Bayesian Decision Theory

SHALA-2020
https://shala2020.github.io/

Most of today's slides are adopted from:
https://www.cc.gatech.edu/~hic/CS7616/pdf/lecture2.pdf

Outline

- Bayes Theorem for Machine Learning
- Bayesian Decision
- Risk
- Discriminant function
- Gaussian distribution
- Decision boundary for Gaussian
- Sufficient Statistics
- Maximum Likelihood estimate of parameters

Bayesian Decision Theory

- Design classifiers to recommend decisions that minimize some total expected ”risk”.
  - The simplest risk is the classification error (i.e., costs are equal).
  - Typically, the risk includes the cost associated with different decisions.

Terminology

- State of nature ω *(random variable):*
  - 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 (cont'd)

- 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

e.g., lightness distributions between salmon/sea-bass populations

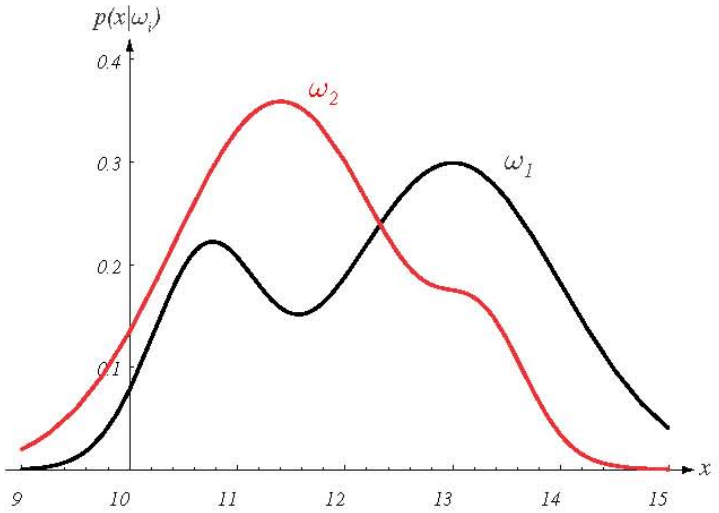


FIGURE 2.1. Hypothetical class-conditional probability density functions show the probability density of measuring a particular feature value *x* given the pattern is in category ω_i. If *x* represents the lightness of a fish, the two curves might describe the difference in lightness of populations of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

Terminology (cont'd)

- Conditional probability $P(\omega_j/x)$ *(posterior)* :
  - e.g., the probability that the fish belongs to class ω_j given measurement x.

Decision Rule Using Prior Probabilities

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

$$P(error) = \begin{cases} P(\omega_1) \ \ if \ we \, decide \, \omega_2 \\ P(\omega_2) \ if \ we \, decide \, \omega_1 \end{cases}$$

or $P(error) = min[P(\omega_1), P(\omega_2)]$

- 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 Conditional Probabilities

- Using Bayes' rule, the posterior probability of category ω_j given measurement x is given by:

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

where $p(x) = \sum_{j=1}^{2} p(x/\omega_j)P(\omega_j)$ (i.e., scale factor – sum of probs = 1)

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

Decision Rule Using Conditional pdf (cont'd)

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

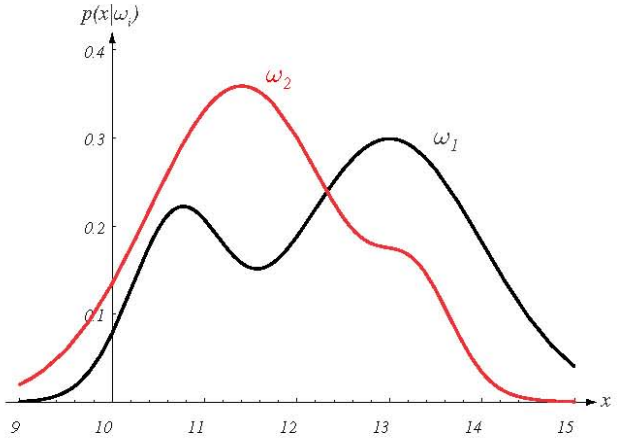


FIGURE 2.1. Hypothetical class-conditional probability density functions show the probability density of measuring a particular feature value x given the pattern is in category ω_i. If x represents the lightness of a fish, the two curves might describe the difference in lightness of populations of two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

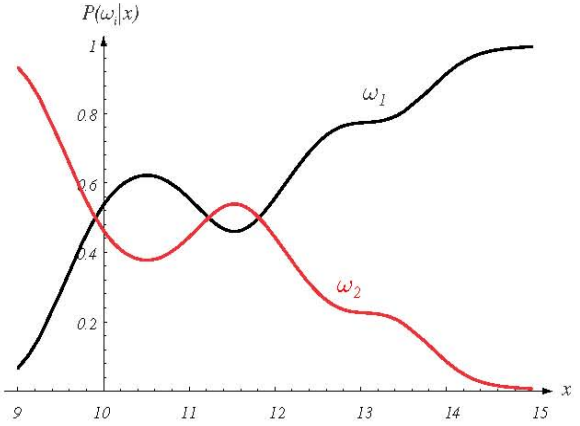


FIGURE 2.2. Posterior probabilities for the particular priors $P(\omega_1) = 2/3$ and $P(\omega_2) = 1/3$ for the class-conditional probability densities shown in Fig. 2.1. Thus in this case, given that a pattern is measured to have feature value $x = 14$, the probability it is in category ω_2 is roughly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

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

or $P(error/x) = min[P(\omega_1/x), P(\omega_2/x)]$

- 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 to this question:

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

  (2) **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 $50K:
  - Classes: C_1 if price > $50K, C_2 if price <= $50K
  - Features: x, the height of a car

