EEL-5840/EEL-4930 Elements of Machine Intelligence

Bayesian Decision Theory

Bayesian Decision Theory

- Fundamental statistical approach to problem classification.
- Quantifies the tradeoffs between various *classification decisions* using probabilities and the *costs* associated with such decisions.
  - Each action is associated with a cost or risk.
  - The simplest risk is the classification error.
  - Design classifiers to recommend actions that minimize some total expected risk.

Terminology

(using sea bass – salmon classification example)

- State of nature ω *(random variable):*
  - ω_1 for sea bass, ω_2 for salmon.
- Probabilities $P(\omega_1)$ and $P(\omega_2)$ *(priors)*
  - prior knowledge of how likely is to get a sea bass or a salmon
- Probability density function *p(x) (evidence):*
  - how frequently we will measure a pattern with feature value *x* (e.g., *x* is a lightness measurement)

Note: if *x* and *y* are different measurements, *p(x)* and *p(y)* correspond to different *pdfs*: $p_X(x)$ and $p_Y(y)$

Terminology (cont'd)

(using sea bass – salmon classification example)

- Conditional probability density $p(x/\omega_j)$ (*likelihood*):
  - 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

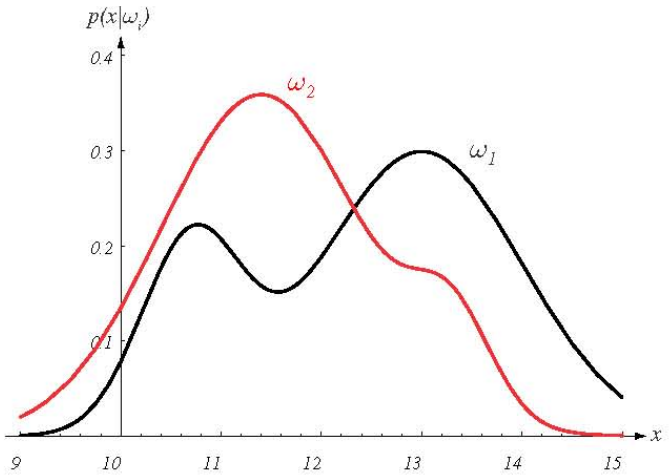


FIGURE 2.1. Hypothetical class-conditional probability density functions show the probability density of measuring a particular feature value *x* given the pattern is in category ω_i. If *x* represents the lightness of a fish, the two curves might describe the difference in lightness of populations of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

(using sea bass – salmon classification example)

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

Note: we will be using an uppercase *P(.)* to denote a probability mass function *(pmf)* and a lowercase *p(.)* to denote a probability density function *(pdf)*.

Decision Rule Using Priors Only

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

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

- Favours the most likely class … (optimum if no other info is available).
- This rule would be making the same decision all the times!
- Makes sense to use for judging just one fish …

Decision Rule Using Conditional pdf

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

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

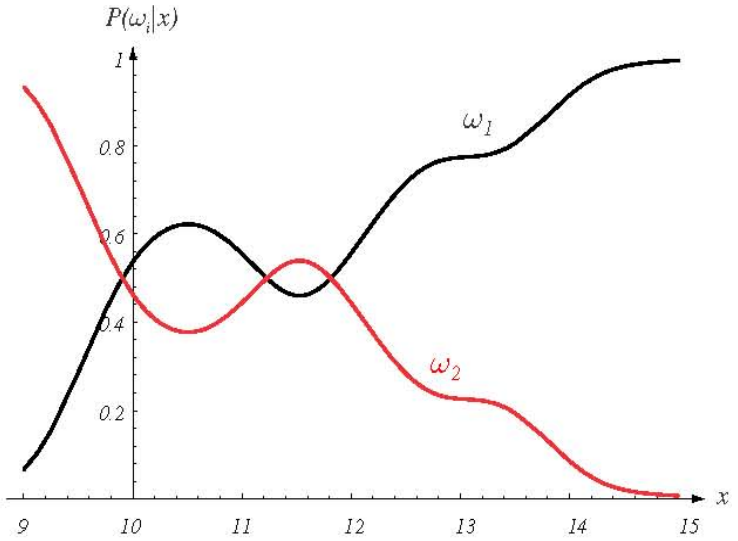
where $p(x) = \sum_{j=1}^{2} p(x / \omega_j)P(\omega_j)$ (scale factor – sum of probs = 1)

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)



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

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:

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

- The average probability error is given by:

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

- The Bayes rule is *optimum*, that is, it minimizes the average probability error since:

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

Where do Probabilities Come From?

- The Bayesian rule is optimal if the *pmf* or *pdf* is known.
- There are two competitive answers to the above 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 as well as experiments.

Example

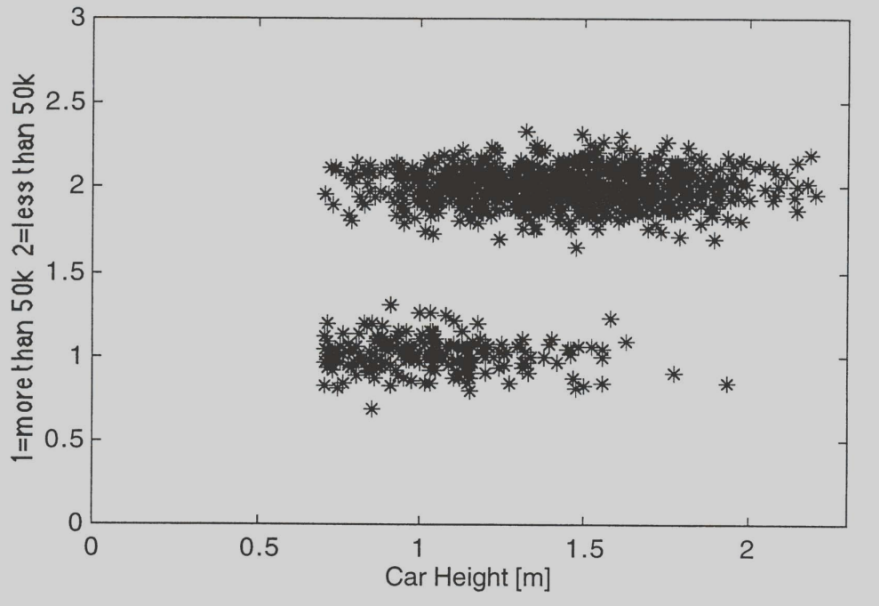
- Classify cars on UF's campus whether they are more or less than $50K:
  - C1: price > $50K
  - C2: price < $50K
  - Feature x: height of car
- From Bayes' rule, we know how to compute the posterior probabilities:

