

# Bayesian Decision Theory

Selim Aksoy

Department of Computer Engineering  
Bilkent University  
saksoy@cs.bilkent.edu.tr

CS 551, Spring 2011



# Bayesian Decision Theory

- ▶ Bayesian Decision Theory is a fundamental statistical approach that quantifies the tradeoffs between various decisions using probabilities and costs that accompany such decisions.
- ▶ First, we will assume that all probabilities are known.
- ▶ Then, we will study the cases where the probabilistic structure is not completely known.



# Fish Sorting Example Revisited

- ▶ State of nature is a random variable.
- ▶ Define  $w$  as the type of fish we observe (state of nature, *class*) where
  - ▶  $w = w_1$  for sea bass,
  - ▶  $w = w_2$  for salmon.
  - ▶  $P(w_1)$  is the *a priori probability* that the next fish is a sea bass.
  - ▶  $P(w_2)$  is the a priori probability that the next fish is a salmon.



# Prior Probabilities

- ▶ Prior probabilities reflect our knowledge of how likely each type of fish will appear before we actually see it.
- ▶ How can we choose  $P(w_1)$  and  $P(w_2)$ ?
  - ▶ Set  $P(w_1) = P(w_2)$  if they are equiprobable (*uniform priors*).
  - ▶ May use different values depending on the fishing area, time of the year, etc.
- ▶ Assume there are no other types of fish

$$P(w_1) + P(w_2) = 1$$

(exclusivity and exhaustivity).



# Making a Decision

- ▶ How can we make a decision with only the prior information?

$$\text{Decide } \begin{cases} w_1 & \text{if } P(w_1) > P(w_2) \\ w_2 & \text{otherwise} \end{cases}$$

- ▶ What is the *probability of error* for this decision?

$$P(\text{error}) = \min\{P(w_1), P(w_2)\}$$

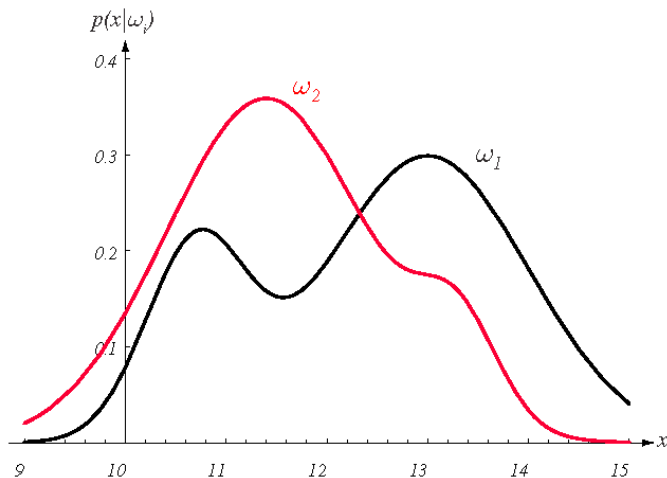


# Class-Conditional Probabilities

- ▶ Let's try to improve the decision using the lightness measurement  $x$ .
- ▶ Let  $x$  be a continuous random variable.
- ▶ Define  $p(x|w_j)$  as the *class-conditional probability density* (probability of  $x$  given that the state of nature is  $w_j$  for  $j = 1, 2$ ).
- ▶  $p(x|w_1)$  and  $p(x|w_2)$  describe the difference in lightness between populations of sea bass and salmon.



# Class-Conditional Probabilities



**Figure 1:** Hypothetical class-conditional probability density functions for two classes.



# Posterior Probabilities

- ▶ Suppose we know  $P(w_j)$  and  $p(x|w_j)$  for  $j = 1, 2$ , and measure the lightness of a fish as the value  $x$ .
- ▶ Define  $P(w_j|x)$  as the *a posteriori probability* (probability of the state of nature being  $w_j$  given the measurement of feature value  $x$ ).
- ▶ We can use the *Bayes formula* to convert the prior probability to the posterior probability

$$P(w_j|x) = \frac{p(x|w_j)P(w_j)}{p(x)}$$

where  $p(x) = \sum_{j=1}^2 p(x|w_j)P(w_j)$ .



