Bayesian Decision Theory

Chapter 2 (Duda et al.) – Sections 2.1-2.10

Bayesian Decision Theory

- Design classifiers to make decisions subject to minimizing an expected ”risk”.
  - The simplest risk is the classification error.
  - When misclassification errors are not equally important, the risk can include the cost associated with different misclassification errors.

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 (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

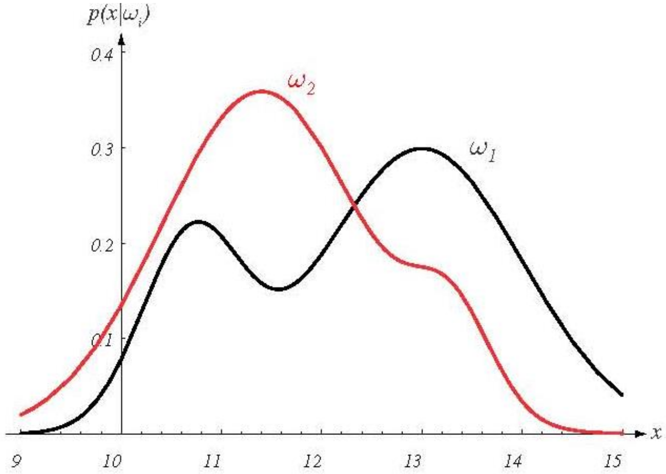


FIGURE 2.1. Hypothetical class-conditional probability density functions show the probability density of measuring a particular feature value x given the pattern is in category ω_i. If x represents the lightness of a fish, the two curves might describe the difference in lightness of populations of two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

Decision Rule Using Prior Probabilities Only

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:

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

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

or

Decide ω_1 if $\underbrace{p(x/\omega_1)/p(x/\omega_2)}_{\text{likelihood ratio}} > \underbrace{P(\omega_2)/P(\omega_1)}_{\text{threshold}}$; otherwise **decide** ω_2

Decision Rule Using Conditional Probabilities (cont'd)

$p(x/\omega_j)$

$$P(\omega_1) = \frac{2}{3} \qquad P(\omega_2) = \frac{1}{3}$$

$P(\omega_j/x)$

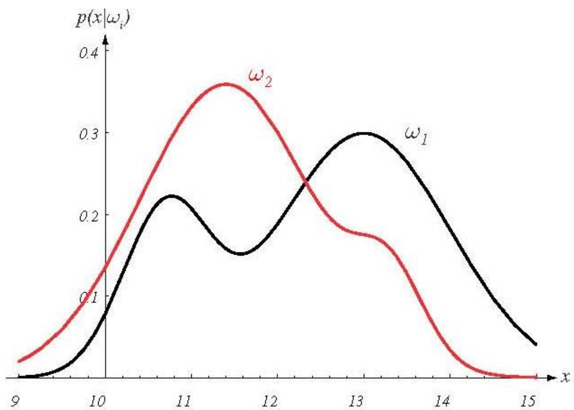


FIGURE 2.1. Hypothetical class-conditional probability density functions show the probability density of measuring a particular feature value x given the pattern is in category ω_i. If x represents the lightness of a fish, the two curves might describe the difference in lightness of populations of two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

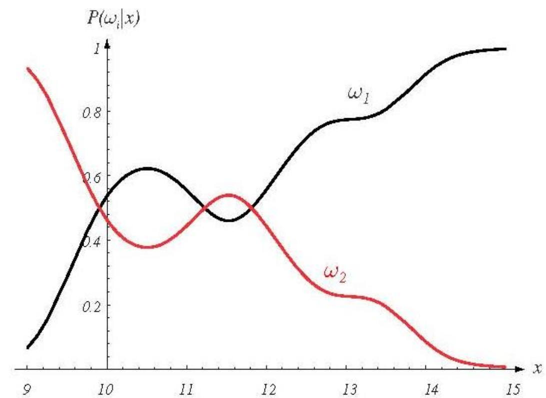


FIGURE 2.2. Posterior probabilities for the particular priors $P(\omega_1) = 2/3$ and $P(\omega_2) = 1/3$ for the class-conditional probability densities shown in Fig. 2.1. Thus in this case, given that a pattern is measured to have feature value $x = 14$, the probability it is in category ω_2 is roughly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

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

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

- 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:

  (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

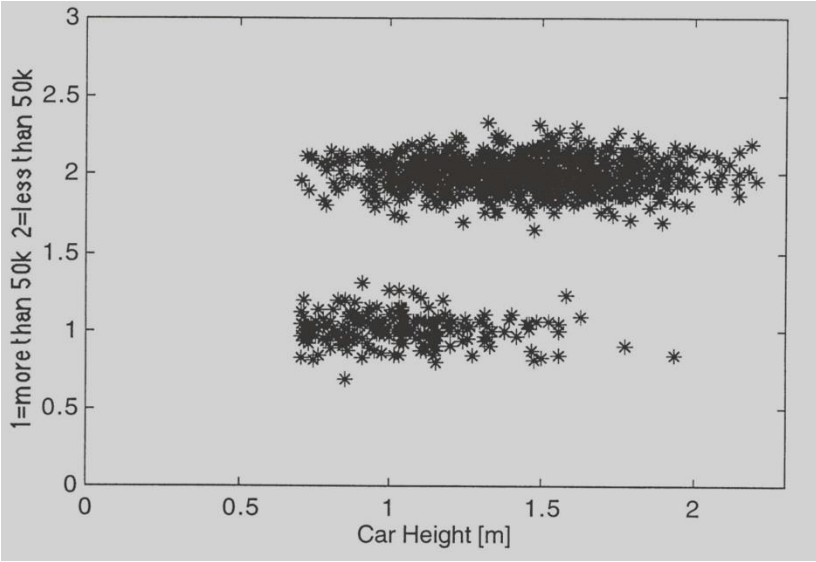
- Use the Bayes' rule to compute the posterior probabilities:

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

- 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



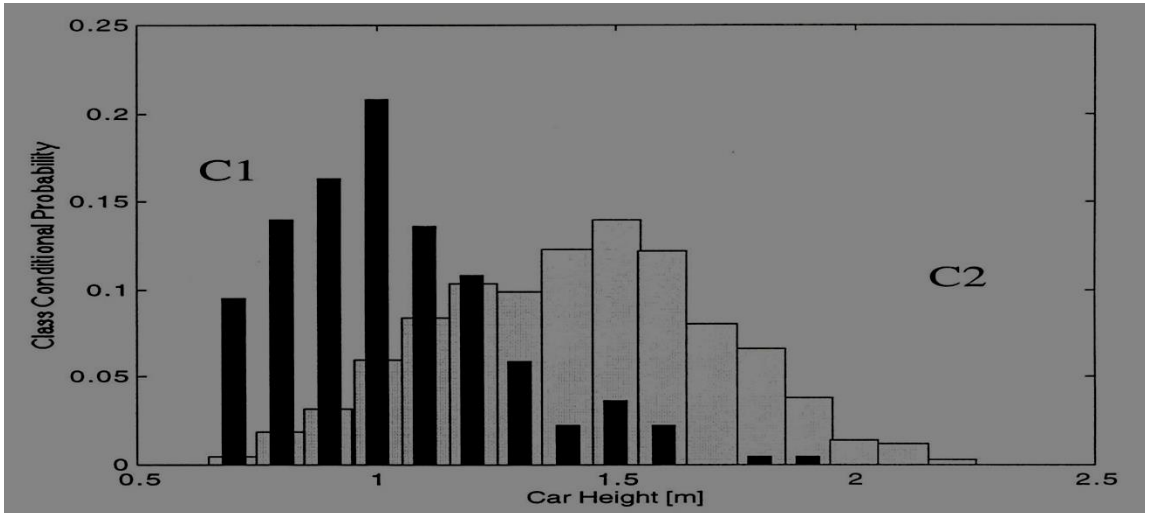
$$P(C_1) = \frac{221}{1209} = 0.183$$

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

Example (cont'd)