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

- Need to compute $p(x/C_1)$, $p(x/C_2)$, $P(C_1)$, $P(C_2)$

Example (cont'd)

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

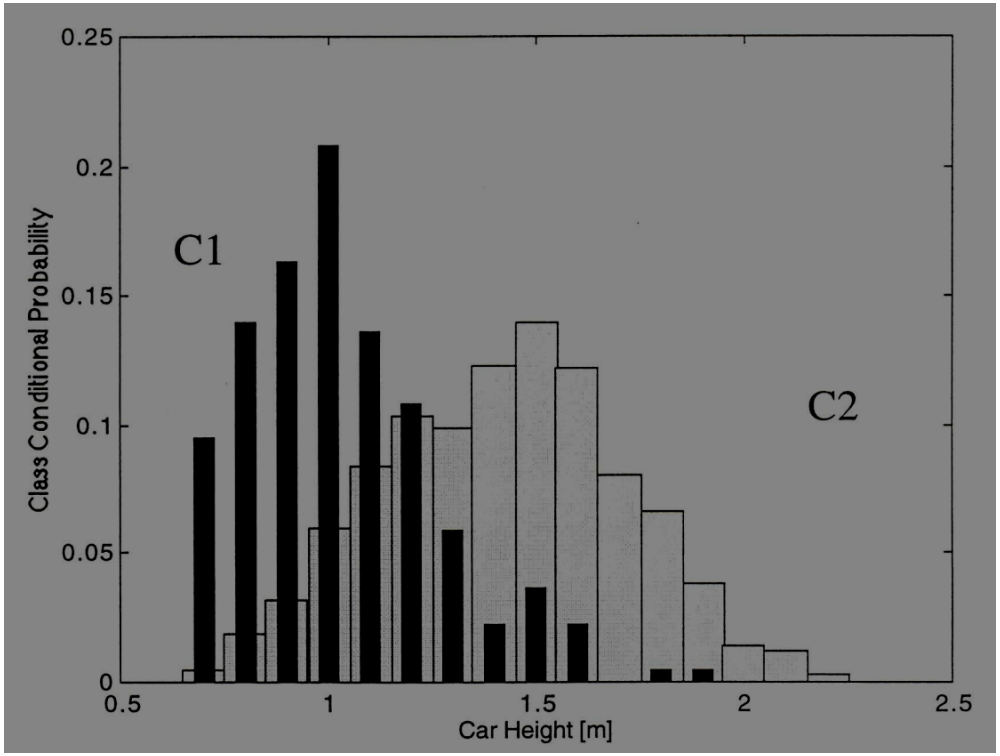


$$P(C_1) = \frac{221}{1209} = 0.183$$

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

Example (cont'd)

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

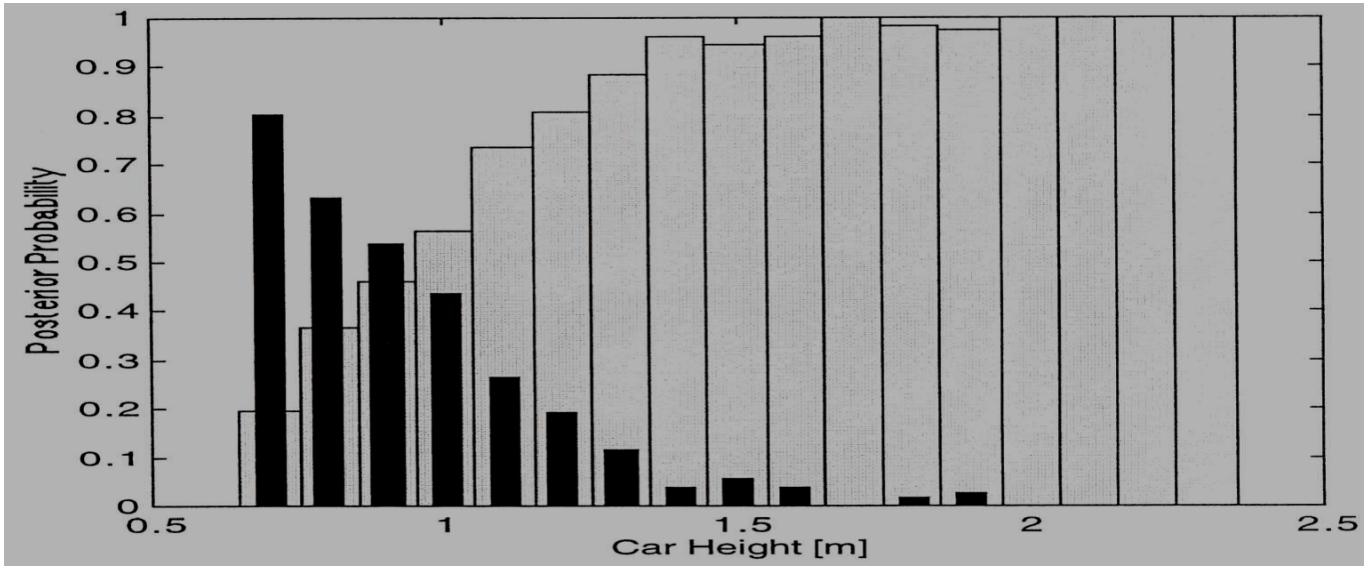


Example (cont'd)

- Calculate the posterior probability for each bin:

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

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



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).
- Introduce a more general error function
  - *loss* function (i.e., associate “costs” with actions)

Terminology

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

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

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

Expected Loss (Conditional Risk)

- **Expected loss** (or **conditional risk**) with taking action α_i :

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

- The expected loss can be minimized by selecting the action that minimizes the conditional risk.