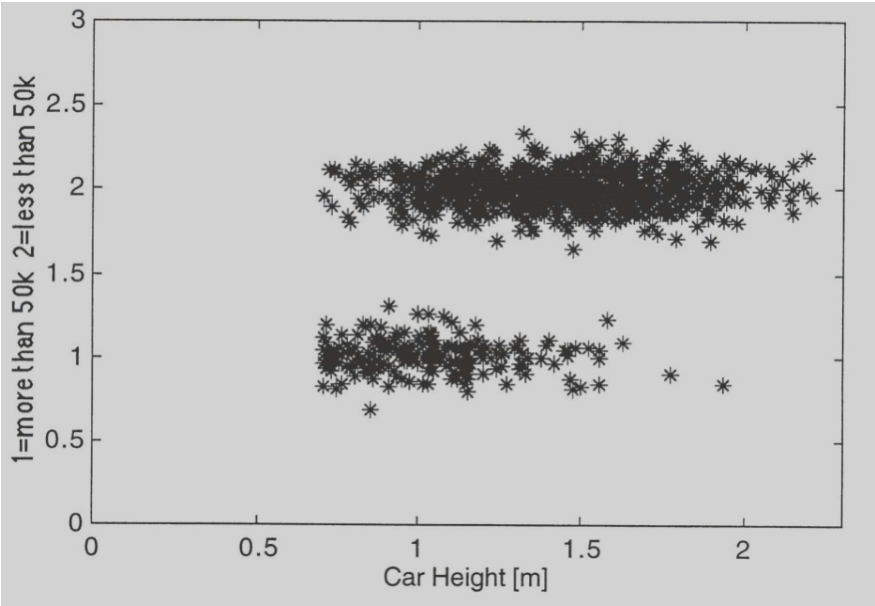
- Use the Bayes' rule to compute the posterior probabilities:

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

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

Example (cont'd)

- 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



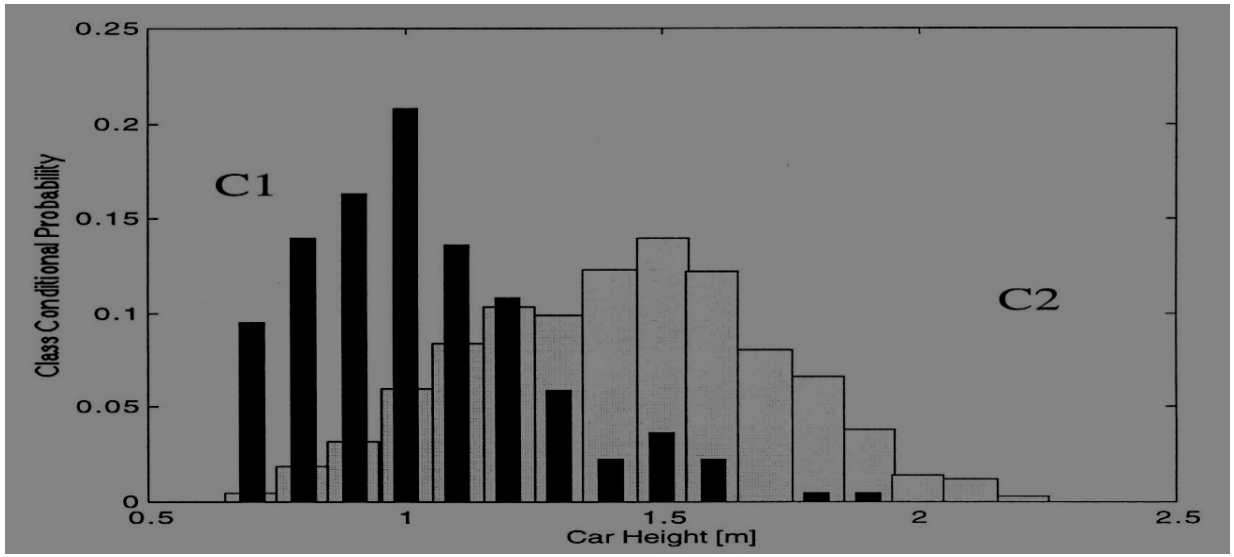
$$P(C_1) = \frac{221}{1209} = 0.183$$

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

Example (cont'd)

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

$p(x / C_i)$



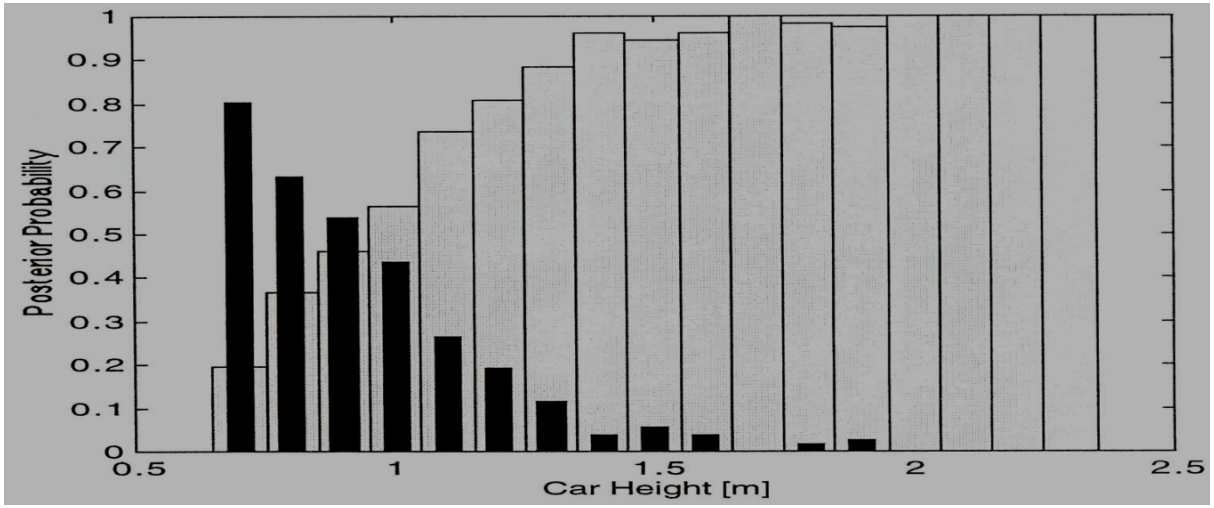
Example (cont'd)

- Calculate the posterior probability for each bin:

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

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

$P(C_i / x)$



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., "risk" function) by associating a "cost" ("loss" function) with each error (i.e., wrong action).