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

$p(x / C_i)$



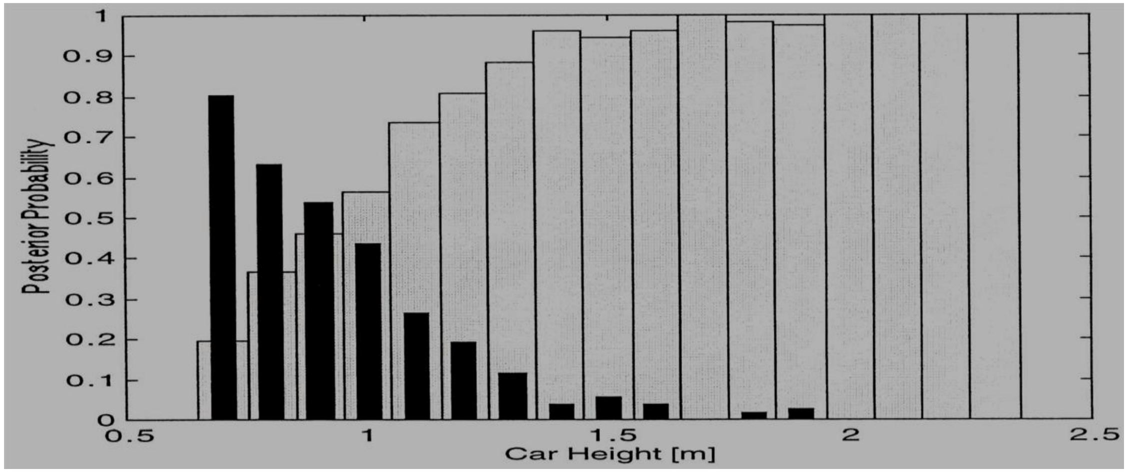
Example (cont'd)

- Calculate the posterior probability for each bin, e.g.:

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

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

$P(C_i / x)$



Example (subjective approach)

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

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

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

- $p(x/C_1) \sim \mathrm{N}(\mu_1,\Sigma_1)$
- $p(x/C_2) \sim \mathrm{N}(\mu_2,\Sigma_2)$
- $P(C_1) = P(C_2) = 0.5$

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 $\boldsymbol{l}$ 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

Bayes rule (using vector notation):

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

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

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:

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

Overall Risk

- The overall risk is defined as:

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

where *α(**x**)* is a general decision rule that determines which action $\alpha_{1,}\ \alpha_{2,\ \dots,}\ \alpha_l$ to take for every **x**.

- The optimum decision rule is the *Bayes rule*

Overall Risk (cont'd)

- 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 performance that can be achieved:

$$R^* = \min R$$

Example: Two-category classification

- Define
  - α_1: decide ω_1
  - $\alpha_{2:}$ decide ω_2
  - $\lambda_{ij} = \lambda(\alpha_i / \omega_j)$

- The conditional risks are:

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

$$\Downarrow$$

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

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

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

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 becomes the average probability error!

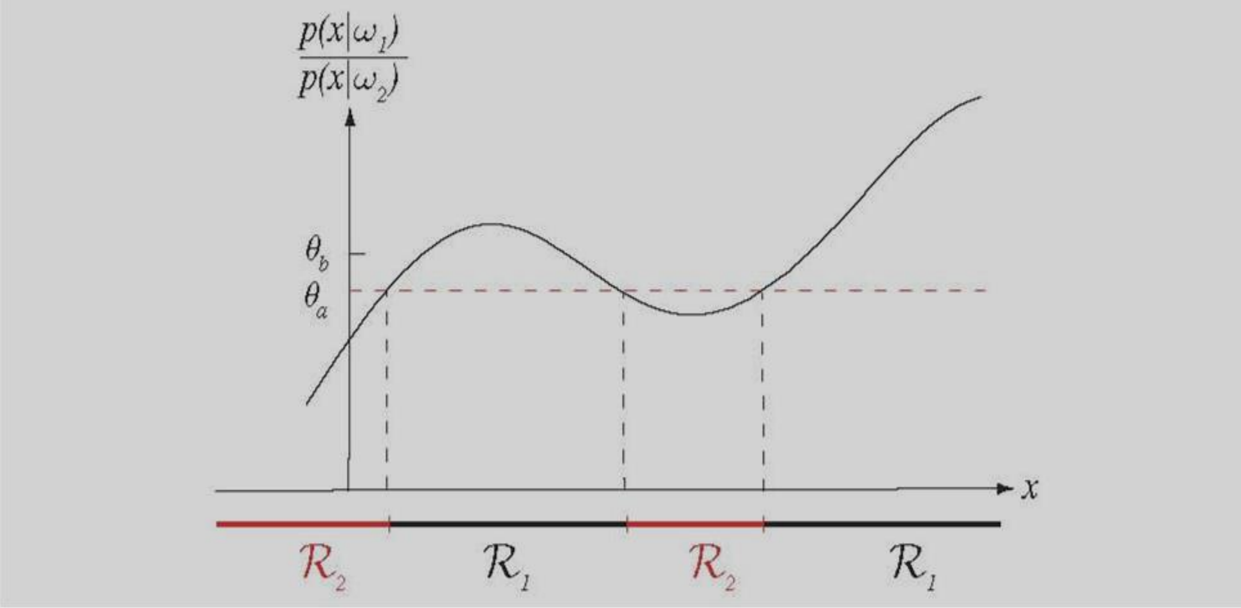
Example

Assuming general loss:

Decide ω_1 if $\dfrac{p(\mathbf{x}/\omega_1)}{p(\mathbf{x}/\omega_2)} > \dfrac{(\lambda_{12}-\lambda_{22})}{(\lambda_{21}-\lambda_{11})}\dfrac{P(\omega_2)}{P(\omega_1)}$; otherwise decide ω_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: $\lambda_{12} > \lambda_{21}$

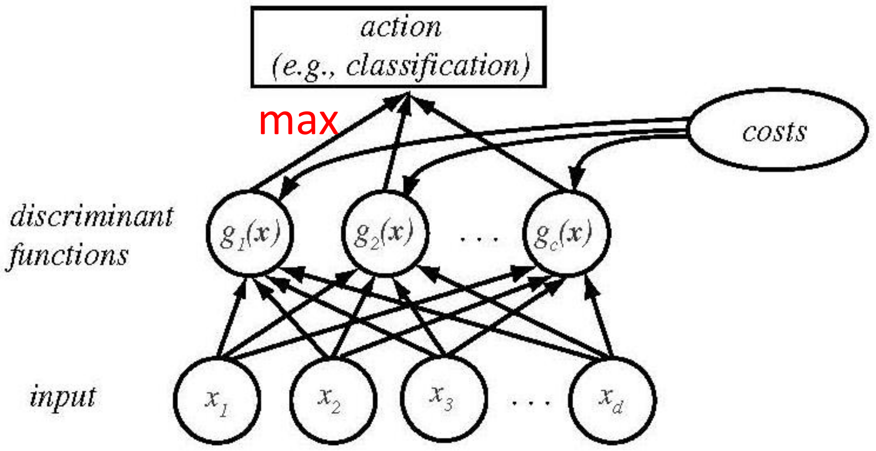
Discriminant Functions

- Represent a classifier is through discriminant functions

$$g_i(x), i = 1, \ldots, c$$

- 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.

$g_i(x)=P(\omega_i/x)$ $\Rightarrow$

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

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

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

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

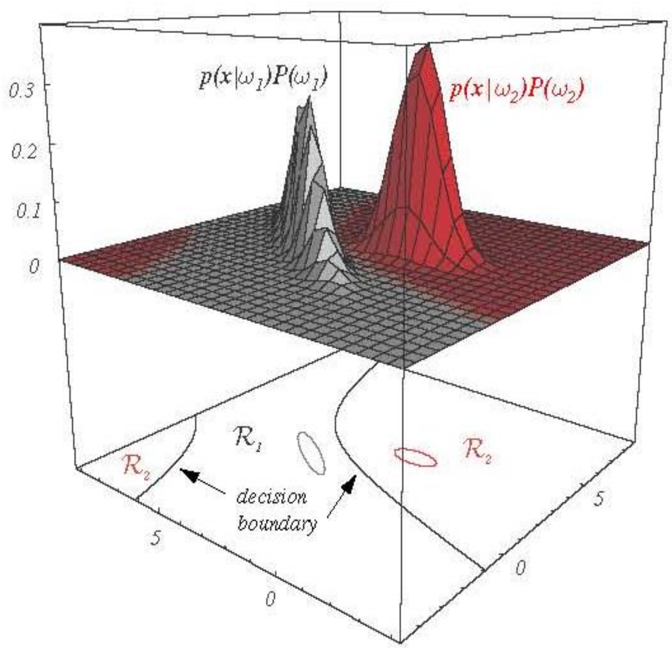
- Examples:

