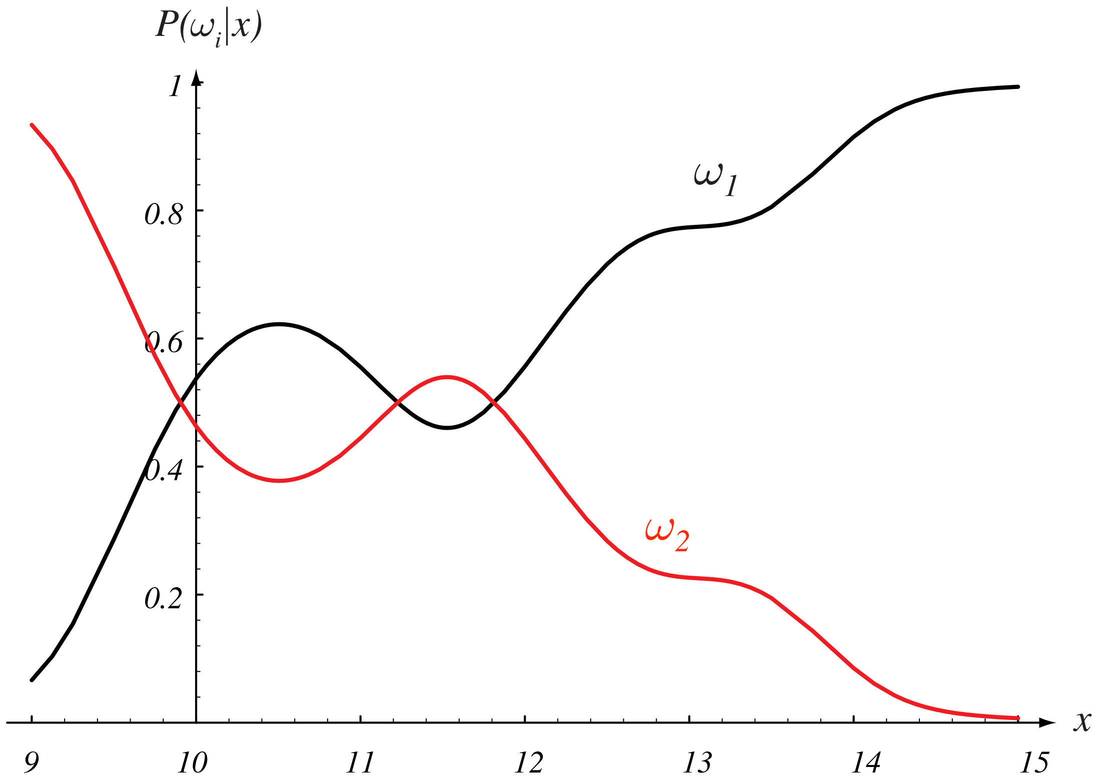


FIGURE 2.2. Posterior probabilities for the particular priors $P(\omega_1) = 2/3$ and $P(\omega_2) = 1/3$ for the class-conditional probability densities shown in Fig. 2.1. Thus in this case, given that a pattern is measured to have feature value $x = 14$, the probability it is in category ω_2 is roughly 0.08, and that it is in ω_1 is 0.92. At every x, the posteriors sum to 1.0. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*.

Choosing the Most Likely Class

- What happens if we do the following?

$$\text{Decide } \omega_1 \text{ if } P(\omega_1|x) > P(\omega_2|x)\text{; otherwise decide } \omega_2$$

- A. We minimize the average probability of error. Consider the two-class case from previous slide

$$P(error|x) = \begin{cases} P(\omega_1|x) & \text{if we choose } \omega_2 \\ P(\omega_2|x) & \text{if we choose } \omega_1 \end{cases}$$

$$P(error) = \int_{-\infty}^{\infty} P(error|x)p(x)\,dx \quad \text{(average error)}$$

Expected Loss or Conditional Risk of an Action

$$R(\alpha_i|\mathbf{x}) = \sum_{j=1}^{c} \lambda(\alpha_i|\omega_j)P(\omega_j|\mathbf{x})$$

