CpE 646 Pattern Recognition and Classification

Prof. Hong Man

Department of Electrical and Computer Engineering
Stevens Institute of Technology

STEVENS Institute of Technology

Visual Information Environment Laboratory

Bayesian Decision Theory

Chapter 2 (Section 2.1 – 2.4) Outline:

- Introduction
- Bayesian Decision Theory – Continuous Features
- Minimum-Error-Rate Classification
- Classifiers, Discriminant Functions and Decision Surfaces

Introduction

- Consider the sea bass/salmon example
- State of nature is a random variable ω
  - $\omega = \omega_1 \Rightarrow$ sea bass
  - $\omega = \omega_2 \Rightarrow$ salmon
- A priori probability (or prior) $P(\omega_1)$ for sea bass, and $P(\omega_2)$ for salmon
  - The prior can be obtained through observation
  - If the catch of salmon and sea bass is equiprobable
    - $P(\omega_1) = P(\omega_2)$ (uniform priors)
    - $P(\omega_1) + P(\omega_2) = 1$ (exclusivity – no more, and exhaustivity – no less)

Introduction

- Decision rule with only the prior information (no observation)
  - If $P(\omega_1) > P(\omega_2)$, always decide $\omega=\omega_1$
  - Otherwise, always decide $\omega=\omega_2$
  - The probability of error is $min[P(\omega_1), P(\omega_2)]$

Introduction

- Decision with feature observation
  - Feature is a random variable or random variable vector x, e.g. the lightness of a fish.
  - Feature value distribution depends on the state of nature $\omega \Rightarrow$ the class-conditional probability density $p(x|\omega)$.
  - $p(x|\omega_1)$ and $p(x|\omega_2)$ describe the difference in lightness between populations of sea bass and salmon

Introduction

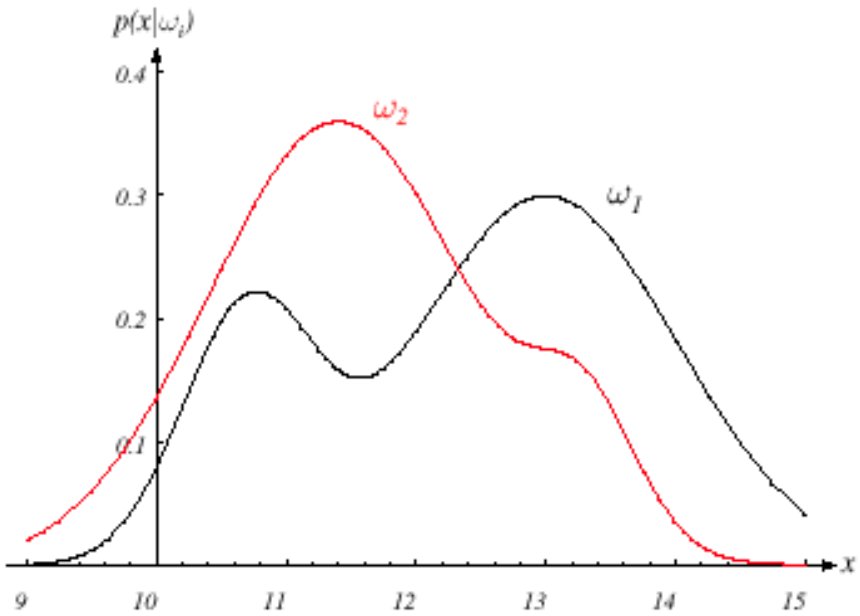

FIGURE 2.1. Hypothetical class-conditional probability density functions show the probability density of measuring a particular feature value x given the pattern is in category ω_i. If x represents the lightness of a fish, the two curves might describe the difference in lightness of populations of two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

Introduction

- Joint probability density

$$p(\omega_j,x) = P(\omega_j|x)p(x) = p(x|\omega_j)P(\omega_j)$$

- Bayes formula

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

  - Posterior = (Likelihood × Prior) / Evidence
  - Where in case of two categories

