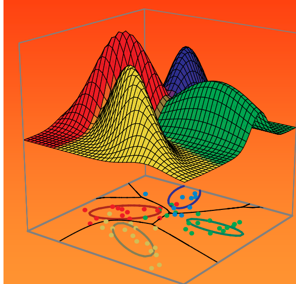
Pattern classification

Biometrics
CSE 190-a
Lecture 3



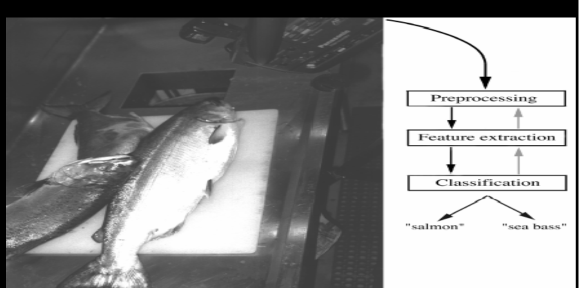
Pattern Classification

An Example

- "Sorting incoming Fish on a conveyor according to species using optical sensing"

Species — Sea bass
Species — Salmon

- Problem Analysis
  - Set up a camera and take some sample images to extract features
    - Length
    - Lightness
    - Width
    - Number and shape of fins
    - Position of the mouth, etc…
    - This is the set of all suggested features to explore for use in our classifier!

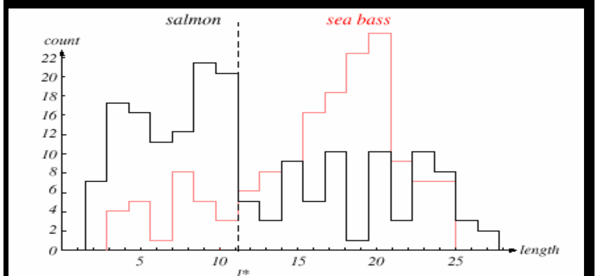


- Classification
  - Select the length of the fish as a possible feature for discrimination

7

Classification

Select the length of the fish as a possible feature for discrimination

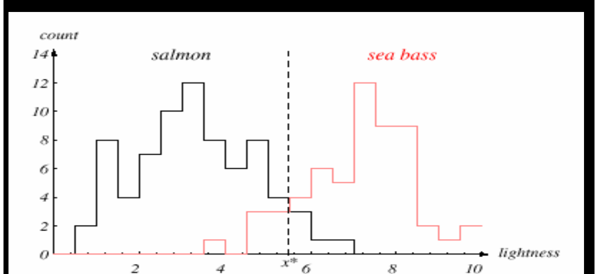


8

The length is a poor feature alone!

Select the lightness as a possible feature.

9



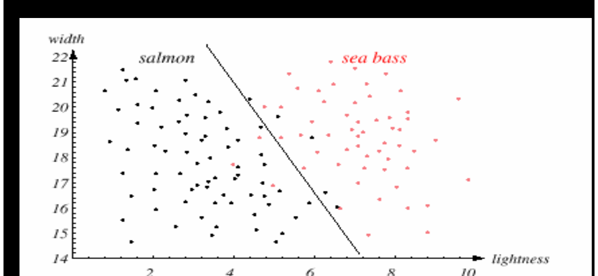
10

- Adopt the lightness and add the width of the fish

Fish → $x^T = [x_1, x_2]$

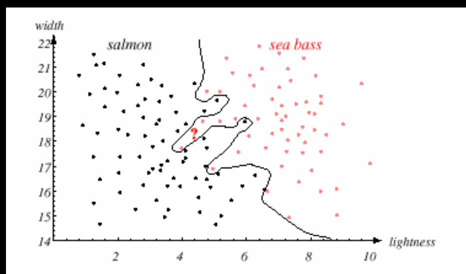
Lightness   Width

11



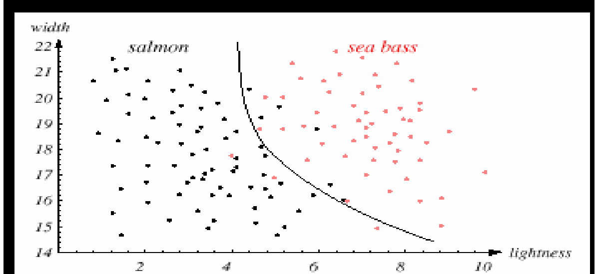
12

- We might add other features that are not correlated with the ones we already have. A precaution should be taken not to reduce the performance by adding such “noisy features”
- Ideally, the best decision boundary should be the one which provides an optimal performance such as in the following figure:



- However, our satisfaction is premature because the central aim of designing a classifier is to correctly classify novel input

Issue of generalization!