Overall Risk

- **Overall risk** *R*

conditional risk

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

where $\alpha(\mathbf{x})$ determines which action $\alpha_1, \alpha_2, \ldots, \alpha_l$ to take for every **x** (i.e., $\alpha(\mathbf{x})$ is a decision rule).

- To minimize *R*, find a decision rule $\alpha(\mathbf{x})$ that chooses the action with the minimum conditional risk $R(a_i/\mathbf{x})$ for every **x**.

- This rule yields *optimal* performance…

Bayes Decision Rule

- The Bayes decision rule minimizes *R* by:
  - Computing $R(\alpha_i / \mathbf{x})$ for every α_i given an **x**
  - Choosing the action α_i with the minimum $R(\alpha_i / \mathbf{x})$
- *Bayes risk* (i.e., resulting minimum) is the best performance that can be achieved.

$$R^* = \min R$$

Example: Two-category classification

- Two possible actions
  - α_1 corresponds to deciding ω_1
  - α_2 corresponds to deciding ω_2

- Notation:

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

- The conditional risks are:

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

where λ_{ij} is equivalent to $\lambda(a_i/\omega_i)$

Example: Two-category classification

- 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)

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

(what is the sign of $(\lambda_{21} - \lambda_{11})$ and $(\lambda_{12} - \lambda_{22})$?)

Special Case: Zero-One Loss Function

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

- What is the overall risk in this case?
  (*answer*: average probability error)

Example

- θ_a was determined assuming zero-one loss function
- θ_b was determined assuming $\lambda_{12} > \lambda_{21}$

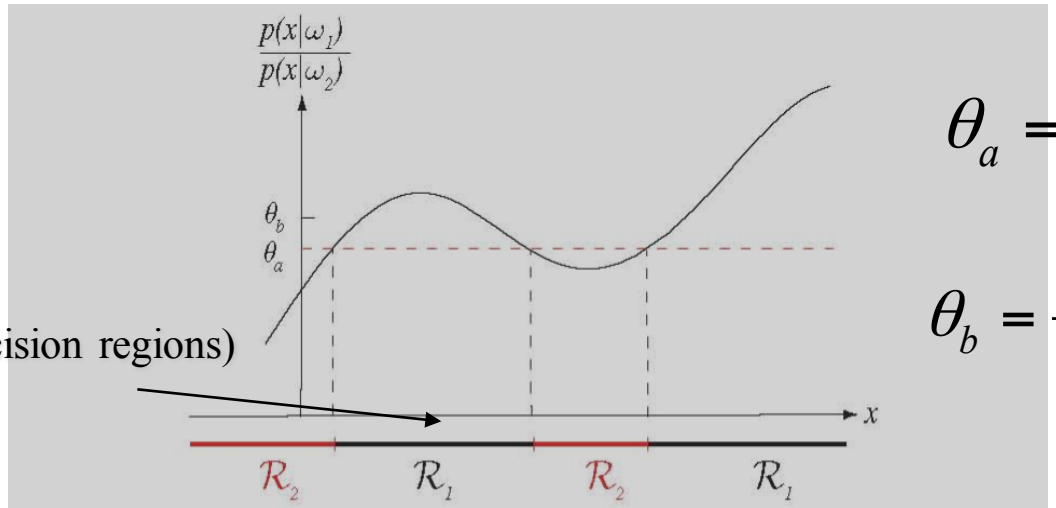


FIGURE 2.3. The likelihood ratio $p(x|\omega_1)/p(x|\omega_2)$ for the distributions shown in Fig. 2.1. If we employ a zero-one or classification loss, our decision boundaries are determined by the threshold θ_a. If our loss function penalizes miscategorizing ω_2 as ω_1 patterns more than the converse, we get the larger threshold θ_b, and hence $\mathcal{R}_1$ becomes smaller. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

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

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

Discriminant Functions

- Functional structure of a general statistical classifier

Assign x to ω_i if: $g_i(\mathbf{x}) > g_j(\mathbf{x})$ for all $j \neq i$

(*discriminant functions*)

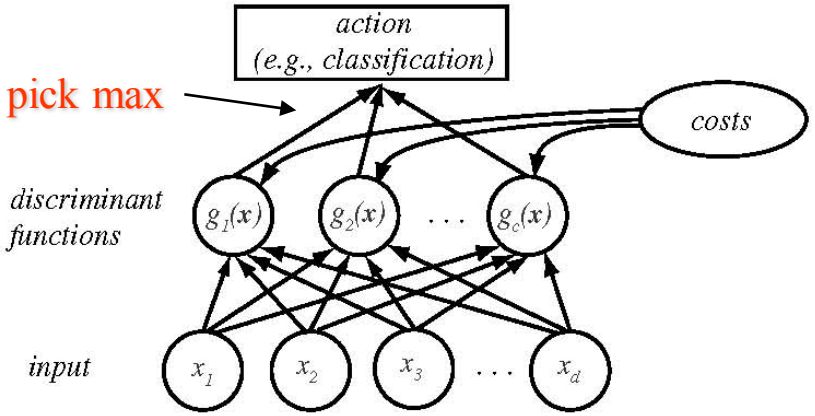


FIGURE 2.5. The functional structure of a general statistical pattern classifier which includes d inputs and c discriminant functions $g_i(\mathbf{x})$. A subsequent step determines which of the discriminant values is the maximum, and categorizes the input pattern accordingly. The arrows show the direction of the flow of information, though frequently the arrows are omitted when the direction of flow is self-evident. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

Discriminants for Bayes Classifier

- Using risks:

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

- Using zero-one loss function (i.e., min error rate):

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

- 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.

