EE5907R: Pattern Recognition

Lecture 2: Bayesian Decision Theory (II)

National University of Singapore

Recap

- Bayesian decision theory (BDT)
  - For each class j, $p(x \mid \omega_j)$ and $P\,(\omega_j)$ are known.
  - Posterior = (Likelihood * Prior) / Evidence

    Posterior $\propto$ (Likelihood * Prior)
- Design optimal classifier using BDT
  - Loss function determines what is “optimal”
  - With zero-one loss function, the optimal classifier achieves minimum error rate
    - ✓ MAP classifier: $P(\omega_j \mid x\ )$
    - ✓ ML classifier (equal priors): $p(x \mid \omega_j)$

Outline

- Classifiers and discriminant functions
- Bayes classifiers for normally (Gaussian) distributed classes
  - Univariate
  - Multivariate
    - ✓ Case 1: $\Sigma_i=\sigma^2 I$
    - ✓ Case 2: $\Sigma_i=\Sigma$
    - ✓ Case 3: $\Sigma_i=\sigma_i^2 I$
    - ✓ Case 4: Σ_i=arbitrary (general case)
- Numerical examples
- Summary

Classifiers and Discriminant Functions

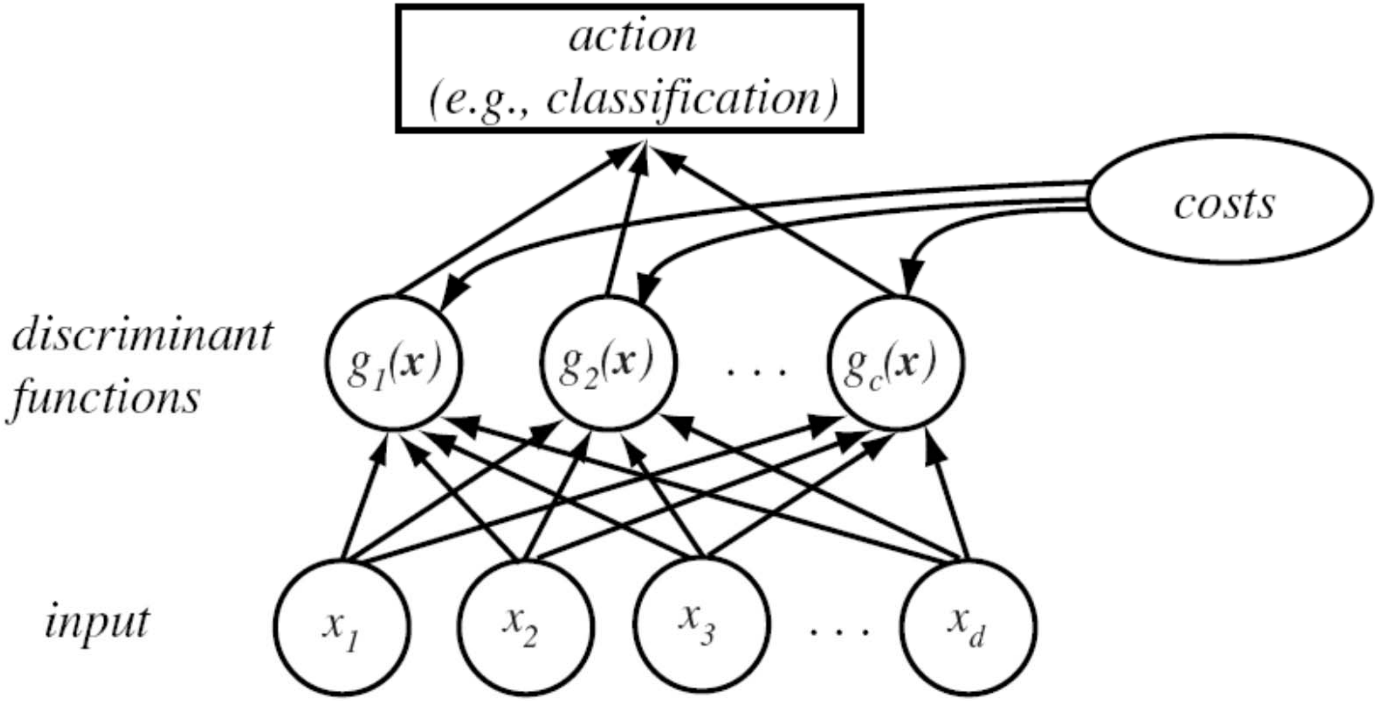
- ***Classifiers*** can be represented in terms of a set of ***discriminant functions*** $g_i(x)$, for $i = 1,\dots, c$
  - The classifier assigns a feature vector x to class ω_i if:

$$g_i(x) > g_j(x) \; \forall j \neq i$$

  - A network that computes *c* discriminant functions
  - Many types: linear, non-linear, high-order, parametric, non-parametric, etc.

A General Statistical Pattern Classifier

The discriminant classifier can be viewed as a network (for *c* classes and a *d*-dimensional input vector)



From: R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Recognition.* Copyright © 2001 by John Wiley & Sons, Inc.

Discriminant Functions

- Let $g_i(x) = - R(\alpha_i \mid x)$
  (max. discriminant corresponds to min. risk!)

- For the minimum error rate, we take

$$g_i(x) = P(\omega_i \mid x)$$

(max. discrimination corresponds to max. posterior!)

Not unique:

$$g_i(x) \equiv p(x \mid \omega_{\mathbf{i}})\, P(\omega_i)$$

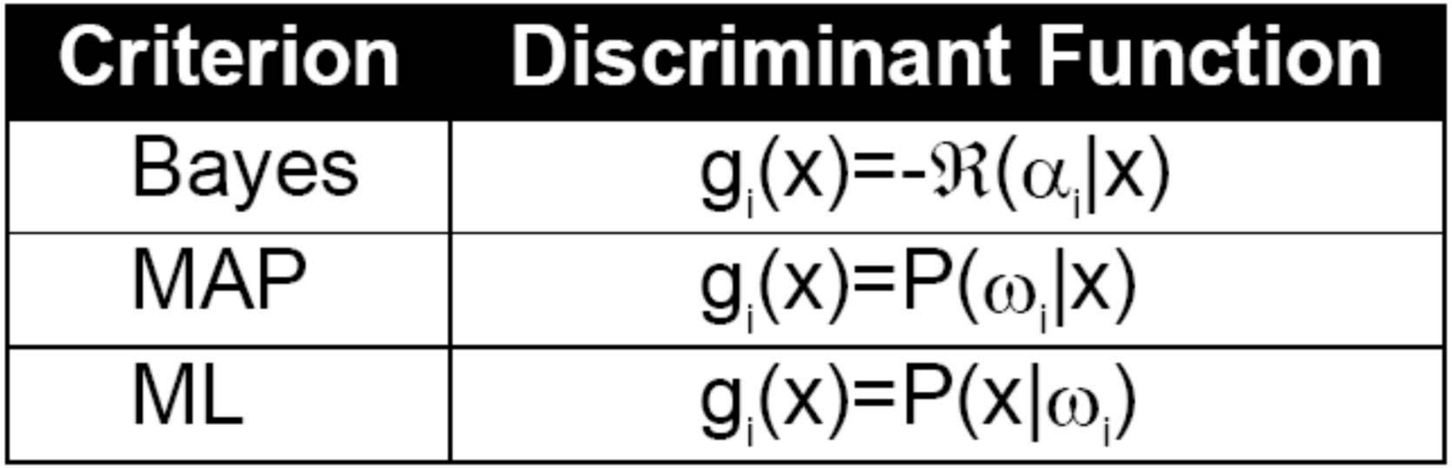
$$g_i(x) = \mathit{ln}\, p(x \mid \omega_i) + \mathit{ln}\, P(\omega_i)$$

(*ln*: natural logarithm)

Three Basic Decision Rules

- Decision rule: assign a feature vector x to ω_i

  if $g_i(x) > g_j(x) \; \forall \; j \neq i$



Criterion	Discriminant Function
Bayes	$g_i(x)=-\Re(\alpha_i\|x)$
MAP	$g_i(x)=P(\omega_i\|x)$
ML	$g_i(x)=P(x\|\omega_i)$

What is the relationship between the three criteria?

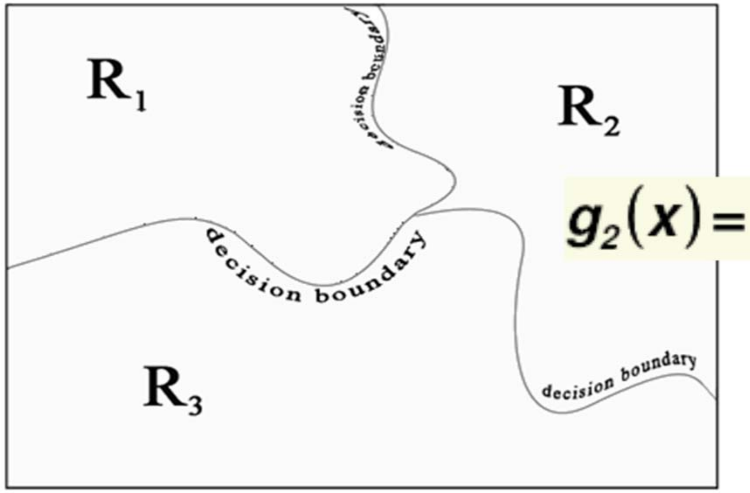
Decision Regions

- Feature space divided into *c decision regions* R_i:

  if $g_i(x) > g_j(x)\ \forall\ j \neq i$ then x is in R_i

  ⇓

  assign x to ω_i



$g_2(x) = max\{g_i\}$

- R_i separated by decision boundaries

Two-Category: Dichotomizer

- Single discriminant function

Let $g(x) \equiv g_1(x) - g_2(x)$

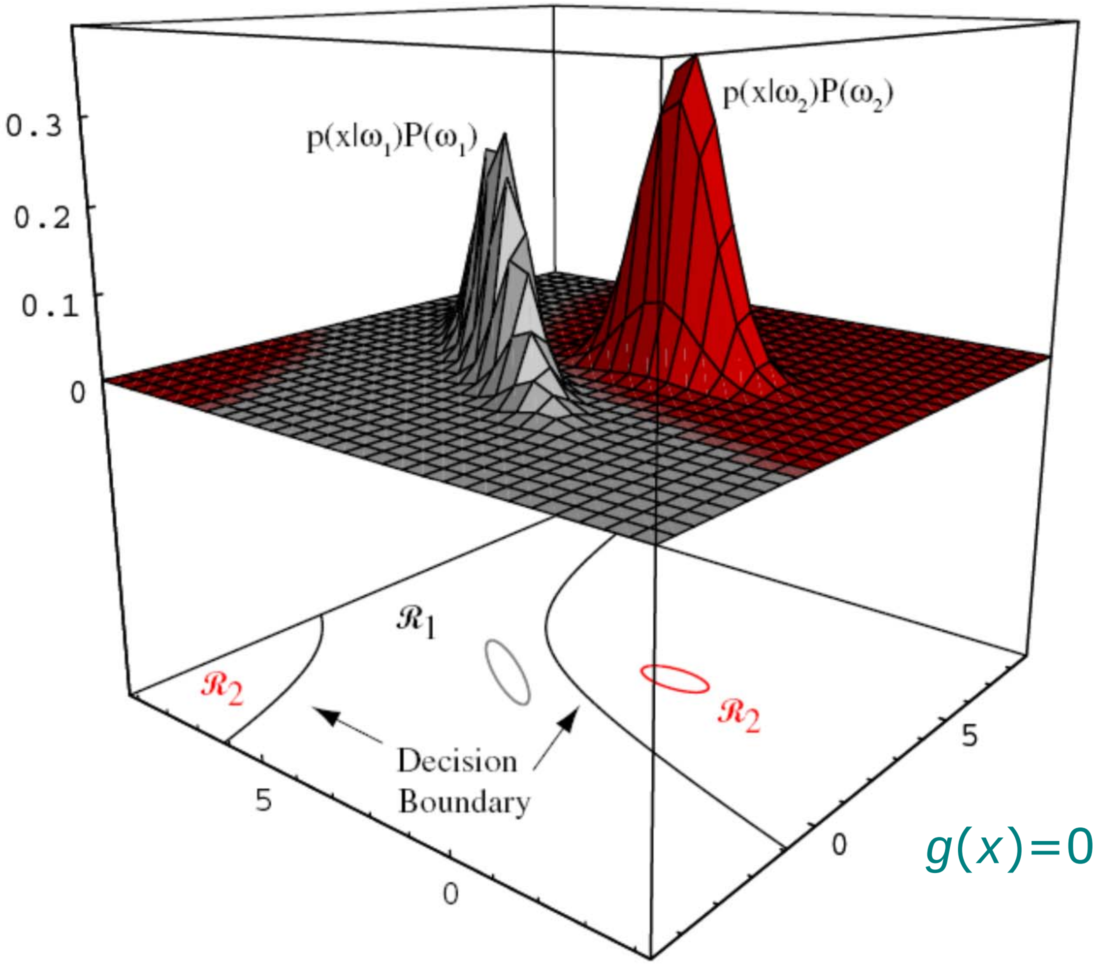
Decide ω_1 if $g(x) > 0$; Otherwise decide ω_2

- Minimum-error-rate discriminant function:

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

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

Two-Category: Gaussian



$g(x)=0$

From: R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Recognition.* Copyright © 2001 by John Wiley & Sons, Inc.

