Bayesian Decision Theory

Dr. Shuang LIANG

School of Software Engineering
TongJi University
Fall, 2012

Today's Topics

- Bayesian Decision Theory
- Bayesian classification for normal distributions
- Error Probabilities and Integrals

Today's Topics

- ***Bayesian Decision Theory (Continuous Features)***
  - ***Minimum-Error-Rate Classification***
  - Minimum-Risk Classification
  - Discriminant Functions
- Bayesian classification for normal distributions
- Error Probabilities and Integrals

Minimum-Error-Rate Classification

- Goal
  - Minimizing the classification error probability

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

- Probability of error in Bayesian

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

- The Bayesian classifier is *optimal* with respect to *minimizing the classification error probability*

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

Make a Decision

- The bayesian decision theory is referred as the minimum-error-rate classification in general cases

- Make a decision?

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

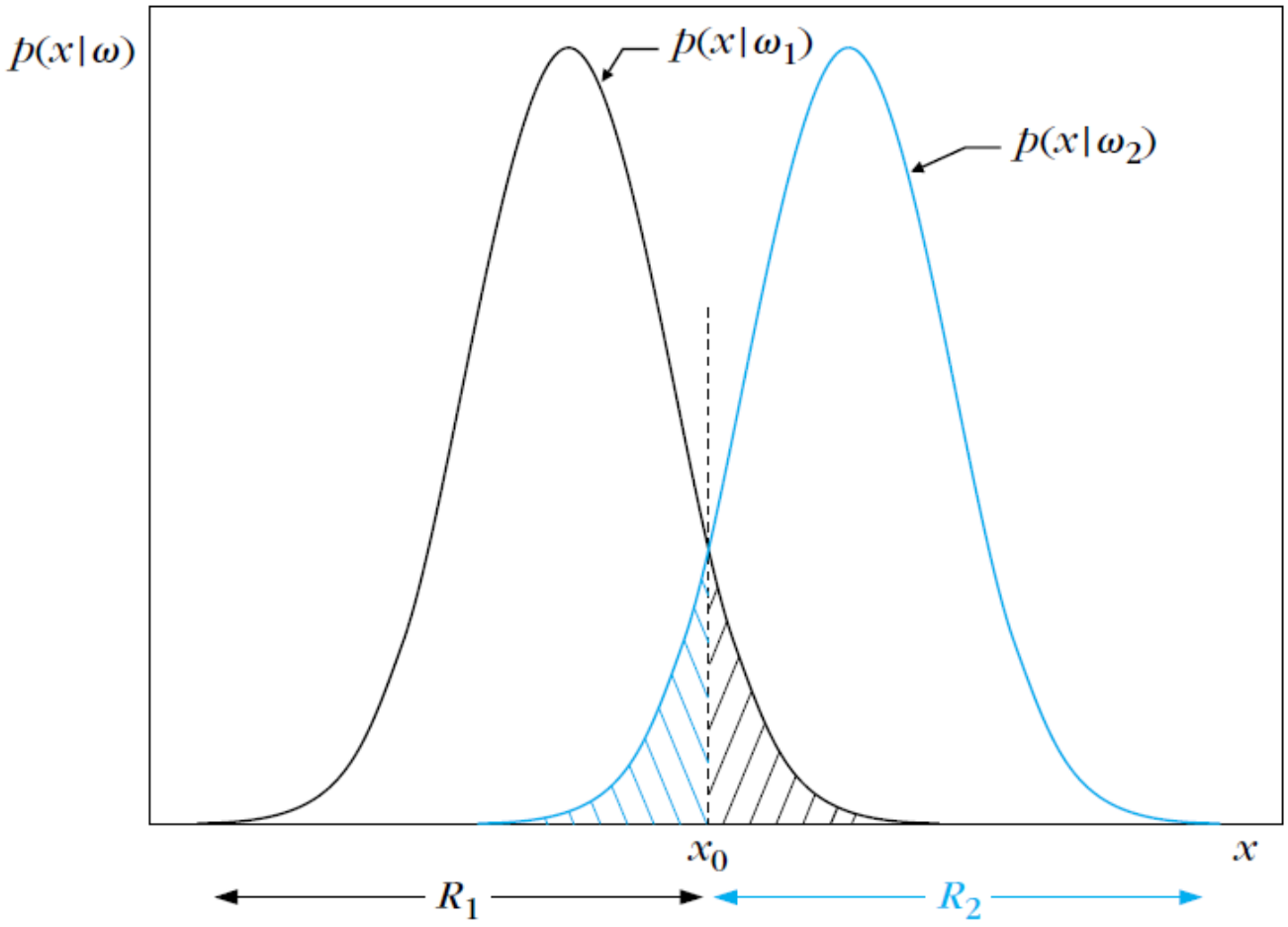
Other equivalence?