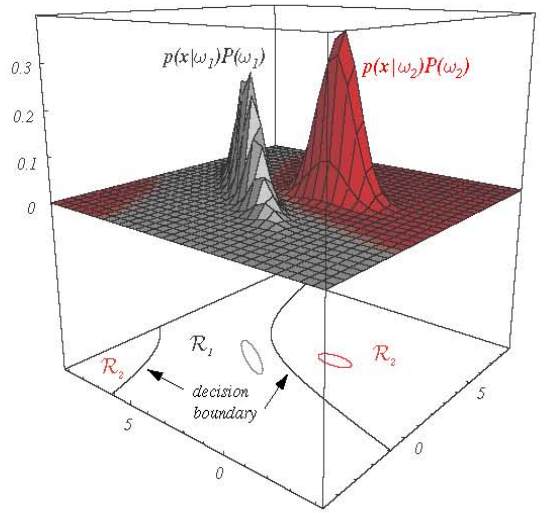
$$g_i(\mathbf{x}) = \frac{p(\mathbf{x}/\omega_j)P(\omega_j)}{p(\mathbf{x})}$$

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

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

Decision Regions and Boundaries

- Decision rules divide the feature space in *decision regions* $R_1, R_2, \ldots, R_c$
- The boundaries of the decision regions are the *decision boundaries*.



$g_1(\mathbf{x})=g_2(\mathbf{x})$ at the decision boundaries

FIGURE 2.6. In this two-dimensional two-category classifier, the probability densities are Gaussian, the decision boundary consists of two hyperbolas, and thus the decision region $\mathcal{R}_2$ is not simply connected. The ellipses mark where the density is $1/e$ times that at the peak of the distribution. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

Case of two categories

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

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

$$\textbf{Decide } \omega_1 \text{ if } g(\mathbf{x}) > 0; \text{ otherwise decide } \omega_2$$

- Examples of *dichotomizers*:

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

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

Discriminant Function for Multivariate Gaussian

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

- Assume the following discriminant function:

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

- 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

- Assumption: $\Sigma_i=\sigma^2$
  - Features are statistically independent
  - Each feature has the same variance

favors the a-priori more likely category

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

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

threshold or bias

- 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)

- Comments about this hyperplane:
  - 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 more likely mean.
  - If σ is very small, the position of the boundary is insensitive to $P(\omega_i)$ and $P(\omega_j)$

Multivariate Gaussian Density: Case I (cont'd)

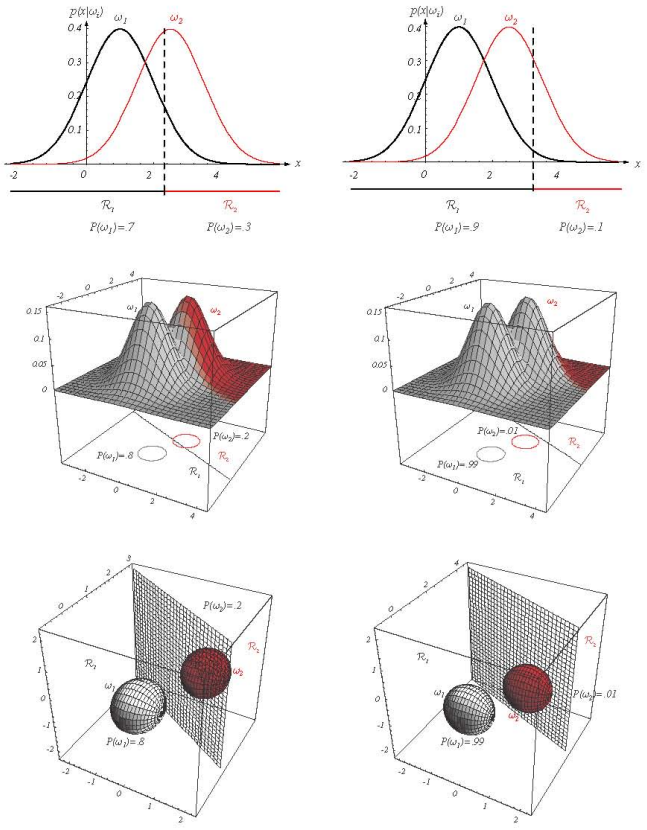


FIGURE 2.11. As the priors are changed, the decision boundary shifts; for sufficiently disparate priors the boundary will not lie between the means of these one-, two- and three-dimensional spherical Gaussian distributions. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

Multivariate Gaussian Density: Case I (cont'd)

- Minimum distance classifier
  - When $P(\omega_i)$ is the same for each of the c classes

$$g_i(\mathbf{x}) = -\| \mathbf{x} - \boldsymbol{\mu}_i \|^2$$

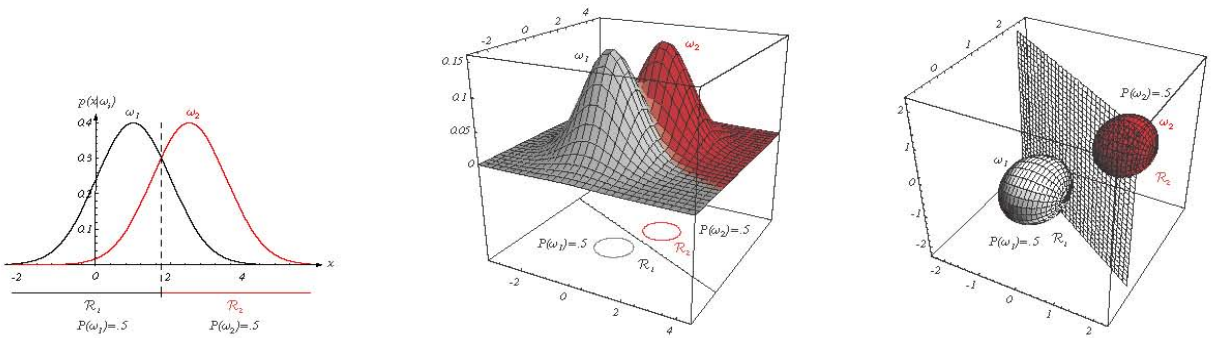


FIGURE 2.10. If the covariance matrices for two distributions are equal and proportional to the identity matrix, then the distributions are spherical in d dimensions, and the boundary is a generalized hyperplane of $d-1$ dimensions, perpendicular to the line separating the means. In these one-, two-,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

Multivariate Gaussian Density: Case II

- Assumption: $\Sigma_i = \Sigma$

  - 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:

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

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

- Comments about this hyperplane:
  - 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 more likely mean.

