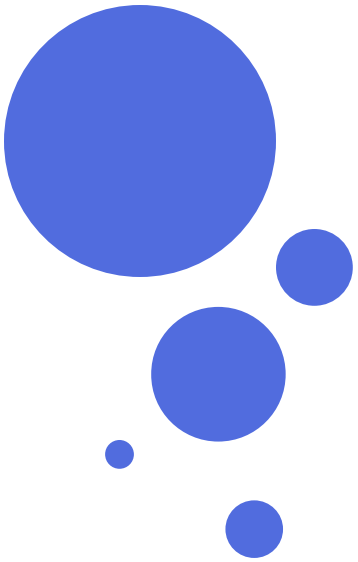
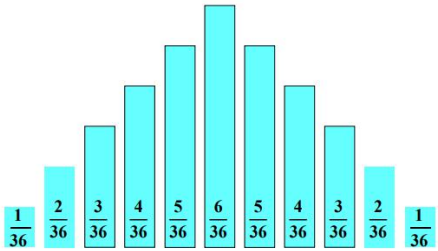
Rensselaer

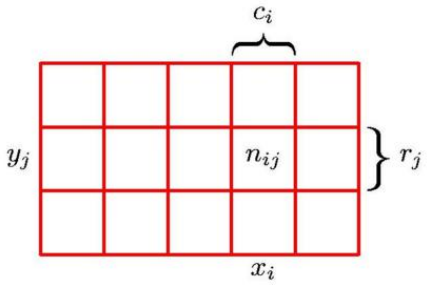


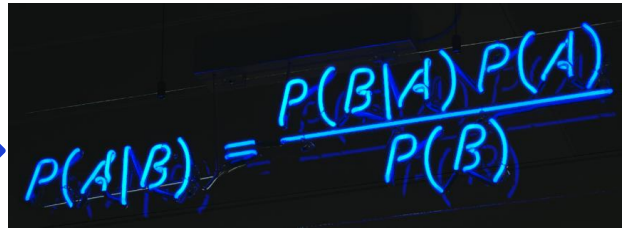
Lecture 3: Bayesian Decision Theory

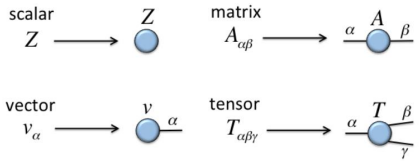
Dr. Chengjiang Long
Computer Vision Researcher at Kitware Inc.
Adjunct Professor at RPI.
Email: **longc3@rpi.edu**

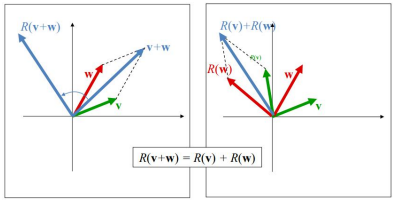
Recap Previous Lecture



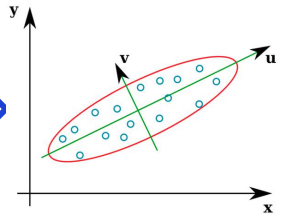








$$tr(M) = \sum \lambda_i$$
$$|M| = \prod \lambda_i$$



Outline

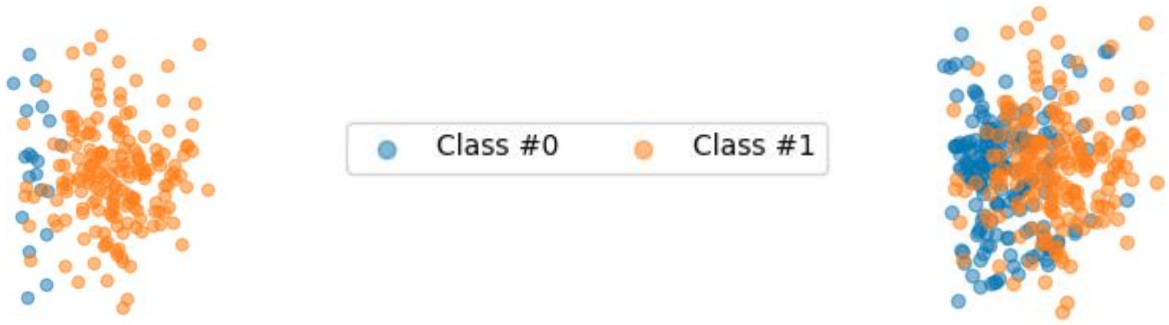
- What's Beyesian Decision Theory?
- A More General Theory
- Discriminant Function and Decision Boundary
- Multivariate Gaussian Density

Outline

- **What's Beyesian Decision Theory?**
- A More General Theory
- Discriminant Function and Decision Boundary
- Multivariate Gaussian Density

Bayesian Decision Theory

- Design classifiers to make decisions subject to minimizing an expected "risk".
  - The simplest risk is the classification error (*i.e.*, assuming that misclassification costs are equal).
  - When misclassification costs are not equal, the risk can include the cost associated with different misclassifications.



Terminology

- State of nature ω *(class label):*
  - e.g., ω_1 for sea bass, ω_2 for salmon

- Probabilities $P(\omega_1)$ and $P(\omega_2)$ *(priors):*
  - e.g., prior knowledge of how likely is to get a sea bass or a salmon

- Probability density function $p(x)$ *(evidence):*
  - e.g., how frequently we will measure a pattern with feature value x (e.g., x corresponds to lightness)

Terminology

- Conditional probability density $p(x/\omega_j)$ (*likelihood*) :
  - e.g., how frequently we will measure a pattern with feature value *x* given that the pattern belongs to class ω_j

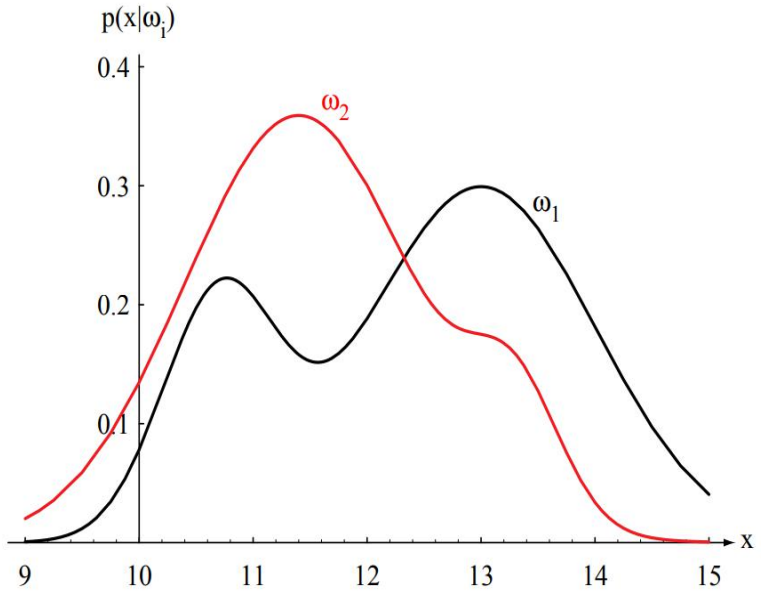
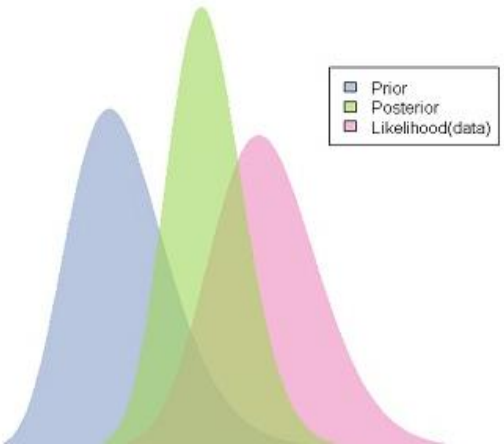


Figure 2.1: Hypothetical class-conditional probability density functions show the probability density of measuring a particular feature value x given the pattern is in category ω_i. If x represents the length of a fish, the two curves might describe the difference in length of populations of two types of fish. Density functions are normalized, and thus the area under each curve is 1.0.