- How to get posteriori probability?

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

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

Probability of Error



Example

- Cancer cell recognition
  - Two classes
    - Normal: ω_1
    - Abnormal: ω_2
    - Prior probability known as
      - $P(\omega_1)$=0.9
      - $P(\omega_2)$=0.1
  - Cell to be classified
    - Feature x
    - Class-conditional probability density
      - $p(x|\omega_1)$=0.2
      - $p(x|\omega_2)$=0.4
  - Which class does x belong to?

Extension to Multiple classes

- The principle is the same as the two-category cases

- Decision rule

$$\text{if} \quad P(\omega_i|x) = \max_{j=1,2,..c} P(\omega_j|x), \qquad \text{then} \quad x \in \omega_i$$

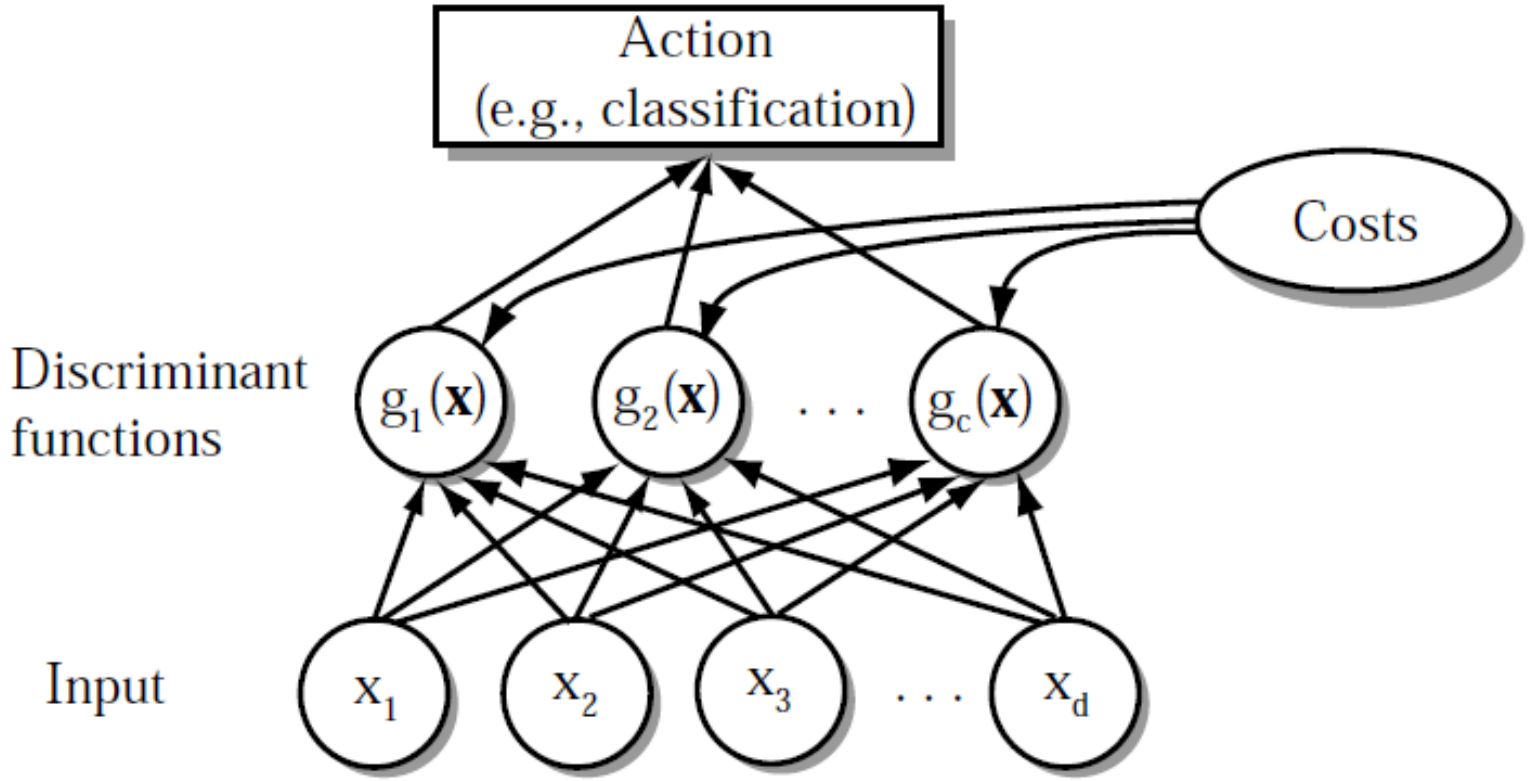
Or

Discriminant Function $g_i(x)$

$$\text{if} \quad p(x|\omega_i)P(\omega_i) = \max_{j=1,2,..c} p(x|\omega_j)P(\omega_j), \qquad \text{then} \quad x \in \omega_i$$

Extension to Multiple classes (cont.)

- Decision is made by comparing the discriminant functions of each class



Probability of Error Revisit

- Assume that the feature space is divided into regions as R1,R2,…Rc
- How to calculate the probability of error?
  - P(e)=?
- Any other solutions?

  P(e)=1-P(c)

Today's Topics

- ***Bayesian Decision Theory***
  - Minimum-Error-Rate Classification
  - ***Minimum-Risk Classification***
  - Discriminant Functions
- Bayesian classification for normal distributions
- Error Probabilities and Integrals

A More General 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
  - introduce a more general error function