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

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

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:

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

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

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

Multivariate Gaussian Density: Case I

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

- **$\Sigma_i = \sigma^2 I$** (diagonal matrix)
  - Features are statistically independent
  - Each feature has the same variance

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

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

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

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

- 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 $\mathbf{x}_0 = \dfrac{1}{2}(\mu_i + \mu_j) - \dfrac{\sigma^2}{\|\mu_i - \mu_j\|^2}\ln\dfrac{P(\omega_i)}{P(\omega_j)}(\mu_i - \mu_j)$

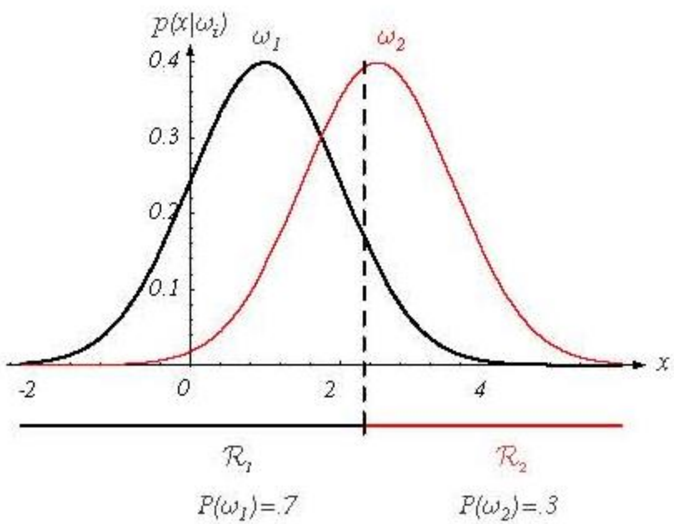
Multivariate Gaussian Density: Case I (cont'd)

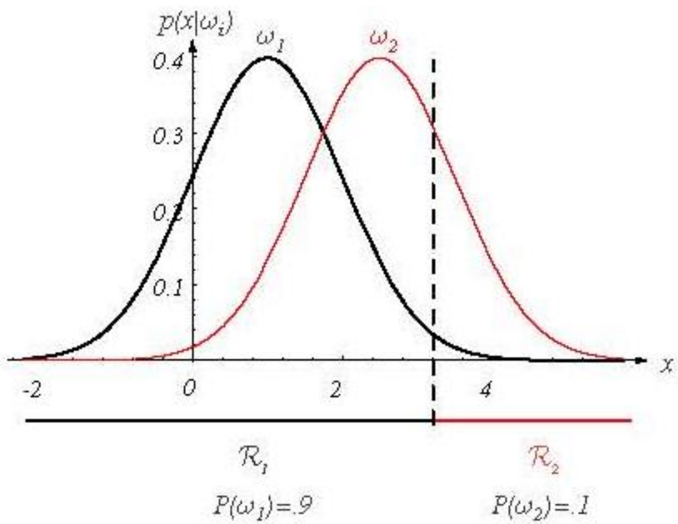
- Properties of decision boundary:
  - It passes through $\mathbf{x_0}$
  - It is orthogonal to the line linking the means.
  - What happens when $P(\omega_i)= P(\omega_j)$?
  - If $P(\omega_i) \neq P(\omega_j)$, then $\mathbf{x_0}$ shifts away from the most likely category.
  - If σ is very small, the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

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

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

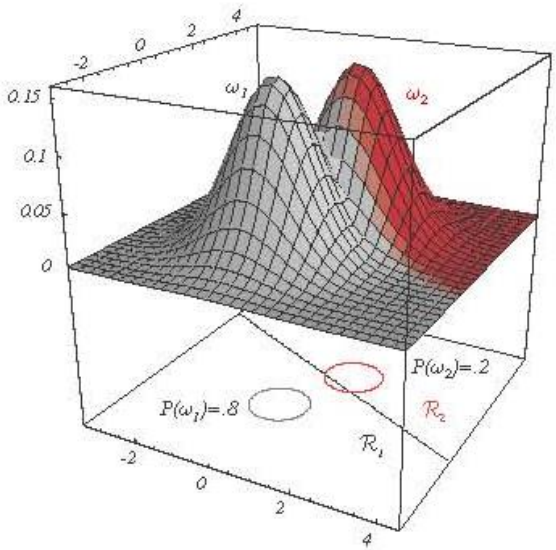
Multivariate Gaussian Density: Case I (cont'd)

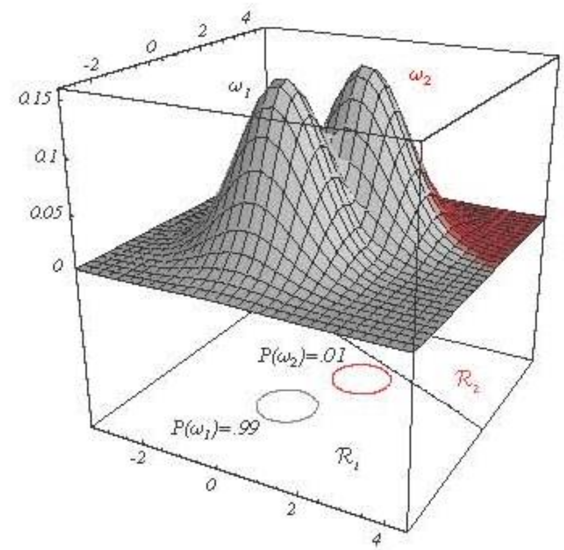




If $P(\omega_i) \neq P(\omega_j)$, then $\mathbf{x_0}$ shifts away from the most likely category.

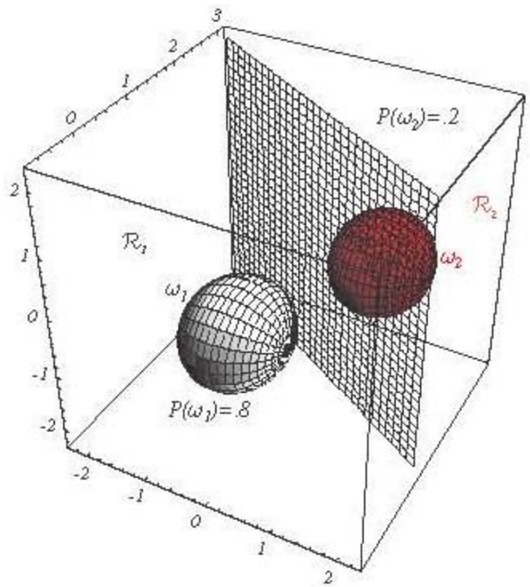
Multivariate Gaussian Density: Case I (cont'd)

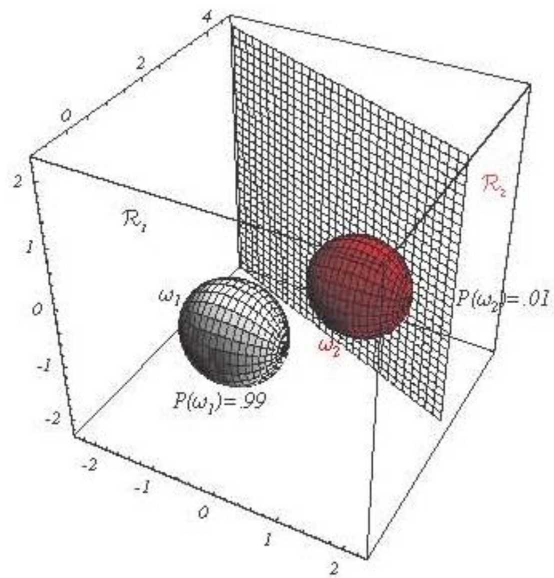




If $P(\omega_i) \neq P(\omega_j)$, then $\mathbf{x_0}$ shifts away from the most likely category.

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 the discriminant becomes:

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

  - This is the Euclidean distance!
  - Assumptions: statistically independent features, same variance!