# Making a Decision

- ▶  $p(x|w_j)$  is called the *likelihood* and  $p(x)$  is called the *evidence*.
- ▶ How can we make a decision after observing the value of  $x$ ?

$$\text{Decide } \begin{cases} w_1 & \text{if } P(w_1|x) > P(w_2|x) \\ w_2 & \text{otherwise} \end{cases}$$

- ▶ Rewriting the rule gives

$$\text{Decide } \begin{cases} w_1 & \text{if } \frac{p(x|w_1)}{p(x|w_2)} > \frac{P(w_2)}{P(w_1)} \\ w_2 & \text{otherwise} \end{cases}$$

- ▶ Note that, at every  $x$ ,  $P(w_1|x) + P(w_2|x) = 1$ .



# Probability of Error

- What is the probability of error for this decision?

$$P(\text{error}|x) = \begin{cases} P(w_1|x) & \text{if we decide } w_2 \\ P(w_2|x) & \text{if we decide } w_1 \end{cases}$$

- What is the average probability of error?

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

- *Bayes decision rule* minimizes this error because

$$P(\text{error}|x) = \min\{P(w_1|x), P(w_2|x)\}.$$



# Bayesian Decision Theory

- ▶ How can we generalize to
  - ▶ more than one feature?
    - ▶ replace the scalar  $x$  by the feature vector  $\mathbf{x}$
  - ▶ more than two states of nature?
    - ▶ just a difference in notation
  - ▶ allowing actions other than just decisions?
    - ▶ allow the possibility of rejection
  - ▶ different risks in the decision?
    - ▶ define how costly each action is



# Bayesian Decision Theory

- ▶ Let  $\{w_1, \dots, w_c\}$  be the finite set of  $c$  states of nature (*classes, categories*).
- ▶ Let  $\{\alpha_1, \dots, \alpha_a\}$  be the finite set of  $a$  possible *actions*.
- ▶ Let  $\lambda(\alpha_i|w_j)$  be the *loss* incurred for taking action  $\alpha_i$  when the state of nature is  $w_j$ .
- ▶ Let  $\mathbf{x}$  be the  $d$ -component vector-valued random variable called the *feature vector*.



# Bayesian Decision Theory

- ▶  $p(\mathbf{x}|w_j)$  is the class-conditional probability density function.
- ▶  $P(w_j)$  is the prior probability that nature is in state  $w_j$ .
- ▶ The posterior probability can be computed as

$$P(w_j|\mathbf{x}) = \frac{p(\mathbf{x}|w_j)P(w_j)}{p(\mathbf{x})}$$

where  $p(\mathbf{x}) = \sum_{j=1}^c p(\mathbf{x}|w_j)P(w_j)$ .



# Conditional Risk

- ▶ Suppose we observe  $\mathbf{x}$  and take action  $\alpha_i$ .
- ▶ If the true state of nature is  $w_j$ , we incur the loss  $\lambda(\alpha_i|w_j)$ .
- ▶ The expected loss with taking action  $\alpha_i$  is

$$R(\alpha_i|\mathbf{x}) = \sum_{j=1}^c \lambda(\alpha_i|w_j)P(w_j|\mathbf{x})$$

which is also called the *conditional risk*.



# Minimum-Risk Classification

- ▶ The general *decision rule*  $\alpha(\mathbf{x})$  tells us which action to take for observation  $\mathbf{x}$ .
- ▶ We want to find the decision rule that minimizes the overall risk

$$R = \int R(\alpha(\mathbf{x})|\mathbf{x}) p(\mathbf{x}) d\mathbf{x}.$$

- ▶ Bayes decision rule minimizes the overall risk by selecting the action  $\alpha_i$  for which  $R(\alpha_i|\mathbf{x})$  is minimum.
- ▶ The resulting minimum overall risk is called the *Bayes risk* and is the best performance that can be achieved.



