Pattern Classification

Biometrics
CSE 190-a
Lecture 4

Announcements

• Readings on E-reserves
• HW1 will be posted shortly
• Project description on web page

Key Probabilities

• $ω_j$ – class label
• $X$ – feature vector
• $P(ω_j)$ – Prior probability of class $j$
• $P(x)$ – Probability distribution function of feature values.
• $P(x | ω_j)$ – Class conditional density function or likelihood of feature $x$ given $ω_j$
• $P(ω_j | x)$ – posterior probability density function of the class given the feature.

• Posterior, likelihood, evidence

  $P(ω_j | x) = \frac{P(x | ω_j) * P(ω_j)}{P(x)}$  \hspace{1cm} \text{(BAYES RULE)}

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

  Where in case of two categories

  $P(x) = \sum_{j=1}^{2} P(x | ω_j)P(ω_j)$

• Intuitive decision rule given the posterior probabilities:
  Given $x$:
  if $P(ω_1 | x) > P(ω_2 | x)$ \hspace{1cm} \text{True state of nature $= ω_1$}
  if $P(ω_1 | x) < P(ω_2 | x)$ \hspace{1cm} \text{True state of nature $= ω_2$}

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

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

Select the action $\alpha_i$ for which $R(\alpha_i | x)$ is minimum.

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

Likelihood ratio:

The preceding rule is equivalent to the following rule:

$$\begin{cases} 
    \frac{P(x | \omega_1)}{P(x | \omega_2)} > \frac{\lambda_{11} - \lambda_{21}}{\lambda_{22} - \lambda_{12}} \cdot \frac{P(\omega_1)}{P(\omega_2)} \\
    \frac{P(x | \omega_2)}{P(x | \omega_1)} > \frac{\lambda_{22} - \lambda_{12}}{\lambda_{11} - \lambda_{21}} \cdot \frac{P(\omega_2)}{P(\omega_1)}
\end{cases}$$

Then take action $\alpha_1$ (decide $\omega_1$)

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

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:
    $$g_i(x) > g_j(x) \iff i = j$$
  - Designing a classifier amounts to specifying the $g_i(x)$

![Figure 2.5](image-url)
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

- Boundary between decision regions.
  
  $\{x : \exists i, j g_i(x) = g_j(x)\}$

Bayes Risk as discriminant function.

- Let $g(x) = R(\omega_i | x)$ (max. discriminant corresponds to min. risk!)

- For the minimum error rate, discriminant function is:
  
  $g_i(x) = P(\omega_i | x)$
  
  (max. discrimination corresponds to max. posterior!)

- 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.

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

We’ll use this form for Normal distributions

Dichotomizer

- The two-category case

  - A classifier is a “dichotomizer” that has two discriminant functions $g_1$ and $g_2$

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

  Decide $\omega_1$ if $g(x) > 0$; Otherwise decide $\omega_2$

The computation of $g(x)$

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

  $= \ln \frac{P(x | \omega_1)}{P(x | \omega_2)} + \ln \frac{P(\omega_1)}{P(\omega_2)}$

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)

  Where:

  $\mu = \text{mean (or expected value) of } x$

  $\sigma^2 = \text{expected squared deviation or variance}$

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 $p(\mu) = 1/\sqrt{2\pi}$. From: Richard O. Duda, Peter E. Hart, and David G. Stork, Pattern Classification. Copyright © 2001 by John Wiley & Sons, Inc.
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:
- \( x = (x_1, x_2, \ldots, x_d) \) (stands for the transpose vector form)
- \( \mu = (\mu_1, \mu_2, \ldots, \mu_d) \) mean vector
- \( \Sigma = \text{d by d} \) covariance matrix
- \(|\Sigma|\) and \(\Sigma^{-1}\) are determinant and inverse respectively.

### Discriminant Functions for the Normal Density

- We saw that the minimum error-rate classification can be achieved by the discriminant function

\[
g(x) = \ln P(x \mid \omega_i) + \ln P(\omega_i)
\]

- Case of multivariate normal for class condition density (likelihood function)

\[
g_i(x) = -\frac{1}{2}(x - \mu_i)^T \Sigma_i^{-1} (x - \mu_i) + \frac{d}{2} \ln(2\pi) - \frac{1}{2} \ln|\Sigma_i| + \ln P(\omega_i)
\]

### Case \( \Sigma = \sigma^2 I \) (I stands for the identity matrix)

- A classifier that uses linear discriminant functions is called “a linear machine”

- The decision surfaces for a linear machine are pieces of hyperplanes defined by:

\[
g(x) = g(x)
\]