Terminology

- Conditional probability $P(\omega_j / x)$ *(posterior)* :
  - e.g., the probability that the fish belongs to class ω_j given feature x.
- Ultimately, we are interested in computing $P(\omega_j / x)$ for each class ω_j.

Posterior distribution is proportional to data distribution * prior distribution



Decision Rule

Decide ω_1 if $P(\omega_1|x) > P(\omega_2|x)$; otherwise decide ω_2

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

or $P(error|x) = \min\left[P(\omega_1|x), P(\omega_2|x)\right]$

- Favours the most likely class.
- This rule will be making the same decision all times.
  - i.e., optimum if no other information is available

Decision Rule

- Using Bayes' rule:

$$P(\omega_j / x) = \frac{p(x/\omega_j)P(\omega_j)}{p(x)} = \frac{likelihood \times prior}{evidence}$$

where

$$p(x) = \sum_{j=1}^{2} p(x/\omega_j)P(\omega_j)$$

Decide ω_1 if $P(\omega_1/x) > P(\omega_2/x)$; otherwise **decide** ω_2

or

Decide ω_1 if $p(x/\omega_1)P(\omega_1) > p(x/\omega_2)P(\omega_2)$; otherwise **decide** ω_2

or

Decide ω_1 if $p(x/\omega_1)/p(x/\omega_2) > P(\omega_2)/P(\omega_1)$; otherwise **decide** ω_2

Decision Rule

$p(x/\omega_j)$ $P(\omega_1) = \frac{2}{3}$ $P(\omega_2) = \frac{1}{3}$ $P(\omega_j/x)$

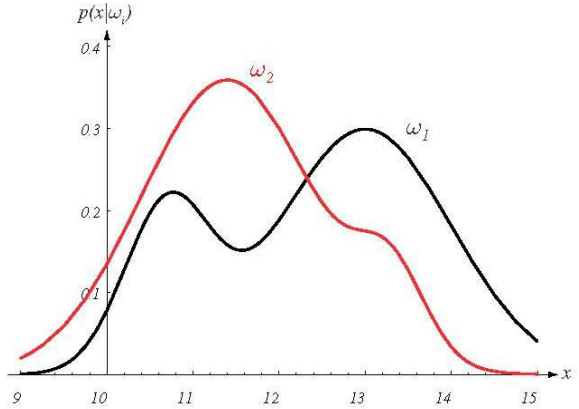


FIGURE 2.1. Hypothetical class-conditional probability density functions show the probability density of measuring a particular feature value x given the pattern is in category ω_i. If x represents the lightness of a fish, the two curves might describe the difference in lightness of populations of two 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,

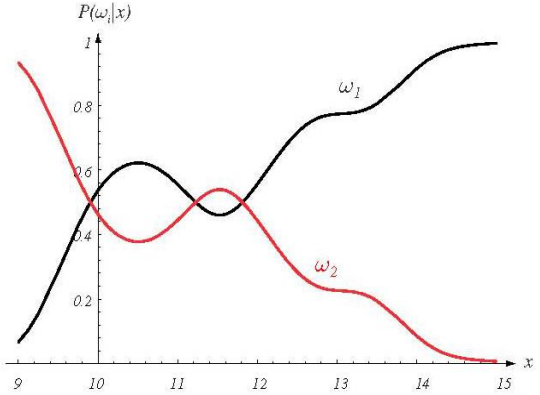


FIGURE 2.2. Posterior probabilities for the particular priors $P(\omega_1) = 2/3$ and $P(\omega_2) = 1/3$ for the class-conditional probability densities shown in Fig. 2.1. Thus in this case, given that a pattern is measured to have feature value $x = 14$, the probability it is in category ω_2 is roughly 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*. Copyright © 2001 by John Wiley & Sons, Inc.

Probability of Error

- The probability of error is defined as:

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

or $P(error|x) = \min\left[P(\omega_1|x), P(\omega_2|x)\right]$

- What is the average probability error?

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

- The Bayes rule is optimum, that is, it minimizes the average probability error!

Where do Probabilities come from?

- There are two competitive answers:

✓ **Relative frequency** (objective) approach.
  - Probabilities can only come from experiments.

✓ **Bayesian** (subjective) approach.
  - Probabilities may reflect degree of belief and can be based on opinion.

Example: Objective approach

- Classify cars whether they are more or less than $\%\ 50$K:
  - Classes: C_1 if price >50K, C_2 if price $<= 50$K
  - Features: x, the height of a car

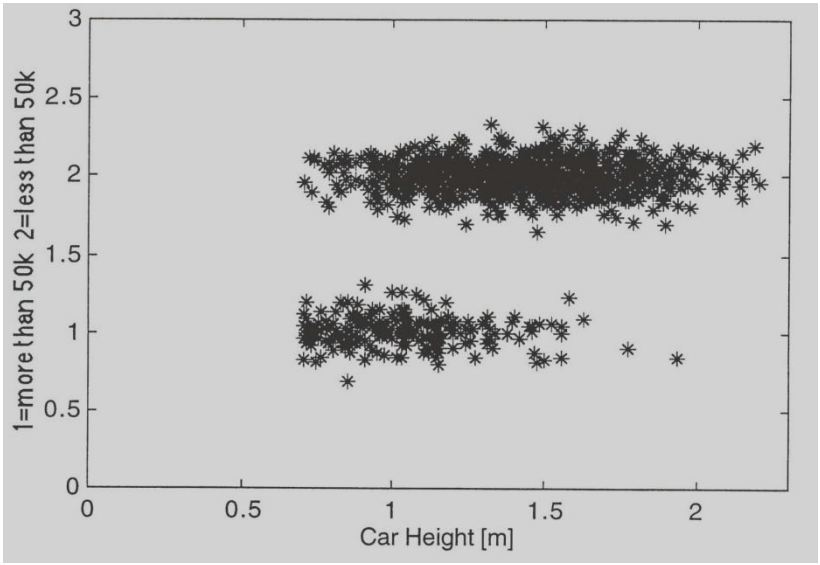
- Use the Bayes' rule to compute the posterior probabilities:

$$P(C_i / x) = \frac{p(x / C_i)P(C_i)}{p(x)}$$

- We need to estimate $p(x/C_1)$, $p(x/C_2)$, $P(C_1)$, $P(C_2)$

Example: Objective approach

- Collect data
  - Ask drivers how much their car was and measure height.
- Determine prior probabilities $P(C_1)$, $P(C_2)$
  - e.g., 1209 samples: #C_1=221 #C_2=988