$$p(x) = \sum_{j=1}^{2} p(x \mid \omega_j)P(\omega_j)$$

Introduction

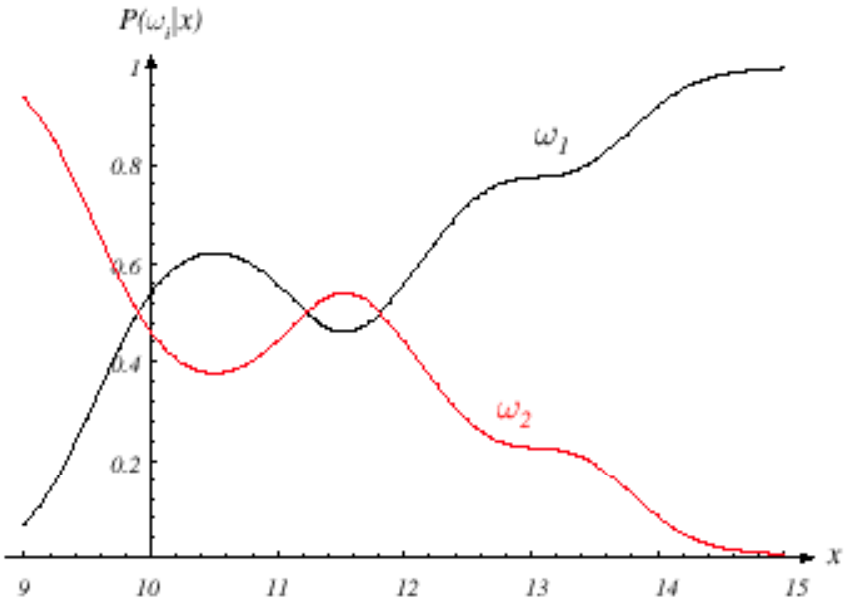

FIGURE 2.2. Posterior probabilities for the particular priors $P(\omega_1) = 2/3$ and $P(\omega_2) = 1/3$ for the class-conditional probability densities shown in Fig. 2.1. Thus in this case, given that a pattern is measured to have feature value $x = 14$, the probability it is in category ω_2 is roughly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

Introduction

- Decision given the posterior probabilities
  - $\boldsymbol{x}$ is an observation for which:

    if $P(\omega_1 | x) > P(\omega_2 | x) \Rightarrow$ true state of nature = ω_1

    if $P(\omega_1 | x) \leq P(\omega_2 | x) \Rightarrow$ true state of nature = ω_2
  - The probability of error is :

    $P(error | x) = P(\omega_1 | x)$ if we decide ω_2

    $P(error | x) = P(\omega_2 | x)$ if we decide ω_1

    The overall probability of error is

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

Introduction

- To minimize the probability of error is to minimize the individual $P(error|x)$. This leads to Bayes decision rule:
  - Decide ω_1 if $P(\omega_1 \mid x) > P(\omega_2 \mid x)$; otherwise decide ω_2
  - Then the probability of error is:

    $$P(error|x) = min\ [P(\omega_1|\ x),\ P(\omega_2|\ x)]$$
- In Bayes decision rule, the evidence $p(x)$ is not important because it is same to all observations, then the rule can be expressed as
  - Decide ω_1 if $p(\mathrm{x}|\omega_1)P(\omega_1) > p(\mathrm{x}|\omega_2)P(\omega_2)$ otherwise decide ω_2
- If the priors $P(\omega_1) = P(\omega_2)$, then the rule becomes
  - Decide ω_1 if $p(x|\omega_1) > p(x|\omega_2)$ otherwise decide ω_2

Bayesian Decision Theory – Continuous Features

- Generalization of the preceding ideas
  - Allowing the use of more than one feature
  - Allowing more than two states of nature
  - Allowing actions and not only decide on the state of nature (classification)
    - For example, allowing the possibility of rejection, i.e. refusing to make a decision in close or bad cases!
  - Introduce a loss of function which is more general than the probability of error
    - The loss function states how costly each action taken will be

Bayesian Decision Theory

- Let $\Omega=\{\omega_1, \omega_2, \ldots, \omega_c\}$ be the set of c states of nature (or "categories", "classes")
- Let $\{\alpha_1, \alpha_2, \ldots, \alpha_a\}$ be the set of possible actions
- Let $\lambda(\alpha_i|\omega_j)$ be the loss incurred for taking action α_i when the state of nature is ω_j
- Let $\boldsymbol{x}$ be the observation feature vector, and $\boldsymbol{x}$ is in a d-dimensional Euclidean space R^d, called the feature space.