Terminology

- Features form a vector $\mathbf{x} \in R^d$
- A finite set of *c* categories $\omega_1, \omega_2, \ldots, \omega_c$
- Bayes rule (i.e., using vector notation):

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

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

- 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

Risk example

Preliminary COVID-19 test using Machine Learning
If detected, then patient is sent for lab test
If not detected, then patient is not sent for lab test

	Infected	Not Infected
Detected	0	0.1
Not Detected	10	0

Conditional Risk (or Expected Loss)

- Suppose we observe **x** and take action α_i
- Suppose that the cost associated with taking action α_i with ω_j being the correct category is $\lambda(\alpha_i / \omega_j)$
- The **conditional risk** (or **expected loss**) with taking action α_i is:

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

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**; then 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 (cont'd)

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

- The resulting minimum overall risk 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 (c=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 (cont'd)

- 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 (cont'd)

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

- In this case, the overall risk is the average probability error!

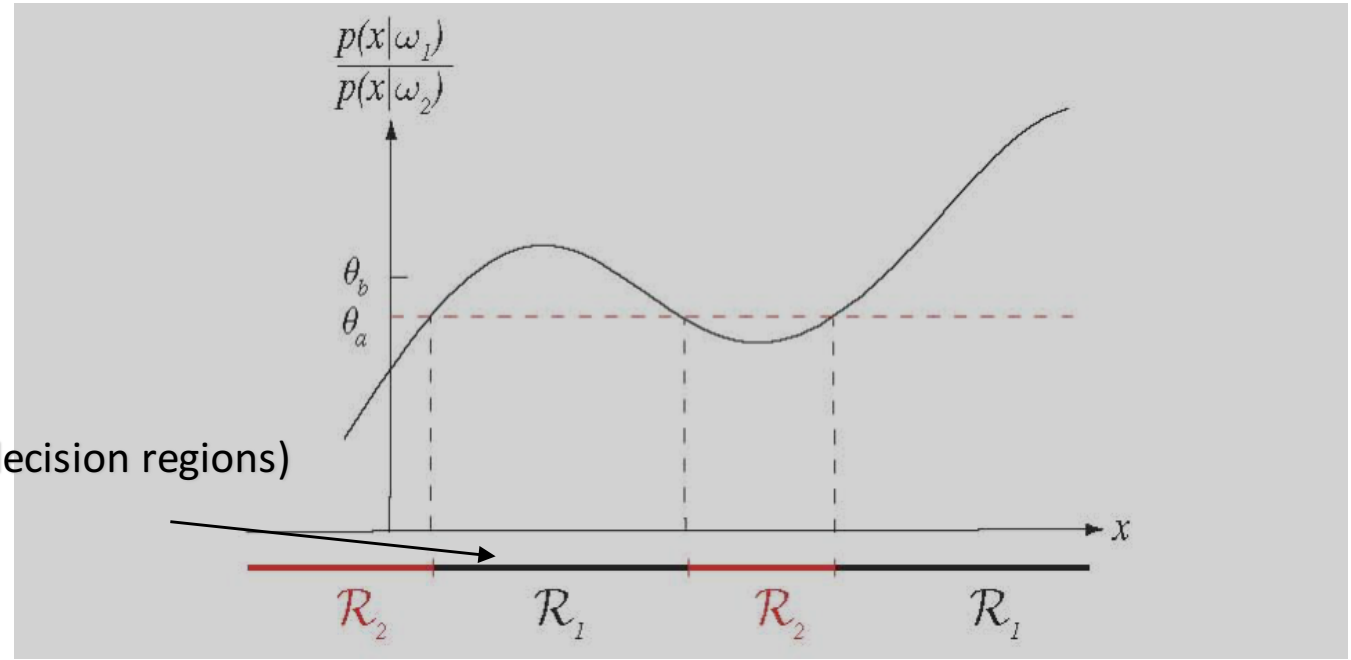
Example

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)

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

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

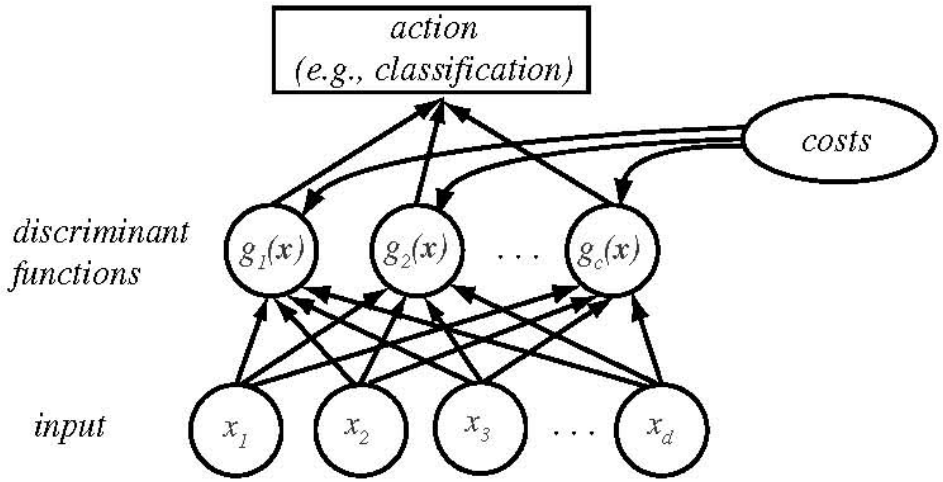
assume: $\lambda_{12} > \lambda_{21}$

Discriminant Functions

- A useful way to represent classifiers 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

- Assuming a general loss function:

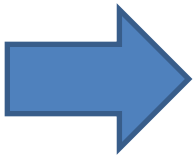
$$g_i(\mathbf{x}) = -R(\alpha_i / \mathbf{x})$$

- Assuming the zero-one loss function:

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

Discriminants for Bayes Classifier (cont'd)

- Is the choice of g_i unique?
  - Replacing $g_i(\mathbf{x})$ with $f(g_i(\mathbf{x}))$, where $f()$ 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 form 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:

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

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

- Recall that the univariate normal distribution, with mean μ and variance σ^2, has the probability density function

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-[(x-\mu)/\sigma]^2/2} \quad -\infty < x < \infty$$

- The term

$$\left(\frac{x-\mu}{\sigma}\right)^2 = (x-\mu)(\sigma^2)^{-1}(x-\mu)$$

- This can be generalized for $p \times 1$ vector **x** of observations on serval variables as

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

The $p \times 1$ vector $\boldsymbol{\mu}$ represents the expected value of the random vector $\boldsymbol{X}$, and the $p \times p$ matrix Σ is the variance-covariance matrix of $\boldsymbol{X}$.

