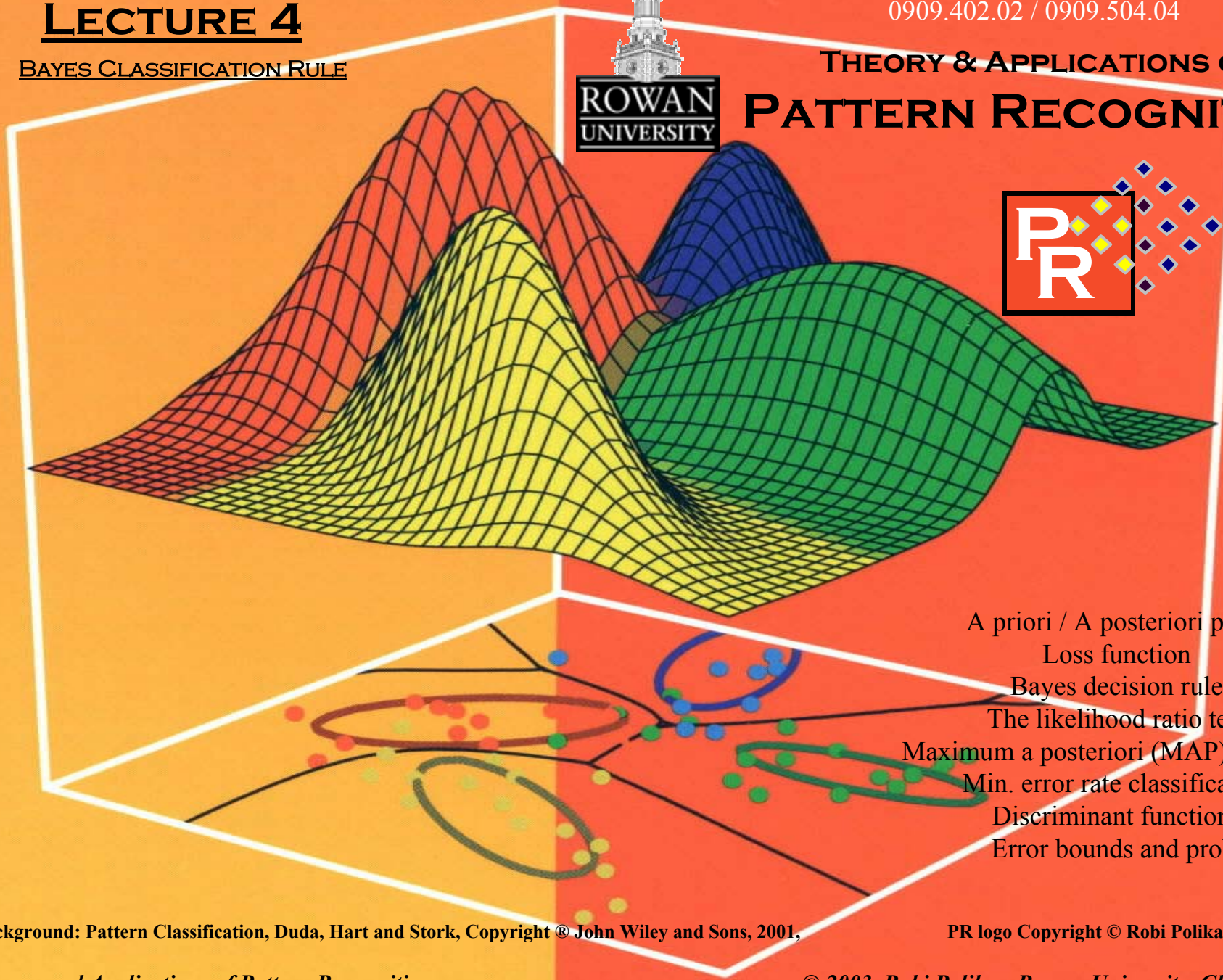
Dept. of Electrical and Computer Engineering
0909.402.02 / 0909.504.04

Lecture 4

Bayes Classification Rule

ROWAN UNIVERSITY

Theory & Applications of
Pattern Recognition



A priori / A posteriori prob.
Loss function
Bayes decision rule
The likelihood ratio test
Maximum a posteriori (MAP) Criterion
Min. error rate classification
Discriminant functions
Error bounds and prob.

TODAY IN PR

- Review of Bayes theorem
- Bayes Decision Theory
  - Bayes rule
  - Loss function & expected loss
  - Minimum error rate classification
- Classification using discriminant functions
- Error bounds & probabilities

BAYES RULE

- Suppose, we know $P(\omega_1)$, $P(\omega_2)$, $P(x|\omega_1)$ and $P(x|\omega_2)$, and that we have observed the value of the feature (a random variable) x
  - How would you decide on the ***"state of nature"*** – type of fish, based on this info?
  - Bayes theory allows us to compute the posterior probabilities from prior and class-conditional probabilities

Likelihood: The (class-conditional) probability of observing a feature value of x, given that the correct class is ω_j. All things being equal, the category with higher class conditional probability is more "likely" to be the correct class.

Prior Probability: The total probability of correct class being class ω_j determined based on prior experience

$$P(\omega_j \mid x) = \frac{P(x \cap \omega_j)}{P(x)} = \frac{P(x \mid \omega_j) \cdot P(\omega_j)}{\sum_{k=1}^{C} P(x \mid \omega_k) \cdot P(\omega_k)}$$

Posterior Probability: The (conditional) probability of correct class being ω_j, given that feature value x has been observed

Evidence: The total probability of observing the feature value as x

Bayes Decision Rule

Choose ω_i if $P(\omega_i \mid x) > P(\omega_j \mid x)$ for all $i \neq j,\ i,j=1,2,\ldots,c$

If there are multiple features, $\mathbf{x}=\{x_1, x_2, \ldots, x_d\}$ ➔

Choose ω_i if $P(\omega_i \mid \boldsymbol{x}) > P(\omega_j \mid \boldsymbol{x})$ for all $i \neq j$ $i,j=1,2,\ldots,c$

The Loss Function

- Mathematical description of how costly each action (making a class decision) is. Are certain mistakes costlier than others?

$\{\omega_1, \omega_2, \ldots, \omega_c\}$: Set of states of nature (classes)

$\{\alpha_1, \alpha_2, \ldots \alpha_a\}$: Set of possible actions. Note that a need not be same as c. Because we may make more (or less) number of actions than the number of classes. For example, not making a decision (rejection) is also an action.

$\{\lambda_1, \lambda_2, \ldots \lambda_a\}$: Losses associated with each action

$\lambda(\alpha_i | \omega_j\}$: ***The loss function***: Loss incurred by taking action i when the true state of nature is in fact j.

$R(\alpha_i | \mathbf{x})$: Conditional risk - Expected loss for taking action i

$$R(\alpha_i \mid \mathbf{x}) = \sum_{j=1}^{c} \lambda(\alpha_i \mid \omega_j) \cdot P(\omega_j \mid \mathbf{x})$$

Bayes decision takes the action that minimizes the conditional risk !