# Two-Category Classification

- ▶ Define
  - ▶  $\alpha_1$ : deciding  $w_1$ ,
  - ▶  $\alpha_2$ : deciding  $w_2$ ,
  - ▶  $\lambda_{ij} = \lambda(\alpha_i | w_j)$ .
- ▶ Conditional risks can be written as

$$R(\alpha_1 | \mathbf{x}) = \lambda_{11} P(w_1 | \mathbf{x}) + \lambda_{12} P(w_2 | \mathbf{x}),$$

$$R(\alpha_2 | \mathbf{x}) = \lambda_{21} P(w_1 | \mathbf{x}) + \lambda_{22} P(w_2 | \mathbf{x}).$$



# Two-Category Classification

- ▶ The *minimum-risk decision rule* becomes

$$\text{Decide } \begin{cases} w_1 & \text{if } (\lambda_{21} - \lambda_{11})P(w_1|\mathbf{x}) > (\lambda_{12} - \lambda_{22})P(w_2|\mathbf{x}) \\ w_2 & \text{otherwise} \end{cases}$$

- ▶ This corresponds to deciding  $w_1$  if

$$\frac{p(\mathbf{x}|w_1)}{p(\mathbf{x}|w_2)} > \frac{(\lambda_{12} - \lambda_{22})}{(\lambda_{21} - \lambda_{11})} \frac{P(w_2)}{P(w_1)}$$

⇒ comparing the *likelihood ratio* to a threshold that is independent of the observation  $\mathbf{x}$ .



# Minimum-Error-Rate Classification

- ▶ Actions are decisions on classes ( $\alpha_i$  is deciding  $w_i$ ).
- ▶ If action  $\alpha_i$  is taken and the true state of nature is  $w_j$ , then the decision is correct if  $i = j$  and in error if  $i \neq j$ .
- ▶ We want to find a decision rule that minimizes the probability of error.



# Minimum-Error-Rate Classification

- Define the *zero-one loss function*

$$\lambda(\alpha_i | w_j) = \begin{cases} 0 & \text{if } i = j \\ 1 & \text{if } i \neq j \end{cases} \quad i, j = 1, \dots, c$$

(all errors are equally costly).

- Conditional risk becomes

$$\begin{aligned} R(\alpha_i | \mathbf{x}) &= \sum_{j=1}^c \lambda(\alpha_i | w_j) P(w_j | \mathbf{x}) \\ &= \sum_{j \neq i} P(w_j | \mathbf{x}) \\ &= 1 - P(w_i | \mathbf{x}). \end{aligned}$$



# Minimum-Error-Rate Classification

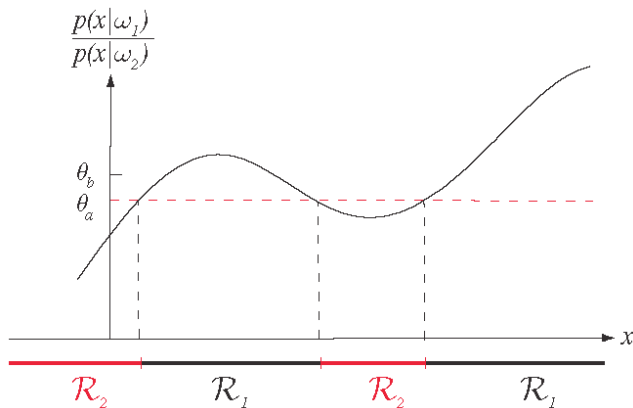
- ▶ Minimizing the risk requires maximizing  $P(w_i|\mathbf{x})$  and results in the *minimum-error decision rule*

Decide  $w_i$  if  $P(w_i|\mathbf{x}) > P(w_j|\mathbf{x}) \quad \forall j \neq i.$

- ▶ The resulting error is called the *Bayes error* and is the best performance that can be achieved.



