Lecture 2: Basics of Pattern Classification

Professor Aly A. Farag
Computer Vision and Image Processing Laboratory
University of Louisville

URL: www.cvip.uofl.edu ; E-mail: aly.farag@louisville.edu
Pattern Classification

Goal: Given measurements (outcomes) of experiments that share common attributes, how to separate these measurements?

Illustration:
An Example: Sorting fish!

Real world approach!!!
No, let’s be Americanized!

“Sorting incoming Fish on a conveyor according to species using optical sensing”

Example:

- Species
  - Sea bass
  - Salmon
Classification

→ Select the length of the fish as a possible feature for discrimination
The **length** is a poor feature alone!

→ Select the **lightness** as a possible feature.

Duda el al., 2001 Pattern Classification, Chap. 1
→ Adopt the lightness and add the width of the fish

\[ x^T = [x_1, x_2] \]

lightness  width

Duda el al., 2001 Pattern Classification, Chap. 1
• We might add other features that are not correlated with the ones we already have.

• Ideally, the best decision boundary should be the one which provides an optimal performance such as in the following figure:

Duda el al., 2001 Pattern Classification, Chap. 1
However, our satisfaction is premature because the central aim of designing a classifier is to correctly classify novel input

Issue of generalization!
Duda el al., 2001 Pattern Classification, Chap. 1
• **Feature extraction**
  – Discriminative features
  – Invariant features with respect to translation, rotation and scale.

• **Classification**
  – Use a feature vector provided by a feature extractor to assign the object to a category

• **Post Processing**
  – Exploit *context* input dependent information other than from the target pattern itself to improve performance
The Design Cycle

• Data collection
• Feature Choice
• Model Choice
• Training
• Evaluation
• Computational Complexity
Bayesian Decision Theory
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

      – \( P(\omega_1) = P(\omega_2) \) (uniform priors)

      – \( P(\omega_1) + P(\omega_2) = 1 \) (exclusivity and exhaustivity)
• Decision rule with only the prior information
  – Decide $\omega_1$ if $P(\omega_1) > P(\omega_2)$ otherwise decide $\omega_2$

• 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 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 \( \omega_j \). 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
  
  – \( P(\omega_j \mid x) = P(x \mid \omega_j) \cdot P(\omega_j) / P(x) \)
  
  – Where in case of two categories
  
  \[
  P(x) = \sum_{j=1}^{j=2} P(x \mid \omega_j)P(\omega_j)
  \]
  
  – Posterior = (Likelihood. Prior) / Evidence
**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 $\omega_2$ is roughly 0.08, and that it is in $\omega_1$ 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.
• Decision given the posterior probabilities

X is an observation for which:

if \( P(\omega_1 \mid x) > P(\omega_2 \mid x) \)

True state of nature = \( \omega_1 \)

if \( P(\omega_1 \mid x) < P(\omega_2 \mid x) \)

True state of nature = \( \omega_2 \)

Therefore:

whenever we observe a particular \( x \), the probability of error is :

\[
P(\text{error} \mid x) = P(\omega_1 \mid x) \text{ if we decide } \omega_2
\]

\[
P(\text{error} \mid x) = P(\omega_2 \mid x) \text{ if we decide } \omega_1
\]
• Minimizing the probability of error

• Decide $\omega_1$ if $P(\omega_1 \mid x) > P(\omega_2 \mid x)$; otherwise decide $\omega_2$

Therefore:

$$P(error \mid x) = \min [P(\omega_1 \mid x), P(\omega_2 \mid x)]$$

(Bayes decision)
Bayesian Decision Theory
(Continuous Features)

Let \( \{\omega_1, \omega_2, \ldots, \omega_c\} \) be the set of \( c \) states of nature (or "categories")

Let \( \{\alpha_1, \alpha_2, \ldots, \alpha_a\} \) be the set of possible actions

Let \( \lambda(\alpha_i \mid \omega_j) \) be the loss incurred for taking action \( \alpha_i \) when the state of nature is \( \omega_j \)
Overall risk

\[ R = \text{Sum of all } R(\alpha_i \mid x) \text{ for } i = 1, \ldots, a \]

Minimizing \( R \) \iff 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 \( i = 1, \ldots, a \)
Select the action $\alpha_i$ for which $R(\alpha_i \mid x)$ is minimum

$R$ is minimum and $R$ in this case is called the Bayes risk = best performance that can be achieved!
• Two-category classification

\( \alpha_1 : \text{deciding } \omega_1 \)

\( \alpha_2 : \text{deciding } \omega_2 \)

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

loss incurred for deciding \( \omega_i \) when the true state of nature is \( \omega_j \)

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:

\[
\text{if } R(\alpha_1 \mid x) < R(\alpha_2 \mid x) \\
\text{action } \alpha_1: \text{“decide } \omega_1 \text{” is taken}
\]

This results in the equivalent rule:

decide \( \omega_1 \) 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 \( \omega_2 \) otherwise
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 \(\alpha_1\) (decide \(\omega_1\))

Otherwise take action \(\alpha_2\) (decide \(\omega_2\))

Optimal decision property:

“If the likelihood ratio exceeds a threshold value independent of the input pattern \(x\), we can take optimal actions”
Exercise

Select the optimal decision where:

\[ \Omega = \{ \omega_1, \omega_2 \} \]

\[
P(x \mid \omega_1) \quad \rightarrow \quad \text{N}(2, 0.5) \quad \text{(Normal distribution)}
\]

\[
P(x \mid \omega_2) \quad \rightarrow \quad \text{N}(1.5, 0.2)
\]

\[
P(\omega_1) = \frac{2}{3}
\]

\[
P(\omega_2) = \frac{1}{3}
\]

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