Bayes Decision Rule Using Conditional Risk

1. Compute conditional risk $R(\alpha_i \mid \mathbf{x}) = \sum_{j=1}^{c} \lambda(\alpha_i \mid \omega_j) \cdot P(\omega_j \mid \mathbf{x})$ for each action taken
2. Select the action that has the minimum conditional risk. Let this be action k
3. The overall risk is then

$$R = \int_{\mathbf{x} \in X} R(\alpha_k \mid \mathbf{x}) \cdot p(\mathbf{x}) d\mathbf{x}$$

Integrated over all possible values of **x**

Probability that **x** will be observed

Conditional risk associated with taking action α(**x**) based on the observation **x**.

4. This is the Bayes Risk, the minimum possible risk that can be taken by any classifier !

Two-Class Special Case

- Definitions:
  - α_1: Decide on ω_1,
  - α_2: Decide on ω_2,
  - λ_{ij}: $\lambda(\alpha_i \mid \omega_j)$ ➔ Loss for deciding on ω_i when the SON is ω_j
- Conditional risk:
  - $R(\alpha_1 \mid \mathbf{x}) = \lambda_{11}P(\omega_1 \mid \mathbf{x}) + \lambda_{12}P(\omega_2 \mid \mathbf{x})$
  - $R(\alpha_2 \mid \mathbf{x}) = \lambda_{21}P(\omega_1 \mid \mathbf{x}) + \lambda_{22}P(\omega_2 \mid \mathbf{x})$
  - Note that λ_{11} and λ_{22} need not be zero, though we expect $\lambda_{11} < \lambda_{12}$, $\lambda_{22} < \lambda_{21}$

Decide on ω_1 if $R(\alpha_1 \mid \mathbf{x}) < R(\alpha_2 \mid \mathbf{x})$, decide on ω_2, otherwise

$$\Lambda(\mathbf{x}) = \frac{p(\mathbf{x} \mid \omega_1)}{p(\mathbf{x} \mid \omega_2)} \underset{\omega_2}{\overset{\omega_1}{\lessgtr}} \frac{(\lambda_{12} - \lambda_{22})}{(\lambda_{21} - \lambda_{11})} \cdot \frac{P(\omega_2)}{P(\omega_1)}$$

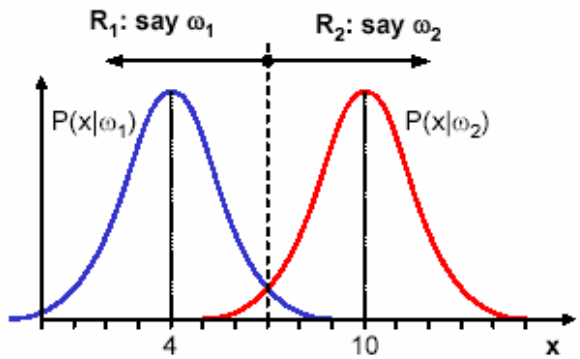
The ***Likelihood Ratio Test (LRT):*** Pick ω_1 if the LRT is greater then a threshold that is independent of $\mathbf{x}$. This rule, which minimizes the Bayes risk, is also called the ***Bayes Criterion***.

EXAMPLE

- **Given a classification problem with the following class conditional densities, derive a decision rule based on the Likelihood Ratio Test (assume equal priors)**

$$P(x\,|\,\omega_1)=\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}(x-4)^2} \qquad P(x\,|\,\omega_2)=\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}(x-10)^2}$$

- **Solution**
  - Substituting the given likelihoods and priors into the LRT expression: $\Lambda(x)=\dfrac{\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}(x-4)^2}}{\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}(x-10)^2}} \underset{\omega_2}{\overset{\omega_1}{\gtrless}} 1$
  - Simplifying the LRT expression: $\Lambda(x)=\dfrac{e^{-\frac{1}{2}(x-4)^2}}{e^{-\frac{1}{2}(x-10)^2}} \underset{\omega_2}{\overset{\omega_1}{\gtrless}} 1$
  - Changing signs and taking logs: $(x-4)^2-(x-10)^2 \underset{\omega_2}{\overset{\omega_1}{\lessgtr}} 0$
  - Which yields: $x \underset{\omega_2}{\overset{\omega_1}{\lessgtr}} 7$
  - This LRT result makes sense from an intuitive point of view since the likelihoods are identical and differ only in their mean value



- **How would the LRT decision rule change if, say, the priors were such that $P(\omega_1)=2P(\omega_2)$?**

From R. Gutierrez @ TAMU

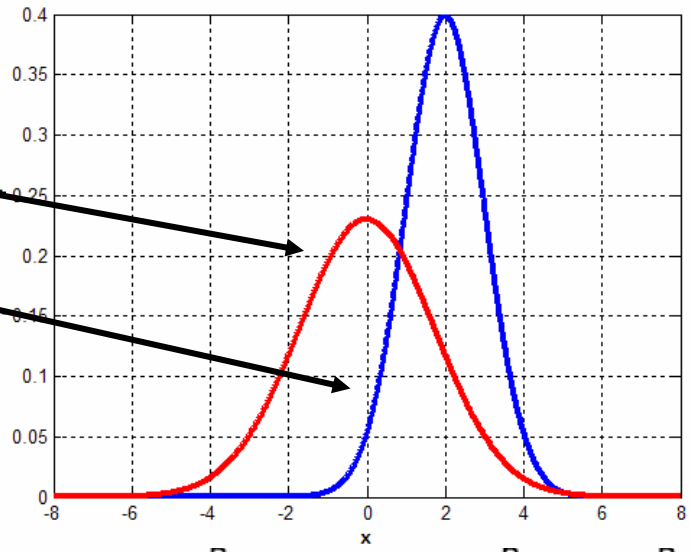
EXAMPLE

(TO BE FULLY SOLVED ON REQUEST ON FRIDAY)

- **Consider a classification problem with two classes defined by the following likelihood functions**

$$P(x \mid \omega_1) = \frac{1}{\sqrt{2\pi}\sqrt{3}} e^{-\frac{1}{2}\frac{x^2}{3}}$$

$$P(x \mid \omega_2) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(x-2)^2}$$



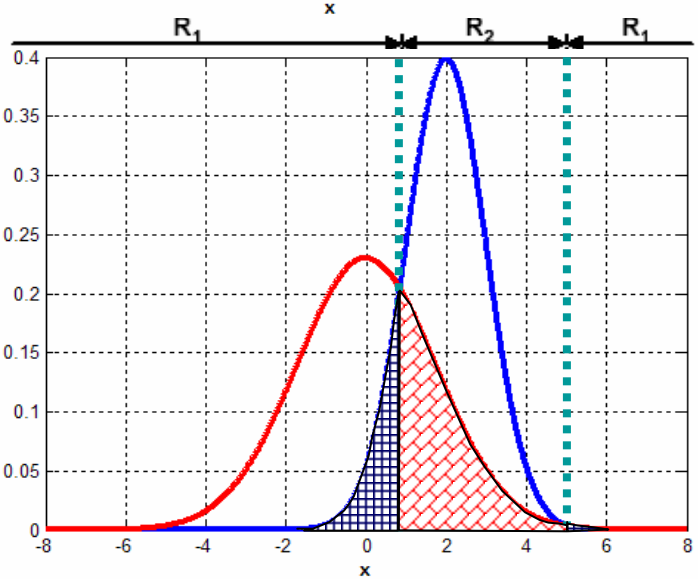
- Sketch the two densities
- What is the likelihood ratio?
- Assume $P[\omega_1]=P[\omega_2]=0.5$, $\lambda_{11}=\lambda_{22}=0$, $\lambda_{12}=1$ and $\lambda_{21}=3^{1/2}$. Determine a decision rule that minimizes the probability of error