Why Gaussian?

- Analytical tractability
  - The parameters (μ, Σ) are **sufficient** to uniquely characterize the distribution
  - If $\boldsymbol{x_i}$'s are mutually **uncorrelated,** then they are also **independent**
  - The *marginal* and *conditional* densities are also Gaussian
  - Any *linear transformation* of any *N* jointly Gaussian RV's results in *N* RV's also Gaussian
- Ubiquity - frequently observed
  - Explained by the central limit theorem

The Normal Density

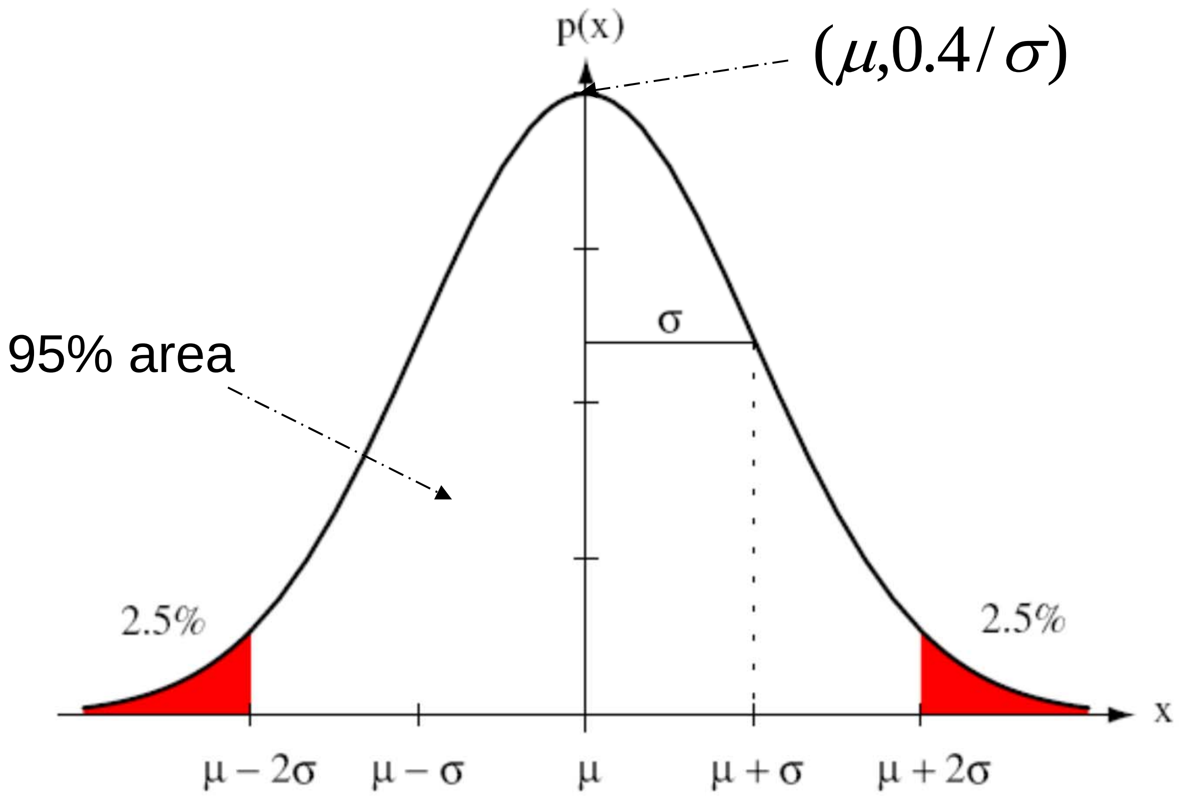
- Univariate density
  - Feature *x* is one dimensional continuous variable
  - Analytically tractable
    - ✓ Completely specified by its mean and variance
  - A lot of processes are asymptotically Gaussian
    - ✓ Handwritten characters, speech sounds

$$p(x) = \frac{1}{\sqrt{2\pi}\,\sigma} \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right], \qquad p(x) \sim N\left(\mu, \sigma^2\right)$$

μ = mean (or expected value) of *x*

σ^2 = expected squared deviation or variance

A Univariate Distribution



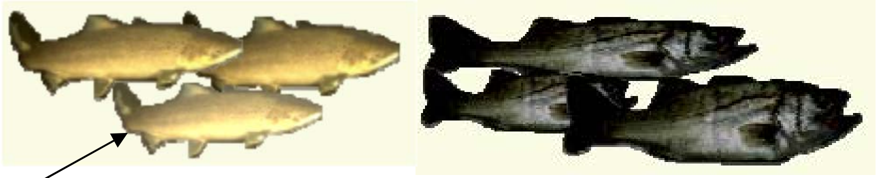
From: R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Recognition.*

Feature Vector

- Easy Representation
  - For each sample, a set of extracted features can be represented as a feature vector.

$$\mathbf{x} = [\text{length}, \quad \text{width}, \quad \text{color}, \ldots]^T$$

  - All samples can be represented as a collection of feature vectors.

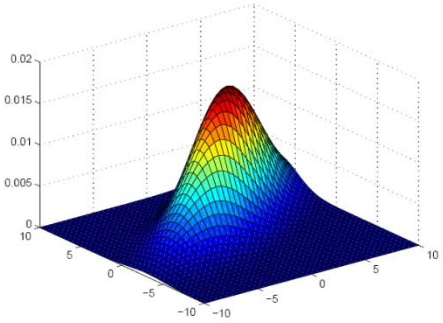


$$\mathbf{x}_1 = [l_1 \quad w_1 \quad c_1 \quad \cdots]^T, \mathbf{x}_2 = [l_2 \quad w_2 \quad c_2 \quad \cdots]^T, \ldots$$

- Linear models are computationally feasible.

Multivariate Gaussian

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



where:

$\boldsymbol{x}$ = $[x_1, x_2, \ldots, x_d]^t$ (t stands for transpose)

$\boldsymbol{\mu}$ = $[\mu_1, \mu_2, \ldots, \mu_d]^t$ mean vector

Σ = $d \times d$ covariance matrix

$|\Sigma|$: determinant

Σ^{-1} : inverse

$$p(\mathbf{x}) \sim N(\boldsymbol{\mu}, \mathbf{\Sigma})$$

$$\mathbf{\Sigma} = \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1d} \\ \sigma_{21} & \sigma_2^2 & \cdots & \sigma_{2d} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{d1} & \sigma_{d2} & \cdots & \sigma_d^2 \end{bmatrix}$$

Covariance Matrix

- From the covariance matrix, we can find out:
  - The variance of each feature x_i
  - Relationship between any two features x_i and x_j
    - ✓ Independent $\sigma_{ij} = 0$
    - ✓ Positive correlation $\sigma_{ij} > 0$
    - ✓ Negative correlation $\sigma_{ij} < 0$
- When Σ is diagonal:

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

$$p(\mathbf{x}) = \prod_{i=1}^{d} \frac{1}{\sigma_i\sqrt{2\pi}} \exp\left[-\frac{(x_i-\mu_i)^2}{2\sigma_i^2}\right]$$

Mahalanobis Distance

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

Squared Mahalanobis Distance:

$$(\mathbf{x}-\boldsymbol{\mu})^t \mathbf{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})$$

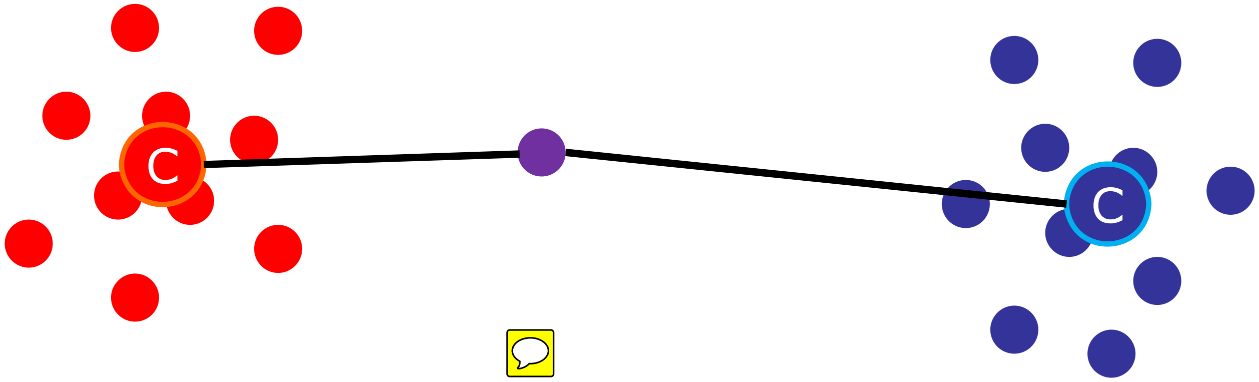
Mahalanobis Distance:

$$\sqrt{(\mathbf{x}-\boldsymbol{\mu})^t \mathbf{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})}$$