Bayesian Decision Theory

- The posterior probability is

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

where the evidence is

$$p(x) = \sum_{j=1}^{c} p(x \mid \omega_j)P(\omega_j)$$

Bayesian Decision Theory

- The expected loss (or risk) associated with taking action α_i (conditional risk) is

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

- Overall risk

$$R = \sum_{i=1}^{a} \int_{\Re_i} R(\alpha_i \mid x)\, p(x) dx$$

- Minimizing $R \Leftrightarrow$ Minimizing $R(\alpha_i|x)$

Bayesian Decision Theory

- To minimize the overall risk,
  - compute the conditional risk $R(\alpha_i|x)$ for every action α_i and $i=1,\ldots,a$,
  - select the action α_i for which $R(\alpha_i|x)$ is minimum
  - R in this case is called the Bayes risk, denoted as R^*, and it is the best performance that can be achieved!

Bayesian Decision Theory

- Two-category classification
  - α_1 : deciding ω_1
  - α_2 : deciding ω_2
  - $\lambda_{ij} = \lambda(\alpha_i \mid \omega_j)$ is the loss incurred for deciding ω_i when the true state of nature is ω_j
  - Conditional risk:

$$R(\alpha_1 \mid x) = \lambda_{11}P(\omega_1 \mid x) + \lambda_{12}P(\omega_2 \mid x)$$

$$R(\alpha_2 \mid x) = \lambda_{21}P(\omega_1 \mid x) + \lambda_{22}P(\omega_2 \mid x)$$

Bayesian Decision Theory

- The rule is:
  - if $R(\alpha_1 \mid x) < R(\alpha_2 \mid x)$ then take action α_1: "decide ω_1"
  - Otherwise take action α_2: "decide ω_2"
- To express the posterior probabilities in terms of likelihoods and priors, the rule becomes
  - if $(\lambda_{21}- \lambda_{11})\, p(x \mid \omega_1)\, P(\omega_1) > (\lambda_{12}- \lambda_{22})\, p(x \mid \omega_2)\, P(\omega_2)$, decide ω_1
  - Otherwise decide ω_2

Bayesian Decision Theory

- Normally the loss of making a mistake is higher than the loss of making a correct action, so (λ_{21}- λ_{11}) and (λ_{12}- λ_{22}) are positive
- The rule is equivalent to:
  - take action α_1 (decide ω_1) if

$$\frac{p(x \mid \omega_1)}{p(x \mid \omega_2)} > \frac{\lambda_{12} - \lambda_{22}}{\lambda_{21} - \lambda_{11}} \cdot \frac{P(\omega_2)}{P(\omega_1)}$$

  - otherwise, take action α_2 (decide ω_2)
- $p(x \mid \omega_1)/p(x \mid \omega_2)$ is called likelihood ratio.

Bayesian Decision Theory

- Optimal decision property
  - If the likelihood ratio exceeds the threshold value independent of the input pattern $\boldsymbol{x}$, we can take optimal actions

Exercise

Select the optimal decision where:

$\Omega=\{\omega_1, \omega_2\}$

$p(x|\omega_1) \Rightarrow N(2, 0.5)$

$p(x|\omega_2) \Rightarrow N(1.5, 0.2)$

where $N(\mu, \sigma^2)$ is Normal distribution, μ is mean and σ^2 is variance

$P(\omega_1) = 2/3$

$P(\omega_2) = 1/3$

$$\lambda = \begin{bmatrix} 0 & 10 \\ 5 & 0 \end{bmatrix}$$

Minimum-Error-Rate Classification

- Actions are decisions on classes

  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 $i \neq j$.

- Seek a decision rule that minimizes the *probability of error* which is the *error rate*

Minimum-Error-Rate Classification

- The loss function of this case is symmetrical or zero-one loss function

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

- The risk corresponding to this loss is the average probability of error,

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

$$= \sum_{j \neq i} P(\omega_j \mid x)$$

$$= 1 - P(\omega_i \mid x)$$

Minimum-Error-Rate Classification

- $P(\omega_i \mid x)$ is the conditional probability that action α_i is correct.
- Minimize the risk requires maximize $P(\omega_i \mid x)$, since $R(\alpha_i \mid x) = 1 - P(\omega_i \mid x)$
- For minimum error rate
  - Decide $\boldsymbol{\omega_i}$ if $P(\boldsymbol{\omega_i} \mid x) > P(\boldsymbol{\omega_j} \mid x) \;\; \forall j \neq i$
- This is the same rule as Bayes decision rule, i.e. Bayes decision gives the minimum error under the zero-one loss.