$$\Lambda(x) = \frac{\frac{1}{\sqrt{2\pi}\sqrt{3}} e^{-\frac{1}{2}\frac{x^2}{3}}}{\frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(x-2)^2}} \underset{\omega_2}{\overset{\omega_1}{\lessgtr}} \frac{1}{\sqrt{3}}$$

$$\frac{e^{-\frac{1}{2}\frac{x^2}{3}}}{e^{-\frac{1}{2}(x-2)^2}} \underset{\omega_2}{\overset{\omega_1}{\lessgtr}} 1$$

$$-\frac{1}{2}\frac{x^2}{3} + \frac{1}{2}(x-2)^2 \underset{\omega_2}{\overset{\omega_1}{\lessgtr}} 0$$

$$2x^2 - 12x + 12 \underset{\omega_2}{\overset{\omega_1}{\lessgtr}} 0 \Rightarrow x = 4.73, 1.27$$



Modified from R. Gutierrez @ TAMU

Minimum Error-Rate Classification: Multiclass Case

- If we associate taking action i as selecting class i, and if all errors are equally likely, we obtain the **zero-one loss** (***symmetrical cost function***)

$$\lambda(\alpha_i \mid \omega_j) = \begin{cases} 0, & \text{if } i = j \\ 1, & \text{if } i \neq j \end{cases}$$

This loss function assigns no loss to correct classification, and assigns 1 to misclassification. The risk corresponding to this loss function is then

$$R(\alpha_i \mid \mathbf{x}) = \sum_{\substack{j \neq i \\ j=1,\ldots,c}} P(\omega_j \mid \mathbf{x}) = 1 - P(\omega_i \mid \mathbf{x})$$

What does this tell us…?

- To minimize this risk (average probability of error), we need to choose the class that maximizes the *posterior probability* $P(\omega_i|\mathbf{x})$

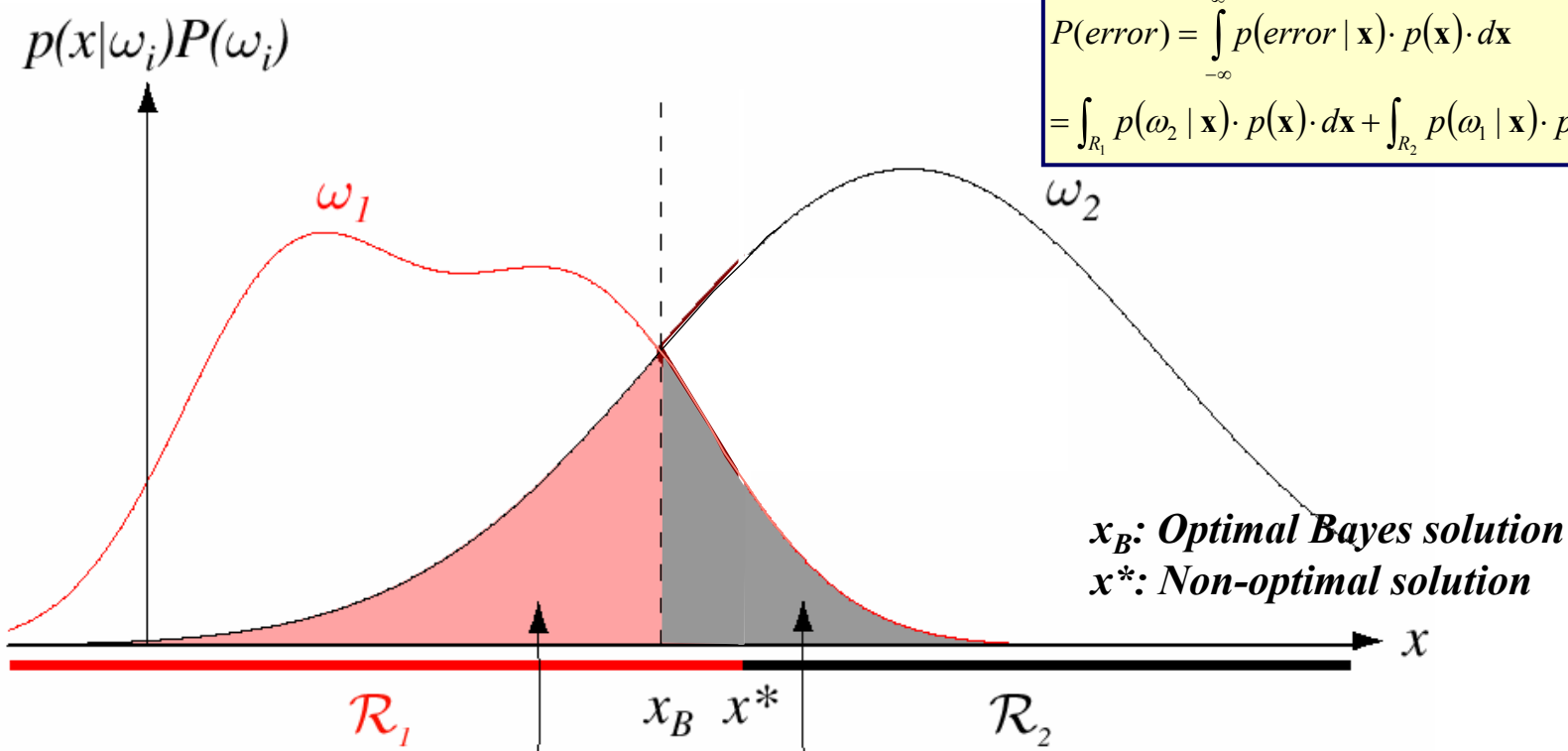
$$\Lambda(\mathbf{x}) = \frac{p(\mathbf{x} \mid \omega_i)}{p(\mathbf{x} \mid \omega_j)} \underset{\omega_j}{\overset{\omega_i}{\gtrless}} \frac{P(\omega_j)}{P(\omega_i)} \Leftrightarrow \frac{P(\omega_i \mid \mathbf{x})}{P(\omega_j \mid \mathbf{x})} \underset{\omega_j}{\overset{\omega_i}{\gtrless}} 1$$

Maximum a posteriori (MAP) criterion
Maximum likelihood criterion for equal priors

Error Probabilities
(Bayes Rule Rules!)