Bayesian Decision Theory
Continuous Features
(Sections 2.1-2.2)

Introduction

- The sea bass/salmon example
  - State of nature, prior
    - State of nature is a random variable
    - The catch of salmon and sea bass is equiprobable
      - $P(\omega_1), P(\omega_2)$ *Prior probabilities*
      - $P(\omega_1) = P(\omega_2)$ (uniform priors)
      - $P(\omega_1) + P(\omega_2) = 1$ (exclusivity and exhaustivity)

- Decision rule with only the prior information
  - Decide ω_1 if $P(\omega_1) > P(\omega_2)$ otherwise decide ω_2
- Use of the class–conditional information
- $P(x \mid \omega_1)$ and $P(x \mid \omega_2)$ describe the difference in lightness between populations of sea-bass and salmon

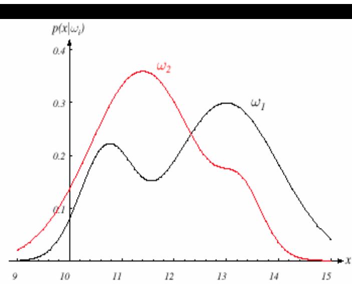


FIGURE 2.1. Hypothetical class-conditional probability density functions show the probability density of measuring a particular feature value x given the pattern is in category ω_i. If x represents the lightness of a fish, the two curves might describe the difference in lightness of populations of two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

- Posterior, likelihood, evidence
  - $P(\omega_j \mid x) = (P(x \mid \omega_j) * P(\omega_j)) / P(x)$ **(BAYES RULE)**
  - In words, this can be said as:
    Posterior = (Likelihood * Prior) / Evidence
  - Where in case of two categories

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

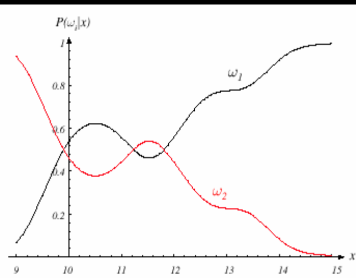
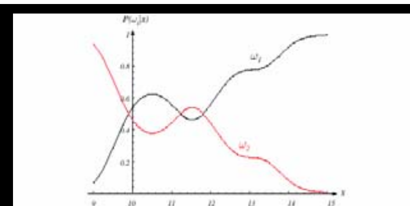


FIGURE 2.2. Posterior probabilities for the particular priors $P(\omega_1) = 2/3$ and $P(\omega_2) = 1/3$ for the class-conditional probability densities shown in Fig. 2.1. Thus in this case, given that a pattern is measured to have feature value $x = 14$, the probability it is in category ω_2 is roughly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



- Intuitive decision rule given the posterior probabilities:
  Given *x*:
  if $P(\omega_1 \mid x) > P(\omega_2 \mid x)$ ➡ True state of nature = ω_1
  if $P(\omega_1 \mid x) < P(\omega_2 \mid x)$ ➡ True state of nature = ω_2

  Why do this?: Whenever we observe a particular *x*, the probability of error is :

  $P(error \mid x) = P(\omega_1 \mid x)$ *if we decide* ω_2
  $P(error \mid x) = P(\omega_2 \mid x)$ *if we decide* ω_1

- Since decision rule is optimal for each feature value X, there is not better rule for all x.

Bayesian Decision Theory – Continuous Features

Generalization of the preceding ideas

- Use of more than one feature
- Use more than two states of nature
- Allowing actions and not only decide on the state of nature
- Introduce a loss of function (more general than the probability of error)
- Allowing actions other than classification primarily allows the possibility of rejection
- Refusing to make a decision in close or bad cases!
- Letting loss function state how costly each action taken is

Bayesian Decision Theory – Continuous Features

- Let **X** be a vector of features.
- Let $\{\omega_1, \omega_2, \ldots, \omega_c\}$ be the set of c states of nature (or "classes")
- Let $\{\alpha_1, \alpha_2, \ldots, \alpha_a\}$ be the set of possible actions
- Let $\lambda(\alpha_i \mid \omega_j)$ be the loss for action α_i when the state of nature is ω_j

What is the Expected Loss for action α_i ?

For any given x the expected loss is

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

$R(\alpha_i \mid x)$ is called the Conditional Risk (or Expected Loss)

Overall risk

R = *Sum of all* $\underbrace{R(\alpha_i \mid x)}_{\textbf{Conditional risk}}$ *for* $i = 1,\ldots,a$

Minimizing R $\Longleftrightarrow$ Minimizing $R(\alpha_i \mid x)$ *for* $i = 1,\ldots, a$

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

for i = 1,…,a

Given a measured feature vector x, which actiion should we take?

Select the action α_i for which $R(\alpha_i \mid x)$ is minimum

$\Rightarrow$ R is minimum and R in this case is called the Bayes risk = best performance that can be achieved!