P.C. Mahalanobis

Mahalanobis Distance: Intuition

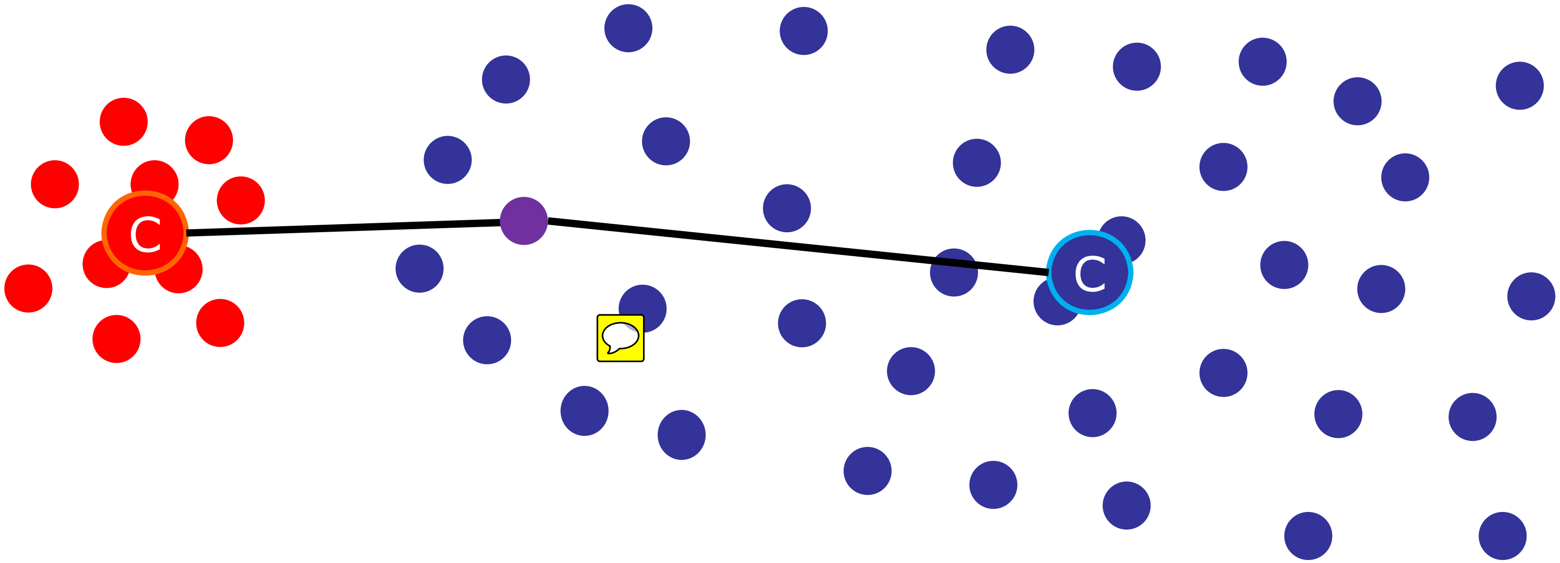
Task: Classify the test point



Classify the test point as being red

Mahalanobis Distance: Intuition

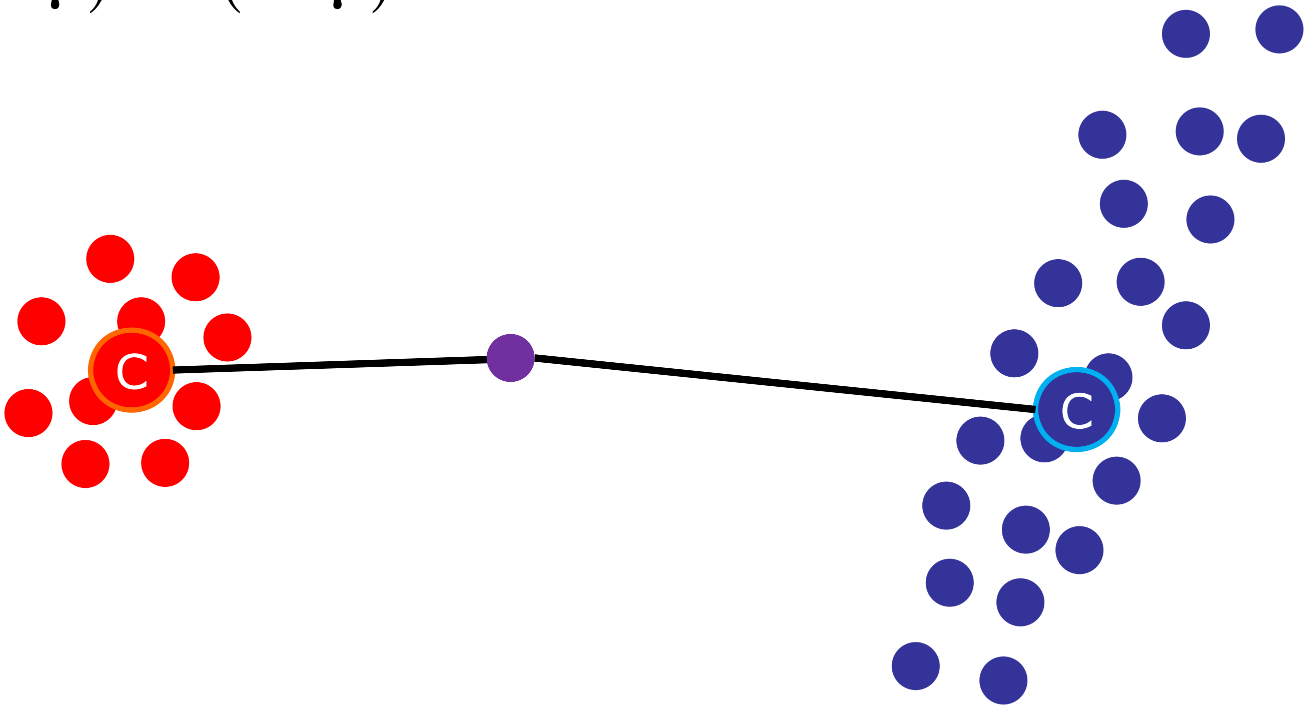
Task: Classify the test point



Classify the test point as being blue

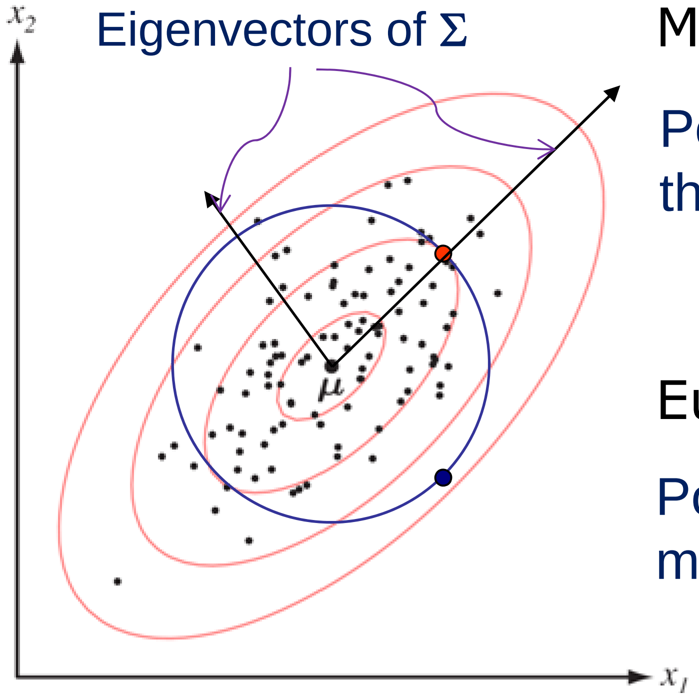
Mahalanobis Distance: Intuition

$$\sqrt{(\mathbf{x}-\boldsymbol{\mu})^t\,\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})}$$



Mahalanobis distance accounts for the covariance between variables and the fact that the variances in each direction are different. It reduces to the Euclidean distance for uncorrelated variables with unit variance.

Mahalanobis Distance vs. Euclidean Distance



Mahalanobis Distance: $\sqrt{(\mathbf{x}-\boldsymbol{\mu})^t \boldsymbol{\Sigma}^{-1} (\mathbf{x}-\boldsymbol{\mu})}$

Points of equal Mahalanobis distance to the mean lie on an ellipse.

Euclidean Distance: $\sqrt{(\mathbf{x}-\boldsymbol{\mu})^t (\mathbf{x}-\boldsymbol{\mu})}$

Points of equal Euclidean distance to the mean lie on a circle.

FIGURE 2.9. Samples drawn from a two-dimensional Gaussian lie in a cloud centered on the mean $\boldsymbol{\mu}$. The ellipses show lines of equal probability density of the Gaussian. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

Discriminant Functions for the Normal Density

- The minimum error-rate classification can be achieved by the discriminant functions

$$g_i(x) = \ln p(x \mid \omega_i) + \ln P(\omega_i)$$

- Case of multivariate normal

$$g_i(\mathbf{x}) = -\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_i)^t \boldsymbol{\Sigma}_i^{-1}(\mathbf{x}-\boldsymbol{\mu}_i) \boxed{-\frac{d}{2}\ln 2\pi} -\frac{1}{2}\ln\left|\boldsymbol{\Sigma}_i\right| + \ln P(\omega_i)$$

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

Case 1: $\Sigma_i = \sigma^2 \mathrm{I}$

All features are *independent* & have the *same* variance for *all* classes.

$$\boxed{g_i(\mathbf{x}) = -\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_i)^t \boldsymbol{\Sigma}_i^{-1}(\mathbf{x}-\boldsymbol{\mu}_i) - \frac{1}{2}\ln|\boldsymbol{\Sigma}_i| + \ln P(\omega_i)}$$

$$\Bigg\downarrow \; \boldsymbol{\Sigma}_i = \sigma^2 \mathbf{I}$$

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

Euclidean norm: $\|\mathbf{x}-\boldsymbol{\mu}_i\|^2 = (\mathbf{x}-\boldsymbol{\mu}_i)^t(\mathbf{x}-\boldsymbol{\mu}_i)$

Case 1: $\Sigma_i = \sigma^2 \mathrm{I}$ (Cont'd)

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

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

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

Linear discriminant function: $g_i(\mathbf{x}) = \mathbf{w}_i^t\mathbf{x} + \mathbf{w}_{i0}$

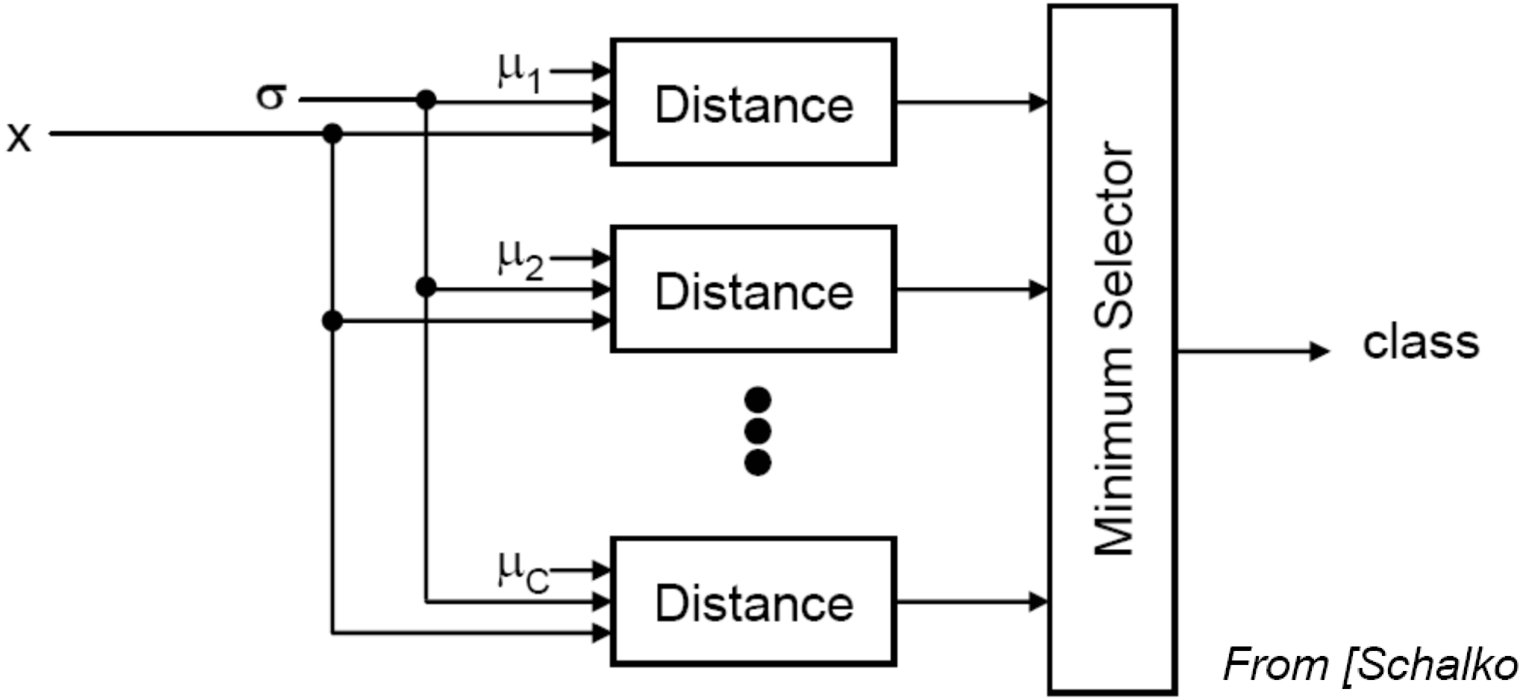
$$\mathbf{w}_i = \frac{\boldsymbol{\mu}_i}{\sigma^2}; \quad \mathbf{w}_{i0} = -\frac{1}{2\sigma^2}\boldsymbol{\mu}_i^t\boldsymbol{\mu}_i + \ln P(\omega_i)$$

w_{i0} is the threshold/bias for the *i*-th category

Minimum Distance Classifier

Assuming equal priors:

$$g_i(\mathbf{x}) = -\frac{1}{2\sigma^2}(\mathbf{x}-\boldsymbol{\mu}_i)^t(\mathbf{x}-\boldsymbol{\mu}_i) \Rightarrow g_i(\mathbf{x}) = -(\mathbf{x}-\boldsymbol{\mu}_i)^t(\mathbf{x}-\boldsymbol{\mu}_i)$$



From [Schalkoff, 1992]

Linear Machine

- A classifier that uses *linear discriminant functions* is called *"a linear machine"*

- The decision surfaces for a linear machine are pieces of hyperplanes defined by:

$$g_i(x) = g_j(x)$$

Decision Surfaces

- The hyperplane separating $\mathcal{R}_i$ and $\mathcal{R}_j$

$$g_i(\mathbf{x}) = g_j(\mathbf{x})$$

$$-\frac{1}{2\sigma^2}(-2\boldsymbol{\mu}_i^t\mathbf{x}+\boldsymbol{\mu}_i^t\boldsymbol{\mu}_i)+\ln P(\omega_i) = -\frac{1}{2\sigma^2}(-2\boldsymbol{\mu}_j^t\mathbf{x}+\boldsymbol{\mu}_j^t\boldsymbol{\mu}_j)+\ln P(\omega_j)$$

$$\boldsymbol{\mu}_i^t\mathbf{x}-\frac{1}{2}\boldsymbol{\mu}_i^t\boldsymbol{\mu}_i+\sigma^2\ln P(\omega_i) = \boldsymbol{\mu}_j^t\mathbf{x}-\frac{1}{2}\boldsymbol{\mu}_j^t\boldsymbol{\mu}_j+\sigma^2\ln P(\omega_j)$$

$$\left(\boldsymbol{\mu}_i-\boldsymbol{\mu}_j\right)^t\mathbf{x}-\frac{1}{2}\left(\boldsymbol{\mu}_i^t\boldsymbol{\mu}_i-\boldsymbol{\mu}_j^t\boldsymbol{\mu}_j\right)+\sigma^2\ln\frac{P(\omega_i)}{P(\omega_j)}=0$$

$$\left(\boldsymbol{\mu}_i-\boldsymbol{\mu}_j\right)^t\mathbf{x}-\frac{1}{2}\left(\boldsymbol{\mu}_i-\boldsymbol{\mu}_j\right)^t\left(\boldsymbol{\mu}_i+\boldsymbol{\mu}_j\right)+\sigma^2\ln\frac{P(\omega_i)}{P(\omega_j)}\frac{\left(\boldsymbol{\mu}_i-\boldsymbol{\mu}_j\right)^t\left(\boldsymbol{\mu}_i-\boldsymbol{\mu}_j\right)}{\left(\boldsymbol{\mu}_i-\boldsymbol{\mu}_j\right)^t\left(\boldsymbol{\mu}_i-\boldsymbol{\mu}_j\right)}=0$$

Decision Surfaces

- The hyperplane separating $\mathcal{R}_i$ and $\mathcal{R}_j$

$$\mathbf{w}^t\left(\mathbf{x}-\mathbf{x}_0\right)=0, \qquad \mathbf{w}=\boldsymbol{\mu}_i-\boldsymbol{\mu}_j$$

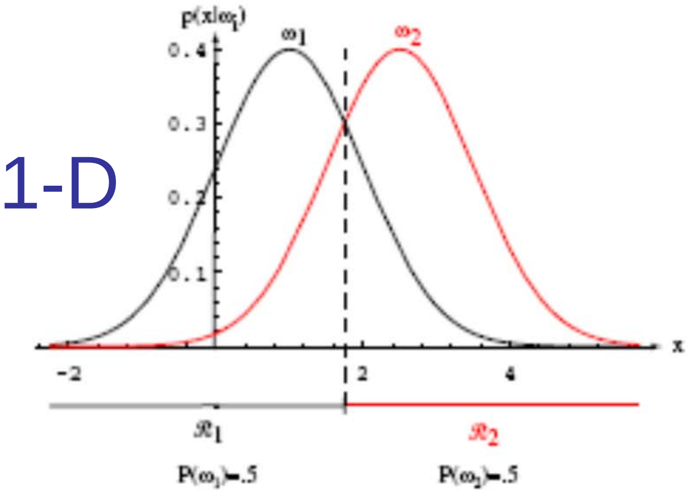
$$\mathbf{x}_0=\frac{1}{2}(\boldsymbol{\mu}_i+\boldsymbol{\mu}_j)-\frac{\sigma^2}{\left\|\boldsymbol{\mu}_i-\boldsymbol{\mu}_j\right\|^2}\ln\frac{P(\omega_i)}{P(\omega_j)}(\boldsymbol{\mu}_i-\boldsymbol{\mu}_j)$$

always orthogonal to the line linking the means!