# Minimum-Error-Rate Classification



**Figure 2:** The likelihood ratio  $p(\mathbf{x}|w_1)/p(\mathbf{x}|w_2)$ . The threshold  $\theta_a$  is computed using the priors  $P(w_1) = 2/3$  and  $P(w_2) = 1/3$ , and a zero-one loss function. If we penalize mistakes in classifying  $w_2$  patterns as  $w_1$  more than the converse, we should increase the threshold to  $\theta_b$ .



# Discriminant Functions

- ▶ A useful way of representing classifiers is through *discriminant functions*  $g_i(\mathbf{x})$ ,  $i = 1, \dots, c$ , where the classifier assigns a feature vector  $\mathbf{x}$  to class  $w_i$  if

$$g_i(\mathbf{x}) > g_j(\mathbf{x}) \quad \forall j \neq i.$$

- ▶ For the classifier that minimizes conditional risk

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

- ▶ For the classifier that minimizes error

$$g_i(\mathbf{x}) = P(w_i|\mathbf{x}).$$



# Discriminant Functions

- ▶ These functions divide the feature space into  $c$  *decision regions* ( $\mathcal{R}_1, \dots, \mathcal{R}_c$ ), separated by *decision boundaries*.
- ▶ Note that the results do not change even if we replace every  $g_i(\mathbf{x})$  by  $f(g_i(\mathbf{x}))$  where  $f(\cdot)$  is a monotonically increasing function (e.g., logarithm).
- ▶ This may lead to significant analytical and computational simplifications.



# The Gaussian Density

- ▶ Gaussian can be considered as a model where the feature vectors for a given class are continuous-valued, randomly corrupted versions of a single typical or prototype vector.
- ▶ Some properties of the Gaussian:
  - ▶ Analytically tractable.
  - ▶ Completely specified by the 1st and 2nd moments.
  - ▶ Has the maximum entropy of all distributions with a given mean and variance.
  - ▶ Many processes are asymptotically Gaussian (Central Limit Theorem).
  - ▶ Linear transformations of a Gaussian are also Gaussian.
  - ▶ Uncorrelatedness implies independence.



# Univariate Gaussian

► For  $x \in \mathbb{R}$ :

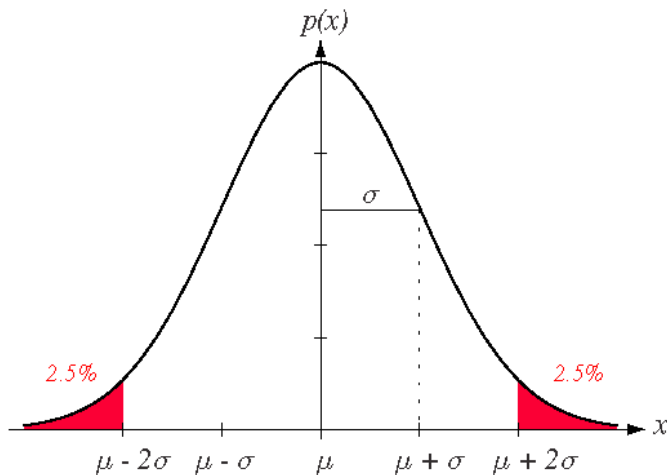
$$\begin{aligned} p(x) &= N(\mu, \sigma^2) \\ &= \frac{1}{\sqrt{2\pi}\sigma} \exp \left[ -\frac{1}{2} \left( \frac{x - \mu}{\sigma} \right)^2 \right] \end{aligned}$$

where

$$\begin{aligned} \mu &= E[x] = \int_{-\infty}^{\infty} x p(x) dx, \\ \sigma^2 &= E[(x - \mu)^2] = \int_{-\infty}^{\infty} (x - \mu)^2 p(x) dx. \end{aligned}$$



# Univariate Gaussian



**Figure 3:** A univariate Gaussian distribution has roughly 95% of its area in the range  $|x - \mu| \leq 2\sigma$ .



# Multivariate Gaussian

► For  $\mathbf{x} \in \mathbb{R}^d$ :