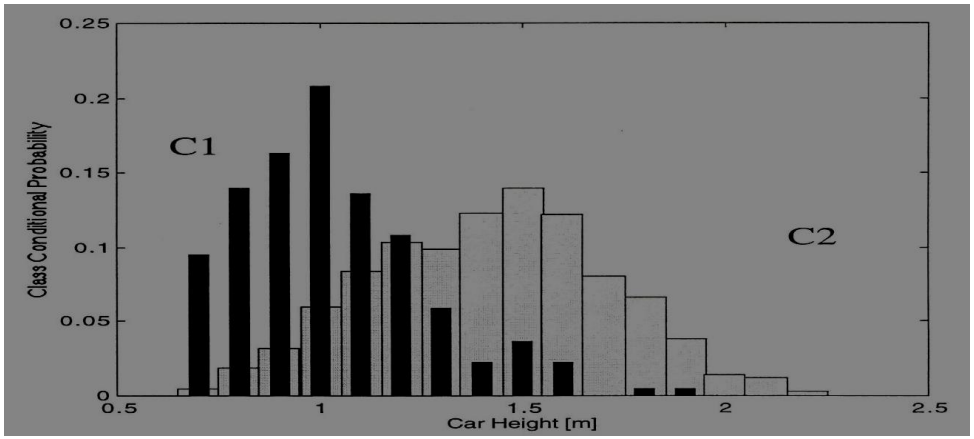
$$P(C_1) = \frac{221}{1209} = 0.183$$

$$P(C_2) = \frac{988}{1209} = 0.817$$

Example: Objective approach

- Determine class conditional probabilities (*likelihood*)
  - Discretize car height into bins and use normalized histogram

$$p(x / C_i)$$



- Calculate the posterior probability for each bin:

$$P(C_1 / x = 1.0) = \frac{p(x = 1.0 / C_1) P(C_1)}{p(x = 1.0 / C_1) P(C_1) + p(x = 1.0 / C_2) P(C_2)} =$$

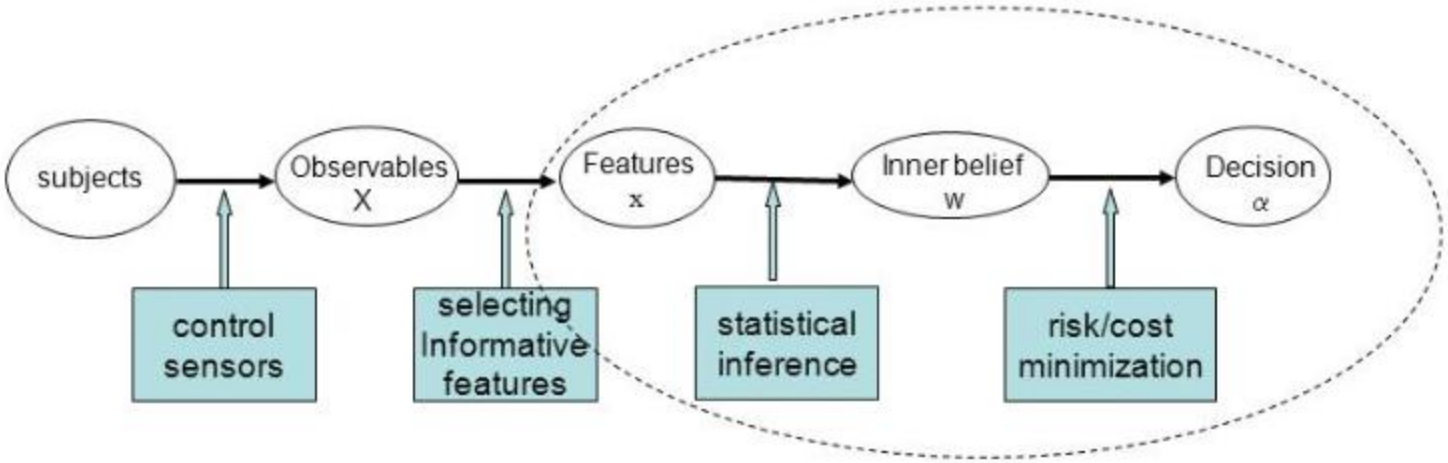
$$= \frac{0.2081 * 0.183}{0.2081 * 0.183 + 0.0597 * 0.817} = 0.438$$

Outline

- What's Beyesian Decision Theory?
- **A More General Theory**
- Discriminant Function and Decision Boundary
- Multivariate Gaussian Density

A More General Theory

- Use more than one features.
- Allow more than two categories.
- Allow actions other than classifying the input to one of the possible categories (e.g., rejection).
- Employ a more general error function (i.e., expected "risk") by associating a "cost" (based on a "loss" function) with different errors.



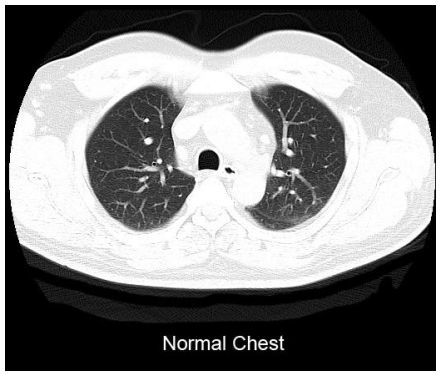
Terminology

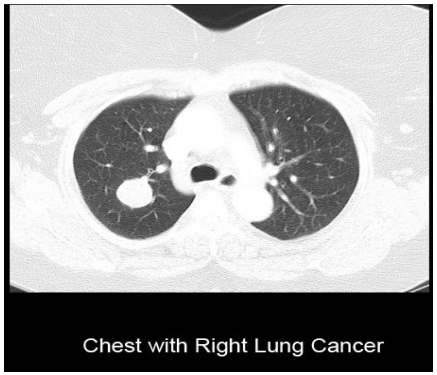
- Features form a vector $\mathbf{x} \in R^d$
- A set of *c* categories $\omega_1, \omega_2, \ldots, \omega_c$
- A finite set of ***l*** actions $\alpha_1, \alpha_2, \ldots, \alpha_l$
- A *loss* function $\lambda(\alpha_i / \omega_j)$
  - the cost associated with taking action α_i when the correct classification category is ω_j

Conditional Risk (or Expected Loss)

- Suppose we observe **x** and take action α_i
- The **conditional risk** (or **expected loss**) with taking action α_i is defined as:

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





$\lambda(\alpha_i|\omega_j)$

	cancer ω_1	normal ω_2
cancer α_1	0	1
normal α_2	100	0

$\omega_1 = \text{cancer}, \quad \omega_2 = \text{normal}$

$\alpha_1 = \text{cancer}, \quad \alpha_2 = \text{normal}$

From a medical image, we want to classify (determine) whether it contains cancer tissues or not.

Overall Risk

- Suppose $\alpha(\mathbf{x})$ is a general decision rule that determines which action $\alpha_1, \alpha_2, \ldots, \alpha_l$ to take for every **x**.

- The overall risk is defined as:

$$R = \int R(a(\mathbf{x}) / \mathbf{x}) p(\mathbf{x}) d\mathbf{x}$$

- The optimum decision rule is the *Bayes rule*

Overall Risk

- The *Bayes rule* minimizes R by:
  (i) Computing $R(\alpha_i / \mathbf{x})$ for every α_i given an $\mathbf{x}$
  (ii) Choosing the action α_i with the minimum $R(\alpha_i / \mathbf{x})$

- The resulting minimum R^* is called *Bayes risk* and is the best (i.e., optimum) performance that can be achieved:

$$R^* = \min R$$

Example: Two-category classification

- Define
  - α_1: decide ω_1
  - α_2: decide ω_2
  - $\lambda_{ij} = \lambda(\alpha_i / \omega_j)$
- The conditional risks are:

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

$$R(a_1/\mathbf{x}) = \lambda_{11} P(\omega_1/\mathbf{x}) + \lambda_{12} P(\omega_2/\mathbf{x})$$
$$R(a_2/\mathbf{x}) = \lambda_{21} P(\omega_1/\mathbf{x}) + \lambda_{22} P(\omega_2/\mathbf{x})$$