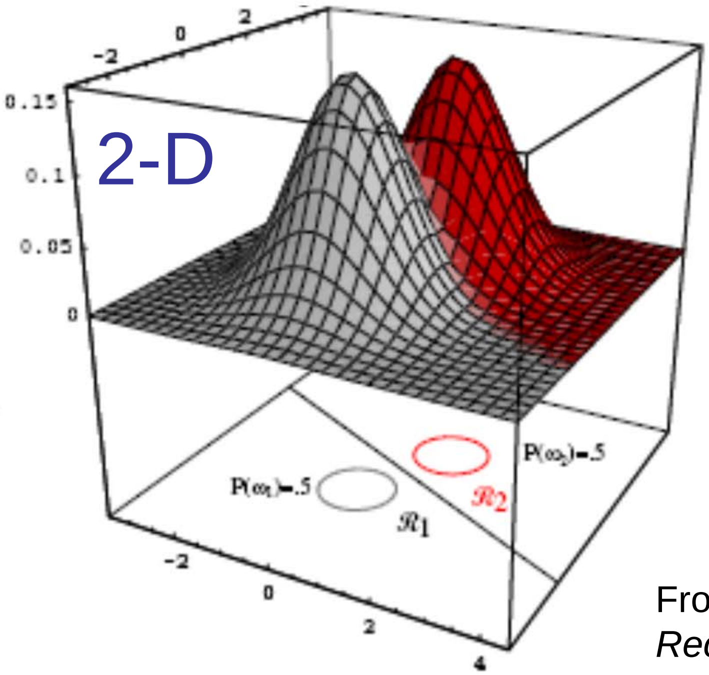
$$\text{if}\quad P(\omega_i)=P(\omega_j)\quad\text{then}\quad \mathbf{x}_0=\frac{1}{2}(\boldsymbol{\mu}_i+\boldsymbol{\mu}_j)$$

Decision Surfaces: Examples

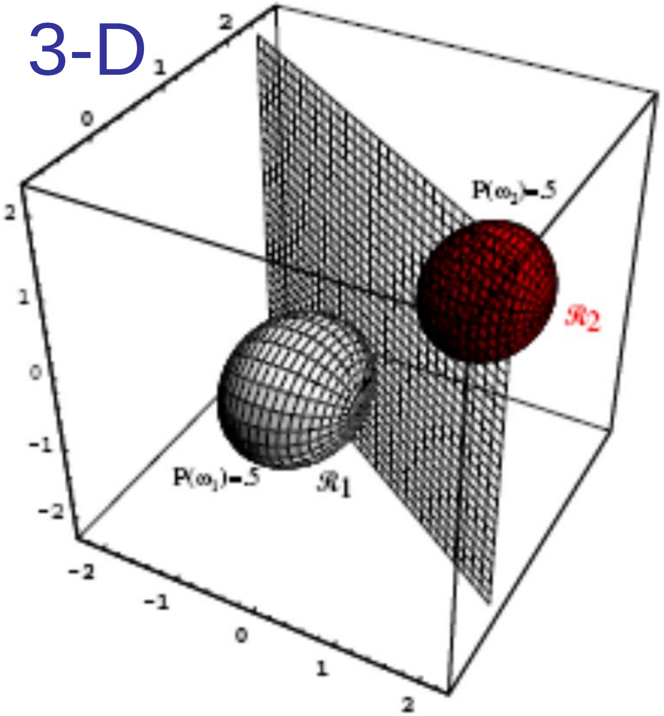
1-D



2-D

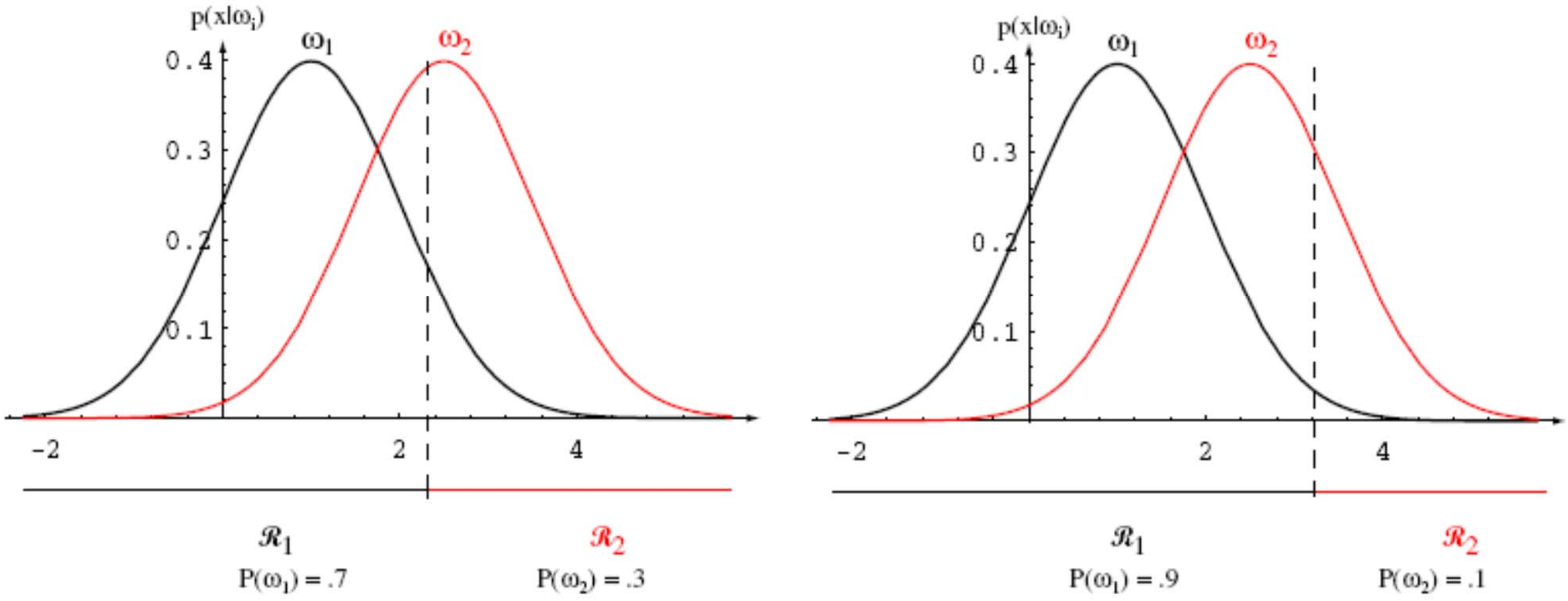


3-D



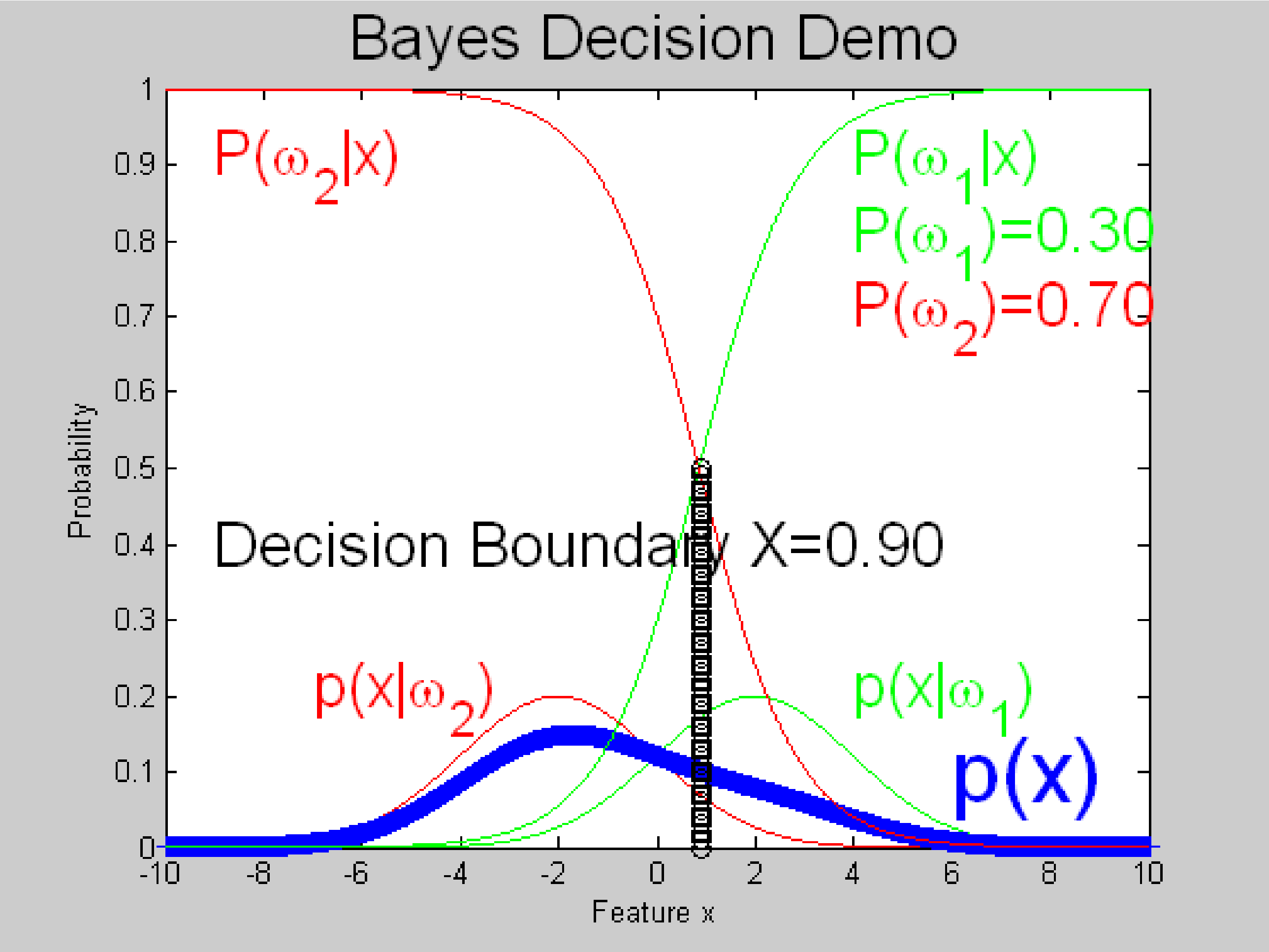
From: R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Recognition.* Copyright © 2001 by John Wiley & Sons, Inc.

Priors and Decision Boundary: 1-D

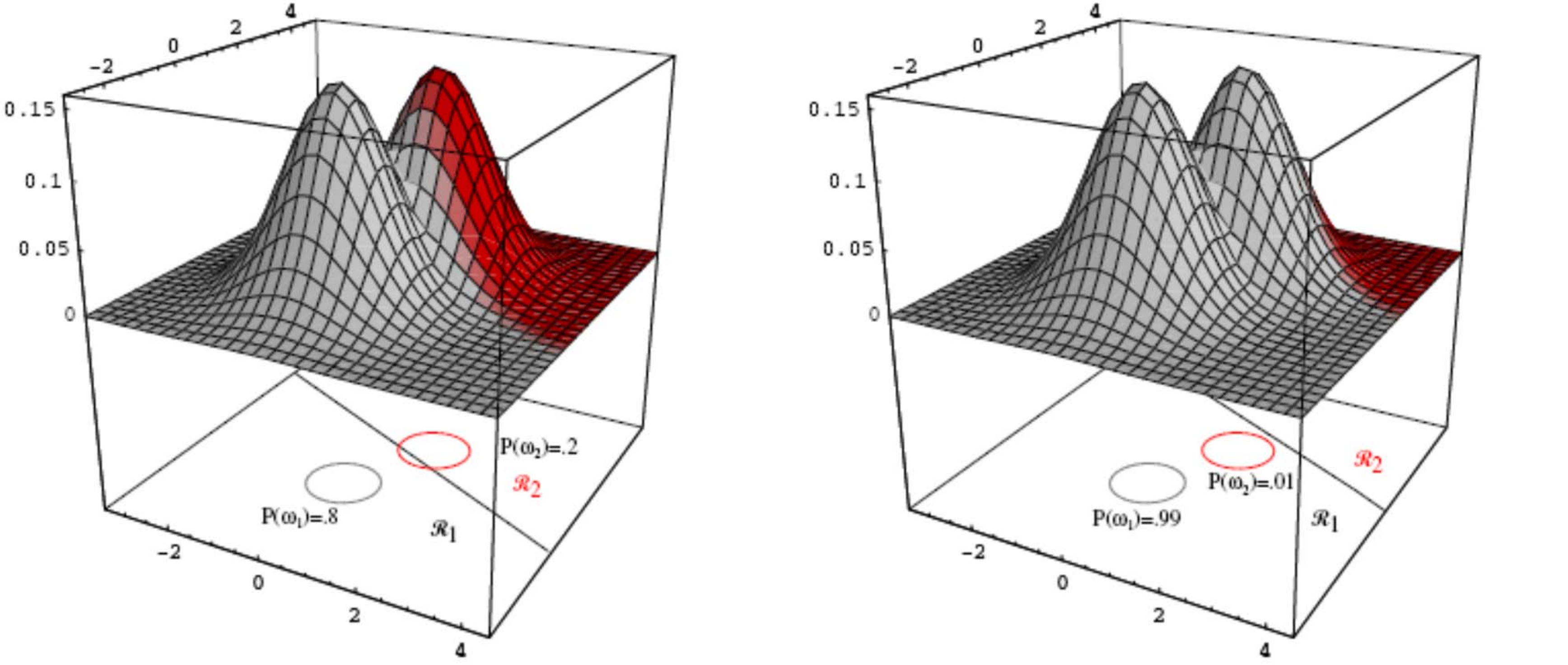


From: R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Recognition.* Copyright © 2001 by John Wiley & Sons, Inc.