In a two class case, there are two sources of error: **x** is in R_1, yet SON is ω_2, or vice versa

$$P(error) = \int_{-\infty}^{\infty} p(error \mid \mathbf{x}) \cdot p(\mathbf{x}) \cdot d\mathbf{x}$$
$$= \int_{R_1} p(\omega_2 \mid \mathbf{x}) \cdot p(\mathbf{x}) \cdot d\mathbf{x} + \int_{R_2} p(\omega_1 \mid \mathbf{x}) \cdot p(\mathbf{x}) \cdot d\mathbf{x}$$



$$\mathbf{P(error)} = \int_{\mathcal{R}_1} p(x|\omega_2)P(\omega_2)\, dx + \int_{\mathcal{R}_2} p(x|\omega_1)P(\omega_1)\, dx$$

$$P(\mathbf{x} \in R_1, \omega_2) = P(\mathbf{x} \in R_1 \mid \omega_2) \cdot P(\omega_2)$$

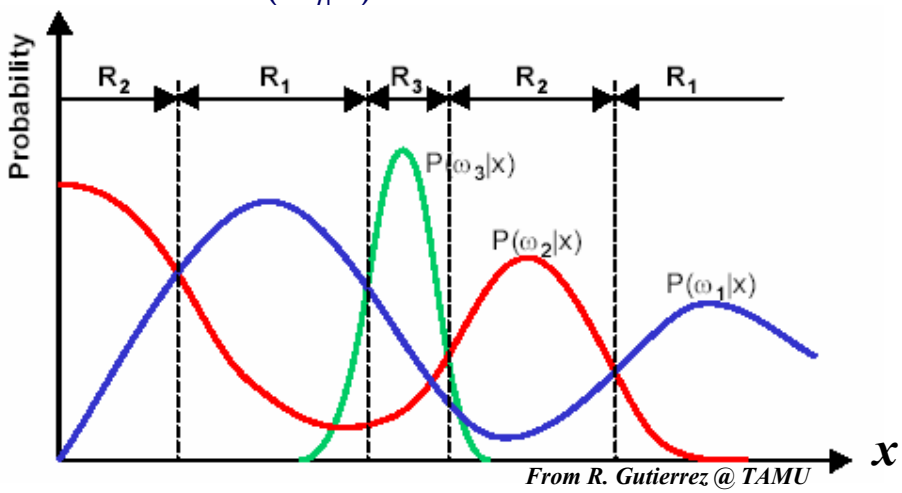
$$P(\mathbf{x} \in R_2, \omega_1) = P(\mathbf{x} \in R_2 \mid \omega_1) \cdot P(\omega_1)$$

Probability of Error

- In multi-class case, there are more ways to be wrong then to be right, so we exploit the fact that $P(error)=1-P(correct)$, where

$$P(correct)=\sum_{i=1}^{C} P(\mathbf{x}\in R_i,\omega_i)=\sum_{i=1}^{C} P(\mathbf{x}\in R_i\mid\omega_i)P(\omega_i)$$
$$=\sum_{i=1}^{C}\int_{\mathbf{x}\in R_i} P(\mathbf{x}\mid\omega_i)P(\omega_i)d\mathbf{x}=\sum_{i=1}^{C}\int_{\mathbf{x}\in R_i} P(\omega_i\mid\mathbf{x})P(\mathbf{x})d\mathbf{x}$$

- Of course, in order to minimize the $P(error)$, we need to maximize $P(correct)$ for which we need to maximize each and every one of the integrals. Note that $P(\mathbf{x})$ is common to all integrals, therefore the expression will be maximized by choosing the decision regions R_i where the posterior probabilities $P(\omega_i|\mathbf{x})$ are maximum:



From R. Gutierrez @ TAMU

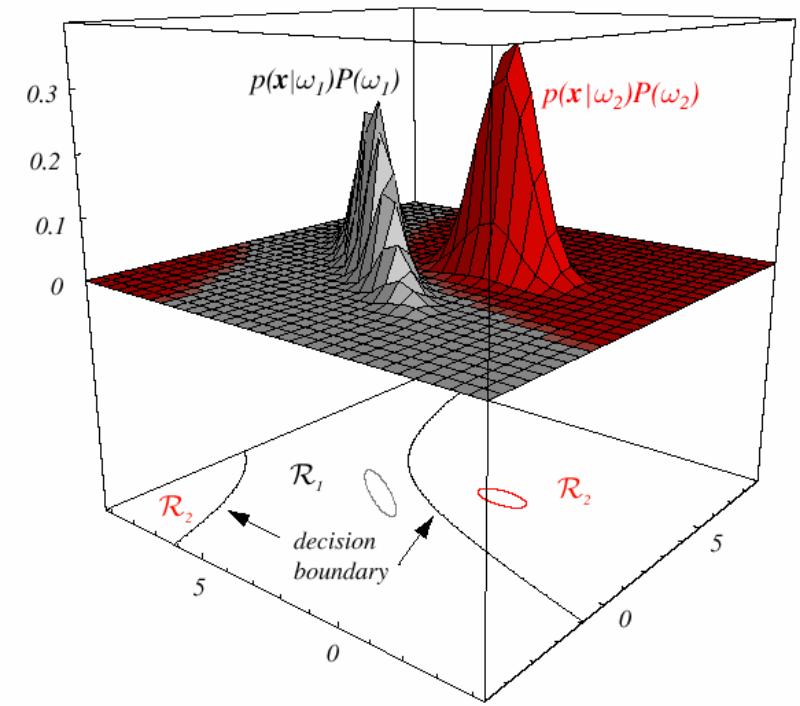
Discriminant Based Classification

- A discriminant is a function g(**x**), that discriminates between classes. This function assigns the input vector to a class according to its definition: Choose class *i* if

$$g_i(\mathbf{x}) > g_j(\mathbf{x}) \quad \forall i \neq j, \quad i, j = 1,2,...,c$$

- Bayes rule can be implemented in terms of discriminant functions

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



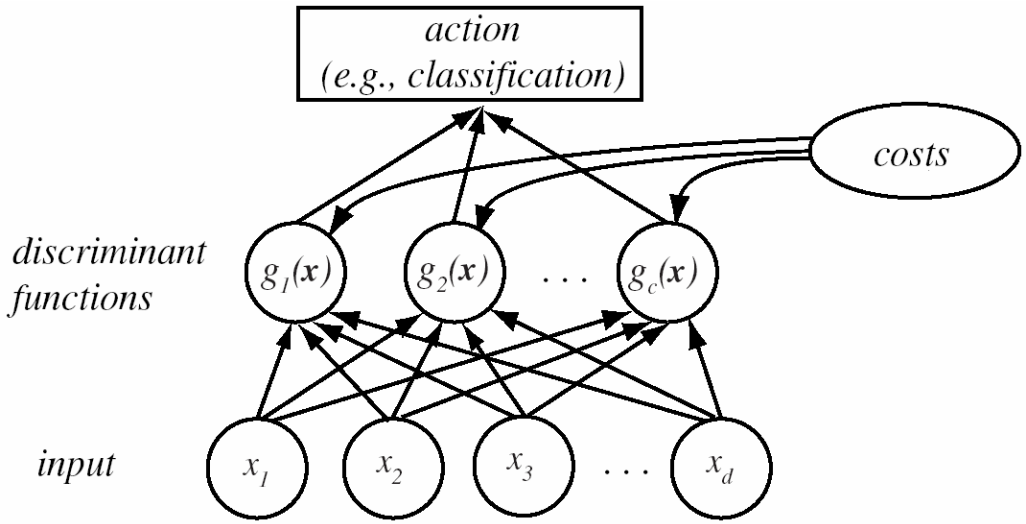
Each discriminant function generates *c* decision regions, $R_1,...,R_c$, which are separated by *decision boundaries*. Decision regions need NOT be contiguous.