Example: Two-category classification

- Minimum risk decision rule:

$$\textbf{Decide } \omega_1 \text{ if } R(a_1/\mathbf{x}) < R(a_2/\mathbf{x}); \text{ otherwise decide } \omega_2$$

or

$$\textbf{Decide } \omega_1 \text{ if } (\lambda_{21} - \lambda_{11})P(\omega_1/\mathbf{x}) > (\lambda_{12} - \lambda_{22})P(\omega_2/\mathbf{x}); \text{ otherwise decide } \omega_2$$

or (i.e., using likelihood ratio)

$$\textbf{Decide } \omega_1 \text{ if } \frac{p(\mathbf{x}/\omega_1)}{p(\mathbf{x}/\omega_2)} > \frac{(\lambda_{12} - \lambda_{22})}{(\lambda_{21} - \lambda_{11})} \frac{P(\omega_2)}{P(\omega_1)}; \text{ otherwise decide } \omega_2$$

likelihood ratio threshold

Special Case: Zero-One Loss Function

- Assign the same loss to all errors:

$$\lambda(a_i/\omega_j) = \begin{cases} 0 & i = j \\ 1 & i \neq j \end{cases}$$

- The conditional risk corresponding to this loss function:

$$R(a_i/\mathbf{x}) = \sum_{j=1}^{c} \lambda(a_i/\omega_j) P(\omega_j/\mathbf{x}) = \sum_{i \neq j} P(\omega_j/\mathbf{x}) = 1 - P(\omega_i/\mathbf{x})$$

Special Case: Zero-One Loss Function

- The decision rule becomes:

$$\textbf{Decide } \omega_1 \text{ if } R(a_1/\mathbf{x}) < R(a_2/\mathbf{x}); \text{ otherwise decide } \omega_2$$

or $$\textbf{Decide } \omega_1 \text{ if } 1 - P(\omega_1/\mathbf{x}) < 1 - P(\omega_2/\mathbf{x}); \text{ otherwise decide } \omega_2$$

or $$\textbf{Decide } \omega_1 \text{ if } P(\omega_1/\mathbf{x}) > P(\omega_2/\mathbf{x}); \text{ otherwise decide } \omega_2$$

- The overall risk turns out to be the average probability error!

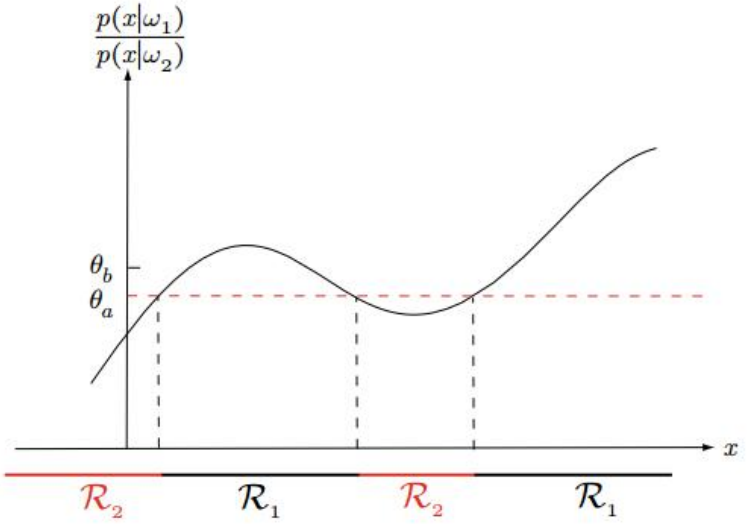
Example

- Assuming general loss:

$$\textbf{Decide } \omega_1 \text{ if } \frac{p(\mathbf{x}/\omega_1)}{p(\mathbf{x}/\omega_2)} > \frac{(\lambda_{12}-\lambda_{22})}{(\lambda_{21}-\lambda_{11})}\frac{P(\omega_2)}{P(\omega_1)}\text{; otherwise decide } \omega_2$$

- Assuming zero–one loss:

Decide ω_1 if $p(x/\omega_1)/p(x/\omega_2) > P(\omega_2)/P(\omega_1)$ otherwise **decide** ω_2



(decision regions)

$$\theta_a = P(\omega_2)/P(\omega_1)$$

$$\theta_b = \frac{P(\omega_2)(\lambda_{12}-\lambda_{22})}{P(\omega_1)(\lambda_{21}-\lambda_{11})}$$

assume: $\lambda_{12} > \lambda_{21}$

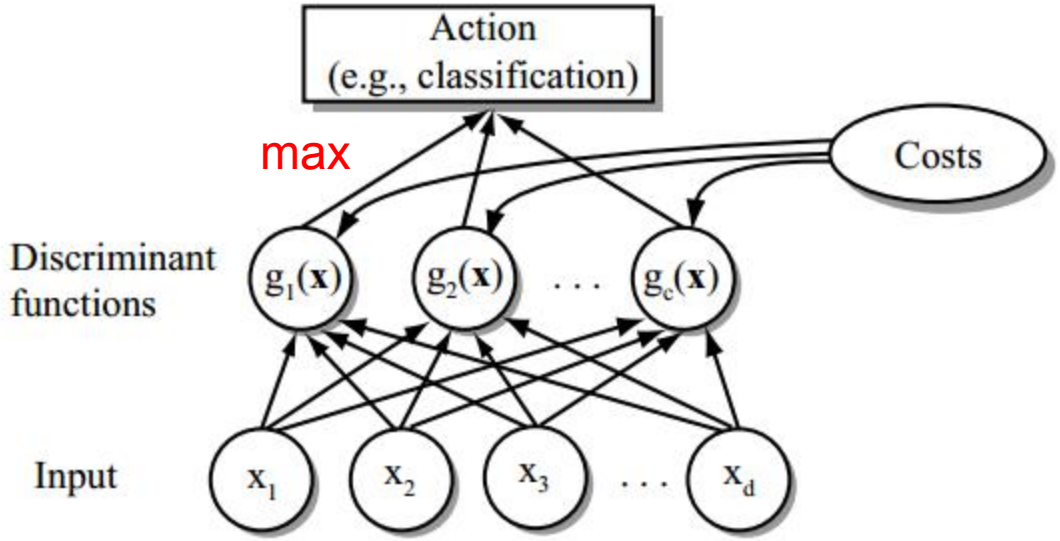
Outline

- What's Beyesian Decision Theory?
- A More General Theory
- **Discriminant Function and Decision Boundary**
- Multivariate Gaussian Density
- Error Bound, ROC, Missing Features and Compound Bayesian Decision Theory
- Summary

Discriminant Functions

- A useful way to represent a classifier is through **discriminant functions** $g_i(x)$, i = 1, . . . , c, where a feature vector x is assigned to class ω_i if

$$g_i(\mathbf{x}) > g_j(\mathbf{x}) \text{ for all } j \neq i$$



Discriminants for Bayes Classifier

- Is the choice of g_i unique?
  - Replacing $g_i(\mathbf{x})$ with $f(g_i(\mathbf{x}))$, where $f(\cdot)$ is monotonically increasing, does not change the classification results.

$g_i(x)=P(\omega_i/x)$ ⇨

$$g_i(\mathbf{x}) = \frac{p(\mathbf{x}/\omega_i)P(\omega_i)}{p(\mathbf{x})}$$

$$g_i(\mathbf{x}) = p(\mathbf{x}/\omega_i)P(\omega_i)$$

$$g_i(\mathbf{x}) = \ln p(\mathbf{x}/\omega_i) + \ln P(\omega_i)$$

we'll use this discriminant extensively!