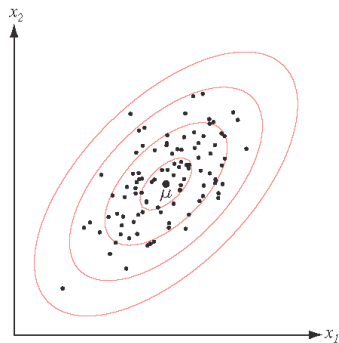
$$\begin{aligned} p(\mathbf{x}) &= N(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \\ &= \frac{1}{(2\pi)^{d/2} |\boldsymbol{\Sigma}|^{1/2}} \exp \left[ -\frac{1}{2} (\mathbf{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}) \right] \end{aligned}$$

where

$$\begin{aligned} \boldsymbol{\mu} &= E[\mathbf{x}] = \int \mathbf{x} p(\mathbf{x}) d\mathbf{x}, \\ \boldsymbol{\Sigma} &= E[(\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})^T] = \int (\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})^T p(\mathbf{x}) d\mathbf{x}. \end{aligned}$$



# Multivariate Gaussian



**Figure 4:** Samples drawn from a two-dimensional Gaussian lie in a cloud centered on the mean  $\mu$ . The loci of points of constant density are the ellipses for which  $(\mathbf{x} - \mu)^T \Sigma^{-1} (\mathbf{x} - \mu)$  is constant, where the eigenvectors of  $\Sigma$  determine the direction and the corresponding eigenvalues determine the length of the principal axes. The quantity  $r^2 = (\mathbf{x} - \mu)^T \Sigma^{-1} (\mathbf{x} - \mu)$  is called the squared *Mahalanobis distance* from  $\mathbf{x}$  to  $\mu$ .



# Linear Transformations

- ▶ Recall that, given  $\mathbf{x} \in \mathbb{R}^d$ ,  $\mathbf{A} \in \mathbb{R}^{d \times k}$ ,  $\mathbf{y} = \mathbf{A}^T \mathbf{x} \in \mathbb{R}^k$ , if  $x \sim N(\boldsymbol{\mu}, \Sigma)$ , then  $y \sim N(\mathbf{A}^T \boldsymbol{\mu}, \mathbf{A}^T \Sigma \mathbf{A})$ .
- ▶ As a special case, the *whitening transform*

$$\mathbf{A}_w = \Phi \Lambda^{-1/2}$$

where

- ▶  $\Phi$  is the matrix whose columns are the orthonormal eigenvectors of  $\Sigma$ ,
  - ▶  $\Lambda$  is the diagonal matrix of the corresponding eigenvalues,
- gives a covariance matrix equal to the identity matrix  $\mathbf{I}$ .



# Discriminant Functions for the Gaussian Density

- ▶ Discriminant functions for minimum-error-rate classification can be written as

$$g_i(\mathbf{x}) = \ln p(\mathbf{x}|w_i) + \ln P(w_i).$$

- ▶ For  $p(\mathbf{x}|w_i) = N(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$

$$g_i(\mathbf{x}) = -\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_i)^T \boldsymbol{\Sigma}_i^{-1}(\mathbf{x} - \boldsymbol{\mu}_i) - \frac{d}{2} \ln 2\pi - \frac{1}{2} \ln |\boldsymbol{\Sigma}_i| + \ln P(w_i).$$



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

- Discriminant functions are

$$g_i(\mathbf{x}) = \mathbf{w}_i^T \mathbf{x} + w_{i0} \quad (\text{linear discriminant})$$

where

$$\mathbf{w}_i = \frac{1}{\sigma^2} \boldsymbol{\mu}_i$$
$$w_{i0} = -\frac{1}{2\sigma^2} \boldsymbol{\mu}_i^T \boldsymbol{\mu}_i + \ln P(w_i)$$

( $w_{i0}$  is the threshold or bias for the  $i$ 'th category).