The decision boundary satisfies $g_i(\mathbf{x}) = g_j(\mathbf{x})$

$$\mathbf{x} \in R_1 \Leftrightarrow g_i(\mathbf{x}) > g_j(\mathbf{x})$$
$$\forall i \neq j, \quad i, j = 1,2,...,c$$

Discriminant Functions

- We may view the classifier as an automated machine that computes c discriminants and selects the category corresponding to the largest discriminant
- A neural network is one such classifier



- for Bayes classifier with non-uniform risks, $R(\alpha_i|\mathbf{x})$: $g_i(\mathbf{x}) = -R(\alpha_i \mid \mathbf{x})$
- for MAP classifier (of uniform risks): $g_i(\mathbf{x}) = P(\omega_i \mid \mathbf{x})$
- for maximum likelihood classifier (of equal priors): $g_i(\mathbf{x}) = P(\mathbf{x} \mid \omega_i)$

Discriminant Functions

- In fact, multiplying every DF with the same constant, or adding/subtracting a constant to all DFs does not change the decision boundary
  - In general every $g_i(\mathbf{x})$ can be replaced by $f(g_i(\mathbf{x}))$, where $f(.)$ is any monotonically increasing function without affecting the actual decision boundary
  - Some linear or non-linear transformations of the previously stated DFs may greatly simplify the design of the classifier

- What examples can you think of…?

Normal Densities

- If likelihood probabilities are normally distributed, then a number of simplifications can be made. In particular, the discriminant function can be written as in this greatly simplified form (!)

$$p(\mathbf{x}\,|\,\omega_i) = \frac{1}{(2\pi)^{d/2}|\boldsymbol{\Sigma}_i|^{1/2}} e^{-\frac{1}{2}\left[(\mathbf{x}-\boldsymbol{\mu}_i)^T \boldsymbol{\Sigma}_i^{-1}(\mathbf{x}-\boldsymbol{\mu}_i)\right]}$$

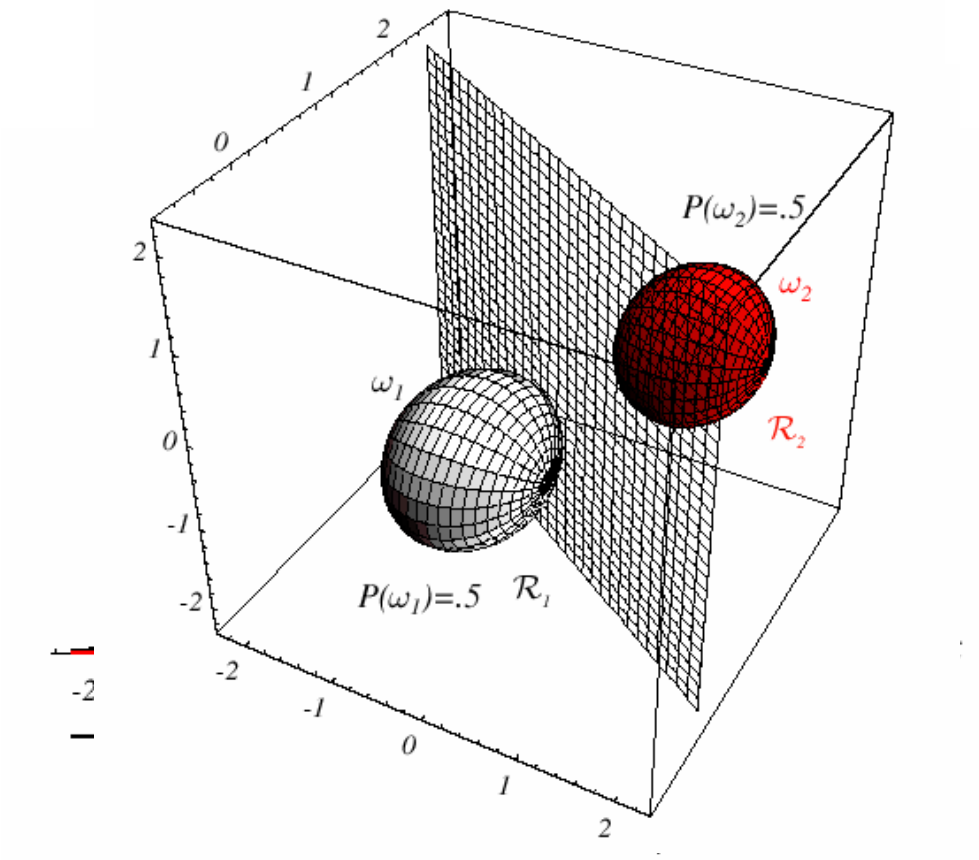
$$p(\mathbf{x}\,|\,\omega_i) \sim N(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$$

$$g_i(\mathbf{x}) = -\frac{1}{2}\left[(\mathbf{x}-\mu)^T \cdot \Sigma_i^{-1} \cdot (\mathbf{x}-\mu)\right] - \frac{d}{2}\ln 2\pi - \frac{1}{2}\ln\left|\Sigma_i^{-1}\right| + \ln P(\omega_i)$$

There are three distinct cases that can occur:

CASE 1: $\Sigma_i = \sigma^2 I$

Features are statistically independent, and all features have the same variance: Dist. are spherical in *d* dimensions, the boundary is a generalized hyperplane (linear discriminant) of *d-1* dimensions, and features create equal sized hyperspherical clusters. Examples of such hyperspherical clusters are:



The general form of the discriminant is then

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

If priors are the same:

$$g_i(\mathbf{x}) = -\frac{(\mathbf{x}-\boldsymbol{\mu})^T(\mathbf{x}-\boldsymbol{\mu})}{2\sigma^2}$$