Multivariate Gaussian Density: Case II (cont'd)

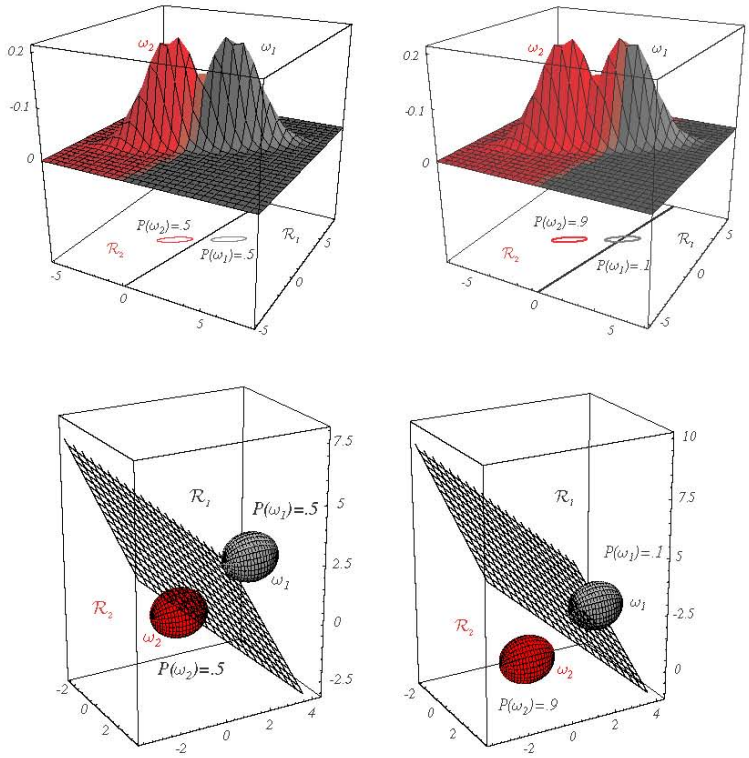


FIGURE 2.12. Probability densities (indicated by the surfaces in two dimensions and ellipsoidal surfaces in three dimensions) and decision regions for equal but asymmetric Gaussian distributions. The decision hyperplanes need not be perpendicular to the line connecting the means. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

Multivariate Gaussian Density: Case II (cont'd)

- Mahalanobis distance classifier
  - When $P(\omega_i)$ is the same for each of the c classes

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

Multivariate Gaussian Density: Case III

- Assumption: Σ_i= arbitrary

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

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

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

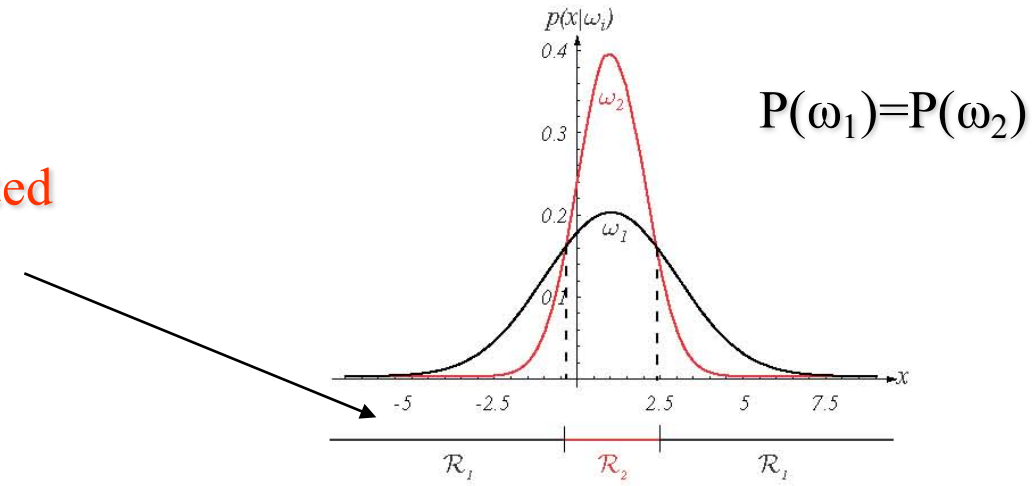
where $\mathbf{W}_i = -\frac{1}{2}\Sigma_i^{-1}$, $\mathbf{w}_i = \Sigma_i^{-1}\mu_i$, and $w_{i0} = -\frac{1}{2}\mu_i^t\Sigma^{-1}\mu_i - \frac{1}{2} ln|\Sigma_i| + \ln P(\omega_i)$

- 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.

Multivariate Gaussian Density: Case III (cont'd)

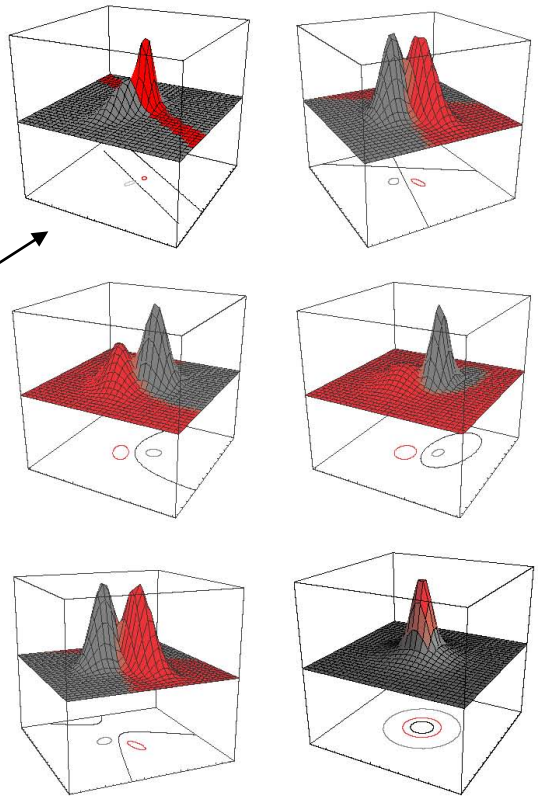
disconnected decision regions



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

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.