- A p-dimensional normal density for the random vector $\boldsymbol{X}' = [X_1, X_2, \ldots, X_p]$ has the form

$$f(\mathbf{x}) = \frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}} e^{-(\mathbf{x}-\boldsymbol{\mu})'\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu})/2}$$

where $-\infty < x_i < \infty, i = 1, 2, \ldots, p$. We should denote this p-dimensional normal density by $N_p(\boldsymbol{\mu}, \Sigma)$.

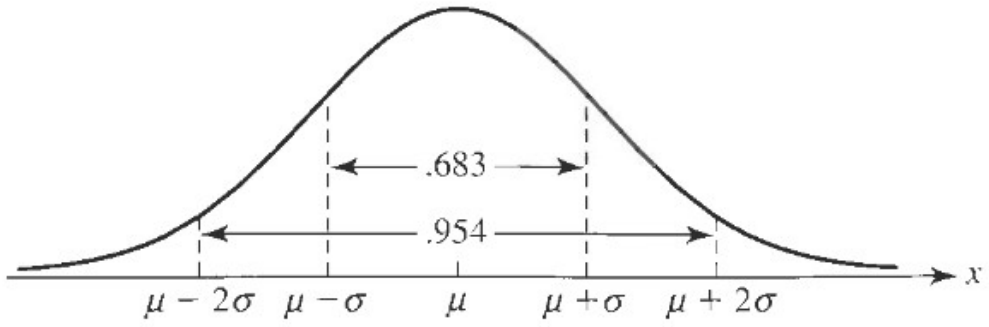


Figure 4.1 A normal density with mean μ and variance σ^2 and selected areas under the curve.

Colab Notebook

https://colab.research.google.com/drive/1ZpeWwfCS3fcn5zodXmrvvzL3gQEt8x8p

Discriminant Function for Multivariate Gaussian Density

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

- Consider the following discriminant function:

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

- If $p(\mathbf{x}/\omega_i) \sim N(\mu_i, \Sigma_i)$, then

$$g_i(\mathbf{x}) = -\frac{1}{2}(\mathbf{x}-\mu_i)^t\Sigma_i^{-1}(\mathbf{x}-\mu_i) - \frac{d}{2}ln2\pi - \frac{1}{2}\ln|\Sigma_i| + \ln P(\omega_i)$$

Multivariate Gaussian Density:

Case I

- $\Sigma_i=\sigma^2$(diagonal)

$$g_i(\mathbf{x}) = -\frac{1}{2}(\mathbf{x}-\mu_i)^t\Sigma_i^{-1}(\mathbf{x}-\mu_i) - \frac{d}{2}ln2\pi - \frac{1}{2}\ln|\Sigma_i| + \ln P(\omega_i)$$

  - Features are statistically independent
  - Each feature has the same variance

- If we disregard $\frac{d}{2}ln2\pi$ and $\frac{1}{2}\ln|\Sigma_i|$ (constants):

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

where $\|\mathbf{x}-\mu_i\|^2 = (\mathbf{x}-\mu_i)^t(\mathbf{x}-\mu_i)$

favours the a-priori more likely category

- Expanding the above expression:

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

Multivariate Gaussian Density: Case I (cont'd)

- Disregarding $\mathbf{x}^t\mathbf{x}$ (constant), we get a linear discriminant:

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

where $\mathbf{w}_i = \frac{1}{\sigma^2}\mu_i$, and $w_{i0} = -\frac{1}{2\sigma^2}\mu_i^t\mu_i + \ln P(\omega_i)$

- Decision boundary is determined by hyperplanes; setting $g_i(\mathbf{x}) = g_j(\mathbf{x})$:

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

where $\mathbf{w} = \mu_i - \mu_j$, and $\mathbf{x}_0 = \frac{1}{2}(\mu_i + \mu_j) - \frac{\sigma^2}{\|\mu_i - \mu_j\|^2}\ln\frac{P(\omega_i)}{P(\omega_j)}(\mu_i - \mu_j)$

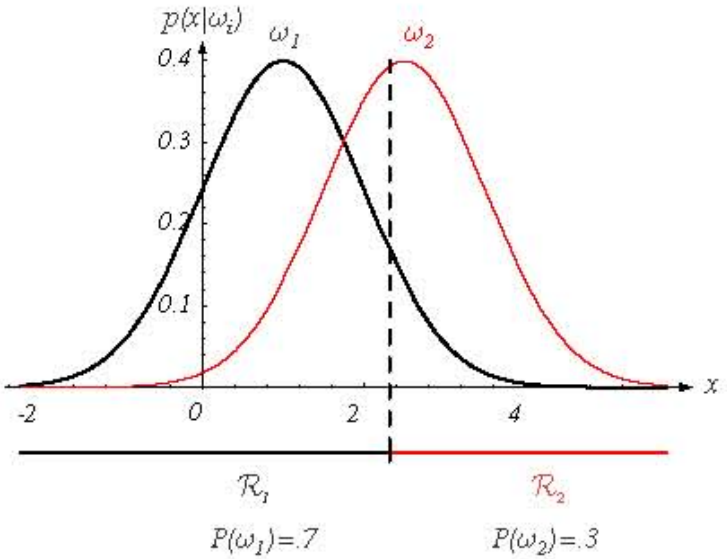
Multivariate Gaussian Density: Case I (cont'd)