Case of two categories

- More common to use a single discriminant function (*dichotomizer*) instead of two:

$$g(\mathbf{x}) = g_1(\mathbf{x}) - g_2(\mathbf{x})$$

Decide ω_1 if $g(\mathbf{x}) > 0$; otherwise decide ω_2

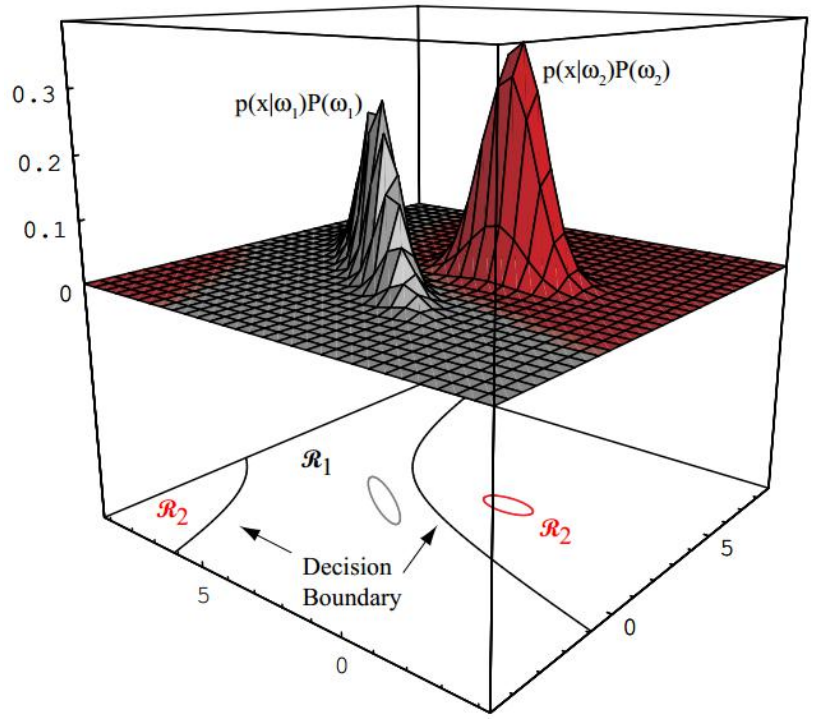
Examples:

$$g(\mathbf{x}) = P(\omega_1 / \mathbf{x}) - P(\omega_2 / \mathbf{x})$$

$$g(\mathbf{x}) = \ln \frac{p(\mathbf{x} / \omega_1)}{p(\mathbf{x} / \omega_2)} + \ln \frac{P(\omega_1)}{P(\omega_2)}$$

Decision Regions and Boundaries

- Discriminants divide the feature space in decision regions $R_1, R_2, \ldots, R_c$, separated by decision boundaries.



Decision boundary is defined by:

$$g_1(x)=g_2(x)$$

Outline

- What's Beyesian Decision Theory?
- A More General Theory
- Discriminant Function and Decision Boundary
- **Multivariate Gaussian Density**

Why are Gaussians so Useful?

- They represent many probability distributions in nature quite accurately. In our case, when patterns can be represented as random variations of an ideal prototype (represented by the mean feature vector)
  - Everyday examples: height, weight of a population

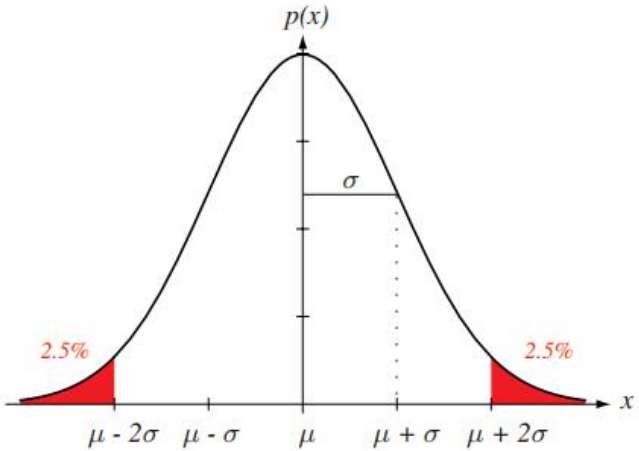


FIGURE 2.7. A univariate normal distribution has roughly 95% of its area in the range $|x - \mu| \leq 2\sigma$, as shown. The peak of the distribution has value $p(\mu) = 1/\sqrt{2\pi}\sigma$. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

Multivariate Gaussian Density

- A normal distribution over two or more variables (d variables/dimensions)

$$p(\mathbf{x}) = \frac{1}{(2\pi)^{d/2}|\mathbf{\Sigma}|^{1/2}} \exp\left[-\frac{1}{2}(\mathbf{x}-\mu)^t \mathbf{\Sigma}^{-1}(\mathbf{x}-\mu)\right]$$

$$\mu = \int_{-\infty}^{\infty} \mathbf{x}\, p(\mathbf{x})\, d\mathbf{x}$$

$$\mathbf{\Sigma} = \int (\mathbf{x}-\mu)(\mathbf{x}-\mu)^t p(\mathbf{x}) d\mathbf{x}$$

The Covariance Matrix

- For our purposes…
  - Assume matrix is positive definite, so the determinant of the matrix is always positive
- Matrix Elements
  - Main diagonal: variances for each individual variable
  - Off–diagonal: covariances of each variable pairing i & j (note: values are repeated, as matrix is symmetric)

$$\Sigma = \begin{bmatrix} \mathrm{E}[(X_1 - \mu_1)(X_1 - \mu_1)] & \mathrm{E}[(X_1 - \mu_1)(X_2 - \mu_2)] & \cdots & \mathrm{E}[(X_1 - \mu_1)(X_n - \mu_n)] \\ \mathrm{E}[(X_2 - \mu_2)(X_1 - \mu_1)] & \mathrm{E}[(X_2 - \mu_2)(X_2 - \mu_2)] & \cdots & \mathrm{E}[(X_2 - \mu_2)(X_n - \mu_n)] \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{E}[(X_n - \mu_n)(X_1 - \mu_1)] & \mathrm{E}[(X_n - \mu_n)(X_2 - \mu_2)] & \cdots & \mathrm{E}[(X_n - \mu_n)(X_n - \mu_n)] \end{bmatrix}$$

Discriminant Function for Multivariate Gaussian Density

- We will consider three special cases for:
  - normally distributed features, and
  - minimum error rate classification $(0-1$ loss$)$

❑ Recall: $$g_i(\mathbf{x}) = \ln p(\mathbf{x}|\omega_i) + \ln P(\omega_i)$$

if $p(x|\omega_i) \sim N(\mu_\mathbf{i}, \Sigma_\mathbf{i})$ then approx. $p(\mathbf{x}|\omega_i)$

using: $$p(\mathbf{x}) = \frac{1}{(2\pi)^{d/2}|\mathbf{\Sigma}|^{1/2}} \exp\left[-\frac{1}{2}(\mathbf{x}-\mu)^t\mathbf{\Sigma}^{-1}(\mathbf{x}-\mu)\right]$$

Minimum Error-Rate Discriminant Function for Multivariate Gaussian Feature Distributions

- ln (natural log) of

$$p(\mathbf{x}) = \frac{1}{(2\pi)^{d/2}|\mathbf{\Sigma}|^{1/2}} \exp\left[-\frac{1}{2}(\mathbf{x}-\mu)^t \mathbf{\Sigma}^{-1}(\mathbf{x}-\mu)\right]$$

gives a general form for our discriminant functions:

$$g_i(\mathbf{x}) = -\frac{1}{2}(\mathbf{x}-\mu_\mathbf{i})^t \mathbf{\Sigma_i}^{-1}(\mathbf{x}-\mu_\mathbf{i}) - \frac{d}{2}\ln 2\pi - \frac{1}{2}\ln|\mathbf{\Sigma_i}| + \ln P(\omega_i)$$