Priors and Decision Boundary: Demo#1

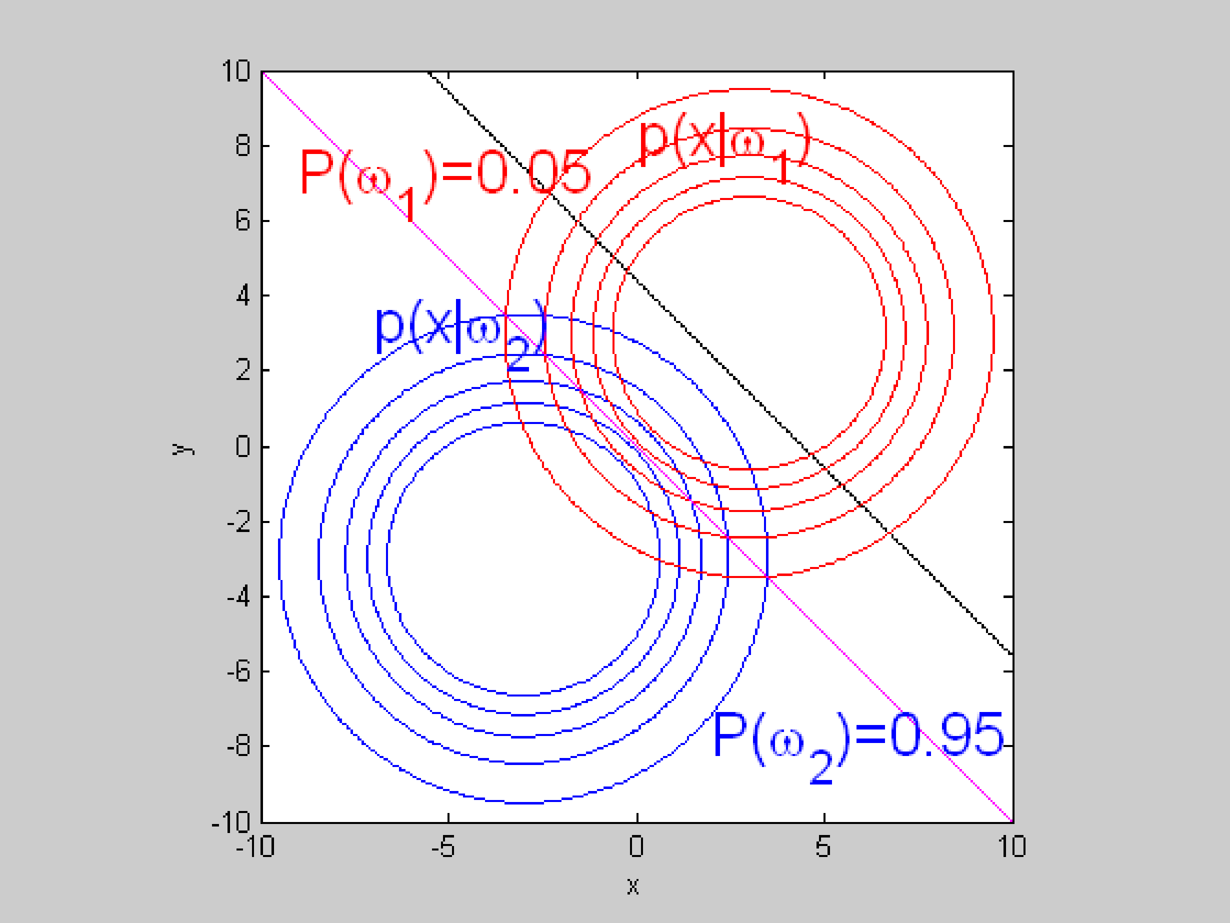


Priors and Decision Boundary: 2-D



From: R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Recognition.* Copyright © 2001 by John Wiley & Sons, Inc.

Priors and Decision Boundary: Demo#2 $\mathbf{\Sigma}_i = \sigma^2\mathbf{I}$

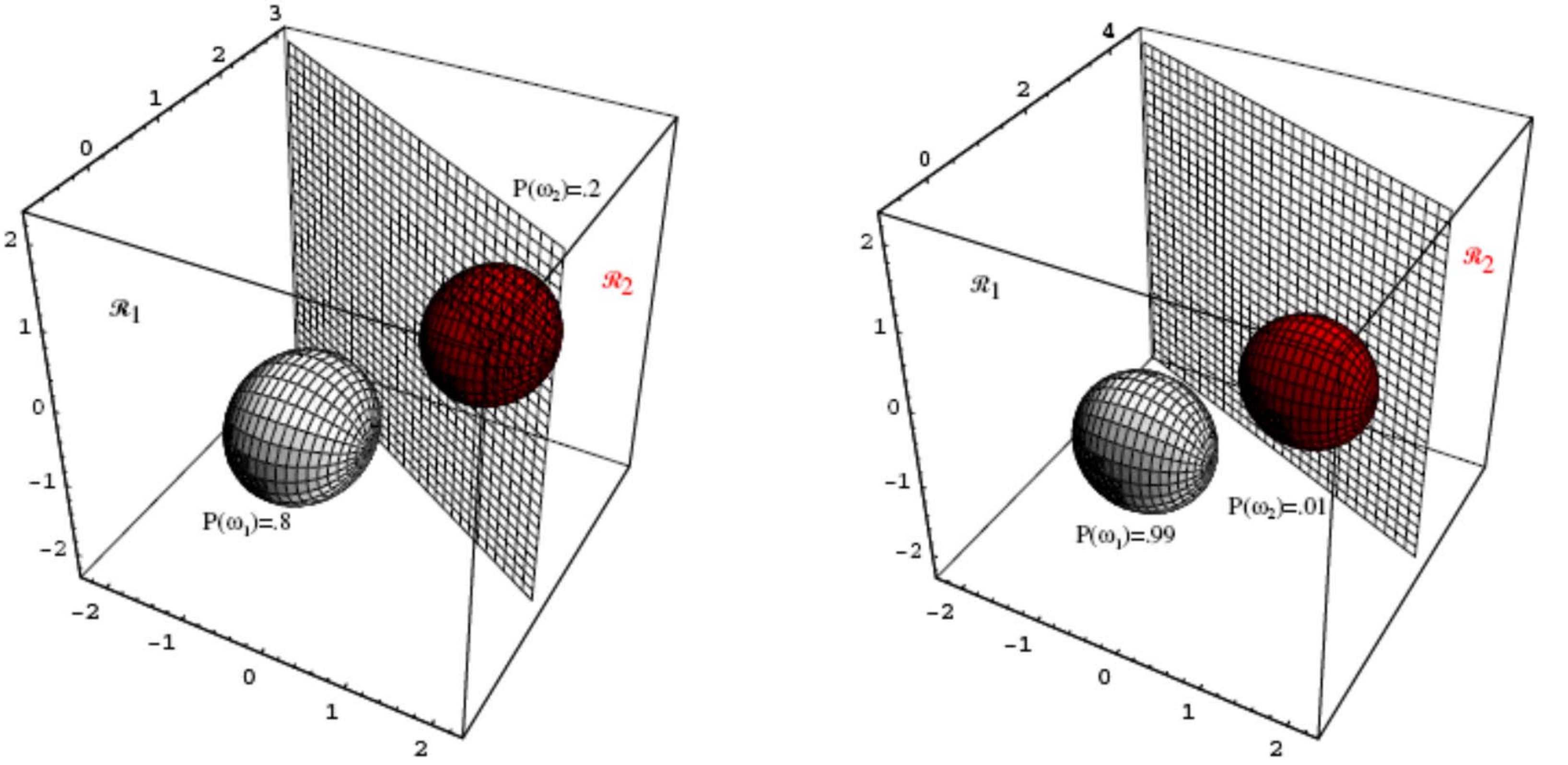


$$\boldsymbol{\mu}_1 = [3 \quad 3]^t$$

$$\boldsymbol{\mu}_2 = [-3 \quad -3]^t$$

$$\sigma = 3$$

Priors and Decision Boundary: 3-D



From: R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Recognition.* Copyright © 2001 by John Wiley & Sons, Inc.

Case 2: Σ$_i$ = Σ

- Covariance of all classes are *identical* but *arbitrary*

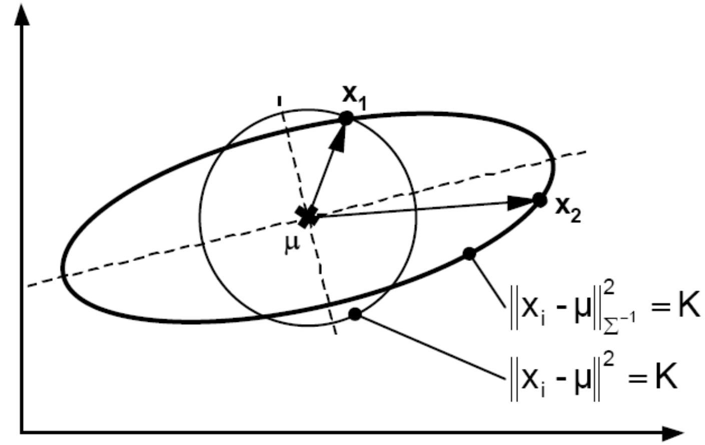
$$\boxed{g_i(\mathbf{x}) = -\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_i)^t \boldsymbol{\Sigma}_i^{-1}(\mathbf{x}-\boldsymbol{\mu}_i) - \frac{1}{2}\ln|\boldsymbol{\Sigma}_i| + \ln P(\omega_i)}$$

$$\boldsymbol{\Sigma}_i = \boldsymbol{\Sigma}$$

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

Squared Mahalanobis Distance:

$$(\mathbf{x}-\boldsymbol{\mu}_i)^t \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}_i)$$



Case 2: $\Sigma_i = \Sigma$ (Cont'd)

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

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

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

Linear discriminant function:

$$g_i(\mathbf{x}) = \mathbf{w}_i^t\mathbf{x} + \mathbf{w}_{i0}$$

$$\mathbf{w} = \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_i \qquad \mathbf{w}_{i0} = -\frac{1}{2}\boldsymbol{\mu}_i^t\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_i + \ln P(\omega_i)$$

Case 2: $\Sigma_i = \Sigma$ (Cont'd)