Minimum-Error-Rate Classification

- Regions of decision and zero-one loss function

Let $\frac{\lambda_{12}-\lambda_{22}}{\lambda_{21}-\lambda_{11}}\cdot\frac{P(\omega_2)}{P(\omega_1)}=\theta_\lambda$

then decide ω_1 **if:** $\frac{p(x\mid\omega_1)}{p(x\mid\omega_2)}>\theta_\lambda$

Minimum-Error-Rate Classification

$$\text{if } \lambda = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \text{ then } \theta_\lambda = \frac{P(\omega_2)}{P(\omega_1)} = \theta_a$$

$$\text{if } \lambda = \begin{pmatrix} 0 & 2 \\ 1 & 0 \end{pmatrix} \text{ then } \theta_\lambda = \frac{2P(\omega_2)}{P(\omega_1)} = \theta_b$$

Minimum-Error-Rate Classification

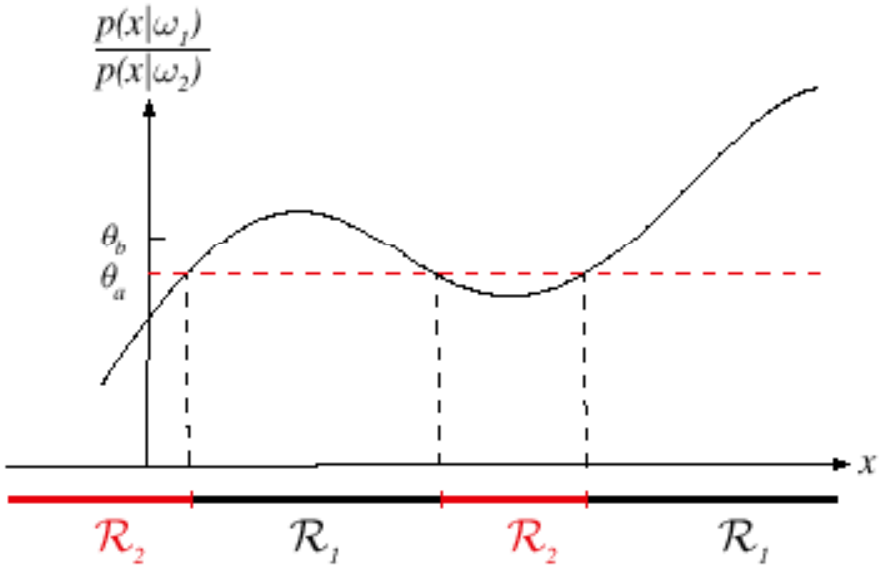

FIGURE 2.3. The likelihood ratio $p(x|\omega_1)/p(x|\omega_2)$ for the distributions shown in Fig. 2.1. If we employ a zero-one or classification loss, our decision boundaries are determined by the threshold θ_a. If our loss function penalizes miscategorizing ω_2 as ω_1 patterns more than the converse, 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.

Classifiers, Discriminant Functions and Decision Surfaces

- A pattern classifier can be represented as a set of discriminant functions $g_i(x)$, $i = 1, \ldots, c$
  - The classifier assigns a feature vector x to class ω_i if $g_i(x) > g_j(x) \;\; \forall j \neq i$
- Such classifier can be viewed as a network or machine that computes c discriminant functions and select the category with the largest discriminant.