Multivariate Gaussian Density: Case III (cont'd)



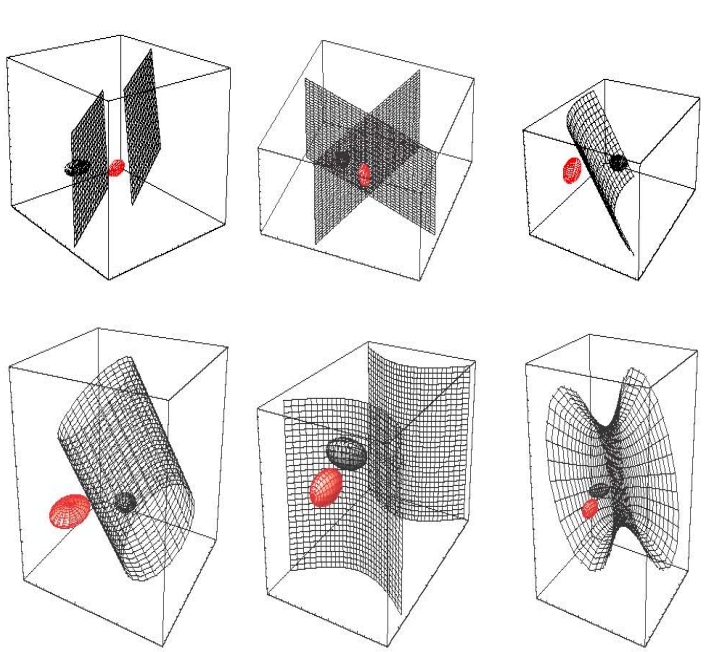
disconnected decision regions

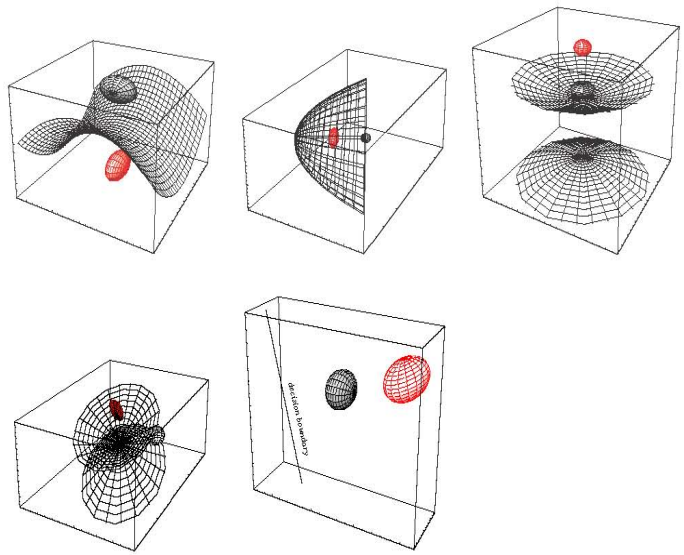
non-linear decision boundaries

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.

Multivariate Gaussian Density: Case III (cont'd)

- More examples (Σ arbitrary)





JRE 2.15. Arbitrary three-dimensional Gaussian distributions yield Bayes decision boundaries that are dimensional hyperquadrics. There are even degenerate cases in which the decision boundary is a line. r: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John ey & Sons, Inc.

Multivariate Gaussian Density: Case III (cont'd)

- A four category example

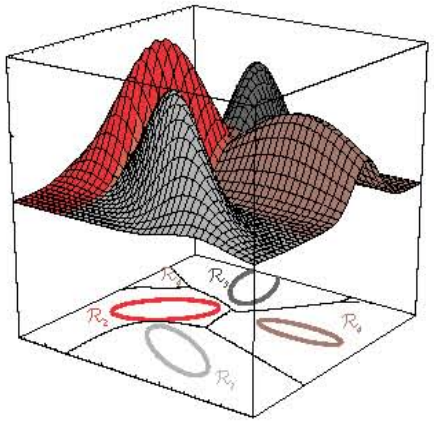


FIGURE 2.16. The decision regions for four normal distributions. Even with such a low number of categories, the shapes of the boundary regions can be rather complex. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

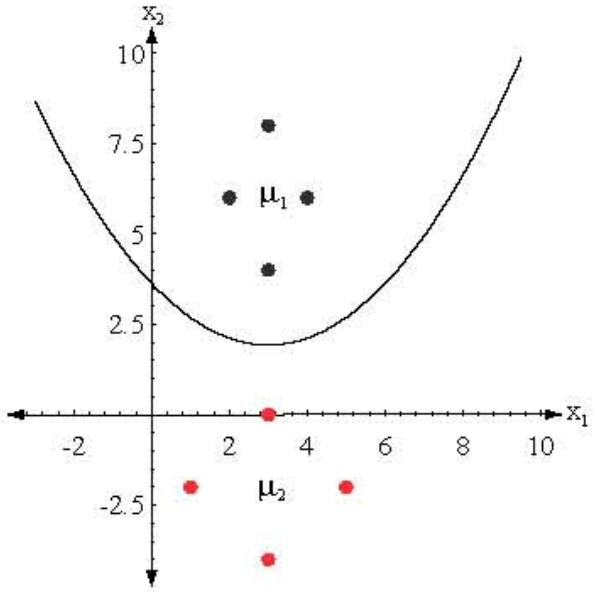
Example - Case III

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

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

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

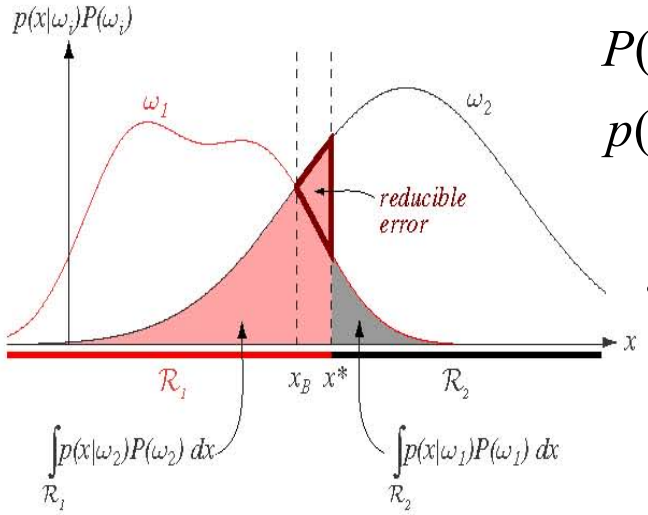
boundary does not pass through midpoint of μ_1,μ_2