Two-Category Classification

α_1 : *deciding* ω_1

α_2 : *deciding* ω_2

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

loss incurred for deciding ω_i when the true state of nature is ω_j

Conditional risk:

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

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

Our rule is the following:

if $R(\alpha_1 \mid x) < R(\alpha_2 \mid x)$

$\lambda_{11} P(\omega_1 \mid x) + \lambda_{12} P(\omega_2 \mid x) < \lambda_{21} P(\omega_1 \mid x) + \lambda_{22} P(\omega_2 \mid x)$

action α_1: "decide ω_1" is taken

This results in the equivalent rule :

decide ω_1 if:

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

and decide ω_2 otherwise

31

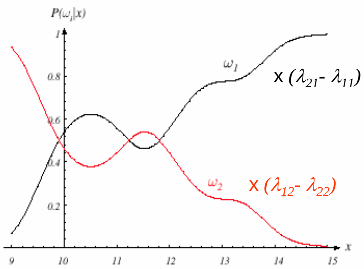


FIGURE 2.2. Posterior probabilities for the particular priors $P(\omega_1) = 2/3$ and $P(\omega_2) = 1/3$ for the class-conditional probability densities shown in Fig. 2.1. Thus in this case, given that a pattern is measured to have feature value $x = 14$, the probability it is in category ω_2 is roughly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

32

Two-Category Decision Theory: Chopping Machine

α_1 = *chop*
α_2 = *DO NOT chop*
ω_1 = *NO hand in machine*
ω_2 = *hand in machine*

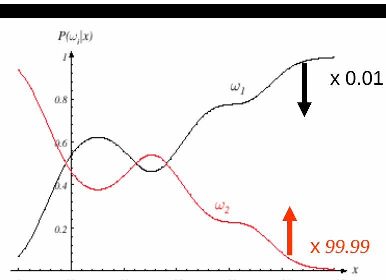
$\lambda_{11} = \lambda(\alpha_1 \mid \omega_1) = \$\ 0.00$
$\lambda_{12} = \lambda(\alpha_1 \mid \omega_2) = \$\ 100.00$
$\lambda_{21} = \lambda(\alpha_2 \mid \omega_1) = \$\ 0.01$
$\lambda_{22} = \lambda(\alpha_1 \mid \omega_1) = \$\ 0.01$

Therefore our rule becomes

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

$0.01\ \ P(x \mid \omega_1)\, P(\omega_1) > 99.99\, P(x \mid \omega_2)\, P(\omega_2)$

33



α_1 = *chop*
α_2 = *DO NOT chop*

ω_1 = *NO hand in machine*
ω_2 = *hand in machine*

34

Exercise to do at home!!

Select the optimal decision where:
$\Omega = \{\omega_1, \omega_2\}$

$P(x \mid \omega_1) \longrightarrow$ N(2, 0.5) (Normal distribution)
$P(x \mid \omega_2) \longrightarrow$ N(1.5, 0.2)

$P(\omega_1) = 2/3$
$P(\omega_2) = 1/3$

$$\lambda = \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix}$$

35

Minimum-Error-Rate Classification revisited

- 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*

36

- Introduction of the zero-one loss function:

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

Therefore, the conditional risk for each action is:

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

$$= \sum_{j \neq i} P(\omega_j \mid x) = 1 - P(\omega_i \mid x)$$ Average Prob. of Error

Since

$$\sum P(\omega_j \mid x) = 1$$

"The risk corresponding to this loss function is the average (or expected) probability error"

37

- 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 ω_i if $P(\omega_i \mid x) > P(\omega_j \mid x)$ $\forall j \neq i$

38

Likelihood ratio:

The preceding rule is equivalent to the following rule:

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

Then take action α_1 (decide ω_1)
Otherwise take action α_2 (decide ω_2)

39

- Regions of decision and zero-one loss function, therefore:

$$\text{Let } \frac{\lambda_{12} - \lambda_{22}}{\lambda_{21} - \lambda_{11}} \cdot \frac{P(\omega_2)}{P(\omega_1)} = \theta_\lambda \text{ then decide } \omega_1 \text{ if : } \frac{P(x \mid \omega_1)}{P(x \mid \omega_2)} > \theta_\lambda$$

- If λ is the zero-one loss function which means:

$$\lambda = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$$

$$\text{then } \theta_\lambda = \frac{P(\omega_2)}{P(\omega_1)} = \theta_a$$

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

40

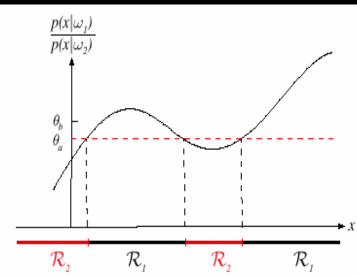


FIGURE 2.3. The likelihood ratio $p(x|\omega_1)/p(x|\omega_2)$ for the distributions shown in Fig. 2.1. If we employ a zero-one or classification loss, our decision boundaries are determined by the threshold θ_a. If our loss function penalizes miscategorizing ω_2 as ω_1 patterns more than the converse,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