Discriminant Functions

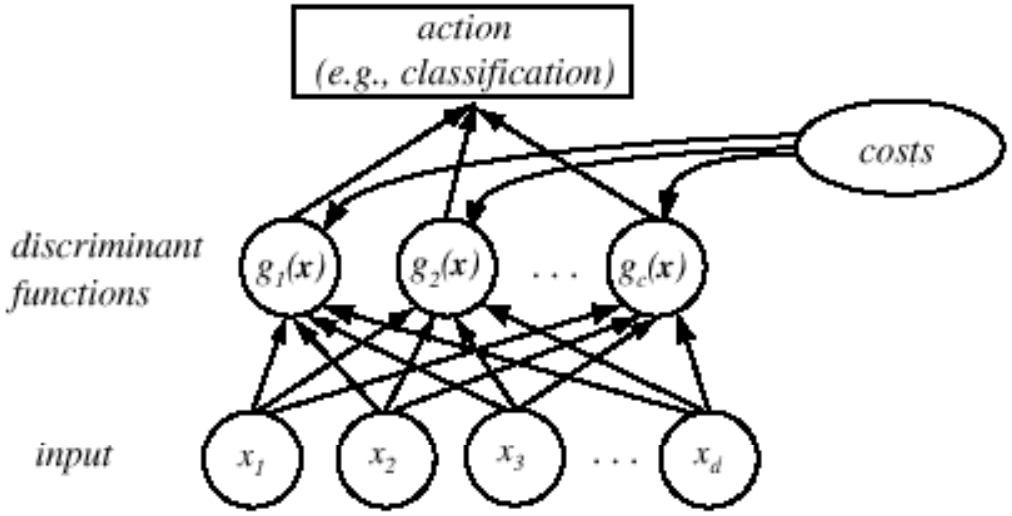

FIGURE 2.5. The functional structure of a general statistical pattern classifier which includes d inputs and c discriminant functions $g_i(\mathbf{x})$. A subsequent step determines which of the discriminant values is the maximum, and categorizes the input pattern accordingly. The arrows show the direction 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.

Discriminant Functions

- To represent a general Bayes classifier, let

$$g_i(x) = - R(\alpha_i \mid x)$$

(max. discriminant corresponds to min. conditional risk)

- To represent a minimum error rate classifier, let

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

(max. discrimination corresponds to max. posterior!)

Discriminant Functions

- The choice of discriminant functions is not unique
- The classification will not change if the discriminant function $g_i(x)$ is replaced by $f(g_i(x))$ where $f(\cdot)$ is a monotonically increasing function.
  - This may lead to simplification in discriminant functions.
  - For minimun error rate classification, the discriminant function can be expressed as

$$g_i(x) = p(x \mid \omega_i)\, P(\omega_i)$$

$$\text{or } g_i(x) = \ln p(x \mid \omega_i) + \ln P(\omega_i)$$

Discriminant Functions

- Any decision rule is to divide a feature space into c decision regions

  if $g_i(x) > g_j(x)\ \ \forall j \neq i$ then x is in R_i

  (R_i means assign x to ω_i)

Discriminant Functions

- The two-category case
  - A classifier is a "dichotomizer" that has two discriminant functions g_1 and g_2

    Let $g(x) = g_1(x) - g_2(x)$, then

    decide ω_1 if $g(x) > 0$; otherwise decide ω_2
- The computation of $g(x)$ for minimum-error-rate classification can be

$$g(x) = P(\omega_1 \mid x) - P(\omega_2 \mid x)$$

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

Discriminant Functions

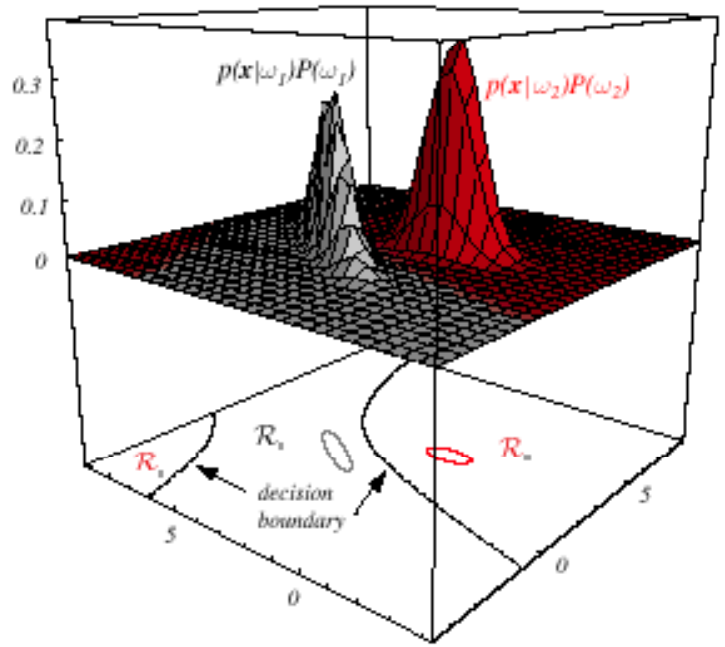

FIGURE 2.6. In this two-dimensional two-category classifier, the probability densities are Gaussian, the decision boundary consists of two hyperbolas, and thus the decision region $\mathcal{R}_2$ is not simply connected. The ellipses mark where the density is $1/e$ times that at the peak of the distribution.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