Pattern classification

Biometrics
CSE 190-a
Lecture 3

An Example

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

Species
- Sea bass
- 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
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.

• Adopt the lightness and add the width of the fish

\[ \mathbf{x}^T = [x_1, x_2] \]

• 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)
  - Use of the class–conditional information
    - $P(x | \omega_1)$ and $P(x | \omega_2)$ describe the difference in lightness between populations of sea-bass and salmon
Pattern Classification, Chapter 1

- Posterior, likelihood, evidence
  \[ P(\omega_j | x) = \frac{P(x | \omega_j) \cdot P(\omega_j)}{P(x)} \]  
  (BAYES RULE)

- In words, this can be said as:
  Posterior = (Likelihood * Prior) / Evidence

- Where in case of two categories
  \[ P(x) = \sum_{j=1}^{2} P(x | \omega_j) P(\omega_j) \]

- Intuitive decision rule given the posterior probabilities:
  Given \( x \):
  - if \( P(\omega_1 | x) > P(\omega_2 | x) \)  
    True state of nature = \( \omega_1 \)
  - if \( P(\omega_1 | x) < P(\omega_2 | x) \)  
    True state of nature = \( \omega_2 \)

  Why do this?: Whenever we observe a particular \( x \), the probability of error is:
  - \( P(\text{error} | x) = P(\omega_2 | x) \) if we decide \( \omega_2 \)
  - \( P(\text{error} | x) = P(\omega_1 | x) \) if we decide \( \omega_1 \)

- 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 \(\lambda(\alpha_i | \omega_j)\) be the loss for action \(\alpha_i\) when the state of nature is \(\omega_j\)

What is the Expected Loss for action \(\alpha_i\)?

For any given \(x\) the expected loss is

\[
R(\alpha_i | x) = \sum_{j=1}^{c} \lambda(\alpha_i | \omega_j) P(\omega_j | x)
\]

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

Overall risk

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

Minimizing \(R\) ↔ Minimizing \(R(\alpha_i | x)\) for \(i = 1, \ldots, a\)

Two-Category Classification

- \(\alpha_1\) : deciding \(\omega_1\)
- \(\alpha_2\) : deciding \(\omega_2\)
- \(\lambda_{ij} = \lambda(\alpha_i | \omega_j)\)

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

Conditional risk:

\[
R(\alpha_1 | x) = \lambda_{11} P(\omega_1 | x) + \lambda_{12} P(\omega_2 | x)
\]

\[
R(\alpha_2 | x) = \lambda_{21} P(\omega_1 | x) + \lambda_{22} P(\omega_2 | x)
\]

Our rule is the following:

if \(R(\alpha_1 | x) < R(\alpha_2 | x)\)

\[\lambda_{21} P(\omega_1 | x) + \lambda_{22} P(\omega_2 | x) < \lambda_{11} P(\omega_1 | x) + \lambda_{12} P(\omega_2 | x)\]

action \(\alpha_1\): "decide \(\omega_1\)" is taken

This results in the equivalent rule:

decide \(\omega_1\) if:

\[
(\lambda_{21} - \lambda_{11}) P(x | \omega_1) P(\omega_1) > (\lambda_{12} - \lambda_{22}) P(x | \omega_2) P(\omega_2)
\]

and decide \(\omega_2\) otherwise
Two-Category Decision Theory: Chopping Machine

\(\alpha_1 = \text{chop} \quad \alpha_2 = \text{DO NOT chop} \quad \omega_1 = \text{NO hand in machine} \quad \omega_2 = \text{hand in machine}\)

\[
\lambda_{11} = \lambda(\alpha_1 | \omega_1) = 0.00 \\
\lambda_{12} = \lambda(\alpha_1 | \omega_2) = 100.00 \\
\lambda_{21} = \lambda(\alpha_2 | \omega_1) = 0.01 \\
\lambda_{22} = \lambda(\alpha_1 | \omega_1) = 0.01
\]

Therefore our rule becomes

\[
(\lambda_{21} - \lambda_{11}) P(x | \omega_1) P(\omega_1) > (\lambda_{12} - \lambda_{22}) P(x | \omega_2) P(\omega_2)
\]

Exercise to do at home!!

Select the optimal decision where:

\(\Omega = \{\omega_1, \omega_2\}\)

\[
P(x | \omega_1) = N(2, 0.5) \quad N(1.5, 0.2)
\]

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

Minimum-Error-Rate Classification revisited

- Actions are decisions on classes
  - If action \(\alpha_i\) is taken and the true state of nature is \(\omega_j\) 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
• Minimize the risk requires maximize \( P(\omega_i | x) \)
(since \( R(\omega_i | x) = 1 - P(\omega_i | x) \))

• For Minimum error rate
  • Decide \( \omega_i \) if \( P(\omega_i | x) > P(\omega_j | x) \) \( \forall j \neq i \)

Likelihood ratio:

The preceding rule is equivalent to the following rule:

\[
\frac{P(x | \omega_i)}{P(x | \omega_j)} > \frac{\lambda_{22} - \lambda_{21}}{\lambda_{12} - \lambda_{11}} \cdot \frac{P(\omega_i)}{P(\omega_j)}
\]

Then take action \( \omega_i \) (decide \( \omega_i \))
Otherwise take action \( \omega_j \) (decide \( \omega_j \))

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

Let \( \frac{\lambda_{11} - \lambda_{22}}{\lambda_{22} - \lambda_{21}} \cdot \frac{P(\omega_i)}{P(\omega_i)} = \theta_i \) then decide \( \omega_i \)

\[
\text{if } \frac{P(x | \omega_i)}{P(x | \omega_j)} > \theta_i \text{ then } \omega_i = \omega_i
\]

If \( \lambda \) is the zero-one loss function which means:

\[
\lambda = \frac{0}{1} \quad \text{or} \quad \frac{2}{1}
\]

\[
\text{then } \theta_i = \frac{P(\omega_i)}{P(\omega_i)} = \theta_i
\]

\[
\text{if } \lambda = \frac{2}{1} \text{ then } \theta_i = 2 \cdot \frac{P(\omega_i)}{P(\omega_i)} = \theta_i
\]

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 \( \omega_i \) if:
      \( g_i(x) > g_j(x) \) \( \forall j \neq i \)
    • Designing a classifier amounts to specifying the \( g_i(x) \)
Decision Regions

• Feature space divided 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 \( \omega_i \))

Decision surfaces

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

Bayes Risk as discriminant function.

• Let \( g(x) = P(\omega | x) \) (max. discriminant corresponds to min. risk!)

For the minimum error rate, discriminant function is:

\( g(x) = P(\omega | x) \)

(max. discrimination corresponds to max. posterior!)

\( g(x) = P(\omega | x) P(\omega) \)

Any function \( F(r) \) which is monotonic over \( r > 0 \) when applied to a set of discriminant functions, yields a new discriminant function with the same decision regions/boundaries.

\( g(x) = \ln P(x | \omega_i) + \ln P(\omega_i) \) (ln: natural logarithm!)

We'll use this form with Normal distributions.

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)

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

  Where:
  \( \mu = \text{mean (or expected value) of } x \)
  \( \sigma^2 = \text{expected squared deviation or variance} \)

On to higher dimensions!

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)
  \( \mu = [\mu_1, \mu_2, \ldots, \mu_d] \) mean vector
  \( \Sigma = d \times d \) covariance matrix
  \( |\Sigma| \) and \( \Sigma^{-1} \) are determinant and inverse respectively