Multivariate Gaussian Density: Case II

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

- $\Sigma_i = \Sigma$

- 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):

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

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

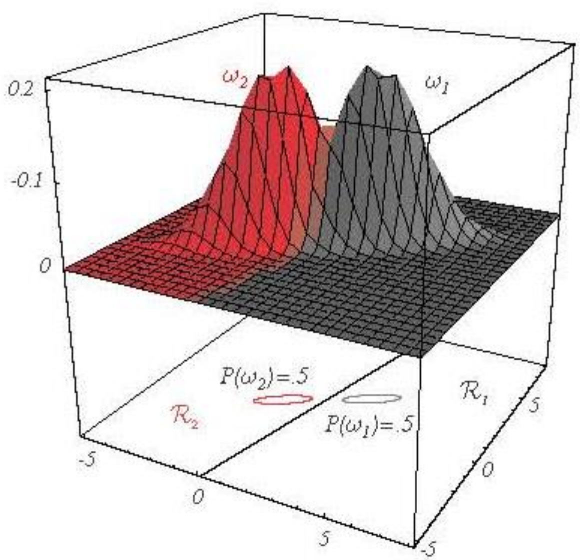
Multivariate Gaussian Density: Case II (cont'd)

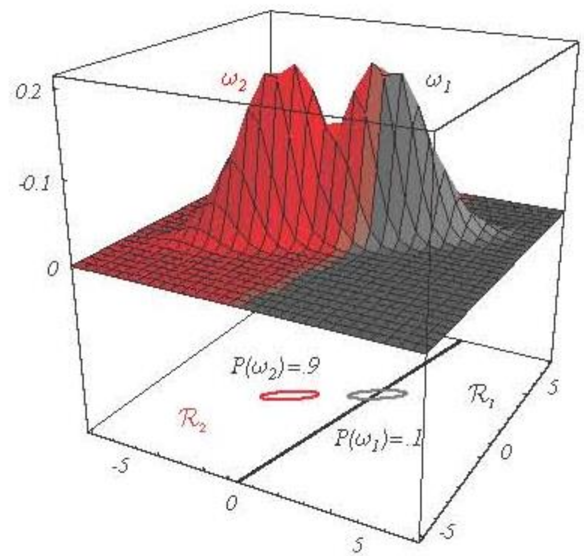
- Properties of hyperplane (decision boundary):
  - It passes through $\mathbf{x_0}$
  - It is not orthogonal to the line linking the means.
  - What happens when $P(\omega_i)= P(\omega_j)$?
  - If $P(\omega_i) \neq P(\omega_j)$, then $\mathbf{x_0}$ shifts away from the most likely category.

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

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

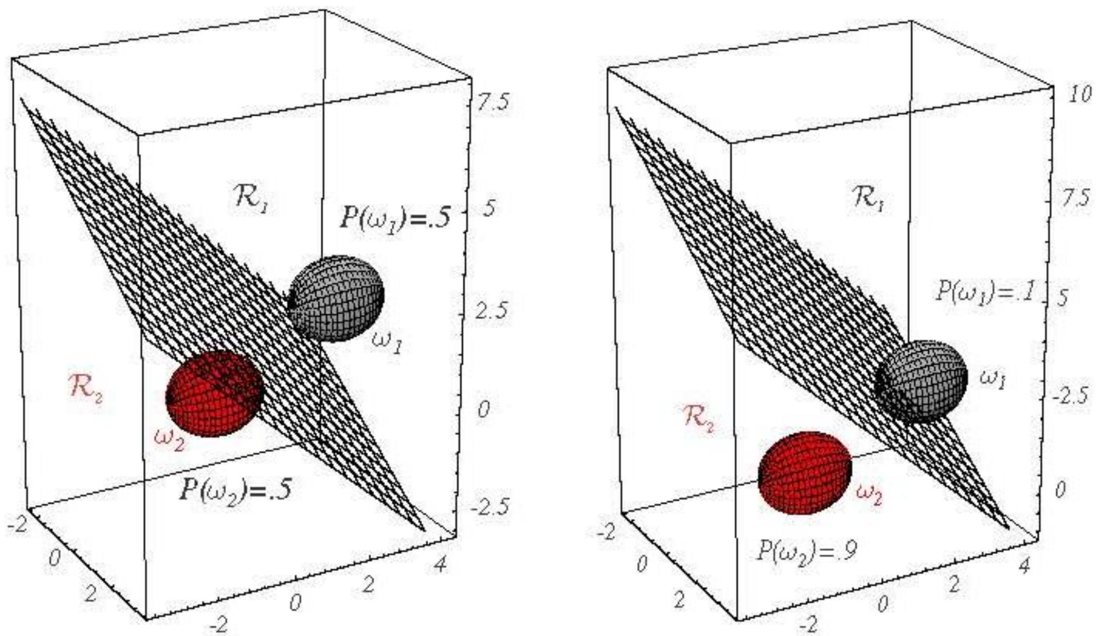
Multivariate Gaussian Density: Case II (cont'd)





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 II (cont'd)

- Mahalanobis distance classifier
  - When $P(\omega_i)$ are equal, then the discriminant becomes:

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

$$\Downarrow$$

$$g_i(\mathbf{x}) = -\frac{1}{2}(\mathbf{x} - \mu_\mathbf{i})^t \Sigma^{-1} (\mathbf{x} - \mu_\mathbf{i})$$

Multivariate Gaussian Density: Case III

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

- Σ_i= arbitrary

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

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

$$\underbrace{g_i(x) = x^t W_i x + w_i^t x + w_{i0}}_{(quadratic\ discriminant)}$$

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

- 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.

Multivariate Gaussian Density: Case III (cont'd)

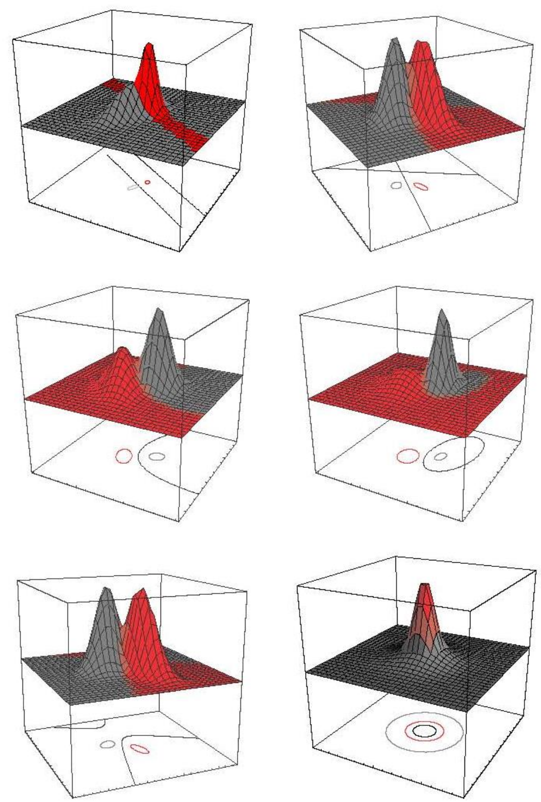
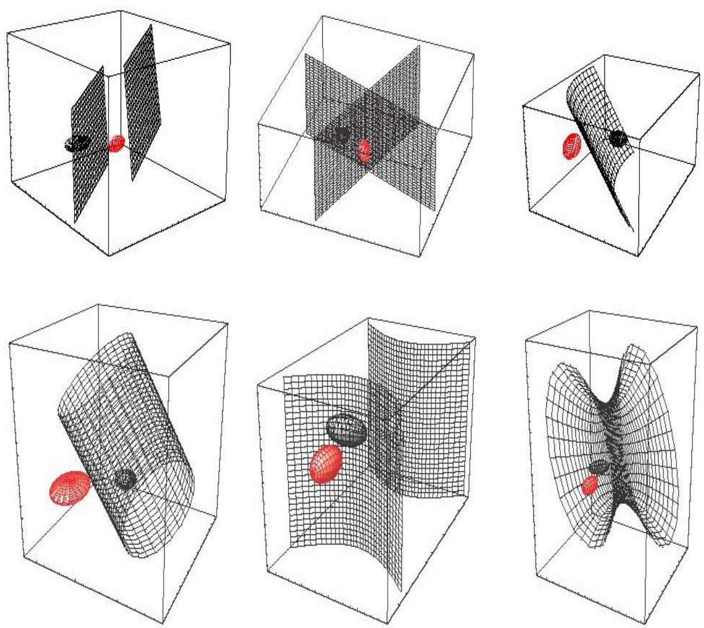