- Decision boundaries are hyperplanes

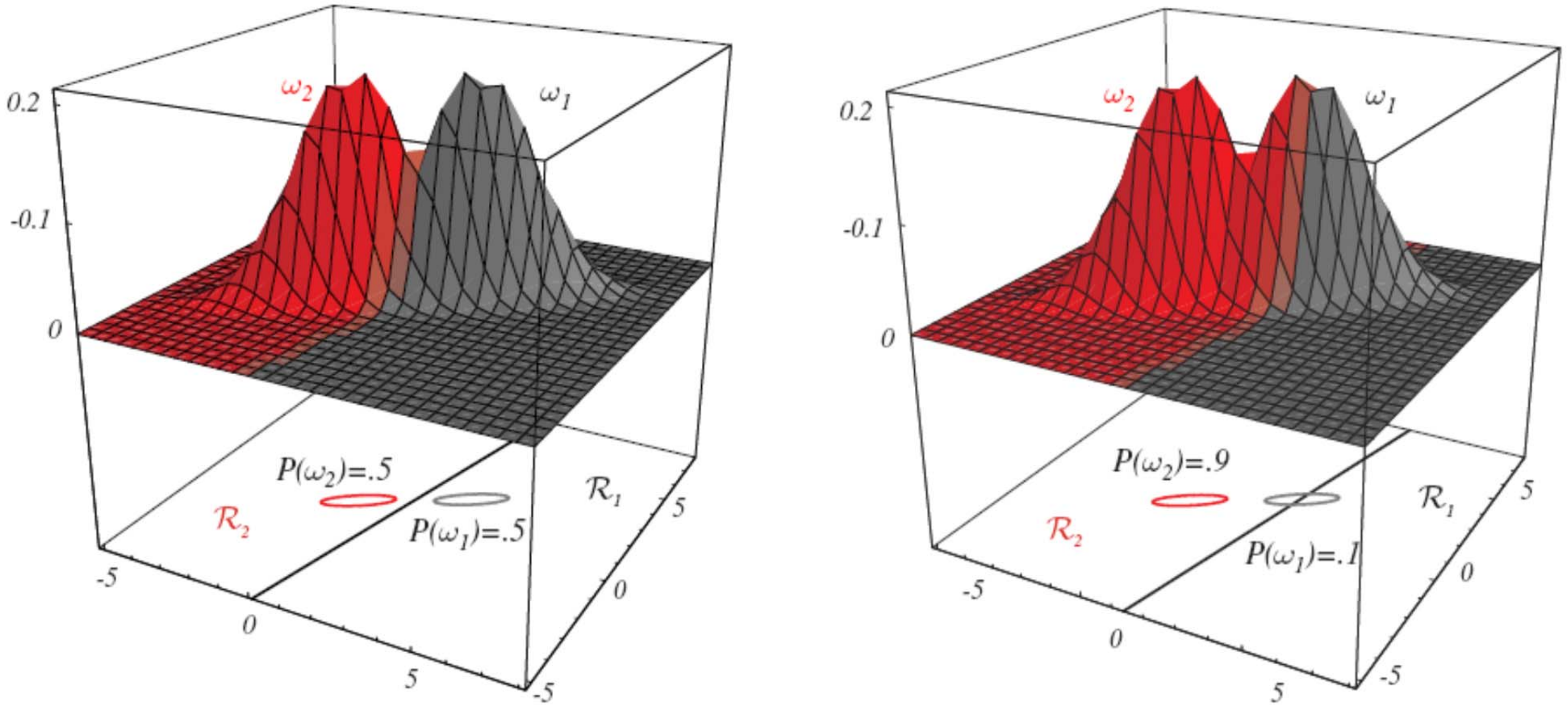
$$\mathbf{w}^t(\mathbf{x}-\mathbf{x}_0)=0, \qquad \mathbf{w}=\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_i-\boldsymbol{\mu}_j)$$

$$\mathbf{x}_0=\frac{1}{2}(\boldsymbol{\mu}_i+\boldsymbol{\mu}_j)-\frac{\ln\left[P(\omega_i)/P(\omega_j)\right]}{(\boldsymbol{\mu}_i-\boldsymbol{\mu}_j)^t\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_i-\boldsymbol{\mu}_j)}.(\boldsymbol{\mu}_i-\boldsymbol{\mu}_j)$$

- Hyperplane separating $\mathcal{R}_i$ and $\mathcal{R}_j$

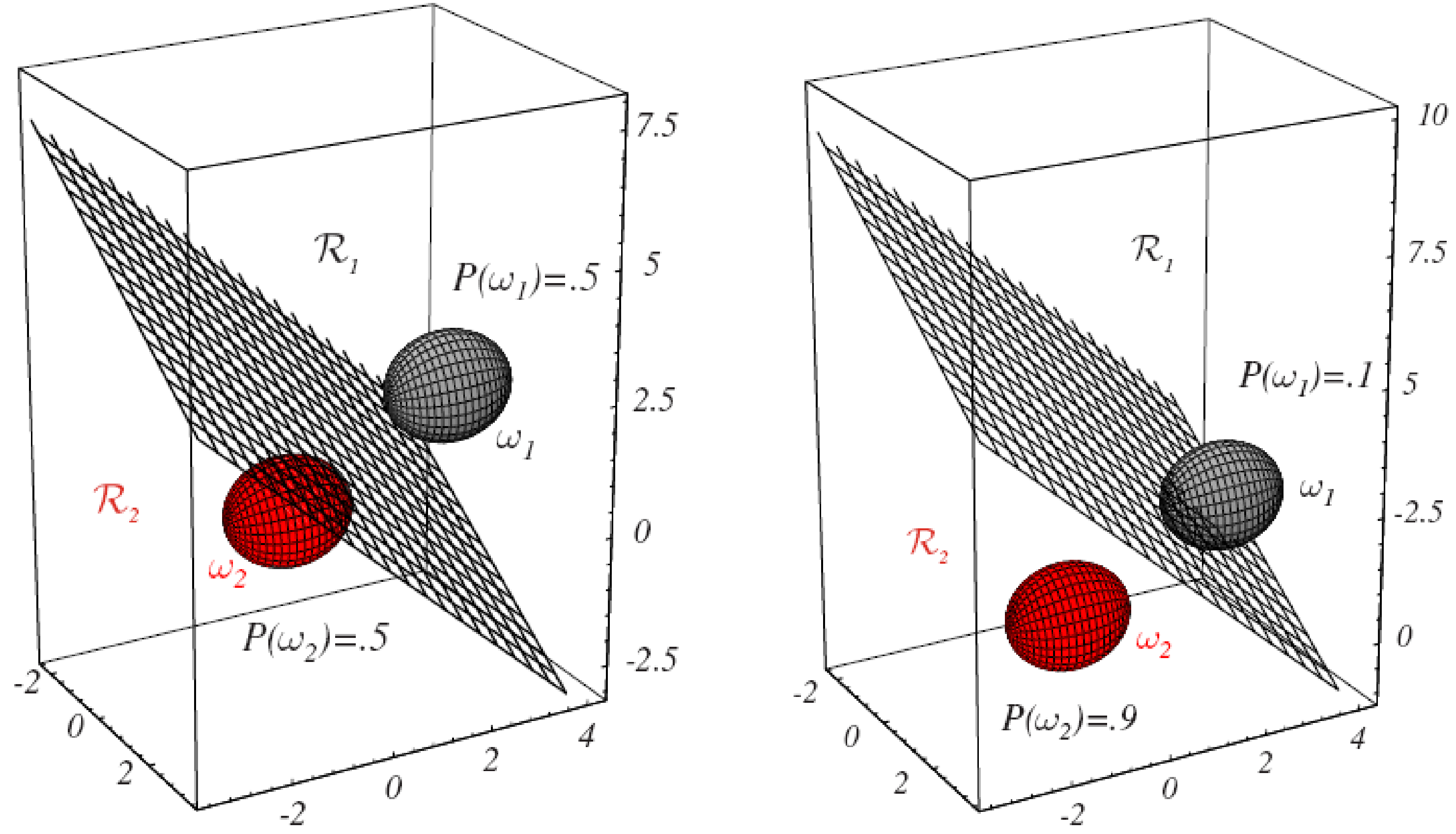
  generally not orthogonal to the line between the means!

Decision Boundaries: 2-D



From: R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Recognition.* Copyright © 2001 by John Wiley & Sons, Inc.

Decision Boundaries: 3-D



From: R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Recognition.*
Copyright © 2001 by John Wiley & Sons, Inc.

Case 3: $\Sigma_i = \sigma_i^2 \mathrm{I}$

Each class has a *different* covariance matrix, which is proportional to the *identity* matrix

$$\boxed{g_i(\mathbf{x}) = -\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_i)^t \boldsymbol{\Sigma}_i^{-1}(\mathbf{x}-\boldsymbol{\mu}_i) - \frac{1}{2}\ln|\boldsymbol{\Sigma}_i| + \ln P(\omega_i)}$$

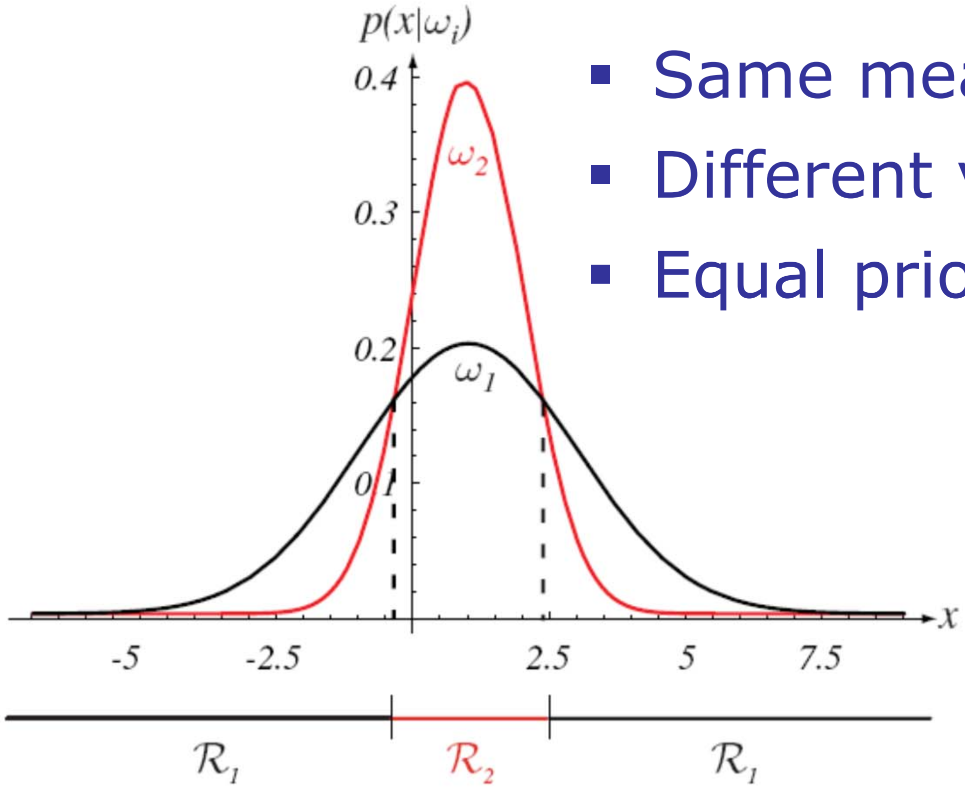
$$\Big\downarrow \; \boldsymbol{\Sigma}_i = \sigma_i^2 \mathbf{I}$$

Quadratic discriminant:

$$g_i(\mathbf{x}) = -\frac{1}{2\sigma_i^2}(\mathbf{x}-\boldsymbol{\mu}_i)^t(\mathbf{x}-\boldsymbol{\mu}_i) - \frac{1}{2} d \ln\left(\sigma_i^2\right) + \ln P(\omega_i)$$

- The decision boundaries are quadratic: hyper-ellipses
- The loci of constant probability are hyper-spheres aligned with the feature axis.

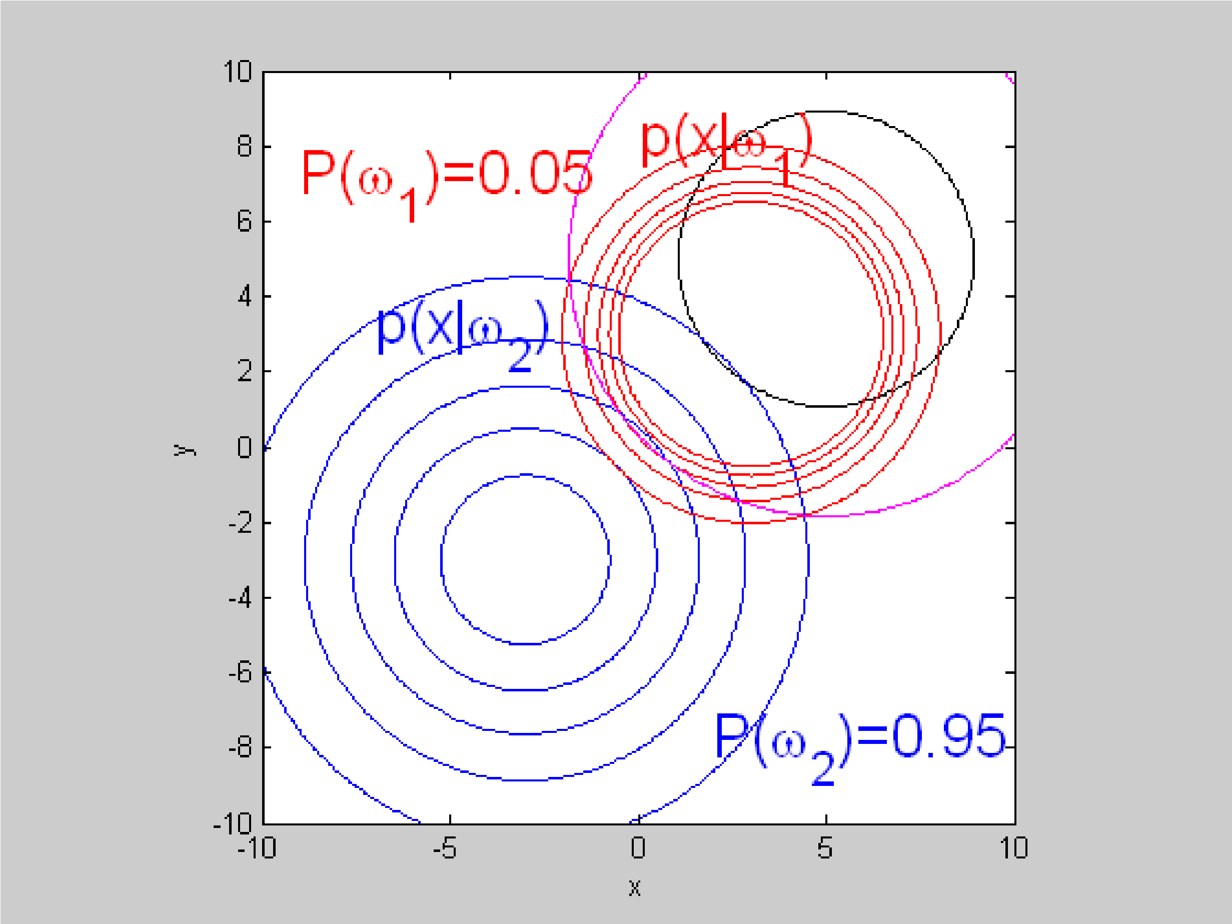
1-D Example



- Same mean
- Different variance
- Equal priors

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.