41

Classifiers, Discriminant Functions and Decision Surfaces

- Discriminant Functions: A generalization
- The multi-category case
  - Consider a set of c discriminant functions $g_i(x), i = 1,\ldots, c$
  - The classifier assigns a feature vector x to class ω_i if:
    $g_i(x) > g_j(x)$ $\forall j \neq I$
  - *Designing a classifer amounts to specifying the $g_i(x)$*

42

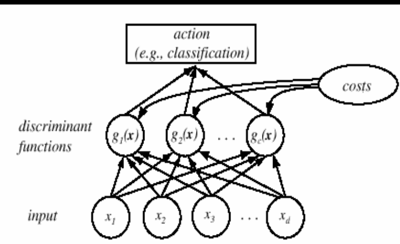


FIGURE 2.5. The functional structure of a general statistical pattern classifier which includes d inputs and c discriminant functions $g_i(\mathbf{x})$. A subsequent step determines which of the discriminant values is the maximum, and categorizes the input pattern accordingly. The arrows show the direction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

43

Decision Regions

- Feature space divided into c decision region

  if $g_i(x) > g_j(x)$ $\forall j \neq i$ then x is in $\mathcal{R}_i$

  ($\mathcal{R}_i$ means assign x to ω_i)

Decision surfaces

$\{x: \exists\, i,j\ g_i(x) = g_j(x)\}$

44

- *Bayes Risk as discriminant function.*
  - *Let $g_i(x) = - R(\alpha_i \mid x)$*

    (max. discriminant corresponds to min. risk!)
- For the minimum error rate, discriminant function is:

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

  (max. discrimination corresponds to max. posterior!)

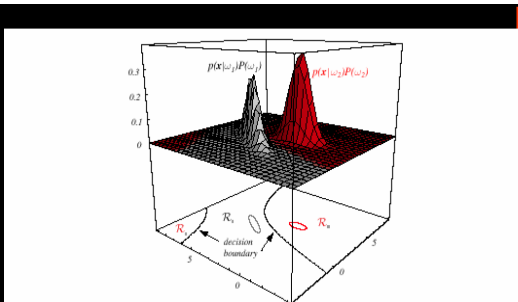
  $g_i(x) \equiv P(x \mid \omega_i)\, P(\omega_i)$
- *Any function F(r) which is monotonic over r>0 when applied to a set of discriminant functions, yields new discriminant function with the same decision regions/boundaries.*

$g_i(x) = \ln P(x \mid \omega_i) + \ln P(\omega_i)$

(ln: natural logarithm!)

We'll see this form with Normal distributions

47



On to higher dimensions!

48

The Normal Density

- Univariate density
  - Density which is analytically tractable
  - Continuous density
  - A lot of processes are asymptotically Gaussian
  - Handwritten characters, speech sounds are ideal or prototype corrupted by random process (central limit theorem)

$$P(x) = \frac{1}{\sqrt{2\pi}\,\sigma} \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right],$$

Where:

μ = mean (or expected value) of x

σ^2 = expected squared deviation or variance

49

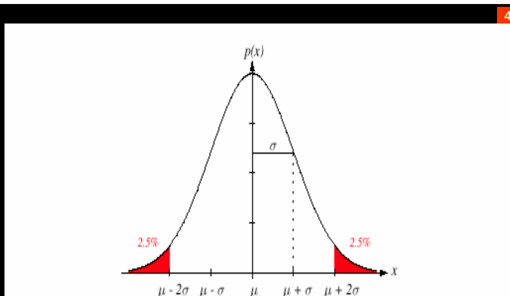
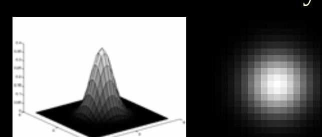


FIGURE 2.7. A univariate normal distribution has roughly 95% of its area in the range $|x - \mu| \leq 2\sigma$, as shown. The peak of the distribution has value $p(\mu) = 1/\sqrt{2\pi}\sigma$. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

50

Multivariate density



- Multivariate normal density in d dimensions is:

$$P(x) = \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} \exp\left[-\frac{1}{2}(x-\mu)^t \Sigma^{-1}(x-\mu)\right]$$

where:

$x = (x_1, x_2, \ldots, x_d)^t$ (t stands for the transpose vector form)

$\mu = (\mu_1, \mu_2, \ldots, \mu_d)^t$ mean vector

$\Sigma = d$ by d covariance matrix

$|\Sigma|$ and Σ^{-1} are determinant and inverse respectively