- Explanation
  - The expected ("average") loss for taking an action (choosing a class) given an input vector, for a given conditional loss function (lambda)

- For M=2
  - Define the **loss matrix** $L = \begin{pmatrix} \lambda_{11} & \lambda_{12} \\ \lambda_{21} & \lambda_{22} \end{pmatrix}$
  - λ_{12} penalty term for deciding class ω_2, although the pattern belongs to ω_1, etc.
- Risk with respect to ω_1

$$r_1 = \lambda_{11} \int_{R_1} p(\underline{x}|\omega_1) d\underline{x} + \boxed{\lambda_{12} \int_{R_2} p(\underline{x}|\omega_1) d\underline{x}}$$

- Risk with respect to ω_2

$$r_2 = \boxed{\lambda_{21} \int_{R_1} p(\underline{x}|\omega_2) d\underline{x}} + \lambda_{22} \int_{R_2} p(\underline{x}|\omega_2) d\underline{x}$$

$\Rightarrow$ Probabilities of wrong decisions, weighted by the penalty terms

- Average risk

$$r = r_1 P(\omega_1) + r_2 P(\omega_2)$$

- Choose R_1 and R_2 so that r **is minimized**

- Then assign $\underline{x}$ to ω_i if

$$\ell_1 \equiv \lambda_{11} p(\underline{x}|\omega_1)P(\omega_1) + \lambda_{21} p(\underline{x}|\omega_2)P(\omega_2) \quad <$$

$$\ell_2 \equiv \lambda_{12} p(\underline{x}|\omega_1)P(\omega_1) + \lambda_{22} p(\underline{x}|\omega_2)P(\omega_2)$$

- Equivalently:

assign $\underline{x}$ in $\omega_1(\omega_2)$ if

$$\boxed{\ell_{12} \equiv \frac{p(\underline{x}|\omega_1)}{p(\underline{x}|\omega_2)} > (<) \frac{P(\omega_2)}{P(\omega_1)} \frac{\lambda_{21} - \lambda_{22}}{\lambda_{12} - \lambda_{11}}}$$

ℓ_{12} : likelihood ratio

❖ If $P(\omega_1) = P(\omega_2) = \dfrac{1}{2}$ and $\lambda_{11} = \lambda_{22} = 0$

$$\boxed{\begin{aligned} &\underline{x} \to \omega_1 \ \text{if} \ P(\underline{x}|\omega_1) > P(\underline{x}|\omega_2)\frac{\lambda_{21}}{\lambda_{12}} \\ &\underline{x} \to \omega_2 \ \text{if} \ P(\underline{x}|\omega_2) > P(\underline{x}|\omega_1)\frac{\lambda_{12}}{\lambda_{21}} \end{aligned}}$$

if $\lambda_{21} = \lambda_{12} \Rightarrow$ Minimum classification error probability

Decision Function and Overall Risk

$$R = \int R(\alpha(x)|x)p(x)\ dx$$

- Decision Function or Decision Rule
  - (alpha(x)): takes on the value of exactly one action for each input vector x
- Overall Risk
  - The expected (average) loss associated with a decision rule

Bayes Decision Rule

- Idea
  - Minimize the overall risk, by choosing the action with the least conditional risk for input vector x
- Bayes Risk (R*)
  - The resulting overall risk produced using this procedure. This is the best performance that can be achieved given available information.

Bayes Decision Rule: Two Category Case

- Bayes Decision Rule
  - For each input, select class with least conditional risk, i.e. choose class one if:

$$R(\alpha_1|\mathbf{x}) < R(\alpha_2|\mathbf{x})$$

  - where

$$\lambda_{ij} = \lambda(\alpha_i|\omega_j)$$

$$R(\alpha_1|\mathbf{x}) = \lambda_{11}P(\omega_1|\mathbf{x}) + \lambda_{12}P(\omega_2|\mathbf{x})$$

$$R(\alpha_2|\mathbf{x}) = \lambda_{21}P(\omega_1|\mathbf{x}) + \lambda_{22}P(\omega_2|\mathbf{x})$$