Bayes Decision Boundaries: 2D Demo $\mathbf{\Sigma}_i = \sigma_i^2 \mathbf{I}$



$$\boldsymbol{\mu}_1 = \begin{bmatrix} 3 & 3 \end{bmatrix}^t$$

$$\sigma_1 = 2$$

$$\boldsymbol{\mu}_2 = \begin{bmatrix} -3 & -3 \end{bmatrix}^t$$

$$\sigma_2 = 4$$

Analytical Solution to 2D Demo

$$\boldsymbol{\Sigma}_i = \sigma_i^2 \mathbf{I} \qquad \boldsymbol{\mu}_1 = [3 \quad 3]^t, \sigma_1 = 2 \qquad \boldsymbol{\mu}_2 = [-3 \quad -3]^t, \ \sigma_2 = 4$$

Decision boundary: $g_1(\mathbf{x}) = g_2(\mathbf{x})$

$$P(\omega_1 \mid \mathbf{x}) = P(\omega_2 \mid \mathbf{x})$$

$$\ln P(\omega_1 \mid \mathbf{x}) = \ln P(\omega_2 \mid \mathbf{x})$$

$$\ln[P(\mathbf{x} \mid \omega_1) P(\omega_1)] = \ln[P(\mathbf{x} \mid \omega_2) P(\omega_2)]$$

$$-\frac{1}{2\sigma_1^2}(\mathbf{x} - \boldsymbol{\mu}_1)^t(\mathbf{x} - \boldsymbol{\mu}_1) - \ln(\sigma_1^2) + \ln P(\omega_1) = -\frac{1}{2\sigma_2^2}(\mathbf{x} - \boldsymbol{\mu}_2)^t(\mathbf{x} - \boldsymbol{\mu}_2) - \ln(\sigma_2^2) + \ln P(\omega_2)$$

$$\text{Let } r = P(\omega_1)/P(\omega_2), a_1 = -\frac{1}{2\sigma_1^2}, a_2 = -\frac{1}{2\sigma_2^2} \qquad a_2/a_1 = \sigma_1^2/\sigma_2^2 = 1/4$$

$$a_1(\mathbf{x} - \boldsymbol{\mu}_1)^t(\mathbf{x} - \boldsymbol{\mu}_1) - a_2(\mathbf{x} - \boldsymbol{\mu}_2)^t(\mathbf{x} - \boldsymbol{\mu}_2) - \ln(a_2/a_1) + \ln(r) = 0$$

Analytical Solution (Cont’d)

$$(a_1 - a_2)\mathbf{x}^t\mathbf{x} - 2\left(a_1\boldsymbol{\mu}_1{}^t - a_2\boldsymbol{\mu}_2{}^t\right)\mathbf{x} + \left(a_1\boldsymbol{\mu}_1{}^t\boldsymbol{\mu}_1 - a_2\boldsymbol{\mu}_2{}^t\boldsymbol{\mu}_2\right) + \ln\left(ra_1/a_2\right) = 0$$

$$\mathbf{x}^t\mathbf{x} - 2\frac{\left(a_1\boldsymbol{\mu}_1{}^t - a_2\boldsymbol{\mu}_2{}^t\right)}{(a_1 - a_2)}\mathbf{x} + \frac{\left(a_1\boldsymbol{\mu}_1{}^t\boldsymbol{\mu}_1 - a_2\boldsymbol{\mu}_2{}^t\boldsymbol{\mu}_2\right) + \ln\left(ra_1/a_2\right)}{(a_1 - a_2)} = 0$$

$$\boldsymbol{\mu}_1 = [3 \quad 3]^t \ \sigma_1 = 2 \qquad \boldsymbol{\mu}_2 = [-3 \quad -3]^t \ \sigma_2 = 4$$

$$\mathbf{x}^t\mathbf{x} - 2\begin{bmatrix}5\\5\end{bmatrix}^t\mathbf{x} + 18 + \frac{\ln(4r)}{(-3/32)} = 0$$

$$\left(\mathbf{x} - \begin{bmatrix}5\\5\end{bmatrix}\right)^t\left(\mathbf{x} - \begin{bmatrix}5\\5\end{bmatrix}\right) - 50 + 18 + \frac{\ln(4r)}{(-3/32)} = 0$$

$$\left(\mathbf{x} - \begin{bmatrix}5\\5\end{bmatrix}\right)^t\left(\mathbf{x} - \begin{bmatrix}5\\5\end{bmatrix}\right) = 32 + \frac{\ln(4r)}{(3/32)} = 32\left(1 + \frac{\ln(4r)}{3}\right)$$

What happens when *r* is very small?

Case 4: Σ_i = arbitrary

- The covariance matrices are *different* for each category

$$\boxed{g_i(\mathbf{x}) = -\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_i)^t \boldsymbol{\Sigma}_i^{-1}(\mathbf{x}-\boldsymbol{\mu}_i) - \frac{1}{2}\ln|\boldsymbol{\Sigma}_i| + \ln P(\omega_i)}$$

Quadratic discriminant:

$$g_i(\mathbf{x}) = \mathbf{x}^t \mathbf{W}_i \mathbf{x} + \mathbf{w}_i^t \mathbf{x} + w_{i0}$$

where :

$$\mathbf{W}_{\mathrm{i}} = -\frac{1}{2}\boldsymbol{\Sigma}_i^{-1}$$

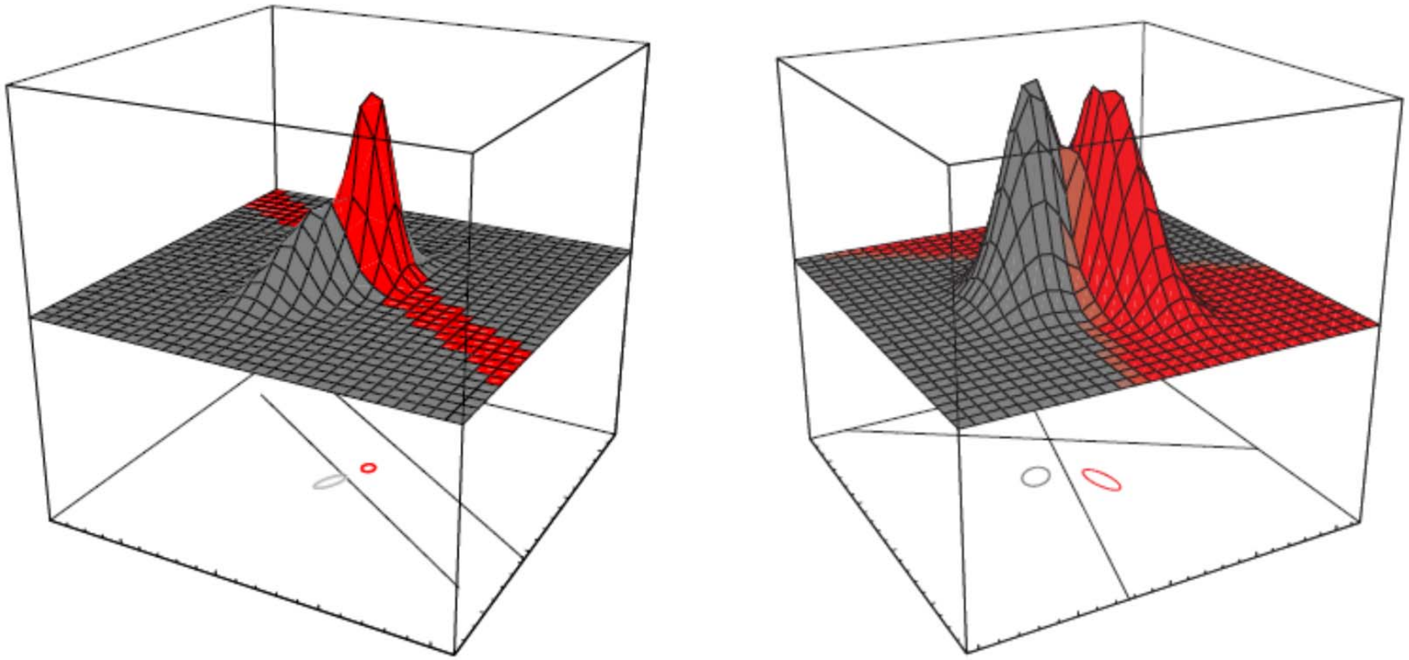
$$\mathbf{w}_{\mathrm{i}} = \boldsymbol{\Sigma}_i^{-1}\boldsymbol{\mu}_i$$

$$\mathbf{w}_{i0} = -\frac{1}{2}\boldsymbol{\mu}_i^t \boldsymbol{\Sigma}_i^{-1}\boldsymbol{\mu}_i - \frac{1}{2}\ln|\boldsymbol{\Sigma}_i| + \ln P(\omega_i)$$

Case 4: Σ_i = arbitrary (Cont'd)

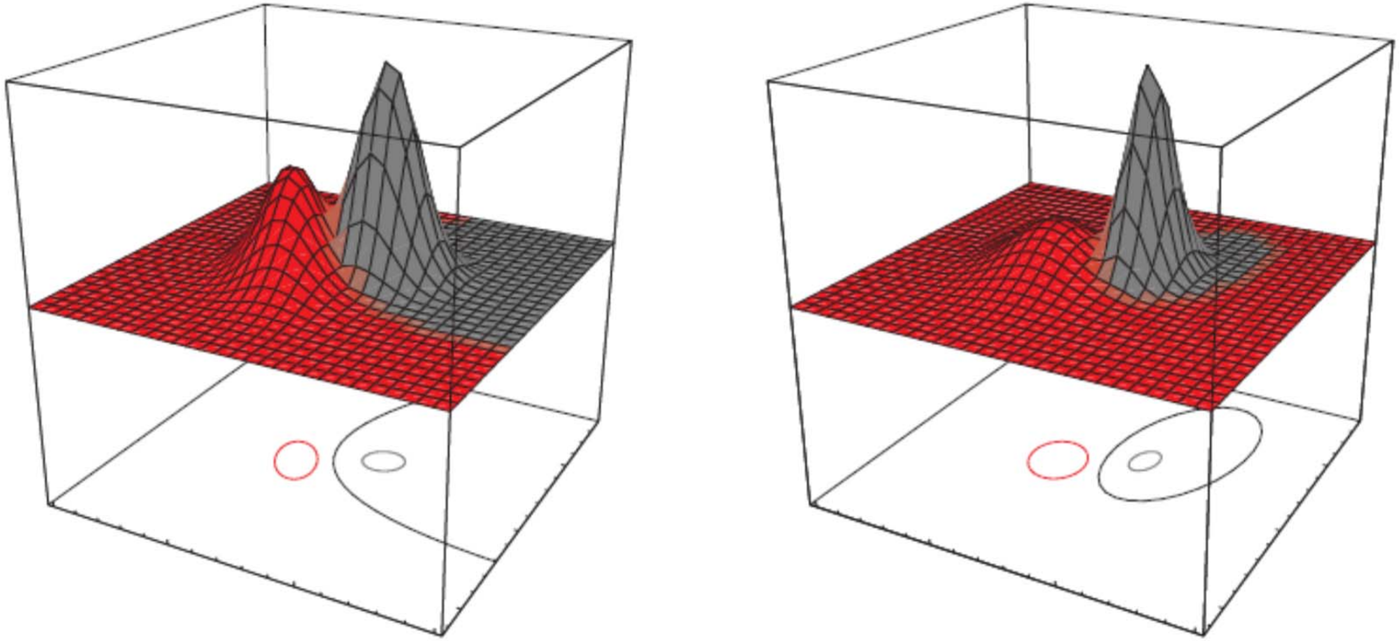
- The loci of constant probability for each class are hyper-ellipses, oriented with the eigenvectors of Σ_i for that class
- The decision boundaries are hyperquadrics:
  - hyperplanes, pairs of hyperplanes, hyperspheres, hyperellipsoids, hyperparaboloids, hyperhyperboloids
- The quadratic term in the discriminant is proportional to the Mahalanobis distance using the class-conditional covariance Σ_i