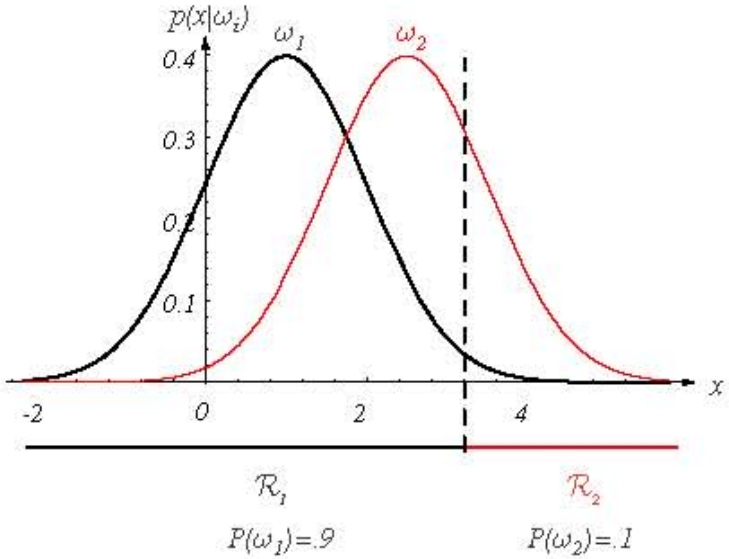
- Properties of decision boundary:
  - It passes through $\mathbf{x_0}$
  - It is orthogonal to the line linking the means.
  - What happens when $P(\omega_i) = P(\omega_j)$?
  - If $P(\omega_i) \neq P(\omega_j)$, then $\mathbf{x_0}$ shifts away from the most likely category.
  - If σ is very small, the position of the boundary is insensitive to $P(\omega_i)$ and $P(\omega_j)$

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

$$\text{where } \mathbf{w} = \mu_i - \mu_j, \text{ and } \mathbf{x}_0 = \frac{1}{2}(\mu_i + \mu_j) - \frac{\sigma^2}{\|\mu_i - \mu_j\|^2} \ln \frac{P(\omega_i)}{P(\omega_j)}(\mu_i - \mu_j)$$

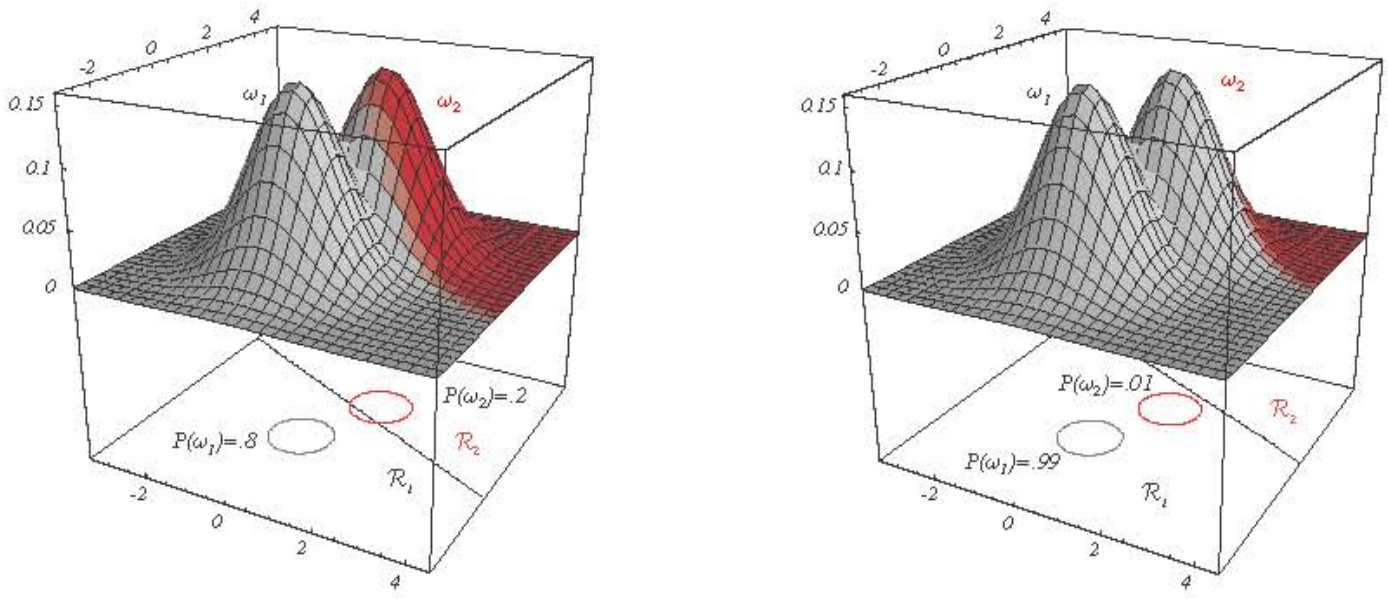
Multivariate Gaussian Density: Case I (cont'd)





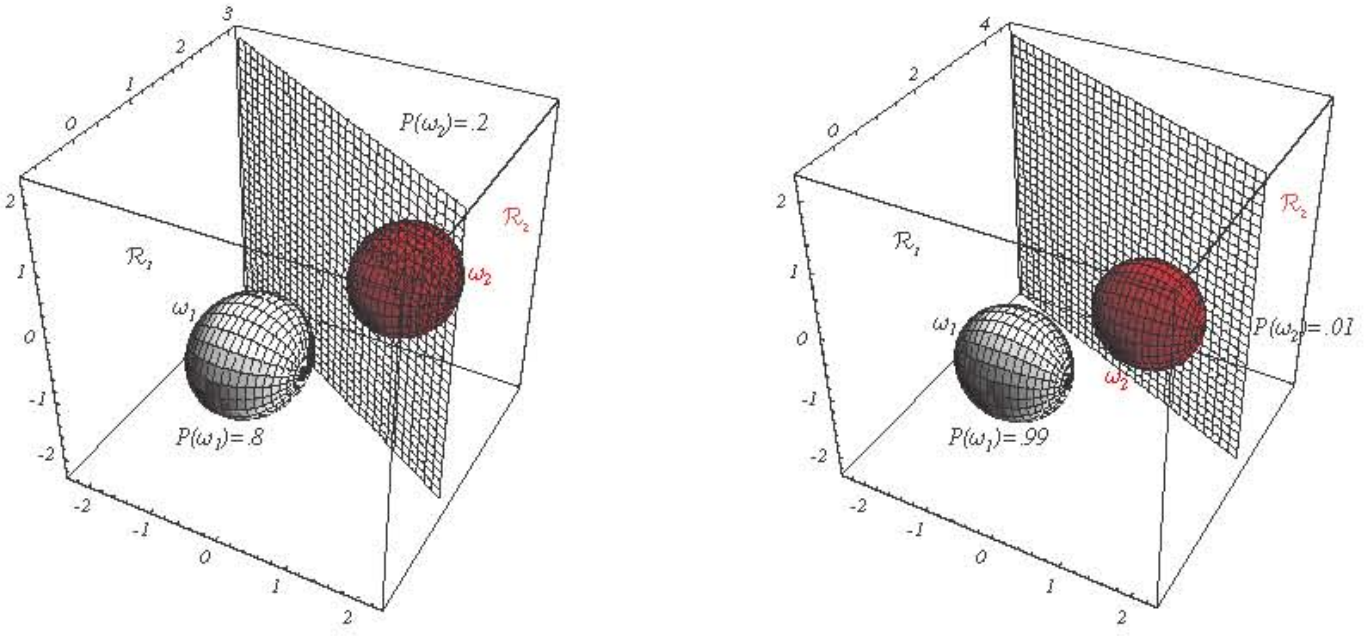
If $P(\omega_i) \neq P(\omega_j)$, then $\mathbf{x_0}$ shifts away from the most likely category.