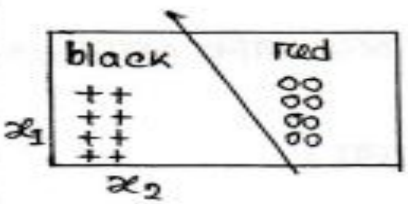
FIGURE 2.14. Arbitrary Gaussian distributions lead to Bayes decision boundaries that are general hyperquadrics. Conversely, given any hyperquadric, one can find two Gaussian distributions whose Bayes decision boundary is that hyperquadric. These variances are indicated by the contours of constant probability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

non-linear decision boundaries

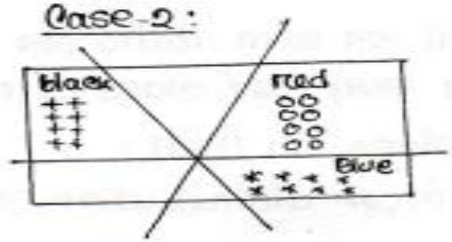
Multivariate Gaussian Density: Case III (cont'd)



Case-1 :



Case-2 :



Discriminant function : $g_i(x) = P(x|w_i)\ P(w_i)$

multivariance টা বসিয়ে দেখবো $w^Tx + w_0$ এই equation টা পাই কিনা

$w^Tx + w_0 =$ linear equation .

unknown point আসলে bayesian দিয়ে probability বের করবো।
সেদিকে যাবে যার probability বেশি।

These lines are ratio between:

$$① \log \frac{P(x|y=\text{blue})\ P(y=\text{blue})}{P(x|y=\text{red})\ P(y=\text{red})}$$

$$② \log \frac{P(x|y=\text{blue})\ P(y=\text{blue})}{P(x|y=\text{black})\ P(y=\text{black})}$$

$$③ \log \frac{P(x|y=\text{red})\ P(y=\text{red})}{P(x|y=\text{black})\ P(y=\text{black})}$$

এই ৩ টা ratio এর যেখানে বেশি value পাবো সেখানেই যাবে।

- if $P(x|y=i) = N(\mu_i, \Sigma)$
and $P(x|y=j) = N(\mu_j, \Sigma)$

| i = any class
| N = Normal distribution
| μ_i = mean
| Σ = Co-variance

ধরি, same variance Σ

Target: প্রতিটা x point এর জন্য ratio বের করবো।
যে point টার জন্য যেই class এ ratio সবচেয়ে বেশি পাবো সেই point কে ঐ class এ দিবো।
probability টাকে distribution এ নিয়ে যাবো।

We know,

$$P(x|y=i) = N(\mu_i, \Sigma) = \frac{1}{\sqrt{(2\pi)^m|\Sigma|}} \exp\left(-\frac{1}{2}(x-\mu_i)^T \Sigma^{-1} (x-\mu_i)\right)$$

So, what's the log of this equation:

গুণ ছিল তাই বিয়োগ

$$\log(P(x)|y=1) = -\frac{1}{2}(x-\mu_1)^T \Sigma^{-1}(x-\mu_1) - \frac{m}{2}\log(2\pi) - \frac{1}{2}\log|\Sigma|$$

$$= -\frac{1}{2}\left(x^T\Sigma^{-1}x + \mu_1^T\Sigma^{-1}\mu_1 - 2\mu_1^T\Sigma^{-1}x\right) - \frac{m}{2}\log(2\pi) - \frac{1}{2}\log(|\Sigma|)$$

same equation টাই হবে, শুধু μ_1 এর জায়গায় μ_2 হবে:

$$\log(P(x)|y=2) = \frac{1}{2}\left(x^T\Sigma^{-1}x + \mu_2^T\Sigma^{-1}\mu_2 - 2\mu_2^T\Sigma^{-1}x\right) - \frac{m}{2}\log(2\pi) - \frac{1}{2}\log(|\Sigma|)$$

likelihood পেলাম, এখন prior দেখবো

Let $P(y=1) = \alpha_1$
$P(y=2) = \alpha_2$

So, posterior probability of being class $y=1$

$$P(y=1|x) = \frac{P(x|y=1)\,P(y=1)}{P(x)}$$

To find the ratio of the posterior probability,

$$\log \frac{P(y=1|x)}{P(y=2|x)} = \log \frac{P(x|y=1)\,P(y=1)}{P(x|y=2)\,P(y=2)}$$

$$= \log P(x|y=1) + \overbrace{\log P(y=1)}^{\log \alpha_1} - \log P(x|y=2)$$
$$- \underbrace{\log P(y=2)}_{\log \alpha_2}$$

prior probability কে সমান ধরলে prior ২ টা কাটাকাটি যায়।

$$\therefore \boxed{\log P(x|y=1) - \log P(x|y=2)}$$ এটাই ratio

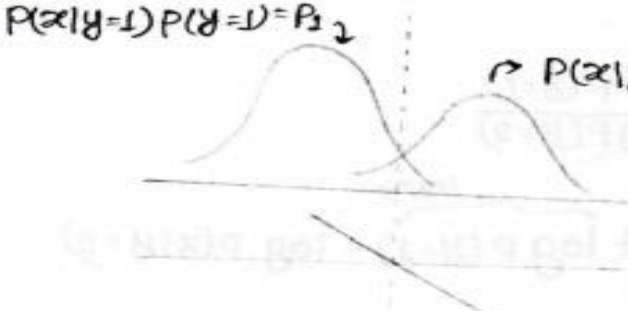
এখন equation টা বসায় দিবো (minus করলে)

$$\therefore \log \frac{P(x|y=1)P(y=1)}{P(x|y=2)P(y=2)} = -\frac{1}{2}\left(2\mu_1^T \Sigma^{-1} x - 2\mu_2^T \Sigma^{-1} x + \mu_1^T \Sigma^{-1} \mu_1 - \mu_2^T \Sigma^{-1} \mu_2\right) + \log \alpha_1 - \log \alpha_2$$

$$= -(\mu_1^T \Sigma^{-1} - \mu_2^T \Sigma^{-1})\, x + w_0$$
$$= W^T x + w_0$$

x এর উপর dependent গুলোকে আলাদা করে, constant গুলো আলাদা করলাম।

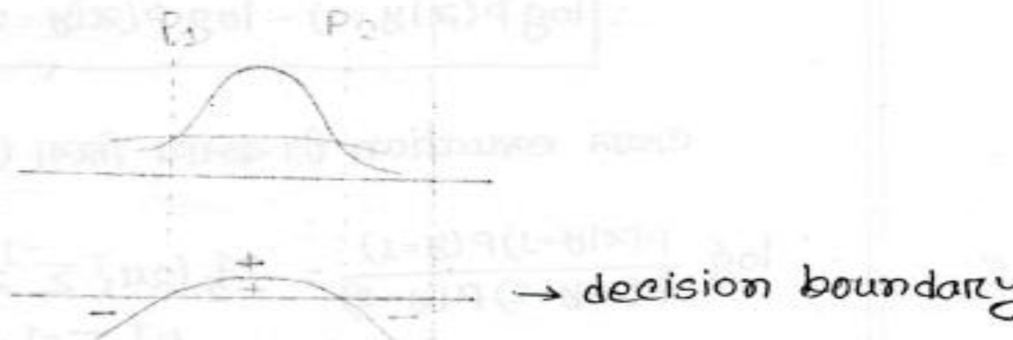
Same variance & same prior assumption করছি।



variance সমান, কিন্তু length সমান না ২টা distribution এর।

$\log \frac{P_1}{P_2}$ হলে decision boundary হবে নিচের টা। (ratio নিলাম)

variance different হলে,



আজকে দেখেছি linear কিভাবে এসেছে।

Next দিন দেখবো quadratic কিভাবে আসে।

$$P(x|y=i) = N(\mu_i, \Sigma) = \frac{1}{\sqrt{(2\pi)^m|\Sigma|}} \exp\left(-\frac{1}{2}(x-\mu_i)^T \Sigma^{-1}(x-\mu_i)\right)$$

$$\ln (P(x)|y=1) = -\frac{1}{2}(x-\mu_1)^T \Sigma^{-1}(x-\mu_1) - \frac{m}{2}\log(2\pi) - \frac{1}{2}\log|\Sigma)$$

$$\boxed{g_i(x) = \ln P(x|w_i) + \ln P(w_i)}$$

$$\underbrace{\ln P(x|w_i)}_{\text{likelihood}} \qquad \underbrace{\ln P(w_i)}_{\text{prior}}$$