Alternate Equivalent Expressions of Bayes Decision Rule ("Choose Class 1 if … ")

- Posterior Class Probabilities

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

- Class Priors and Conditional Densities
  - Produced by applying Bayes Formula to the above, multiplying both sides by p(x)

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

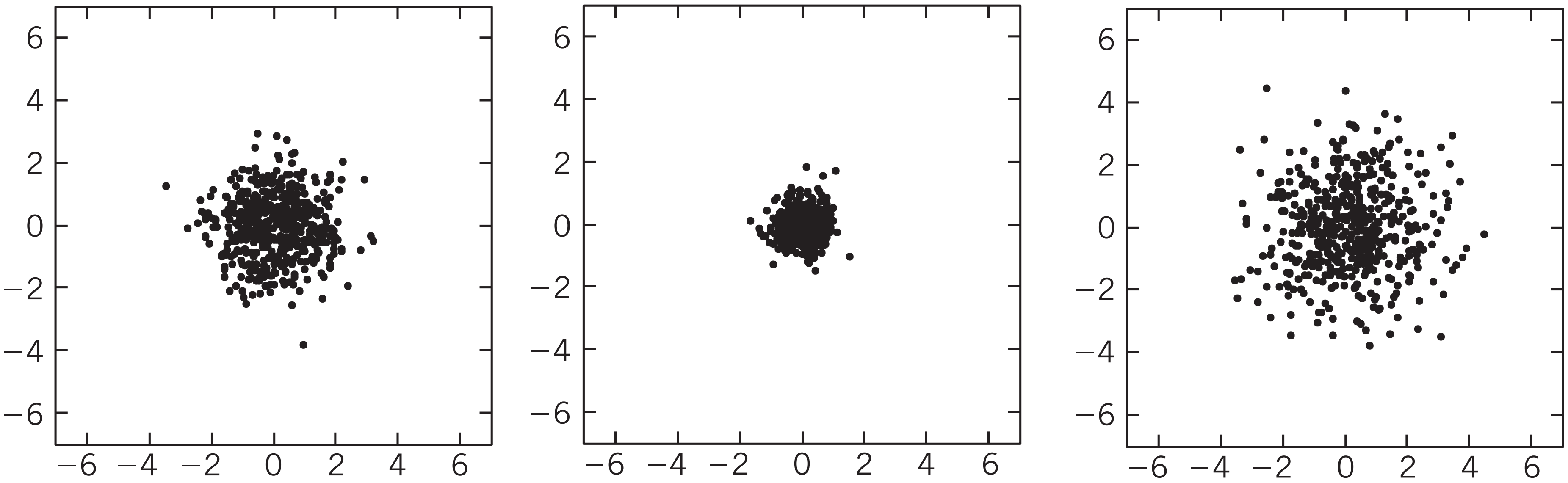
- Likelihood Ratio

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

Gaussian Distributions for $S = \begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{bmatrix}$ and $m = [0,\ 0]^T$

Spherically Shaped Data:

When the two coordinates of *x* are uncorrelated ($\sigma_{12} = 0$) and their variances are equal,



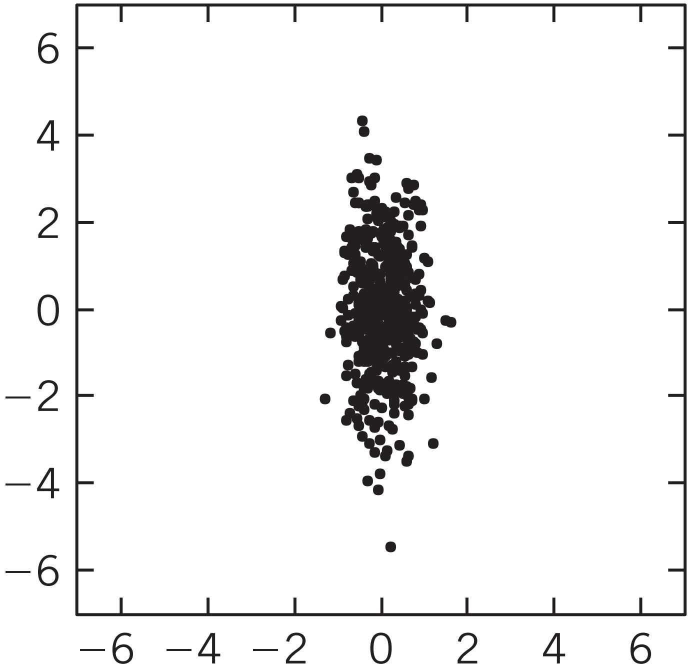
$\sigma_1^2 = \sigma_2^2 = 1,\ \sigma_{12} = 0$ $\sigma_1^2 = \sigma_2^2 = 0.2,\ \sigma_{12} = 0$ $\sigma_1^2 = \sigma_2^2 = 2,\ \sigma_{12} = 0$