Bayes Decision Boundaries for Arbitrary Gaussian Distributions



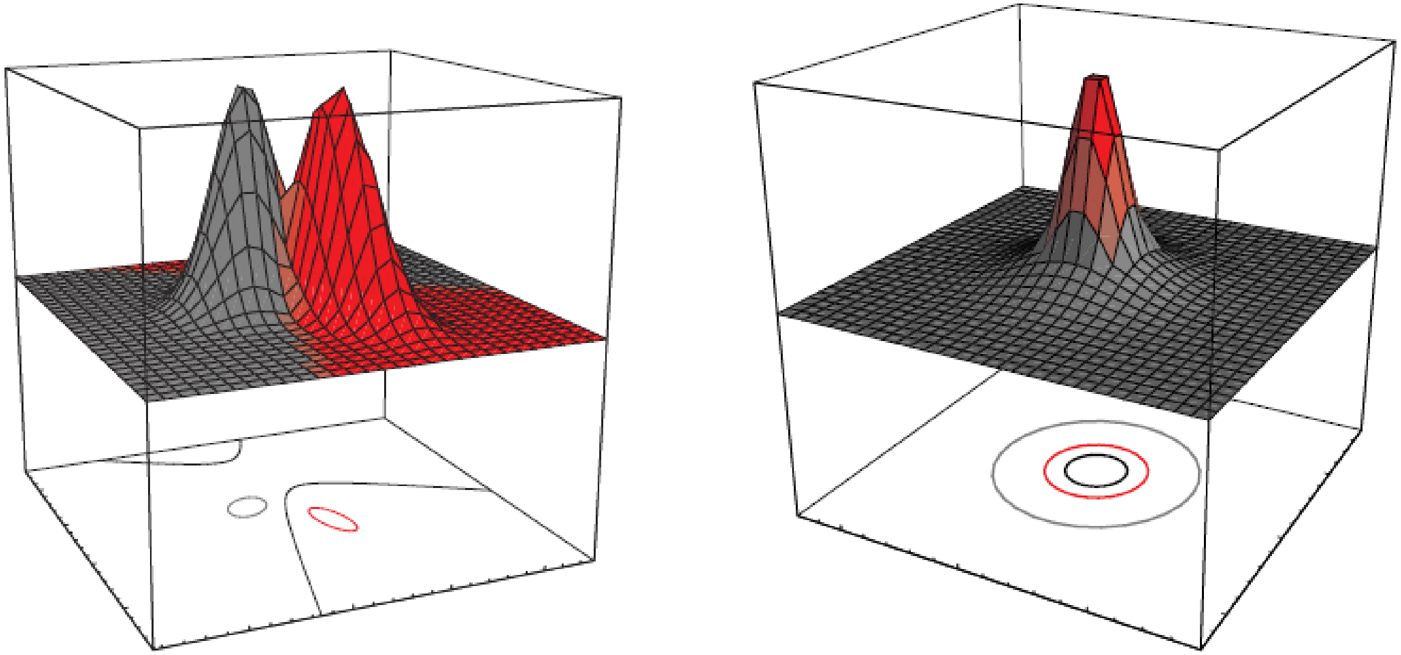
From: R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Recognition.* Copyright © 2001 by John Wiley & Sons, Inc.

Bayes Decision Boundaries for Arbitrary Gaussian Distributions



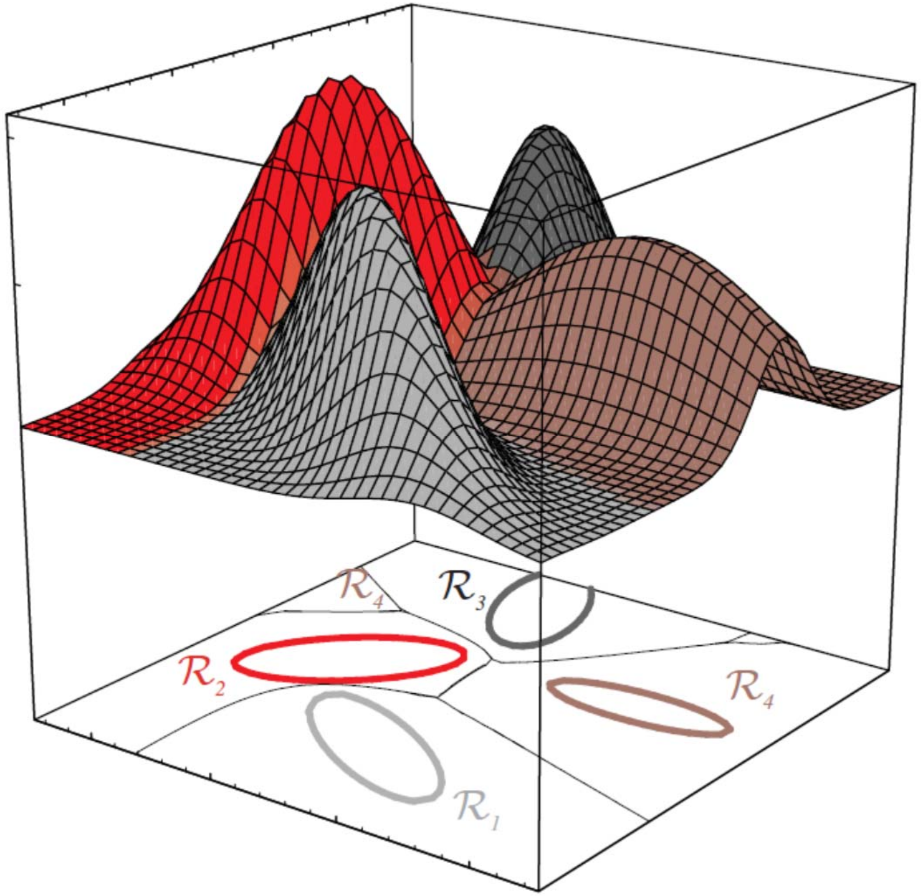
From: R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Recognition.* Copyright © 2001 by John Wiley & Sons, Inc.

Bayes Decision Boundaries for Arbitrary Gaussian Distributions



From: R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Recognition.* Copyright © 2001 by John Wiley & Sons, Inc.

Bayes Decision Boundaries for Four Normal Distributions



From: R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Recognition.*
Copyright © 2001 by John Wiley & Sons, Inc.

Numerical Example #1

a) Derive a linear discriminant function for a two-class 3D classification problem

Gaussian likelihoods:

$$\mu_1 = [0 \quad 0 \quad 0]^T; \ \mu_2 = [1 \quad 1 \quad 1]^T; \qquad \Sigma_1 = \Sigma_2 = \begin{bmatrix} 1/4 & 0 & 0 \\ 0 & 1/4 & 0 \\ 0 & 0 & 1/4 \end{bmatrix};$$

Priors:

$$p(\omega_2) = 2p(\omega_1)$$

b) Classify the test example: $x=[0.1, 0.7, 0.8]^T$

Numerical Example #1: Solution

Discriminant function:

$$\boxed{g_i(\mathbf{x}) = -\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_i)^t \boldsymbol{\Sigma}_i^{-1}(\mathbf{x}-\boldsymbol{\mu}_i) - \frac{1}{2}\ln|\boldsymbol{\Sigma}_i| + \ln P(\omega_i)}$$

$$\boldsymbol{\Sigma}_i = \sigma^2 \mathbf{I} \qquad \Sigma_1 = \Sigma_2 = \begin{bmatrix} 1/4 & 0 & 0 \\ 0 & 1/4 & 0 \\ 0 & 0 & 1/4 \end{bmatrix};$$

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

Linear discriminant:

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

Numerical Example #1: Solution

a) Derive a linear discriminant function

Solution: $$g_i(\mathbf{x}) = -\frac{1}{2\sigma^2}(-2\boldsymbol{\mu}_i{}^t\mathbf{x} + \boldsymbol{\mu}_i{}^t\boldsymbol{\mu}_i) + \ln P(\omega_i)$$

$$\mu_1 = [0 \quad 0 \quad 0]^T; \ \mu_2 = [1 \quad 1 \quad 1]^T; \quad \sigma^2 = \frac{1}{4}; \quad P(\omega_1) = \frac{1}{3}; P(\omega_2) = \frac{2}{3};$$

$$g_1(\mathbf{x}) = \ln\left(\frac{1}{3}\right); \qquad g_2(\mathbf{x}) = -2[-2(x_1 + x_2 + x_3) + 3] + \ln\left(\frac{2}{3}\right)$$

$$g_1(x) \underset{\omega_2}{\overset{\omega_1}{\gtrless}} g_2(x) \quad \Longrightarrow \quad x_1 + x_2 + x_3 \underset{\omega_1}{\overset{\omega_2}{\gtrless}} \frac{6 - \ln 2}{4} = 1.3267$$

Numerical Example #1: Solution

b)Classify the test example: x=[0.1, 0.7, 0.8]^T

Solution:

According to a), the decision rules is:

$$x_1 + x_2 + x_3 \underset{\omega_1}{\overset{\omega_2}{\gtrless}} \frac{6 - \ln 2}{4} = 1.3267$$

By substitution,

$$0.1 + 0.7 + 0.8 = 1.6 \underset{\omega_1}{\overset{\omega_2}{\gtrless}} 1.32$$

$$\mathbf{x} \in \omega_2$$

Numerical Example #2

In a two-class, two-dimensional classification task, the feature vectors are generated by two normal distributions sharing the same covariance matrix:

$$\mathbf{\Sigma} = \begin{bmatrix} 1 & 0 \\ 0 & 2 \end{bmatrix}$$

The mean vectors are: $\boldsymbol{\mu}_1 = [0 \quad 0]^t$ and $\boldsymbol{\mu}_2 = [3 \quad 3]^t$

Assuming that the two classes are equally likely, derive the Bayes decision rule and classify

$$\mathbf{x} = [1.0 \quad 2.1]^t$$

Numerical Example #2: Solution

The Bayes decision rule is:

$$\text{Decide } \omega_1 \text{, if } P(\omega_1 \mid \mathbf{x}) > P(\omega_2 \mid \mathbf{x}) \text{; otherwise decide } \omega_2 \text{ .}$$

It is equivalent to

$$\text{Decide } \omega_1 \text{, if } P(\mathbf{x} \mid \omega_1)P(\omega_1) > P(\mathbf{x} \mid \omega_2)P(\omega_2) \text{; otherwise decide } \omega_2 \text{ .}$$

$$\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \boldsymbol{\Sigma} \qquad \Downarrow \qquad P(\omega_1) = P(\omega_2) = 0.5$$

$$\text{Decide } \omega_1 \text{, if } (\mathbf{x}-\boldsymbol{\mu}_1)^t \boldsymbol{\Sigma}^{-1} (\mathbf{x}-\boldsymbol{\mu}_1) < (\mathbf{x}-\boldsymbol{\mu}_2)^t \boldsymbol{\Sigma}^{-1} (\mathbf{x}-\boldsymbol{\mu}_2) \text{; otherwise decide } \omega_2 \text{ .}$$

$$\boldsymbol{\Sigma} = \begin{bmatrix} 1 & 0 \\ 0 & 2 \end{bmatrix} \qquad \boldsymbol{\Sigma}^{-1} = \begin{bmatrix} 1 & 0 \\ 0 & 1/2 \end{bmatrix}$$

Numerical Example #2: Solution

Decide ω_1, if $(\mathbf{x}-\boldsymbol{\mu}_1)^t\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}_1) < (\mathbf{x}-\boldsymbol{\mu}_2)^t\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}_2)$; otherwise decide ω_2 .

$$\boldsymbol{\mu}_1 = \begin{bmatrix}0 & 0\end{bmatrix}^t \text{ and } \boldsymbol{\mu}_2 = \begin{bmatrix}3 & 3\end{bmatrix}^t \qquad \boldsymbol{\Sigma}^{-1} = \begin{bmatrix}1 & 0\\ 0 & 1/2\end{bmatrix}$$

Given $\mathbf{x} = \begin{bmatrix}1.0 & 2.1\end{bmatrix}^t$,

$$(\mathbf{x}-\boldsymbol{\mu}_1)^t\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}_1) = \begin{bmatrix}1 & 2.1\end{bmatrix}\begin{bmatrix}1 & 0\\ 0 & 1/2\end{bmatrix}\begin{bmatrix}1\\ 2.1\end{bmatrix} = 1 + 2.1\times 2.1/2 < 4$$

$$(\mathbf{x}-\boldsymbol{\mu}_2)^t\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}_2) = \begin{bmatrix}-2 & -0.9\end{bmatrix}\begin{bmatrix}1 & 0\\ 0 & 1/2\end{bmatrix}\begin{bmatrix}-2\\ -0.9\end{bmatrix} = 4 + 0.9\times 0.9/2 > 4$$

Therefore, $\mathbf{x} = \begin{bmatrix}1.0 & 2.1\end{bmatrix}^t$ should be classified as ω_1 .

Summary

- 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.

Key Concepts

- **Classifiers**
  - **Discriminant function**
    - ✓ **Linear discriminant function**
    - ✓ **Quadratic discriminant function**
  - **Minimum distance classifier**
  - **Distance measures**
    - ✓ **Mahalanobis distance**
    - ✓ **Euclidean distance**

- **Normal density function**
  - **Univariate density**
  - **Multivariate density**
  - **Variance, covariance**
  - **Statistically independent**

Next Time

- In PR applications, we rarely have complete knowledge about the probabilistic structure of the problem.
  - Conditional densities $p(x|\omega_j)$ and a priori probabilities $P(\omega_j)$ are unknown.
  - Merely have some vague, general knowledge, together with training data
- Given training data, how to estimate $p(x|\omega_j)$ and $P(\omega_j)$?
  - Parameter estimation (e.g., mean, variance)
  - Non-parametric techniques (e.g., Parzen windows)

Readings

- Chapter 2, Pattern Classification by Duda, Hart, Stork, 2001, Sections 2.5 – 2.6