Minimum Distance Classifier

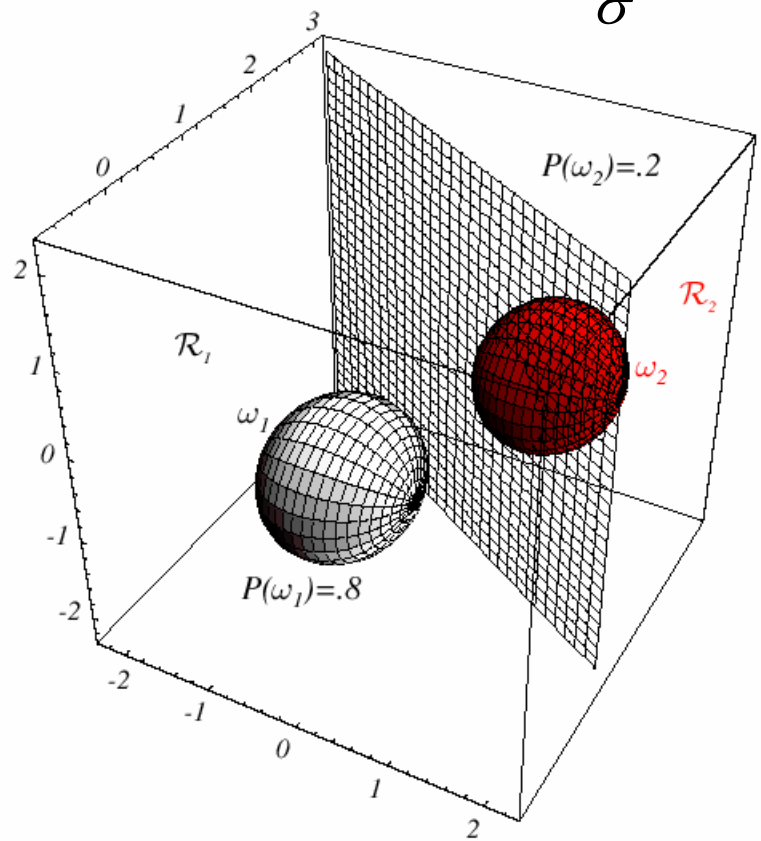
CASE 1: $\Sigma_i = \sigma^2 I$

- This case results in linear discriminants that can be written in the form

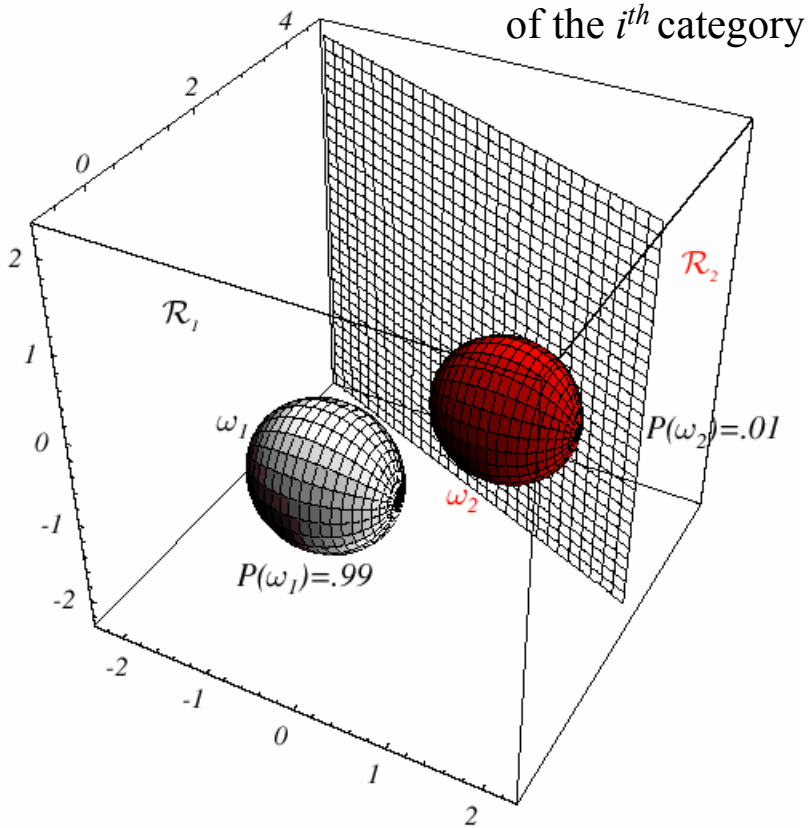
$$g_i(\mathbf{x}) = \mathbf{w}_i^T \mathbf{x} + w_{i0}$$

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

Threshold (***Bias***) of the i^{th} category



3-D case



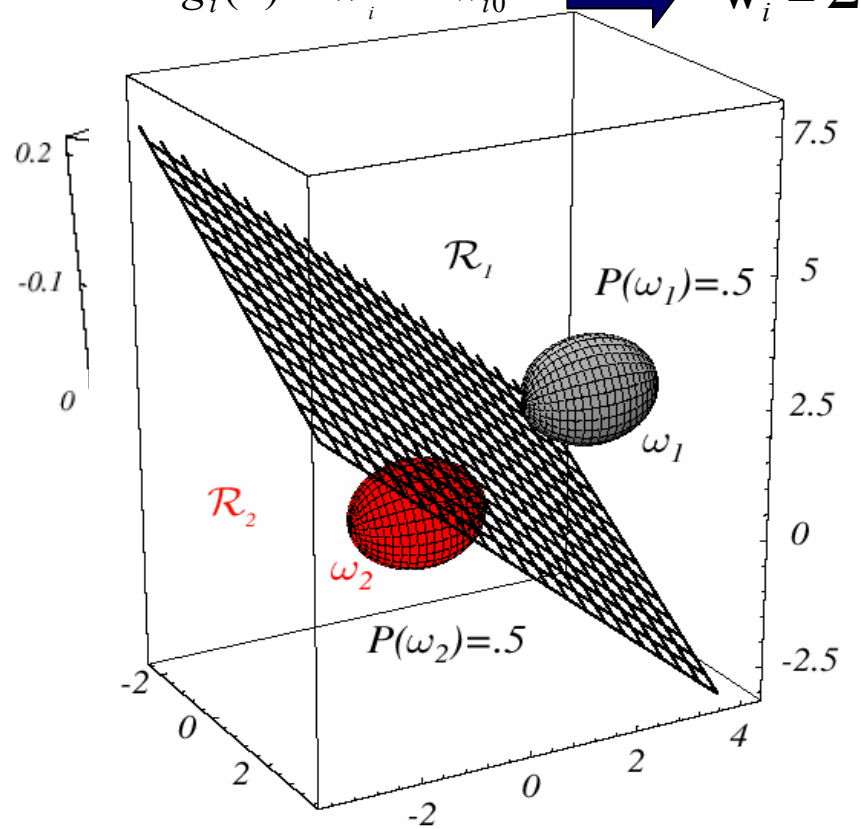
Note how priors shift the discriminant function away from the more likely mean !!!

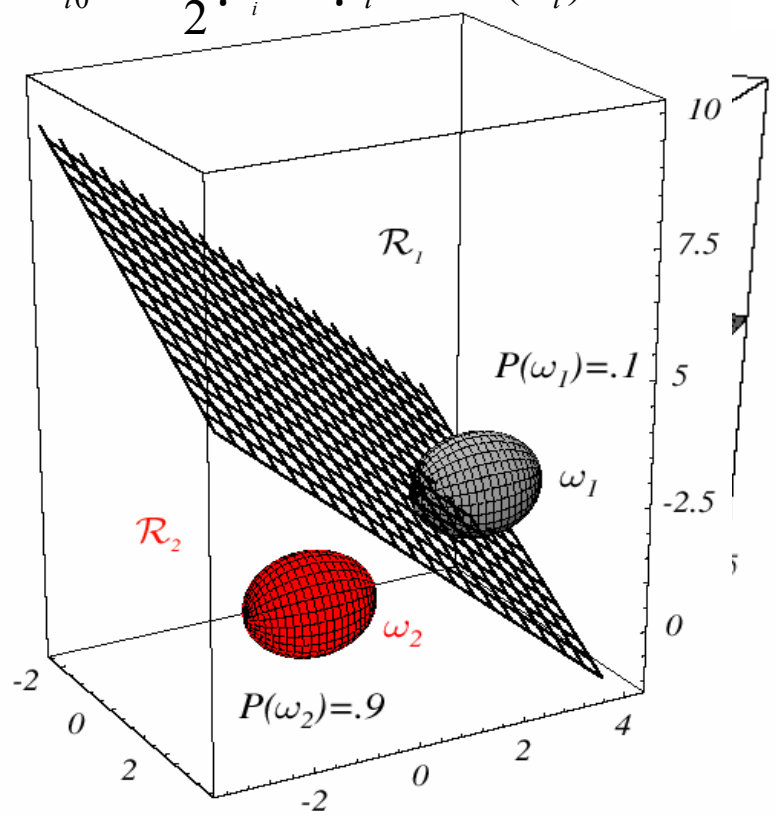
Case 2: $\Sigma_i = \Sigma$

Covariance matrices are arbitrary, but equal to each other for all classes. Features then form hyper-ellipsoidal clusters of equal size and shape. This also results in linear discriminant functions whose decision boundaries are again hyperplanes:

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

$$g_i(\mathbf{x}) = \mathbf{w}_i^T \mathbf{x} + w_{i0} \quad \Rightarrow \quad \mathbf{w}_i = \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_i, \quad w_{i0} = -\frac{1}{2}\boldsymbol{\mu}_i^T \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_i + \ln P(\omega_i)$$





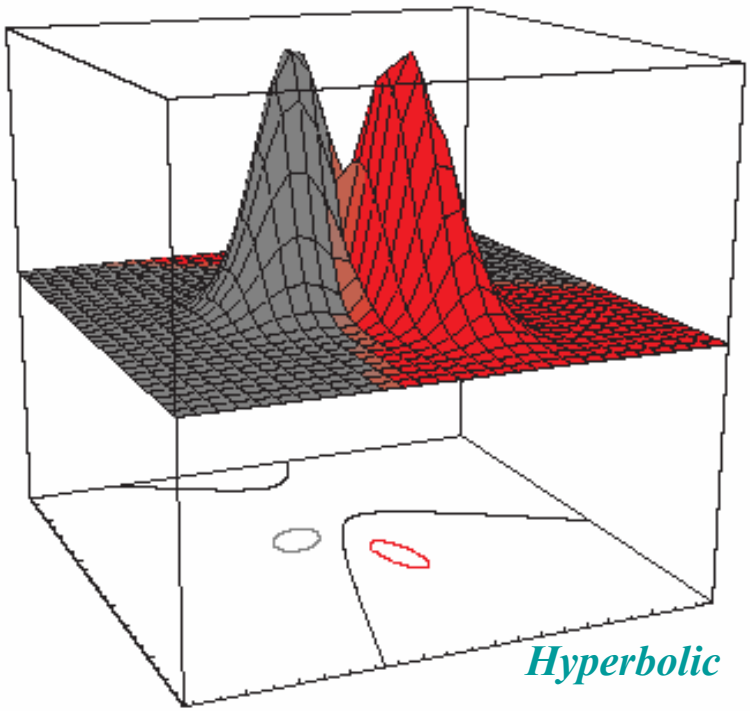
Case 3: Σ_i = Arbitrary

All bets are off !In two class case, the decision boundaries form *hyperquadratics.*
The discriminant functions are now, in general, quadratic (nor linear) and non-contiguous

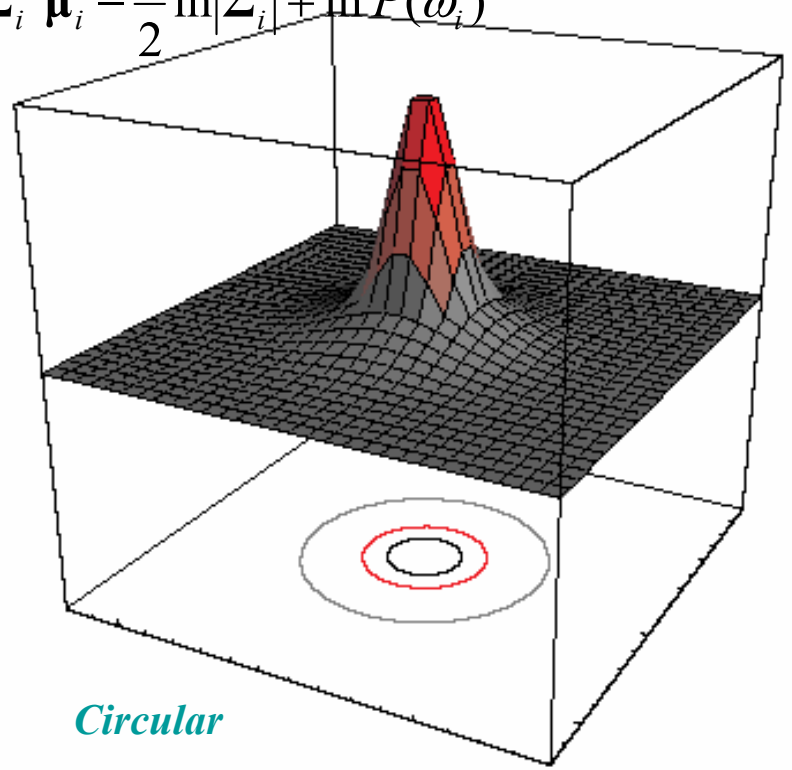
$$g_i(\mathbf{x}) = \mathbf{x}^T \mathbf{W}_i \mathbf{x} + \mathbf{w}_i^T \mathbf{x} + w_{i0}$$

$$\mathbf{W}_i = -\frac{1}{2}\boldsymbol{\Sigma}_i^{-1}, \quad \mathbf{w}_i = \boldsymbol{\Sigma}_i^{-1}\boldsymbol{\mu}_i$$