$$\log \frac{P(x|y=1)\,P(y=1)}{P(x|y=2)\,P(y=2)} = -\frac{1}{2}\left(x^T\Sigma_1^{-1}x - x^T\Sigma_2^{-1}x - 2\mu_1^T\Sigma_1^{-1}x + 2\mu_2^T\Sigma^{-1}x\right.$$

$$\left. + \mu_1^T\Sigma^T\mu_1 - \mu_2^T\Sigma^{-1}\mu_2\right) - \log|\Sigma_1| + \log|\Sigma_2|$$

$$+ \log q_1 - \log q_2$$

$$= -\frac{1}{2}\left(x^T(\Sigma_1^{-1} - \Sigma_2^{-1})x\right) - (\mu_1^T\Sigma_1^{-1} - \mu_2^T\Sigma_2^{-1})x + w_0$$

$$= x^T W_{12} x - w_{12}^T x + w_0$$

$$= \text{quadratic equation.}$$

এক্ষেত্রে variance same না, q_1, q_2 আছে।

তাইলে bayesian থেকে quadratic পেতে পারি।

Bayesian decision theory এই slide এর 31 slide পর্যন্ত পড়ানো হয়েছে।

Example - Case III

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

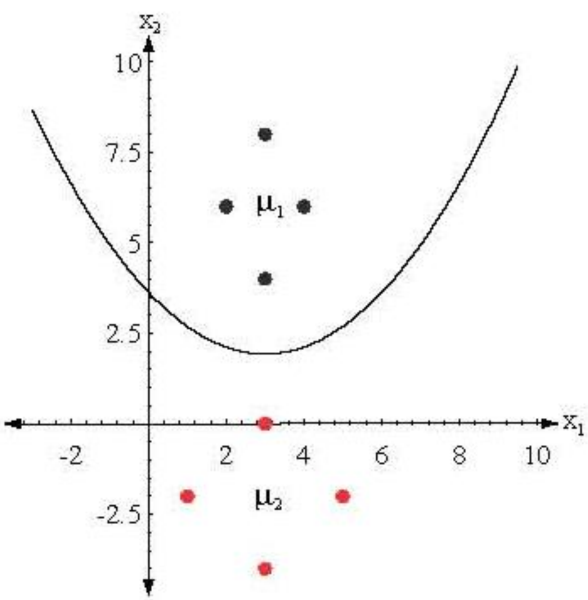
decision boundary: $x_2 = 3.514 - 1.125x_1 + 0.1875x_1^2.$

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

boundary does

not pass through

midpoint of μ_1,μ_2



Error Bounds

- Exact error calculations could be difficult – easier to estimate error bounds!

$$P(error) = \int P(error, \mathbf{x})\, d\mathbf{x} = \int P(error/\mathbf{x}) p(\mathbf{x})\, d\mathbf{x}$$

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

or

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

- Using the inequality:

$$min[a, b] \leq a^{\beta} b^{1-\beta}, \qquad a, b \geq 0, 0 \leq \beta \leq 1$$

$$P(error) = \int min[p(\mathbf{x}/\omega_1)P(\omega_1),\ p(\mathbf{x}/\omega_2)P(\omega_2)]d\mathbf{x} \leq$$

$$P^{\beta}(\omega_1)P^{1-\beta}(\omega_2) \int p^{\beta}(\mathbf{x}/\omega_1)\ p^{1-\beta}(\mathbf{x}/\omega_2)d\mathbf{x}$$

Error Bounds (cont'd)

- If the class conditional distributions are Gaussian, then

$$\int p^{\beta}(\mathbf{x}/\omega_1)\ p^{1-\beta}(\mathbf{x}/\omega_2)d\mathbf{x} = e^{-\kappa(\beta)}$$

where:

$$k(\beta) = \frac{\beta(1-\beta)}{2}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^t\left[(1-\beta)\boldsymbol{\Sigma}_1 + \beta\boldsymbol{\Sigma}_2\right]^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) + \frac{1}{2}\ln\frac{\left|(1-\beta)\boldsymbol{\Sigma}_1 + \beta\boldsymbol{\Sigma}_2\right|}{|\boldsymbol{\Sigma}_1|^{1-\beta}|\boldsymbol{\Sigma}_2|^{\beta}}.$$

Error Bounds (cont'd)

- The *Chernoff* bound is obtained by minimizing $e^{-\kappa(\beta)}$
  - This is a 1-D optimization problem, regardless to the dimensionality of the class conditional densities.

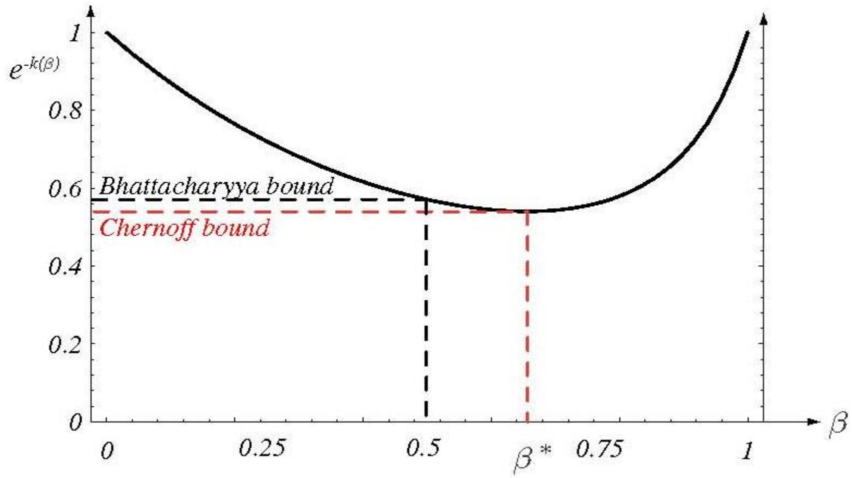


FIGURE 2.18. The Chernoff error bound is never looser than the Bhattacharyya bound. For this example, the Chernoff bound happens to be at $\beta^* = 0.66$, and is slightly tighter than the Bhattacharyya bound ($\beta = 0.5$). From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

Error Bounds (cont'd)

- The *Bhattacharyya* bound is obtained by setting $\beta=0.5$
  - Easier to compute than Chernoff error but looser.

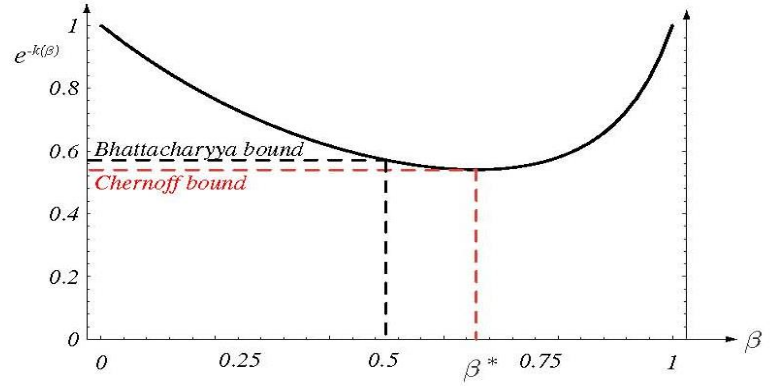
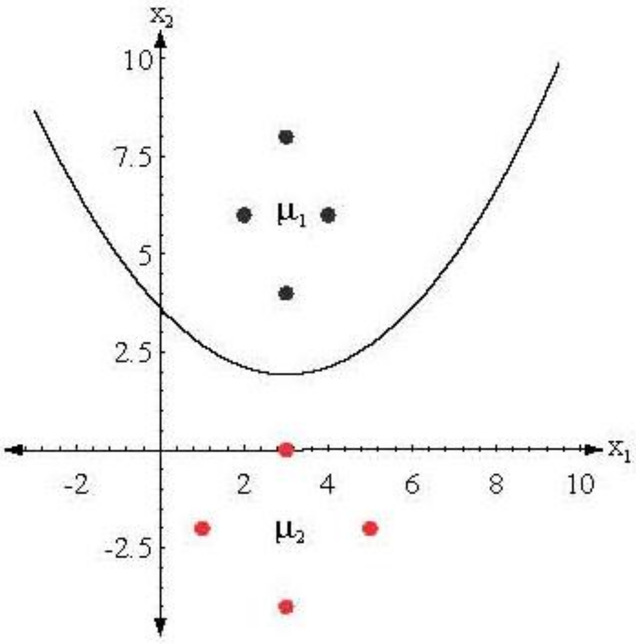


FIGURE 2.18. The Chernoff error bound is never looser than the Bhattacharyya bound. For this example, the Chernoff bound happens to be at $\beta^* = 0.66$, and is slightly tighter than the Bhattacharyya bound ($\beta = 0.5$). From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

- **Note**: the Chernoff and Bhattacharyya bounds will not be good bounds if the densities are not Gaussian.