    - *loss* function (i.e., associate "costs" with actions)

Bayesian Decision Theory (cont.)

- Let $\{\omega_1, \ldots, \omega_c\}$ be the finite set of c states of nature (*classes, categories*)
- Let $\{\alpha_1, \ldots, \alpha_a\}$ be the finite set of a possible actions
- Let $\lambda(\alpha_i | \omega_j)$ be the *loss* incurred for taking action i when the state of nature is ω_j
- Let $\mathbf{x}$ be the d-component vector-valued random variable called the *feature vector* .

Bayesian Decision Theory (cont.)

- $p(\mathbf{x}|\omega_j)$ is the class-conditional probability density function

- $P(\omega_j)$ is the prior probability that nature is in state ω_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 α_i
- If the true state of nature is ω_j , we incur the loss $\lambda(\alpha_i \mid \omega_j)$
- The expected loss with taking action α_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* $\alpha(x)$ 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 α_i for which $R(\alpha_i|x)$ 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
  - α_1: deciding ω_1,
  - α_2: deciding ω_2,
  - $\lambda_{ij} = \lambda(\alpha_i | \omega_j)$.

- 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 ω_1 i

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

Other equivalence?

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

Minimum-Error-Rate Classification

- Actions are decisions on classes (α_i is deciding ω_i).
- If action α_i is taken and the true state of nature is ω_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*

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

(all errors are equally costly).