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

- Decision boundaries are the hyperplanes  $g_i(\mathbf{x}) = g_j(\mathbf{x})$ , and can be written as

$$\mathbf{w}^T (\mathbf{x} - \mathbf{x}_0) = 0$$

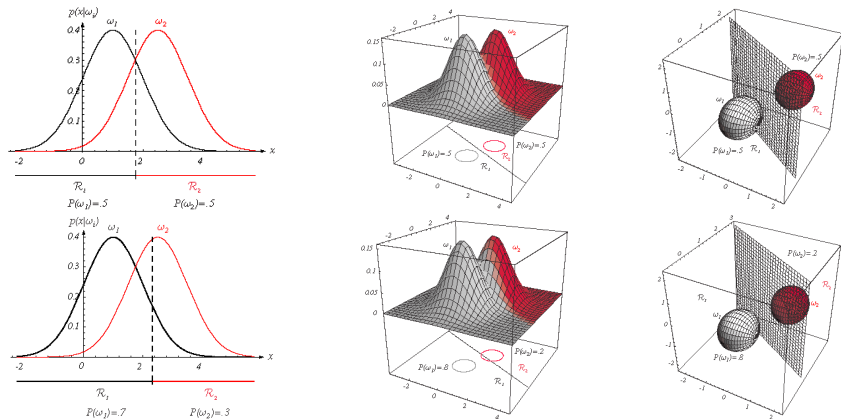
where

$$\begin{aligned}\mathbf{w} &= \boldsymbol{\mu}_i - \boldsymbol{\mu}_j \\ \mathbf{x}_0 &= \frac{1}{2}(\boldsymbol{\mu}_i + \boldsymbol{\mu}_j) - \frac{\sigma^2}{\|\boldsymbol{\mu}_i - \boldsymbol{\mu}_j\|^2} \ln \frac{P(w_i)}{P(w_j)} (\boldsymbol{\mu}_i - \boldsymbol{\mu}_j).\end{aligned}$$

- Hyperplane separating  $\mathcal{R}_i$  and  $\mathcal{R}_j$  passes through the point  $\mathbf{x}_0$  and is orthogonal to the vector  $\mathbf{w}$ .



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



**Figure 5:** If the covariance matrices of 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. The decision boundary shifts as the priors are changed.

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

- Special case when  $P(w_i)$  are the same for  $i = 1, \dots, c$  is the *minimum-distance classifier* that uses the decision rule

assign  $\mathbf{x}$  to  $w_{i^*}$  where  $i^* = \arg \min_{i=1, \dots, c} \|\mathbf{x} - \boldsymbol{\mu}_i\|$ .



## Case 2: $\Sigma_i = \Sigma$

- Discriminant functions are

$$g_i(\mathbf{x}) = \mathbf{w}_i^T \mathbf{x} + w_{i0} \quad (\text{linear discriminant})$$

where

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

$$w_{i0} = -\frac{1}{2} \boldsymbol{\mu}_i^T \Sigma^{-1} \boldsymbol{\mu}_i + \ln P(w_i).$$



## Case 2: $\Sigma_i = \Sigma$

- Decision boundaries can be written as

$$\mathbf{w}^T(\mathbf{x} - \mathbf{x}_0) = 0$$

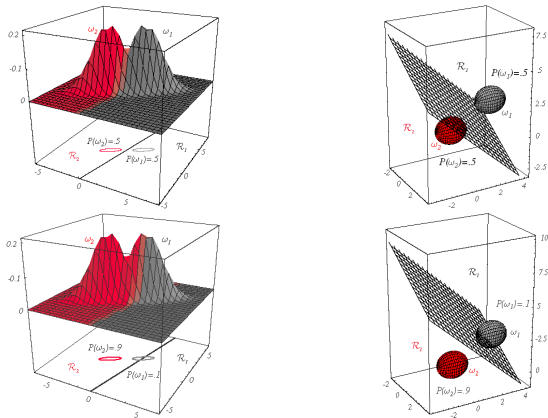
where

$$\begin{aligned}\mathbf{w} &= \Sigma^{-1}(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j) \\ \mathbf{x}_0 &= \frac{1}{2}(\boldsymbol{\mu}_i + \boldsymbol{\mu}_j) - \frac{\ln(P(w_i)/P(w_j))}{(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)^T \Sigma^{-1}(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)}(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j).\end{aligned}$$

- Hyperplane passes through  $\mathbf{x}_0$  but is not necessarily orthogonal to the line between the means.