Example (cont'd)

$$k(\beta) = \frac{\beta(1-\beta)}{2}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^t \left[(1-\beta)\boldsymbol{\Sigma}_1 + \beta\boldsymbol{\Sigma}_2\right]^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) + \frac{1}{2}\ln\frac{\left|(1-\beta)\boldsymbol{\Sigma}_1 + \beta\boldsymbol{\Sigma}_2\right|}{|\boldsymbol{\Sigma}_1|^{1-\beta}|\boldsymbol{\Sigma}_2|^{\beta}}.$$



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

$$\boldsymbol{\mu}_2 = \begin{bmatrix} 3 \\ -2 \end{bmatrix}; \quad \boldsymbol{\Sigma}_2 = \begin{pmatrix} 2 & 0 \\ 0 & 2 \end{pmatrix}.$$

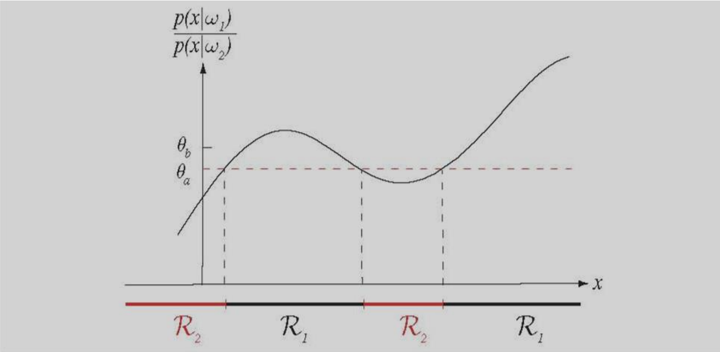
Bhattacharyya error:

k(0.5)=4.06

$$P(error) \leq 0.0087$$

Receiver Operating Characteristic (ROC) Curve

- Every classifier typically employs some kind of a threshold.



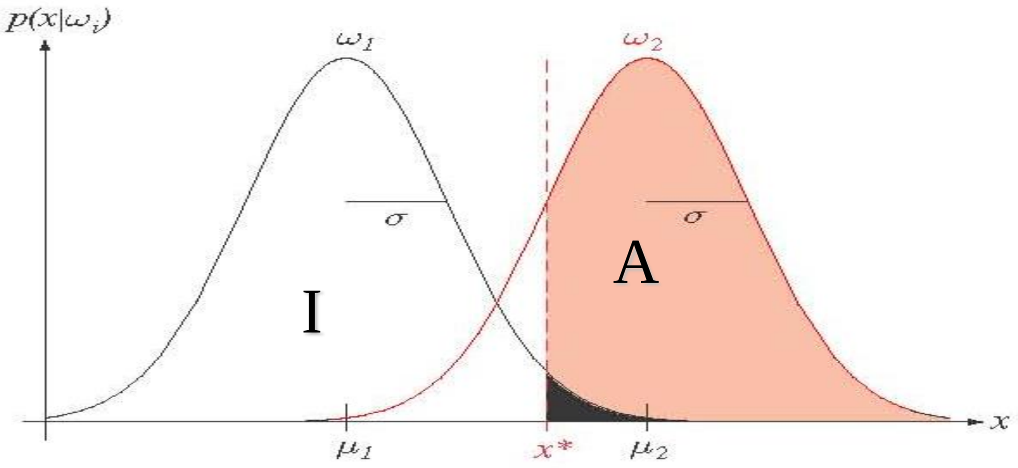
$$\theta_a = P(\omega_2)/P(\omega_1)$$

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

- Changing the threshold can affect the performance of the classifier.
- ROC curves allow us to evaluate/compare the performance of a classifier using different thresholds.

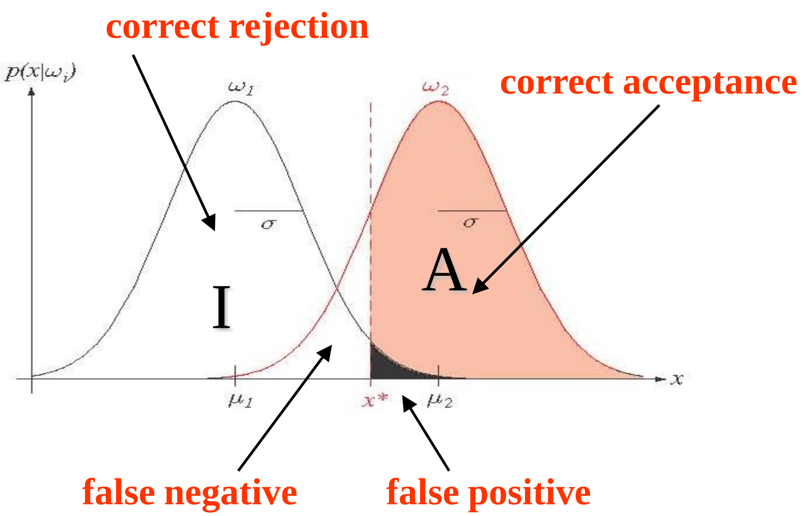
Example: Person Authentication

- Authenticate a person using biometrics (e.g., fingerprints).
- There are two possible distributions (i.e., classes):
  - *Authentic* (A) and *Impostor* (I)



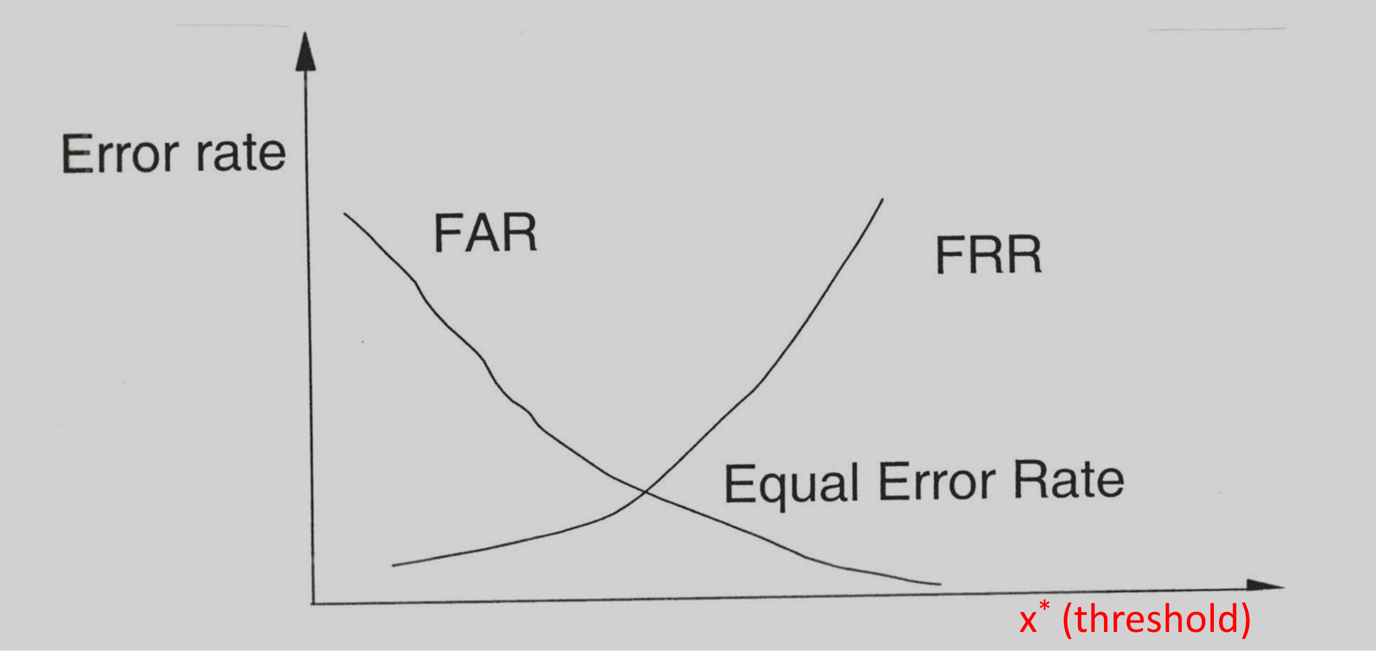
Example: Person Authentication (cont'd)