Special Cases for Binary Classification

- Purpose

  Overview of commonly assumed cases for feature likelihood densities,

  - Goal: eliminate common additive constants in discriminant functions. These do not affect the classification decision (i.e. define $g_i(x)$ providing "just the differences")
  - Also, look at resulting decision surfaces ($g_i(x) = g_j(x)$)
- Three Special Cases
  1. Statistically independent features, identically distributed Gaussians for each class
  2. Identical covariances for each class
  3. Arbitrary covariances

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

$$g_i(\mathbf{x}) = -\frac{1}{2}(\mathbf{x}-\mu_\mathbf{i})^t \mathbf{\Sigma_i}^{-1}(\mathbf{x}-\mu_\mathbf{i}) - \frac{d}{2}\ln 2\pi - \frac{1}{2}\ln|\mathbf{\Sigma_i}| + \ln P(\omega_i)$$

- Satisfiy two conditions: (1) Features are statistically independent and (2) Each feature has the same variance.
- Remove Items in red: same across classes ("unimportant additive constants")
- Inverse of covariance matrix: $\mathbf{\Sigma_i^{-1}} = (1/\sigma^2)\mathbf{I}$
- Only effect is to scale vector product by $1/\sigma^2$
- Discriminant function:

$$g_i(x) = -\frac{(\mathbf{x}-\mu_\mathbf{i})^t(\mathbf{x}-\mu_\mathbf{i})}{2\sigma^2} + \ln P(\omega_i)$$

$$g_i(x) = -\frac{1}{2\sigma^2}[\mathbf{x}^t\mathbf{x} - 2\mu_\mathbf{i}^t\mathbf{x} + \mu_\mathbf{i}{}^t\mu_\mathbf{i}] + \ln P(\omega_i)$$

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

- Linear Discriminant Function
- Produced by factoring the previous form

$$g_i(x) = \mathbf{w_i^t x} + \omega_{i0}$$

$$g_i(x) = \frac{1}{\sigma^2}\mu_i{}^t \mathbf{x} - \frac{1}{2\sigma^2}\mu_i^t\mu_i + \ln P(\omega_i)$$

- Threshold or Bias for Class i: ω_{i0}
- Change in prior translates decision boundary

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

- Decision Boundary: $g_i(x) = g_j(x)$

$$\mathbf{w}^t(\mathbf{x} - \mathbf{x_0}) = 0$$

$$(\mu_i - \mu_j)^t \left(x - \left(\frac{1}{2}(\mu_i + \mu_j) - \frac{\sigma^2}{(\mu_i - \mu_j)^t(\mu_i - \mu_j)} \ln \frac{P(\omega_i)}{P(\omega_j)}(\mu_i - \mu_j)\right)\right)$$

  - Decision boundary goes through x0 along line between means, orthogonal to this line
  - If priors equal, x0 between means (minimum distance classifier), otherwise x0 shifted
  - If variance small relative to distance between means, priors have limited effect on boundary location

Case 1: Statistically Independent Features with Identical Variances

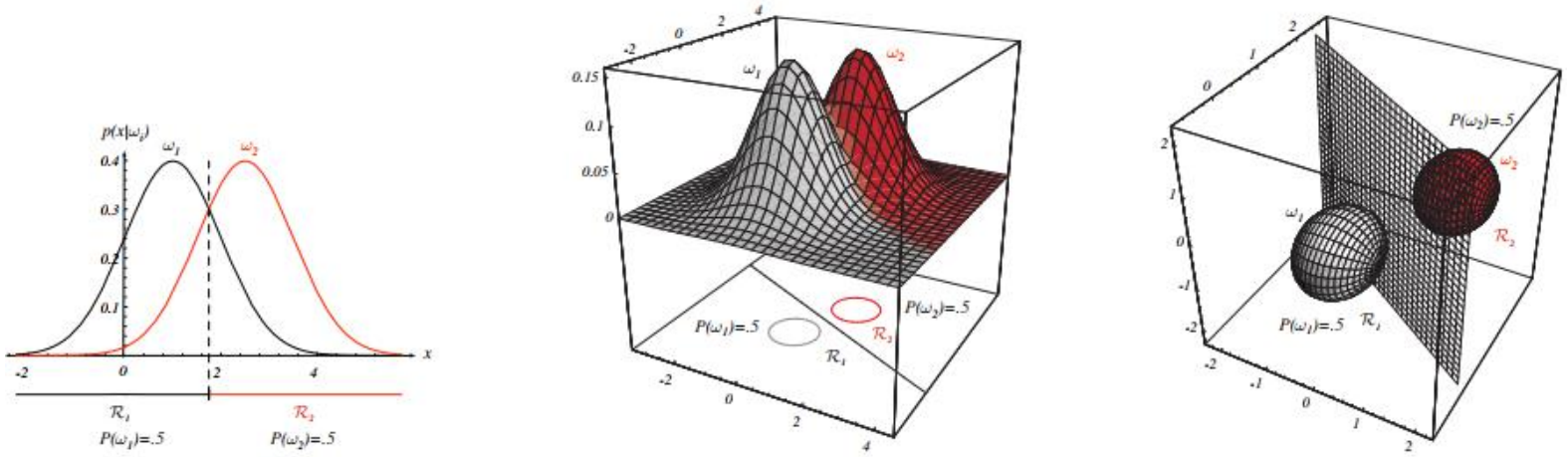
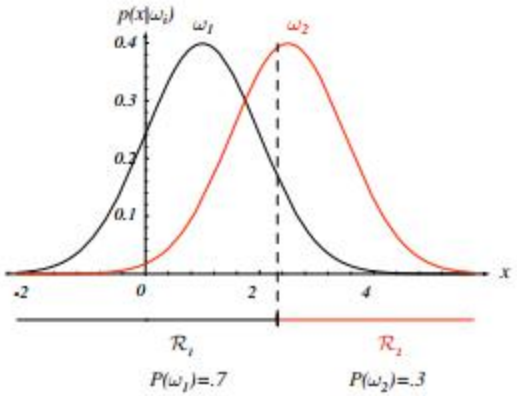
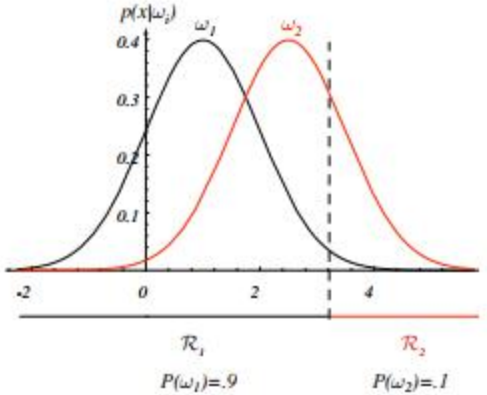
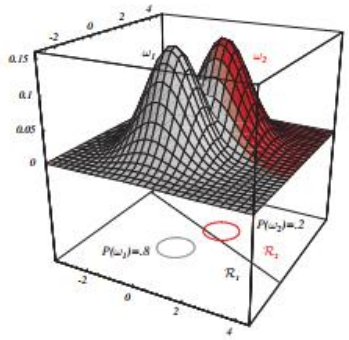