- Conditional risk becomes

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

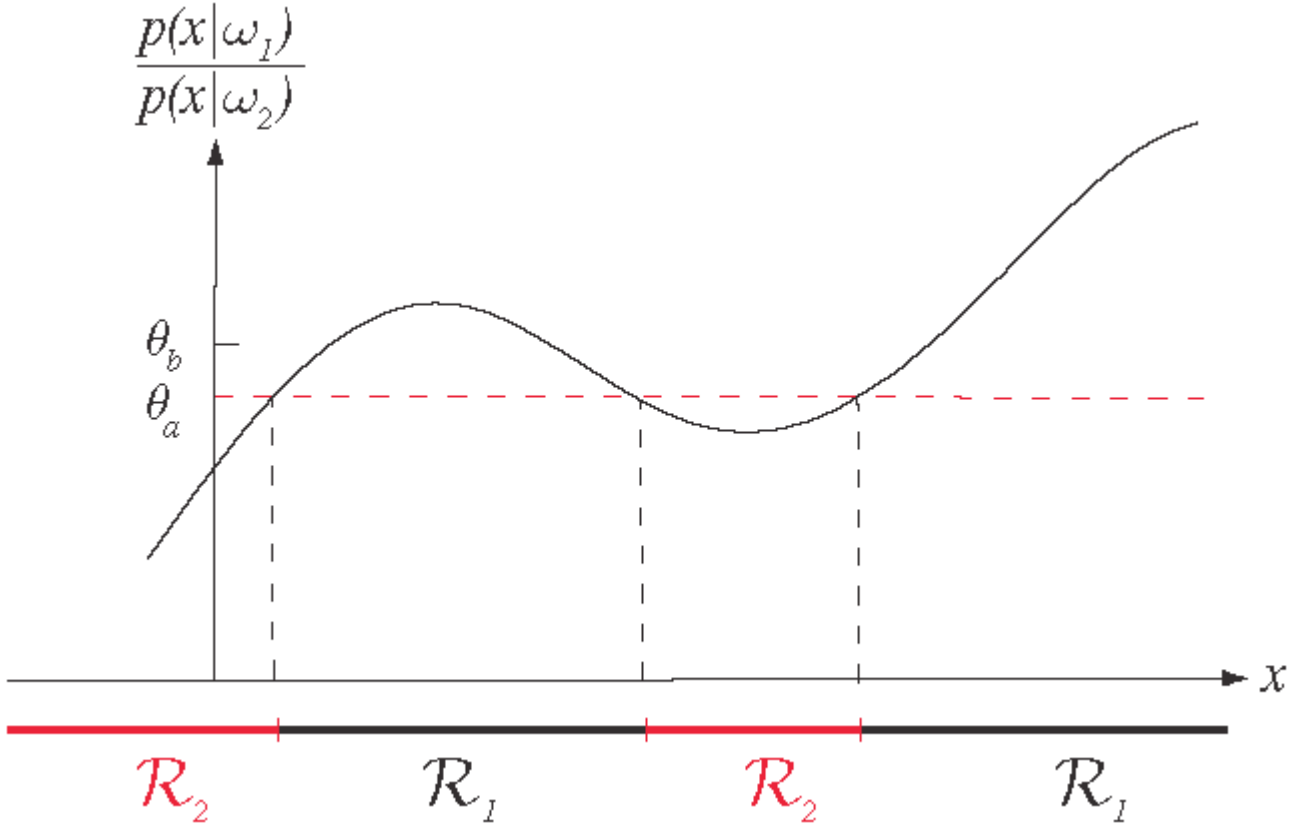
Minimum-Error-Rate Classification

- Minimizing the risk requires maximizing $P(\omega_i|x)$ and results in the *minimum-error decision rule*

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

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

Minimum-Error-Rate Classification



- The likelihood ratio $p(\mathbf{x}|\omega_1)/p(\mathbf{x}|\omega_2)$. The threshold θ_a is computed using the priors $P(\omega_1) = 2/3$ and $P(\omega_2) = 1/3$, and a zero-one loss function. If we penalize mistakes in classifying ω_2 patterns as ω_1 more than the converse, we should increase the threshold to θ_b.

Example Revisited

- Cancer cell recognition
  - Two classes
    - Normal: ω_1
    - Abnormal: ω_2
    - Prior probability known as
      - $P(\omega_1)$=0.9
      - $P(\omega_2)$=0.1
  - Cell to be classified
    - Feature x
    - Class-conditional probability density
      - $p(x|\omega_1)$=0.2
      - $p(x|\omega_2)$=0.4
  - Which class does x belong to?

Loss functions

$\lambda_{11}=0$

$\lambda_{12}=6$

$\lambda_{21}=0$

$\lambda_{22}=0$

Today's Topics

- ***Bayesian Decision Theory***
  - Minimum-Error-Rate Classification
  - Minimum-Risk Classification
  - ***Discriminant Functions***
- Bayesian classification for normal distributions
- Error Probabilities and Integrals

Discriminant Functions

- A useful way of representing classifiers is through *discriminant functions* $g_i(x)$, $i = 1, \ldots, c$, where the classifier assigns a feature vector $\mathbf{x}$ to class ω_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* ($R_1, \ldots, R_c$), separated by *decision boundaries*.

- Note that the results do not change even if we replace every $g_i(x)$ by $f(g_i(x))$ where $f(\cdot)$ is a monotonically increasing function (e.g., logarithm).

- This may lead to significant analytical and computational simplifications.