- Possible decisions:
  - (1) **correct acceptance** (true positive):
    - X belongs to A, and we decide A
  - (2) **incorrect acceptance** (false positive):
    - X belongs to I, and we decide A
  - (3) **correct rejection** (true negative):
    - X belongs to I, and we decide I
  - (4) **incorrect rejection** (false negative):
    - X belongs to A, and we decide I



Error vs Threshold

ROC Curve

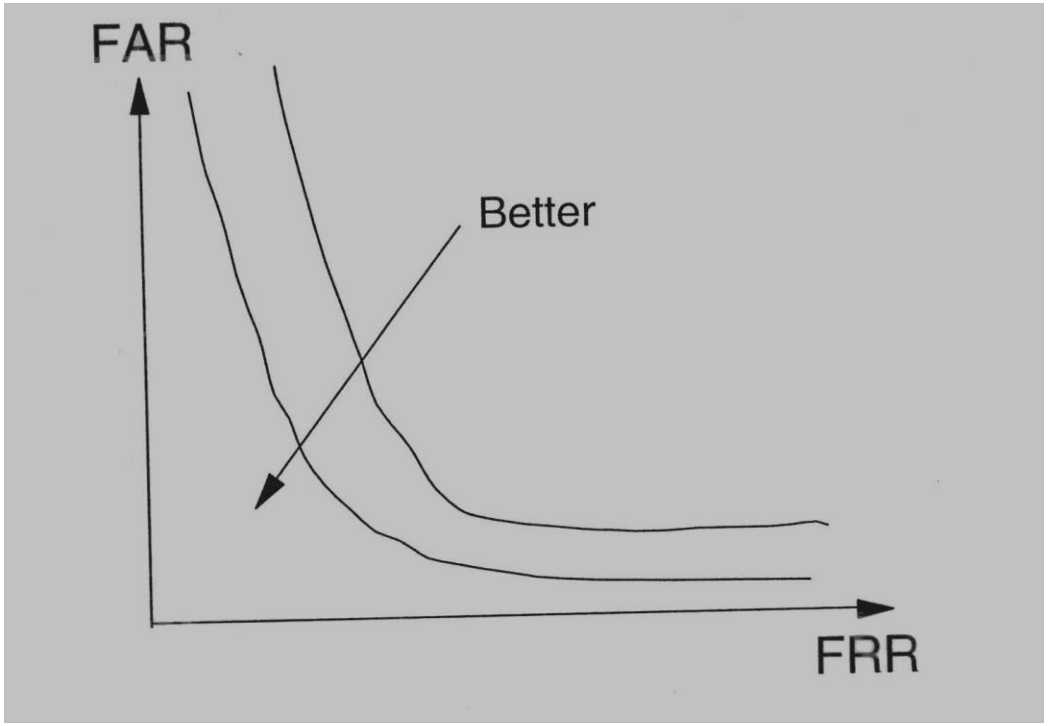


FAR: False Accept Rate (False Positive)

FRR: False Reject Rate (False Negative)

False Negatives vs False Positives

ROC Curve



FAR: False Accept Rate (False Positive)

FRR: False Reject Rate (False Negative)

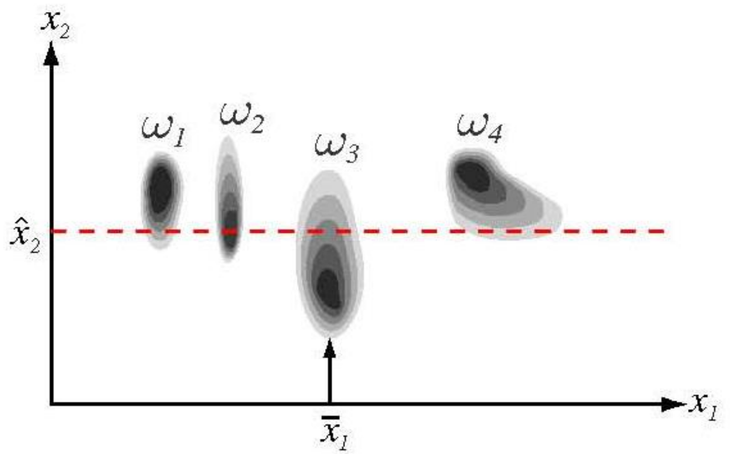
Bayes Decision Theory: Case of Discrete Features

- Replace $\int p(\mathbf{x}/\omega_j)d\mathbf{x}$ with $\sum_{\mathbf{x}} P(\mathbf{x}/\omega_j)$
- See section 2.9

Missing Features

- Suppose $\mathbf{x}=(x_1,x_2)$ is a test vector where x_1 is missing and $x_2 = \hat{x}_2$ - how would we classify it?
  - If we set x_1 equal to the average value, we will classify $\mathbf{x}$ as ω_3
  - But $p(\hat{x}_2 / \omega_2)$ is larger; should we classify $\mathbf{x}$ as ω_2 ?

Example:



Missing Features (cont'd)

- Suppose $\mathbf{x}=[\mathbf{x}_g, \mathbf{x}_b]$ ($\mathbf{x}_g$: good features, $\mathbf{x}_b$: bad features)
- Derive the Bayes rule using the good features:

$$P(\omega_i/\mathbf{x}_g) = \frac{\mathrm{p}(\omega_i, \mathbf{x}_g)}{p(\mathbf{x}_g)} = \frac{\int \mathrm{p}(\omega_i, \mathbf{x}_g, \mathbf{x}_b)d\mathbf{x}_b}{p(\mathbf{x}_g)} =$$

$$\frac{\int P(\omega_i/\mathbf{x}_g, \mathbf{x}_b)p(\mathbf{x}_g, \mathbf{x}_b)d\mathbf{x}_b}{p(\mathbf{x}_g)} = \frac{\int P(\omega_i/\mathbf{x}_g, \mathbf{x}_b)p(\mathbf{x})d\mathbf{x}_b}{\int p(\mathbf{x})d\mathbf{x}_b}$$

marginalize posterior probability over bad features.

Decide ω_1 if $P(\omega_1/\mathbf{x}_g) > P(\omega_2/\mathbf{x}_g)$; otherwise decide ω_2

Compound Bayesian Decision Theory

- Sequential decision

  (1) Decide as each pattern (e.g., fish) emerges.

- Compound decision

  (1) Wait for *n* patterns (e.g., fish) to emerge.

  (2) Make all *n* decisions jointly.

  – Could improve performance when consecutive states of nature are **not** statistically independent.

Compound Bayesian Decision Theory (cont'd)

- Suppose $\mathbf{X}=(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n)$ are **n** observed vectors.
- Suppose $\boldsymbol{\Omega}=(\omega(1), \omega(2), \ldots, \omega(n))$ denotes the **n** states of nature.
  - $\omega(i)$ can take one of c values $\omega_1, \omega_2, \ldots, \omega_c$ (i.e., c categories)
- Suppose $P(\boldsymbol{\Omega})$ is the prior probability of the **n** states of nature.

Compound Bayesian Decision Theory (cont'd)

- Suppose p($\mathbf{X}/\mathbf{\Omega}$) is the conditional probability function for $\mathbf{X}$

$$P(\mathbf{\Omega}/\mathbf{X}) = \frac{p(\mathbf{X}/\mathbf{\Omega})P(\mathbf{\Omega})}{p(\mathbf{X})}$$

- The assumption $p(\mathbf{X}/\mathbf{\Omega}) = \prod_{i=1}^{c} \mathrm{p}(\boldsymbol{x}_i/\omega(\mathrm{i}))$ might be acceptable.
- The assumption $\mathrm{P}(\mathbf{\Omega}) = \prod_{i=1}^{c} \mathrm{P}(\omega(\mathrm{i}))$ is not acceptable!

i.e., consecutive states of nature may **not** be statistically independent!