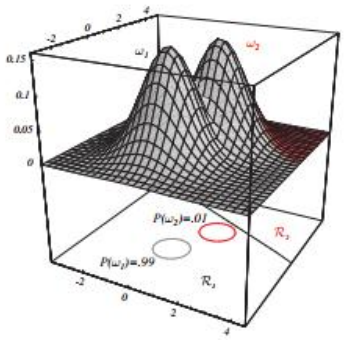
FIGURE 2.10. If the covariance matrices for two distributions are equal and proportional to the identity matrix, then the distributions are spherical in d dimensions, and the boundary is a generalized hyperplane of $d-1$ dimensions, perpendicular to the line separating the means. In these one-, two-, and three-dimensional examples, we indicate $p(\mathbf{x}|\omega_i)$ and the boundaries for the case $P(\omega_1) = P(\omega_2)$. In the three-dimensional case, the grid plane separates $\mathcal{R}_1$ from $\mathcal{R}_2$. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

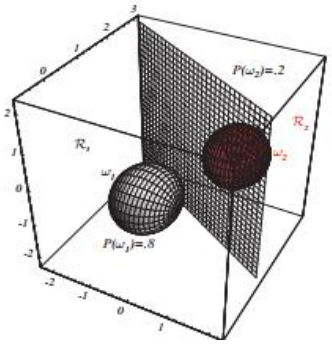
Example: Translation of Decision Boundaries Through Changing Priors

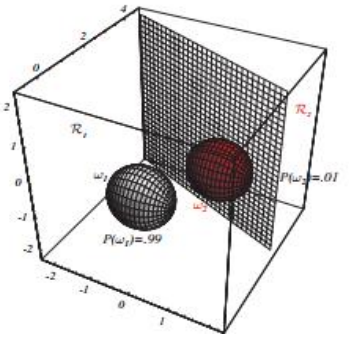












Case II: Identical Covariances $\boldsymbol{\Sigma_i = \Sigma}$

$$g_i(\mathbf{x}) = -\frac{1}{2}(\mathbf{x}-\mu_\mathbf{i})^t\boldsymbol{\Sigma_i}^{-1}(\mathbf{x}-\mu_\mathbf{i}) - \frac{d}{2}\ln 2\pi - \frac{1}{2}\ln|\boldsymbol{\Sigma_i}| + \ln P(\omega_i)$$

- Remove terms in red as in Case I these can be ignored (same across classes)
- Squared Mahalanobis Distance (yellow)
  - Distance from x to mean for class i, taking covariance into account; defines contours of fixed density

Case II: Identical Covariances $\boldsymbol{\Sigma_i = \Sigma}$

- Expansion of squared Mahalanobis distance

$$
\begin{aligned}
&(\mathbf{x}-\mu_i)^t\Sigma^{-1}(\mathbf{x}-\mu_\mathbf{i}) \\
&\quad = \mathbf{x}^\mathbf{t}\Sigma^{-1}\mathbf{x} - \mathbf{x}^\mathbf{t}\Sigma^{-1}\mu_\mathbf{i} - \mu_\mathbf{i}^\mathbf{t}\Sigma^{-1}\mathbf{x} + \mu_\mathbf{i}^\mathbf{t}\Sigma^{-1}\mu_\mathbf{i} \\
&\quad = \mathbf{x}^\mathbf{t}\Sigma^{-1}\mathbf{x} - 2(\Sigma^{-1}\mu_\mathbf{i})^\mathbf{t}\mathbf{x} + \mu_\mathbf{i}^\mathbf{t}\Sigma^{-1}\mu_\mathbf{i}
\end{aligned}
$$

the last step comes from symmetry of the covariance matrix and thus its inverse:

$$\Sigma^t = \Sigma, (\Sigma^{-1})^t = \Sigma^{-1}$$

- Once again, term above in red is an additive constant independent of class, and can be removed

Multivariate Gaussian Density: $\boldsymbol{\Sigma_i = \Sigma}$

- Linear Discriminant Function

$$g_i(x) = \mathbf{w_i^t x} + \omega_{i0}$$

$$g_i(\mathbf{x}) = (\Sigma^{-1}\mu_i)^t \mathbf{x} \quad -\frac{1}{2}\mu_i^t \Sigma^{-1}\mu_i + \ln P(\omega_i)$$

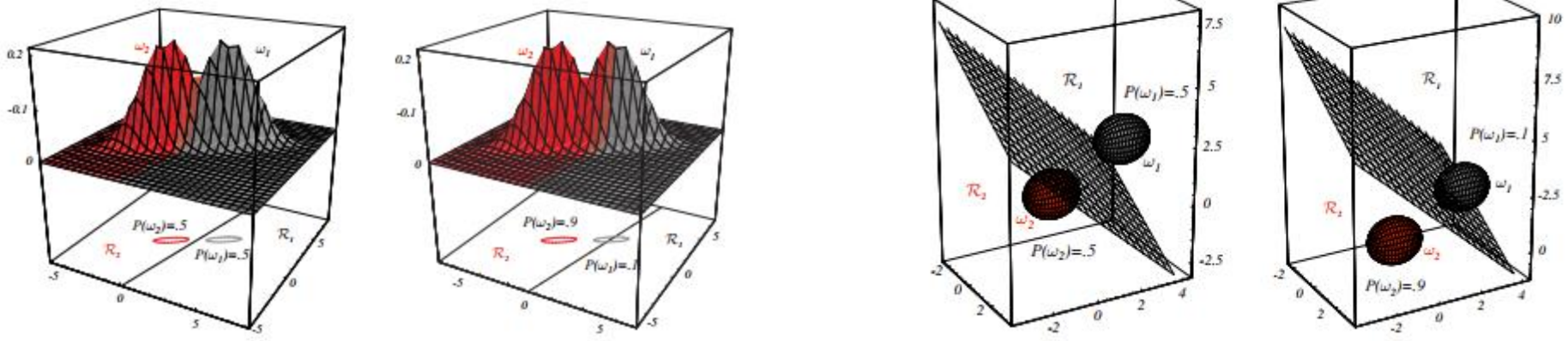
- Decision Boundary: $g_i(x) = g_j(x)$

$$\mathbf{w}^t(\mathbf{x} - \mathbf{x_0}) = 0$$

$$(\Sigma^{-1}(\mu_\mathbf{i} - \mu_\mathbf{j}))^\mathbf{t}(\mathbf{x} - \left(\frac{1}{2}(\mu_\mathbf{i} + \mu_\mathbf{j}) - \frac{\ln[\mathbf{P}(\omega_\mathbf{i})/\mathbf{P}\omega_\mathbf{j})}{(\mu_\mathbf{i} - \mu_\mathbf{j})\Sigma^{-1}(\mu_\mathbf{i} - \mu_\mathbf{j})}(\mu_\mathbf{i} - \mu_\mathbf{j})\right)) = 0$$

Case II: Identical Covariances $\boldsymbol{\Sigma_i = \Sigma}$

- Notes on Decision Boundary
  - As for Case I, passes through point x0 lying on the line between the two class means. Again, x0 in the middle if priors identical
  - Hyperplane defined by boundary generally not orthogonal to the line between the two means



Case III: arbitrary $\boldsymbol{\Sigma_i}$

$$g_i(\mathbf{x}) = -\frac{1}{2}(\mathbf{x}-\mu_\mathbf{i})^t\boldsymbol{\Sigma_i}^{-1}(\mathbf{x}-\mu_\mathbf{i}) - \frac{d}{2}\ln 2\pi - \frac{1}{2}\ln|\boldsymbol{\Sigma_i}| + \ln P(\omega_i)$$

- Can only remove the one term in red aboveSquared
- Discriminant Function (quadratic)

$$g_i(x) = x^t W_i x + w_i^t x + \omega_{i0}$$

$$g_i(x) = x^t\left(-\frac{1}{2}\Sigma_i^{-1}\right)x + (\Sigma_i^{-1}\mu_i)^t x - \frac{1}{2}\mu_i^t\Sigma_i^{-1}\mu_i - \frac{1}{2}\ln|\Sigma_i| + \ln P(\omega_i)$$