Multivariate Gaussian Density: Case I (cont'd)



If $P(\omega_i) \neq P(\omega_j)$, then $\mathbf{x_0}$ shifts away from the most likely category.

Multivariate Gaussian Density: Case I (cont'd)

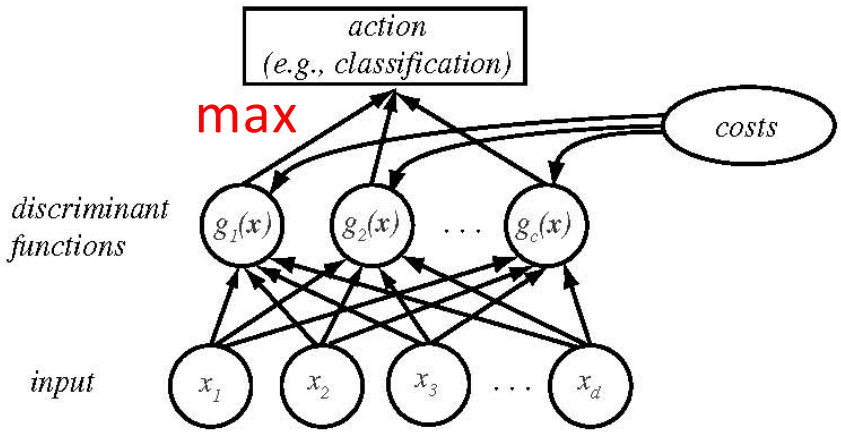


If $P(\omega_i) \neq P(\omega_j)$, then $\mathbf{x_0}$ shifts away from the most likely category.

Multivariate Gaussian Density: Case I (cont'd)

- Minimum distance classifier
  - When $P(\omega_i)$ are equal, then:

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



Multivariate Gaussian Density:

Case II

- $\Sigma_i = \Sigma$ $$g_i(\mathbf{x}) = -\frac{1}{2}(\mathbf{x}-\mu_i)^t\Sigma_i^{-1}(\mathbf{x}-\mu_i) - \frac{d}{2}ln2\pi - \frac{1}{2}\ln|\Sigma_i| + \ln P(\omega_i)$$

- The clusters have hyperellipsoidal shape and same size (centered at μ).

- If we disregard $\frac{d}{2}ln2\pi$ and $\frac{1}{2}\ln|\Sigma_i|$ (constants):

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

- Expanding the above expression and disregarding the quadratic term:

$$\boxed{\begin{array}{c} g_i(\mathbf{x}) = \mathbf{w}_i^t\mathbf{x} + w_{i0} \\ (\textit{linear discriminant}) \end{array}}$$

where $\mathbf{w}_i = \Sigma^{-1}\mu_i$, and $w_{i0} = -\frac{1}{2}\mu_i^t\Sigma^{-1}\mu_i + \ln P(\omega_i)$

Multivariate Gaussian Density: Case II (cont'd)

- Decision boundary is determined by hyperplanes; setting $g_i(\mathbf{x}) = g_j(\mathbf{x})$:

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

where $\mathbf{w} = \Sigma^{-1}(\mu_i - \mu_j)$ and $\mathbf{x}_0 = \frac{1}{2}(\mu_i + \mu_j) - \frac{ln[P(\omega_i)/P(\omega_j)]}{(\mu_i - \mu_j)^t\Sigma^{-1}(\mu_i - \mu_j)}(\mu_i - \mu_j)$

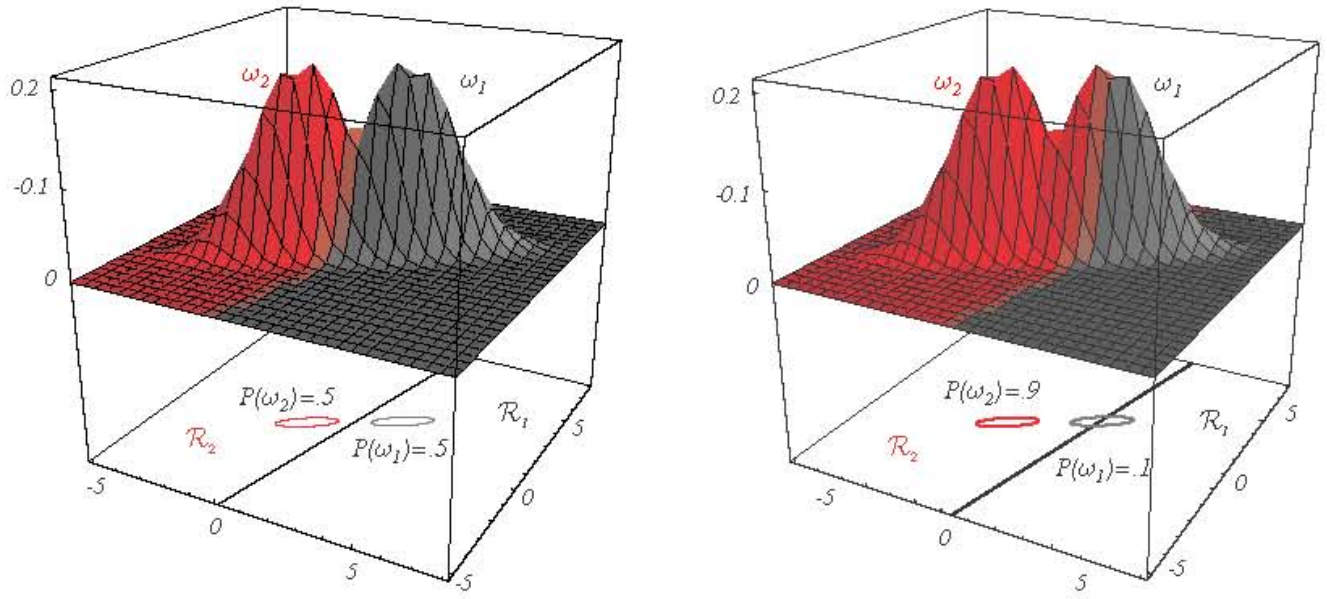
Multivariate Gaussian Density: Case II (cont'd)