Error Probabilities and Integrals

- Case of two categories

Bayes rule minimizes:



$$P(error) = p(\mathbf{x}\in R_2, \omega_1) + p(\mathbf{x}\in R_1, \omega_2)$$

$$p(\mathbf{x}\in R_2 / \omega_1)P(\omega_1) + p(\mathbf{x}\in R_1 / \omega_2)P(\omega_2) =$$

$$\int_{R_2} p(\mathbf{x}/\omega_1)P(\omega_1)d\mathbf{x} + \int_{R_1} p(\mathbf{x}/\omega_2)P(\omega_2)d\mathbf{x} =$$

$$\int_{x^*}^{\infty} p(\mathbf{x}/\omega_1)P(\omega_1)d\mathbf{x} + \int_{-\infty}^{x^*} p(\mathbf{x}/\omega_2)P(\omega_2)d\mathbf{x}$$

FIGURE 2.17. Components of the probability of error for equal priors and (nonoptimal) decision point x^*. The pink area corresponds to the probability of errors for deciding ω_1 when the state of nature is in fact ω_2; the gray area represents the converse, as given in Eq. 70. If the decision boundary is instead at the point of equal posterior probabilities, x_B, then this reducible error is eliminated and the total shaded area is the minimum possible; this is the Bayes decision and gives the Bayes error rate. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

Optimum

using Bays rule: $x^* = x_B$

Error Probabilities and Integrals (cont'd)

- Case of multiple categories
  - Simpler to compute the probability of being correct

$$P(correct) = \sum_{i=1}^{c} p(\mathbf{x} \in R_i, \omega_i) =$$

$$\sum_{i=1}^{c} p(\mathbf{x} \in R_i / \omega_i) P(\omega_i) = \sum_{i=1}^{c} \int_{R_i} p(\mathbf{x}/\omega_i) P(\omega_i) d\mathbf{x}$$

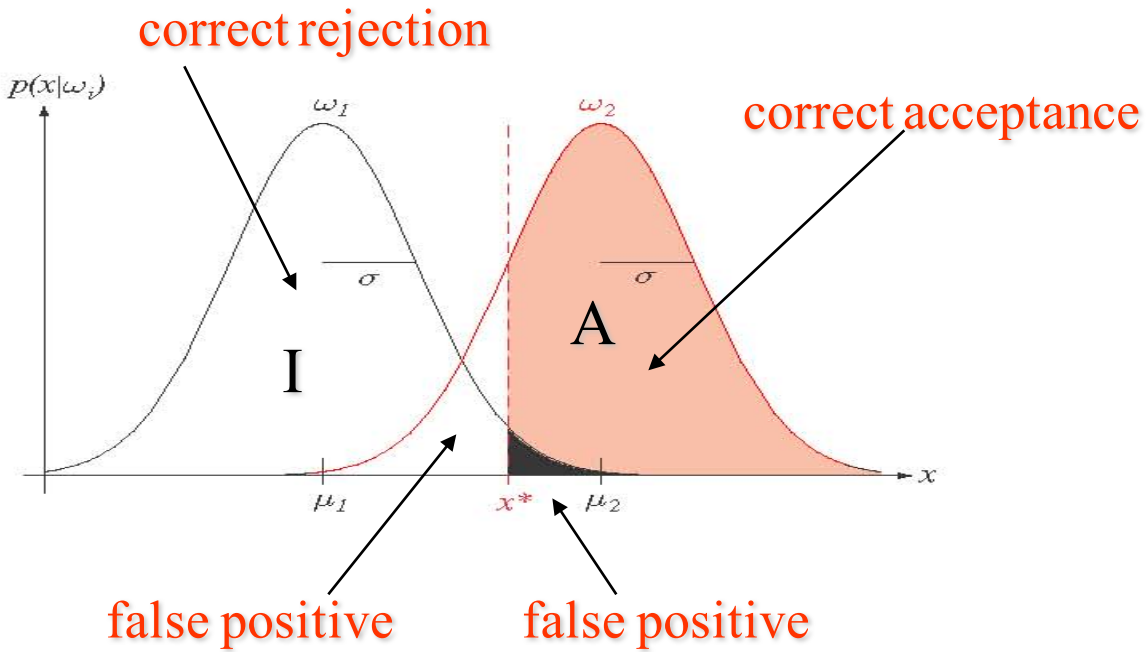
Bayes rule maximizes *P(correct)*

Receiver Operating Characteristic (ROC) Curve

- Every classifier employs some kind of a threshold value.
- Changing the threshold affects the performance of the system.
- ROC curves can help us distinguish between *discriminability* and *decision* bias (i.e., choice of threshold)