Discriminant Functions

- In particular, for minimum-error-rate classification, any of the following choices gives identical classification results, but some can be much simpler to understand or to compute than others

$$g_i(\mathbf{x}) = P(\omega_i|\mathbf{x}) = \frac{p(\mathbf{x}|\omega_i)P(\omega_i)}{\sum_{j=1}^{c} p(\mathbf{x}|\omega_j)P(\omega_j)}$$

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

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

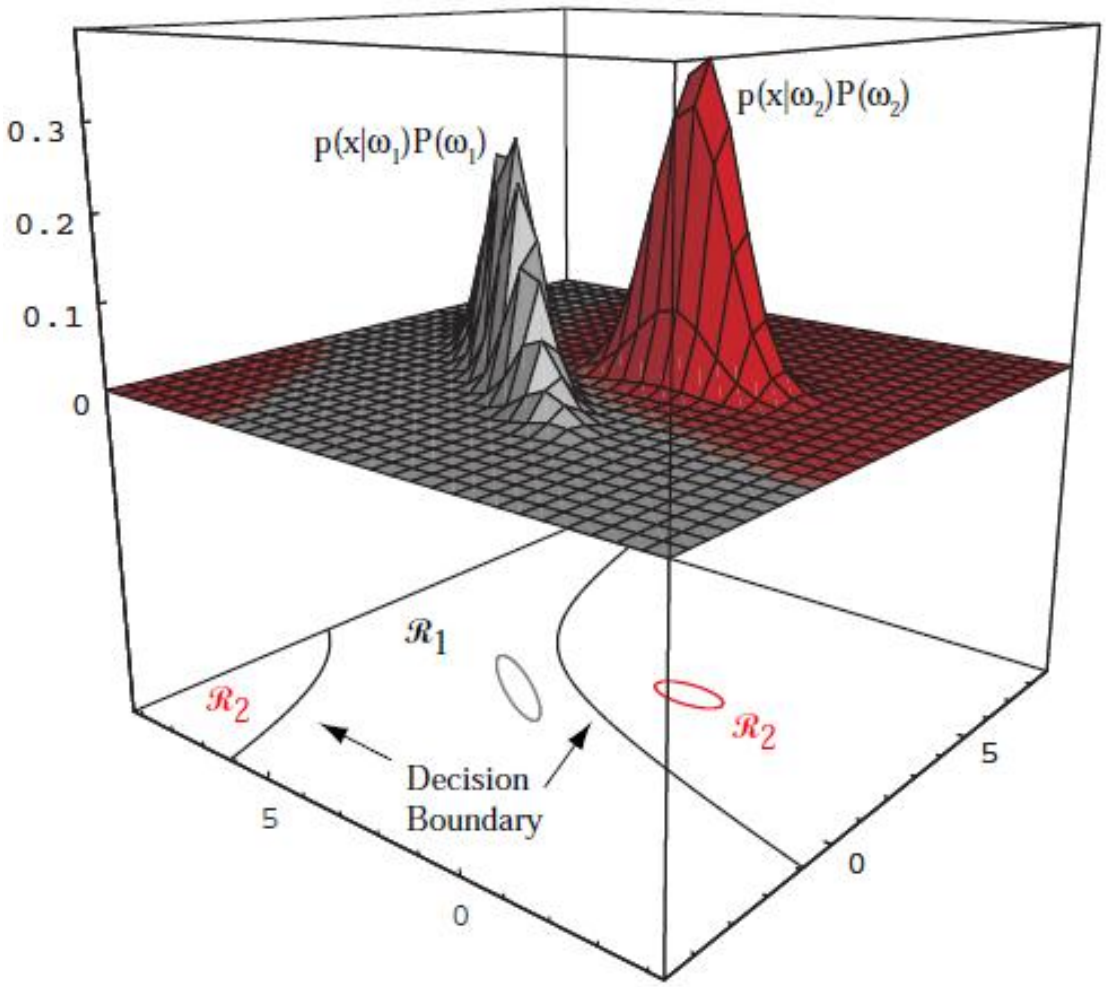
Discriminant Functions

- Even though the discriminant functions can be written in a variety of forms, the decision rules are equivalent

- The effect of any decision rule is to divide the feature space into *c decision regions,* $R_1,\ldots,R_c$. If $g_i(\mathbf{x}) > g_j(\mathbf{x})$ for all $j \neq i$, then **x** is in R_i, and the decision rule calls for us to assign **x to** ω_i.

- The regions are separated by *decision boundaries,* surfaces in feature space where ties occur among the largest discriminant functions

Discriminant Functions



In this two-dimensional two-category classifier, the probability densities are Gaussian, the decision boundary consists of two hyperbolas, and thus the decision region R_2 is not simply connected.

Discriminant Functions

- Instead of using two discriminant functions g_1 *and* g_2 and assigning **x** to ω_1 if $g_1 > g_2$, it is more common to define a single discriminant function

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

- Decide ω_1 if $g(\boldsymbol{x}) > 0$; otherwise decide ω_2.

- Other various forms

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

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

Thank you for all your attention