- Properties of hyperplane (decision boundary):
  - It passes through $\mathbf{x_0}$
  - It is not orthogonal to the line linking the means.
  - What happens when $P(\omega_i) = P(\omega_j)$?
  - If $P(\omega_i) \neq P(\omega_j)$, then $\mathbf{x_0}$ shifts away from the most likely category.

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

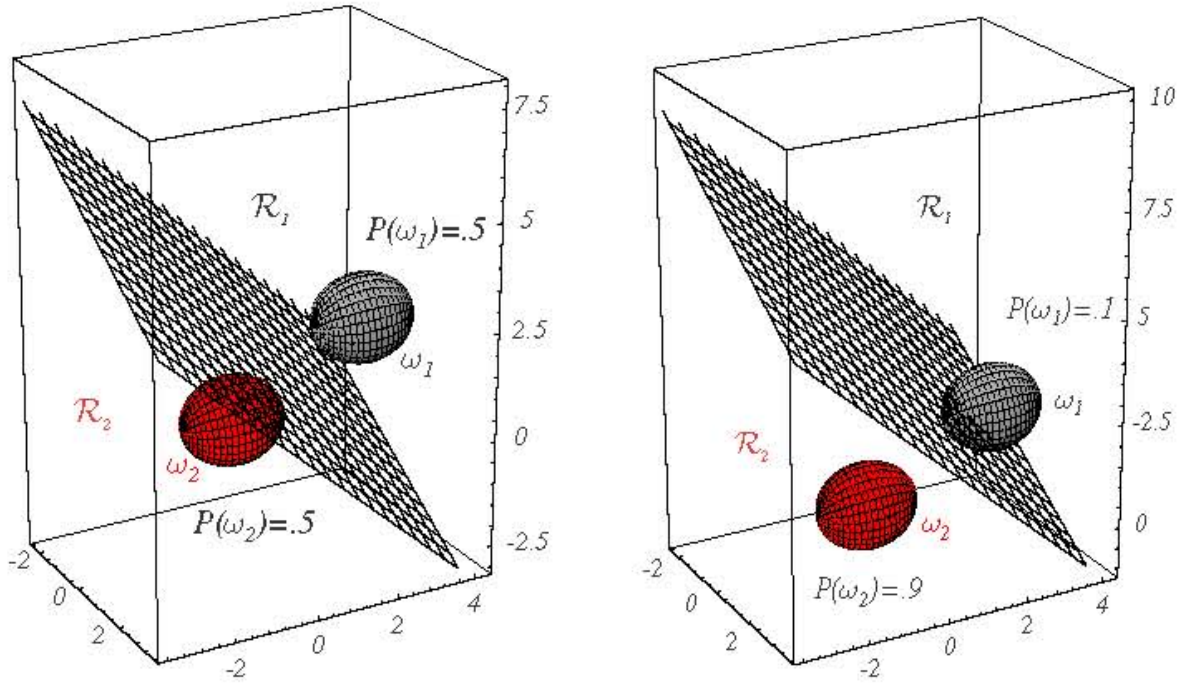
$$\text{where } \mathbf{w} = \Sigma^{-1}(\mu_i - \mu_j) \text{ and } \mathbf{x}_0 = \frac{1}{2}(\mu_i + \mu_j) - \frac{ln[P(\omega_i)/P(\omega_j)]}{(\mu_i - \mu_j)^t \Sigma^{-1}(\mu_i - \mu_j)}(\mu_i - \mu_j)$$

Multivariate Gaussian Density: Case II (cont'd)



If $P(\omega_i) \neq P(\omega_j)$, then $\mathbf{x_0}$ shifts away from the most likely category.

Multivariate Gaussian Density: Case II (cont'd)



If $P(\omega_i) \neq P(\omega_j)$, then $\mathbf{x_0}$ shifts away from the most likely category.

Multivariate Gaussian Density:

Case III

- Σ_i= arbitrary $$g_i(\mathbf{x}) = -\frac{1}{2}(\mathbf{x}-\mu_i)^t\Sigma_i^{-1}(\mathbf{x}-\mu_i) - \frac{d}{2}ln2\pi - \frac{1}{2}\ln|\Sigma_i| + \ln P(\omega_i)$$

  - The clusters have different shapes and sizes (centered at μ).

  - If we disregard $\frac{d}{2}ln2\pi$ (constant):

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

where $\mathbf{W}_i = -\frac{1}{2}\Sigma_i^{-1}$, $\mathbf{w}_i = \Sigma_i^{-1}\mu_i$, and $w_{i0} = -\frac{1}{2}\mu_i^t\Sigma^{-1}\mu_i - \frac{1}{2}ln|\Sigma_i| + \ln P(\omega_i)$

- Decision boundary is determined by hyperquadrics; setting $g_i(\mathbf{x}) = g_j(\mathbf{x})$

  e.g., hyperplanes, pairs of hyperplanes, hyperspheres, hyperellipsoids, hyperparaboloids etc.