## Case 2: $\Sigma_i = \Sigma$



**Figure 6:** Probability densities with equal but asymmetric Gaussian distributions. The decision hyperplanes are not necessarily perpendicular to the line connecting the means.

## Case 3: $\Sigma_i = \text{arbitrary}$

- Discriminant functions are

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

where

$$\mathbf{W}_i = -\frac{1}{2} \Sigma_i^{-1}$$

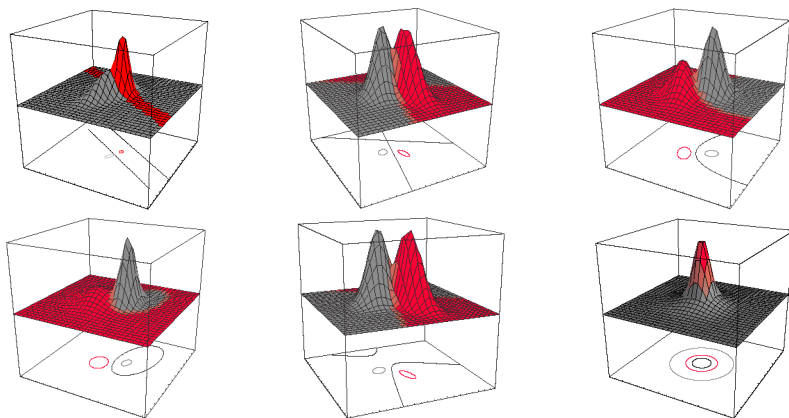
$$\mathbf{w}_i = \Sigma_i^{-1} \boldsymbol{\mu}_i$$

$$w_{i0} = -\frac{1}{2} \boldsymbol{\mu}_i^T \Sigma_i^{-1} \boldsymbol{\mu}_i - \frac{1}{2} \ln |\Sigma_i| + \ln P(w_i).$$

- Decision boundaries are hyperquadrics.

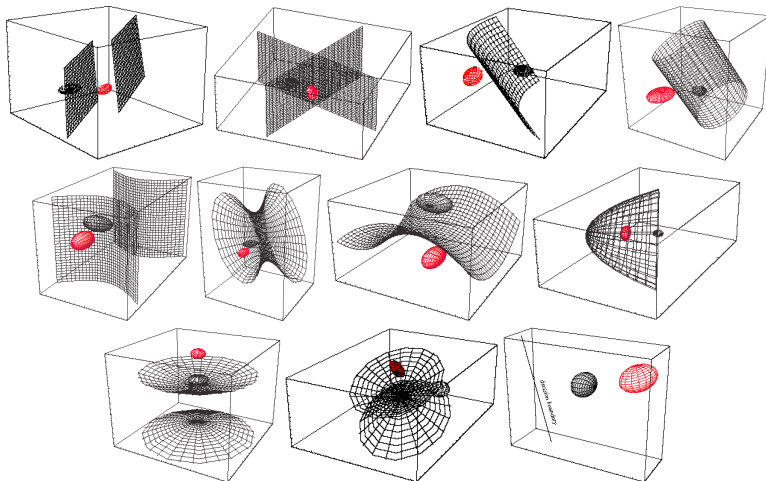


# Case 3: $\Sigma_i = \text{arbitrary}$



**Figure 7:** Arbitrary Gaussian distributions lead to Bayes decision boundaries that are general hyperquadrics.

# Case 3: $\Sigma_i = \text{arbitrary}$



**Figure 8:** Arbitrary Gaussian distributions lead to Bayes decision boundaries that are general hyperquadrics.

# Error Probabilities and Integrals

- For the two-category case

$$\begin{aligned}P(\text{error}) &= P(\mathbf{x} \in \mathcal{R}_2, w_1) + P(\mathbf{x} \in \mathcal{R}_1, w_2) \\&= P(\mathbf{x} \in \mathcal{R}_2|w_1)P(w_1) + P(\mathbf{x} \in \mathcal{R}_1|w_2)P(w_2) \\&= \int_{\mathcal{R}_2} p(\mathbf{x}|w_1) P(w_1) d\mathbf{x} + \int_{\mathcal{R}_1} p(\mathbf{x}|w_2) P(w_2) d\mathbf{x}.\end{aligned}$$