Example: Person Authentication

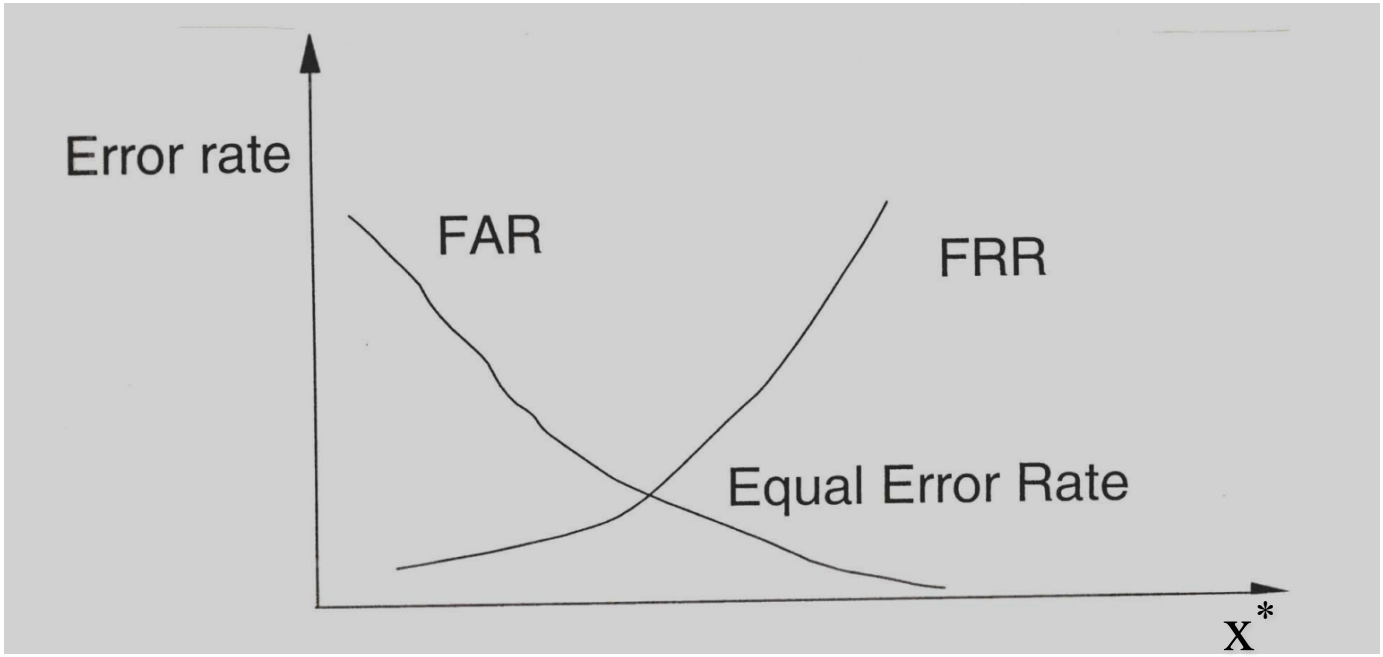
- Authenticate a person using biometrics (e.g., face image).
- There are two possible distributions:
  - *authentic* (A) and *impostor* (I)



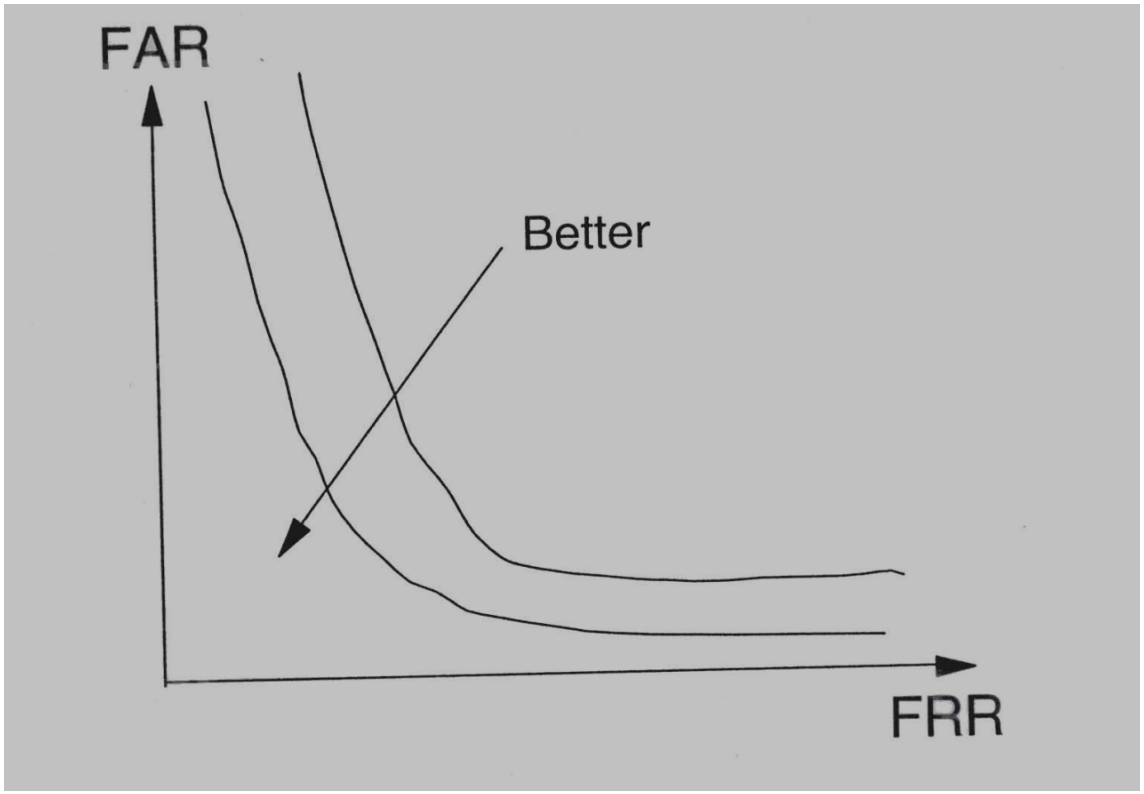
Example: Person Authentication (cont'd)

- Possible cases
  - (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



False Negatives vs Positives



Bayes Decision Theory
Case of Discrete Features

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

Compound Bayesian Decision Theory

- Sequential compound decision
  - Decide as each fish emerges.
- Compound decision
  - Wait for *n* fish to emerge.
  - Make all *n* decisions jointly.

Bayes Rule for Compound Decisions

Decisions are of the form: $\Omega = (\omega(1), \omega(2), \ldots, \omega(n))^t$

Inputs are of the form: $X = (\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n)$

c^n possible vectors

The posteriori probability of Ω is:

c^n possible values

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

Computing $p(X/\Omega)$ and $P(\Omega)$ can be very time consuming.

The assumption $p(X/\Omega) = \prod_{i=1}^{n} p(\mathbf{x}_i/\omega(i))$ might be acceptable

The assumption $p(\Omega) = \prod_{i=1}^{c} p(\omega(i))$ is NOT acceptable in practice.

(consecutive states ω_i are not independent – can lead to better performance)

Questions

Credits

- Prof. George Bebis (UNR)