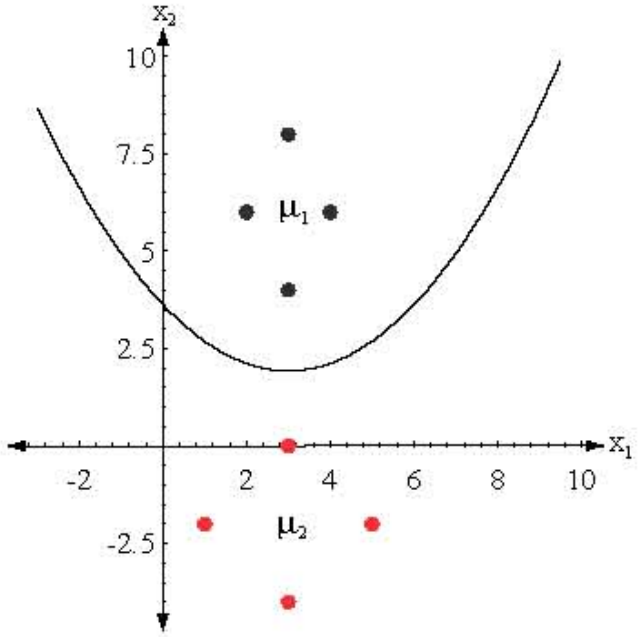
Example - Case III

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

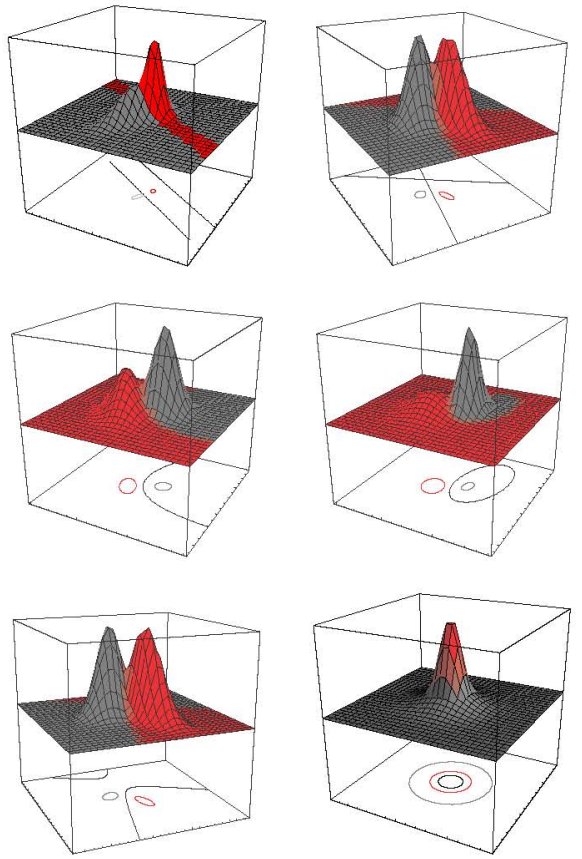
decision boundary: $x_2 = 3.514 - 1.125x_1 + 0.1875x_1^2.$

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

boundary does not pass through midpoint of μ_1,μ_2



Multivariate Gaussian Density: Case III (cont'd)



non-linear decision boundaries

FIGURE 2.14. Arbitrary Gaussian distributions lead to Bayes decision boundaries that are general hyperquadrics. Conversely, given any hyperquadric, one can find two Gaussian distributions whose Bayes decision boundary is that hyperquadric. These variances are indicated by the contours of constant probability density. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

Sufficient Statistics

Suppose we have a random sample$X_1,\cdots, X_n$ taken from a distribution $f(x|\theta)$ which relies on an unknown parameter θ in a parameter space Θ. The purpose of parameter estimation is to estimate the parameter θ from the random sample.

Any real-valued function $T=r(X_1,\cdots, X_n)$ of the observations in the sample is called a statistic. Eg. $T=X_1+X_2+...X_n$

Sufficient Statistics

A statistic is called sufficient statistic if it contains all the necessary information.

Formally, a statistic $T(X1,\cdots, Xn)$ is said to be sufficient for θ if the conditional distribution of $X1,\cdots, Xn$, given $T=t$, does not depend on θ for any value of t.

Example 1: Let $X_1, \cdots, X_n$ be a sequence of independent bernoulli trials with $P(X_i = 1) = \theta$. We will verify that $T = \sum_{i=1}^{n} X_i$ is sufficient for θ.

Proof: We have

$$P(X_1 = x_1, \cdots, X_n = x_n | T = t) = \frac{P(X_1 = x_1, \cdots, X_n = x_n)}{P(T = t)}$$

Bearing in mind that the X_i can take on only the values 0s or 1s, the probability in the numerator is the probability that some particular set of t X_i are equal to 1s and the other $n - t$ are 0s. Since the X_i are independent, the probability of this is $\theta^t(1-\theta)^{n-t}$. To find the denominator, note that the distribution of T, the total number of ones, is binomial with n trials and probability of success θ. Therefore the ratio in the above equation is

$$\frac{\theta^t(1-\theta)^{n-t}}{\binom{n}{t}\theta^t(1-\theta)^{n-t}} = \frac{1}{\binom{n}{t}}$$

The conditional distribution thus does not involve θ at all. Given the total number of ones, the probability that they occur on any particular set of t trials is the same for any value of θ so that set of trials contains no additional information about θ.

Sufficient statistics for Normal Distribution

Suppose that $X1,\cdots, Xn$ form a random sample from a normal distribution for which the mean μ is unknown but the variance σ^2 is known. $T = (1/n)\sum_i X_i$ is a sufficient statistic for μ.

In other words, the sample mean is a sufficient statistic for the true mean if the variance in known.

Maximum Likelihood estimate of Parameters

$$\theta = [\mu, \Sigma, \pi], Z = \sqrt{(2\pi)^D \det(\Sigma)}$$

$$p(x|t) = \frac{1}{Z} \exp\left(-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)\right)$$

$$\log L(\theta) = \log p(x, t|\theta) = \log p(t|\theta) + \log p(x|t, \theta)$$

Maximum Likelihood estimate of Parameters

$$\pi_k = \frac{\sum_{i=1}^{N} \mathbb{1}\left(t^{(i)} = k\right))}{N}$$

$$\mu_k = \frac{\sum_{i=1}^{N} \mathbb{1}\left(t^{(i)} = k\right) x^{(i)}}{\sum_{i=1}^{N} \mathbb{1}\left(t^{(i)} = k\right)}$$

$$\Sigma_k = \frac{\sum_{i=1}^{N} \mathbb{1}\left(t^{(i)} = k\right) \left(x^{(i)} - \mu_k\right) \left(x^{(i)} - \mu_k\right)^T}{\sum_{i=1}^{N} \mathbb{1}\left(t^{(i)} = k\right)}$$

Thank you