Run Example 1.3.3

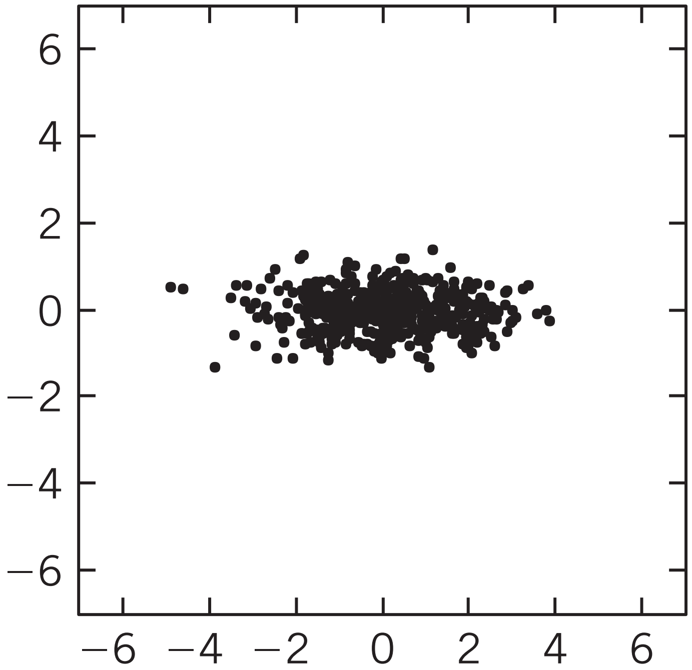
Gaussian Distributions for $S = \begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{bmatrix}$ and $m = [0,\ 0]^T$

Ellipsoidally Shaped Data:

When the two coordinates of *x* are uncorrelated ($\sigma_{12} = 0$) and their variances are UNequal,