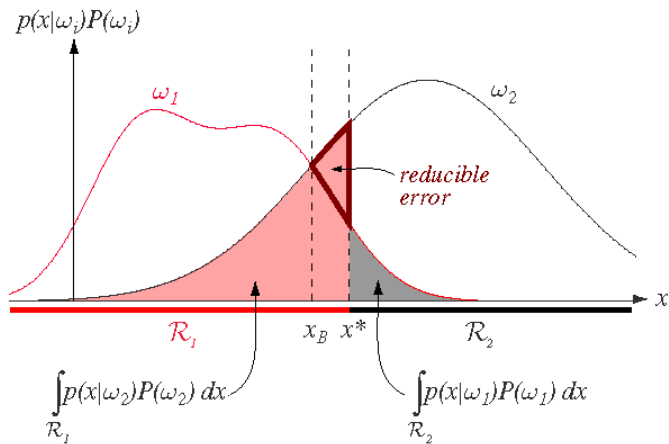
# Error Probabilities and Integrals

- For the multcategory case

$$\begin{aligned}P(\text{error}) &= 1 - P(\text{correct}) \\&= 1 - \sum_{i=1}^c P(\mathbf{x} \in \mathcal{R}_i, w_i) \\&= 1 - \sum_{i=1}^c P(\mathbf{x} \in \mathcal{R}_i | w_i) P(w_i) \\&= 1 - \sum_{i=1}^c \int_{\mathcal{R}_i} p(\mathbf{x} | w_i) P(w_i) d\mathbf{x}.\end{aligned}$$



# Error Probabilities and Integrals



**Figure 9:** Components of the probability of error for equal priors and the non-optimal decision point  $x^*$ . The optimal point  $x_B$  minimizes the total shaded area and gives the Bayes error rate.

# Receiver Operating Characteristics

- ▶ Consider the two-category case and define
  - ▶  $w_1$ : target is present,
  - ▶  $w_2$ : target is not present.

Table 1: *Confusion matrix.*

		Assigned	
		$w_1$	$w_2$
True	$w_1$	correct detection	mis-detection
	$w_2$	false alarm	correct rejection

- ▶ Mis-detection is also called false negative or Type I error.
- ▶ False alarm is also called false positive or Type II error.



# Receiver Operating Characteristics

- If we use a parameter (e.g., a threshold) in our decision, the plot of these rates for different values of the parameter is called the *receiver operating characteristic* (ROC) curve.

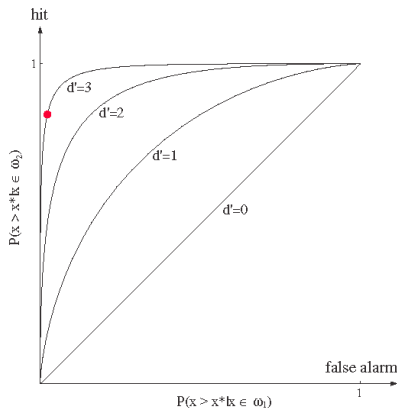


Figure 10: Example receiver operating characteristic (ROC) curves for different settings of the system.



# Summary

- ▶ To minimize the overall risk, choose the action that minimizes the conditional risk  $R(\alpha|\mathbf{x})$ .
- ▶ To minimize the probability of error, choose the class that maximizes the posterior probability  $P(w_j|\mathbf{x})$ .
- ▶ If there are different penalties for misclassifying patterns from different classes, the posteriors must be weighted according to such penalties before taking action.
- ▶ Do not forget that these decisions are the optimal ones under the assumption that the “true” values of the probabilities are known.