Case III: arbitrary Σ_i

- Decision Boundaries
  - Are hyperquadrics: can be hyperplanes,
  - hyperplane pairs, hyperspheres,
  - hyperellipsoids, hyperparabaloids,
  - hyperhyperparabaloids
- Decision Regions
  - Need not be simply connected, even in one dimension (next slide)

Case III: arbitrary Σ_i

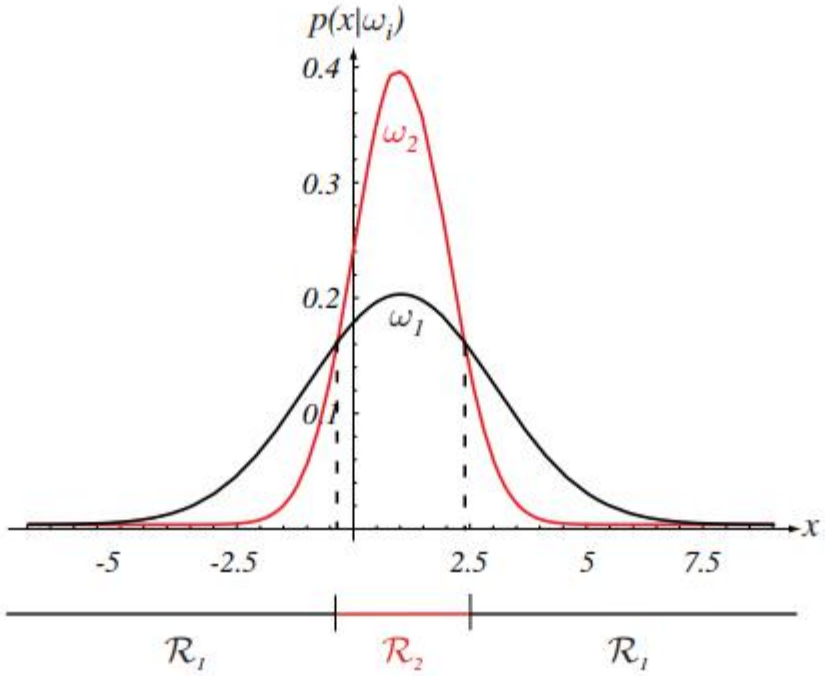
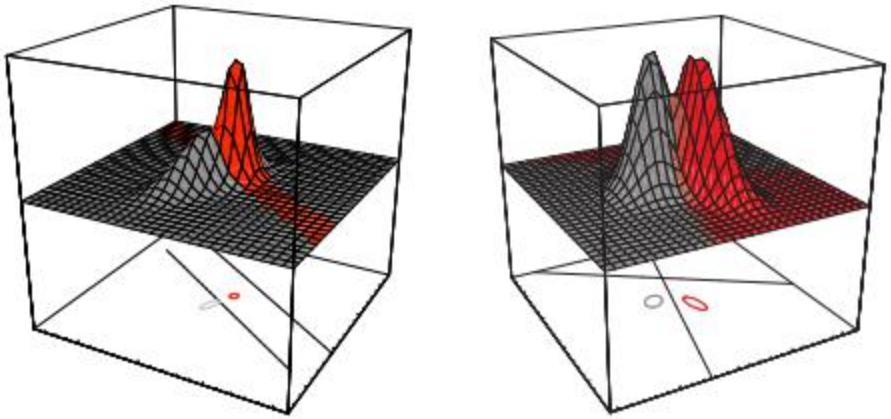
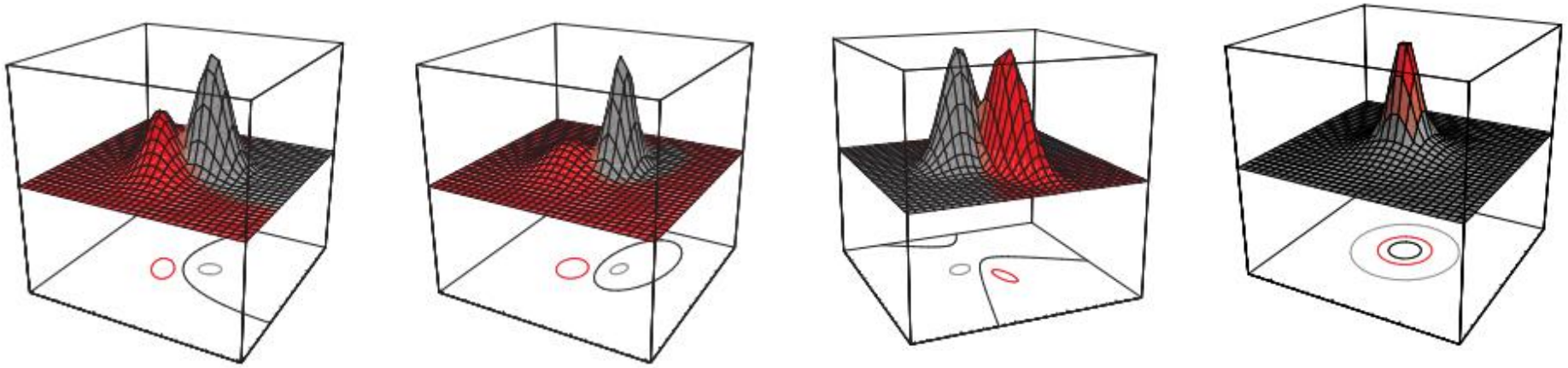


FIGURE 2.13. Non-simply connected decision regions can arise in one dimensions for Gaussians having unequal variance. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

Case III: arbitrary Σ_i



Nonlinear decision boundaries



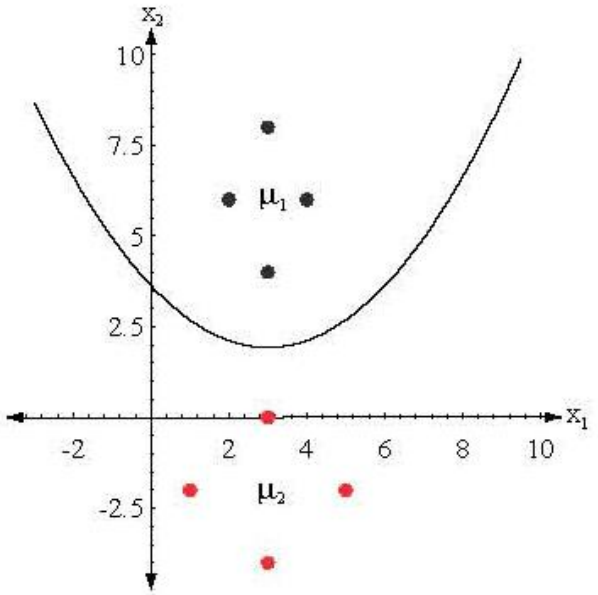
Example: Case III

$$\boldsymbol{\mu}_1 = \begin{bmatrix} 3 \\ 6 \end{bmatrix}; \quad \boldsymbol{\Sigma}_1 = \begin{pmatrix} 1/2 & 0 \\ 0 & 2 \end{pmatrix} \text{ and } \boldsymbol{\mu}_2 = \begin{bmatrix} 3 \\ -2 \end{bmatrix}; \quad \boldsymbol{\Sigma}_2 = \begin{pmatrix} 2 & 0 \\ 0 & 2 \end{pmatrix}.$$

decision boundary: $x_2 = 3.514 - 1.125x_1 + 0.1875x_1^2.$

$P(\omega_1)=P(\omega_2)$

boundary does not pass through midpoint of μ_1,μ_2



Q & A