$\sigma_1^2 = 0.2,\ \sigma_2^2 = 2,\ \sigma_{12} = 0$



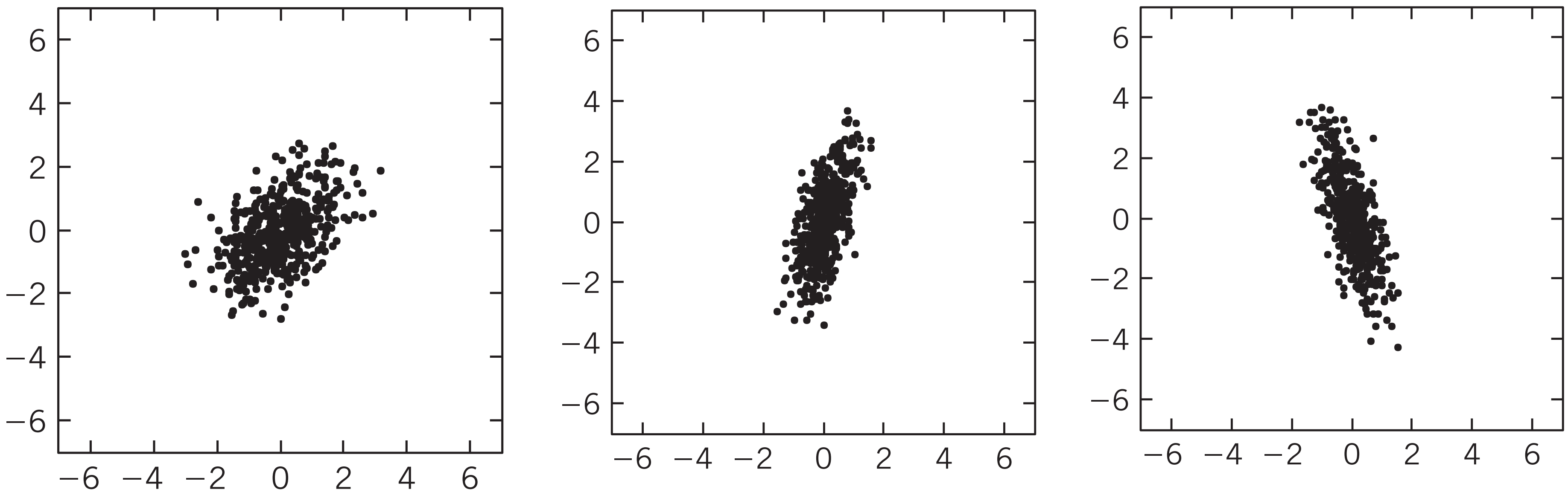
$\sigma_1^2 = 2,\ \sigma_2^2 = 0.2,\ \sigma_{12} = 0$

Run Example 1.3.3

Gaussian Distributions for $S = \begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{bmatrix}$ and $m = [0,\ 0]^T$

Spherically Shaped Data clustered unparalleled to the axes:

When the two coordinates of *x* are correlated ($\sigma_{12} \neq 0$) , The degree of rotation with respect to the axes depends on the value of σ_{12} ,



$\sigma_1^2 = \sigma_2^2 = 1, \sigma_{12} = 0.5$ $\quad$ $\sigma_1^2 = 0.3, \sigma_2^2 = 2, \sigma_{12} = 0.5$ $\quad$ $\sigma_1^2 = 0.3, \sigma_2^2 = 2, \sigma_{12} = -0.5$

Run Example 1.3.3

MINIMUM DISTANCE CLASSIFIERS

- The Euclidean Distance Classifier is the optimal Bayesian Classifier when:
  - The optimal Bayesian classifier is significantly simplified under the following assumptions:
  - The classes are equiprobable.
  - The data in all classes follow Gaussian distributions.
  - The covariance matrix is the same for all classes.
  - The covariance matrix is diagonal and all elements across the diagonal are equal. That is, $S = \sigma^2 I$, where I is the identity matrix.

$$||x - m_i|| \equiv \sqrt{(x - m_i)^T (x - m_i)} < ||x - m_j||, \quad \forall i \neq j$$

MINIMUM DISTANCE CLASSIFIERS

- The Mahalanobis Distance Classifier is the optimal Bayesian Classifier when the covairance matrix is not diagonal with equal elements:
  - The optimal Bayesian classifier is significantly simplified under the following assumptions:
  - The classes are equiprobable.
  - The data in all classes follow Gaussian distributions.
  - The covariance matrix is the same for all classes.

$$\sqrt{(x-m_i)^T S^{-1}(x-m_i)} < \sqrt{(x-m_j)^T S^{-1}(x-m_j)}, \quad \forall j \neq i$$

Run Example 1.4.1

Maximum Likelihood Parameter Estimation of Gaussian pdfs

- The maximum likelihood (ML) is a popular method for the estimation of an unknown mean value and the associated covariance matrix of a known pdf.

- Given N points, $x_i \in R^l$, $i = 1,2,\ldots,N$, which are known to be normally distributed, the ML estimates of the unknown mean value and the associated covariance matrix are given by:

$$m_{ML} = \frac{1}{N}\sum_{i=1}^{N} x_i$$

and

$$S_{ML} = \frac{1}{N}\sum_{i=1}^{N}(x_i - m_{ML})(x_i - m_{ML})^T$$

Run Example 1.4.2