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



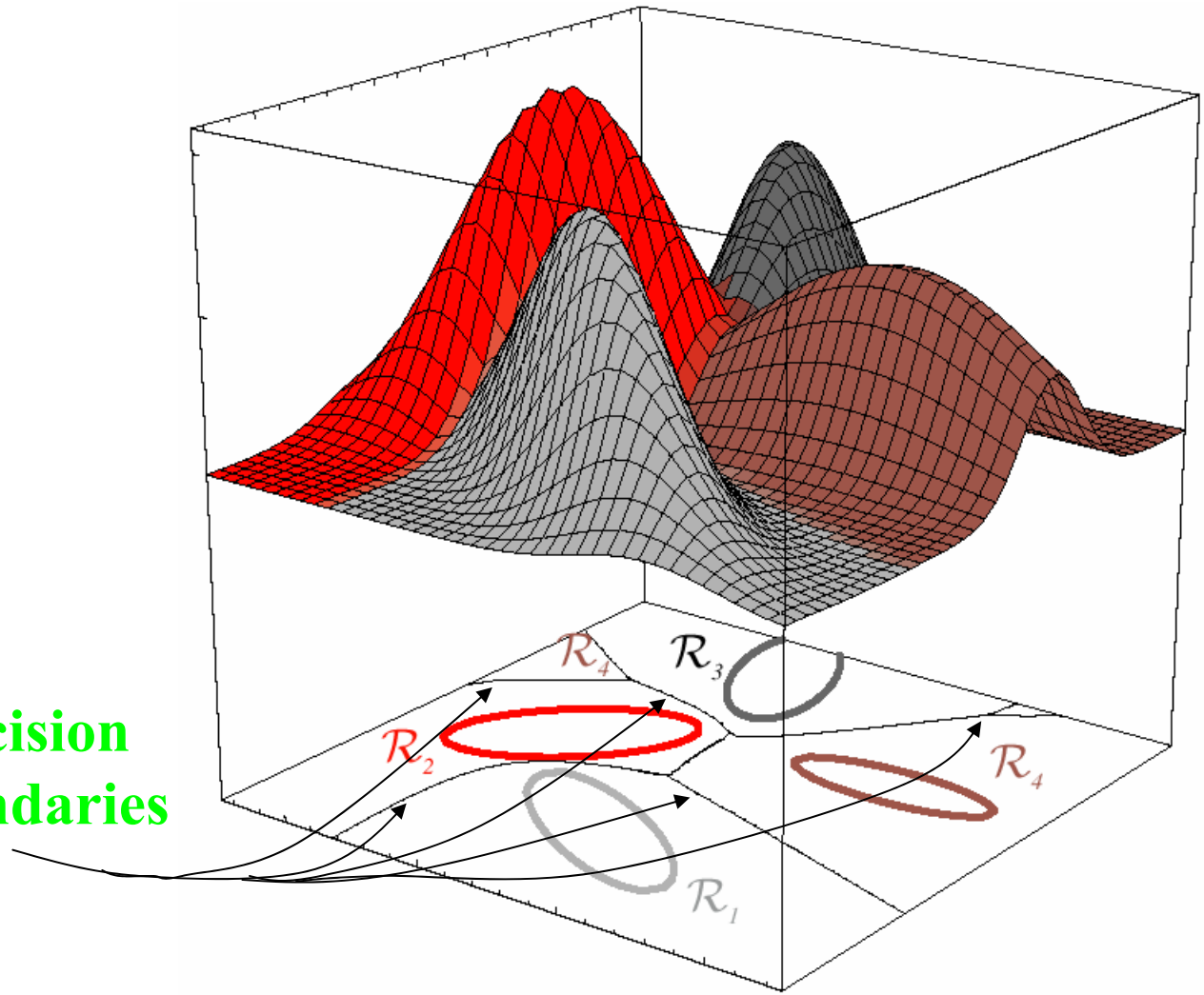
Hyperbolic



Circular

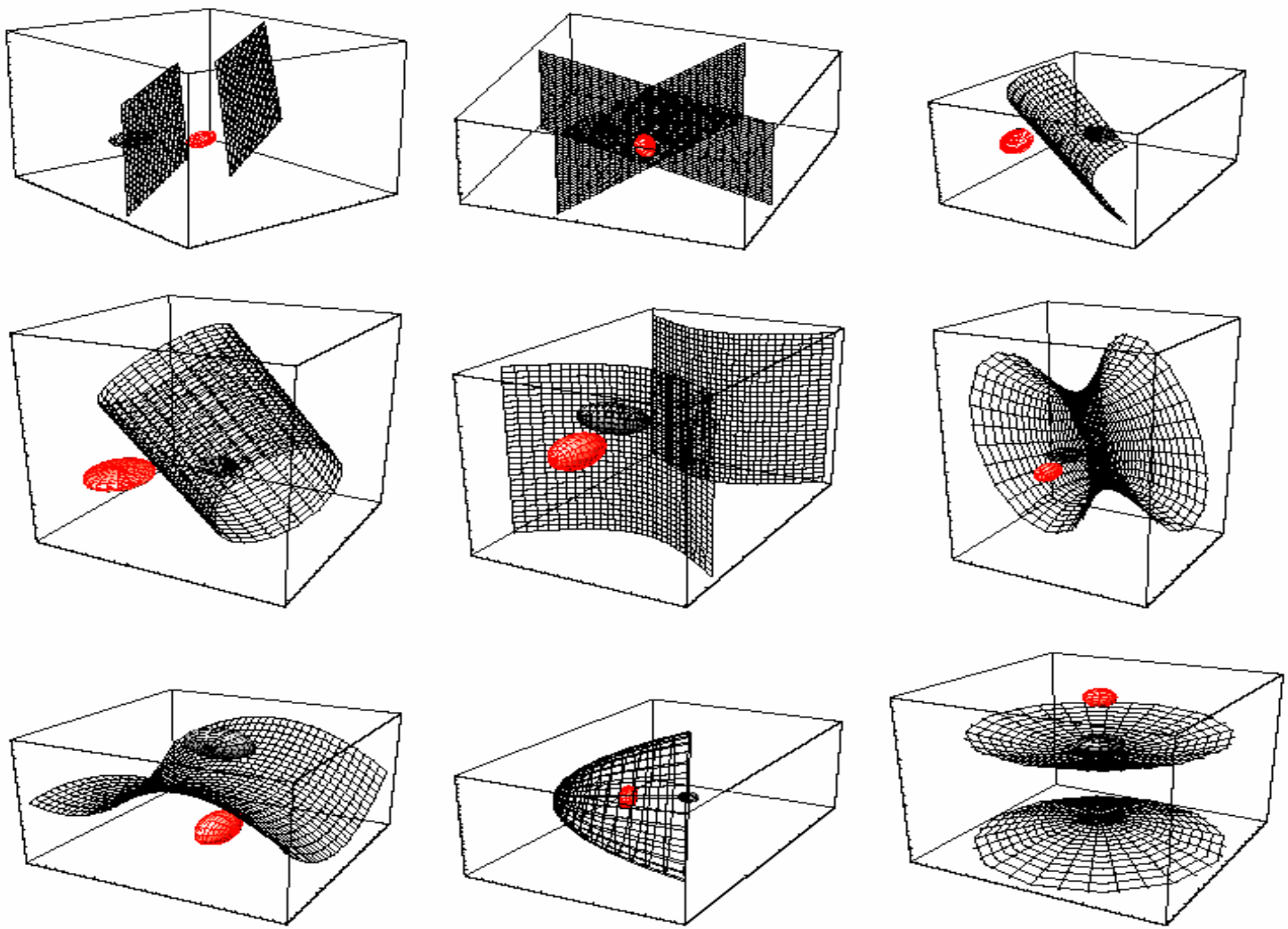
CASE 3: Σ_i = Arbitrary

For the multi class case, the boundaries will look even more complicated. As an example



CASE 3: Σ_i = Arbitrary

In 3-D



CONCLUSIONS

- The Bayes classifier for normally distributed classes is in general a quadratic classifier and can be computed
- The Bayes classifier for normally distributed classes with equal covariance matrices is a linear classifier
  - For normally distributed classes with equal covariance matrices and equal priors is a ***minimum – Mahalanobis distance classifier***
  - For normally distributed classes with equal covariance matrices proportional to the identity matrix and with equal priors is a ***minimum Euclidean distance classifier***
- Note that using a minimum Euclidean or Mahalanobis distance classifier implicitly makes certain assumptions regarding statistical properties of the data, which may or may not – and in general are not – true.
  - However, in many cases, certain simplifications and approximations can be made that warrant making such assumptions even if they are not true. The bottom line in practice in deciding whether the assumptions are warranted is *does the damn thing solve my classification problem…?*

Error Bounds

It is difficult, at best if possible, to analytically compute the error probabilities, Particularly when the decision regions are not contiguous. However, upper bounds for this error can be obtained:

The Chernoff bound and its approximation Bhattacharya bound are two such bounds that are often used. If the distributions are Gaussian, these expressions are relatively easier to compute➔ Often times even non-Gaussian cases are considered as Gaussian.