Supervised Learning with Applications

Chaur-Chin Chen (陳朝欽)
Institute of Information Systems & Applications
Department of Computer Science
National Tsing Hua University
HsinChu (新竹), Taiwan (台灣)
cchen@cs.nthu.edu.tw
March 20, 2009
Mandrill, Koala, Panda, Gorilla
Possum, Dingo, Fox, Wombat
Supervised Learning Problems

The problem of supervised learning can be defined as to design a function which takes the training data $x_i^{(k)}$, $i=1,2,...,n_i$, $k=1,2,...,C$, as input vectors with the output as either a single category or a regression curve.

The unsupervised learning (Cluster Analysis) is similar to that of the supervised learning problem (Pattern Recognition) except that the categories are unknown in the training data.
Distinguish Eggplants from Bananas

1. Features (characteristics)
   - Colors
   - Shapes
   - Size
   - Tree leaves
   - Other quantitative measurements

2. Decision rules: Classifiers

3. Performance Evaluation

4. Classification
Illustration of Supervised Learning

An example of dichotomous problem
Two Chi-Square Distributions

An Example of Two $\chi^2$ Distributions: $r_1=4$, $r_2=8$

The decision boundary: ML=4.899
Introduction to Bayes Decision

(1) \( p(\omega_i) \): a priori probability

(2) \( p(x|\omega_i) \): class conditional density function

(3) \( p(\omega_i|x) \): a posteriori probability

(4) \( \alpha(x) \): an action (a decision)

(5) \( \lambda(\alpha(x)|\omega_j) \): the loss function

\[
\begin{align*}
p(\text{error}) &= \sum_x p(\text{error}|x)p(x) = \sum_x p(\alpha(x) \in \omega_i, x \in \omega_j, i \neq j)p(x) \\
R(\alpha(x)|x) &= \sum_{j=1}^{C} \lambda(\alpha(x)|\omega_j)p(\omega_j|x): \text{conditional risk for pattern} \ x \\
\sum_x R(\alpha(x)|x)p(x): \text{average error of probability (error rate)}
\end{align*}
\]

Bayes Decision Rule

For each \( x \), find \( \alpha(x) \) which minimizes \( R(\alpha(x)|x) \)

For the 0-1 loss function, i.e. \( \lambda(\alpha(x)|\omega_j) = \begin{cases} 0 & \text{if } \alpha(x) = \omega_j, \\ 1 & \text{otherwise} \end{cases} \)

Then the Bayes decision rule can be reduced to

\[
\min \ [R(\alpha(x)|x)] = \min_j [1 - p(\omega_j|x)] = \max_i p(\omega_i|x)
\]

or

Assign \( x \) to class \( \omega_i \) if \( p(\omega_i|x) > p(\omega_j|x) \) for \( j \neq i \)
Bayes vs. Maximum Likelihood

- $X|\omega_1 \sim \chi^2(r_1)$ and $X|\omega_2 \sim \chi^2(r_2)$

$$p(x|\omega_i) = \frac{1}{\Gamma(r_i/2)2^{r_i/2}}x^{(r_i/2)-1}e^{-x/2}, \quad x > 0, \quad i = 1, 2$$

The Maximum Likelihood (ML) decision is to assign $x$ to $\omega_1$ if $p(x|\omega_1) > p(x|\omega_2)$

The Bayes decision is to assign $x$ to $\omega_1$ if $p(x|\omega_1)p(\omega_1) > p(x|\omega_2)p(\omega_2)$

Note that ML is the special case by assuming $p(\omega_1) = p(\omega_2) = \frac{1}{2}$ which need not be true in practical applications. We shall show the effect of $p(\omega_1) = \frac{1}{2}, p(\omega_2) = \frac{3}{2}$ with $r_1 < r_2$.

ML Decision:

$$y \in \omega_1 \text{ if } \frac{\Gamma(r_2/2)}{\Gamma(r_1/2)}x^{(r_2-r_1)/2} > x^{(r_2-r_1)/2}$$

Bayes Decision:

$$y \in \omega_1 \text{ if } \frac{\Gamma(r_2/2)}{\Gamma(r_1/2)}x^{(r_2-r_1)/2}p(\omega_1) > x^{(r_2-r_1)/2}p(\omega_2)$$

The error probability can be computed by

$$Err = p(\omega_1) \int_0^{\infty} \frac{1}{\Gamma(r_2/2)2^{r_2/2}}x^{(r_2-r_1)/2}e^{-x/2}dx + p(\omega_2) \int_0^{\infty} \frac{1}{\Gamma(r_2/2)2^{r_2/2}}x^{(r_2-r_1)/2}e^{-x/2}dx$$

When $r_1 = 4, r_2 = 8$, $p(\omega_1) = 1/7, p(\omega_2) = 6/7$, we have t=4.899 for ML decision, and t=2.0 for Bayes decision, and

\begin{itemize}
  \item $Err_{ML} = 0.2411$ and $Err_{Bayes} = 0.1214$.
\end{itemize}
Illustration of Bayes vs. ML (1)
Illustration of Bayes vs. ML (2)

An Example of Two $\chi^2$ Distributions: $r_1 = 4$, $r_2 = 8$

$\text{Err}_{ML} = 0.2411$, $\text{Err}_{Bayes} = 0.1214$

The decision boundary: $t_{BS} = 4.8990$, $t_{ML} = 2$ with $p_1 = 1/7$, $p_2 = 6/7$
Matlab Code

- \( r_1=4; \ r_2=8; \ p_1=1/7; \ p_2=1-p_1; \)
- \( X_1=0:0.4:12; \ X_2=0:0.4:12; \)
- \( Y_1=X_1.*exp(-X_1/2)/4; \)
- \( Y_2=(X_2.^3).*exp(-X_2/2)/96; \)
- \( t_{ML} = \text{sqrt}(24); \)
- \( t_{BS} = 2; \)
- \( x = t_{ML}; \ y = x.*exp(-x/2); \)
- \( z = t_{BS}; \ w = (z^3).*exp(-z/2)/96; \)
- \( X_3 = \text{zeros}(11)+t_{ML}; \ Y_3 = 0:0.02:0.2; \)
- \( X_5 = \text{zeros}(11)+t_{BS}; \ Y_5 = 0:0.02:0.2; \)
- \( t = t_{ML}; \)
- \( \text{Err}_{ML} = p_1*\text{exp}(-t/2)*(2*t+4)/4+p_2*(1.0-\text{exp}(-t/2))*(2*t^3+12*t^2+48*t+96)/96; \)
- \( t = t_{BS}; \)
- \( \text{Err}_{BS} = p_1*\text{exp}(-t/2)*(2*t+4)/4+p_2*(1.0-\text{exp}(-t/2))*(2*t^3+12*t^2+48*t+96)/96; \)
- \( \text{plot}(X_1,Y_1,'b-',X_2,Y_2,'r-',X_3,Y_3,'g--',X_5,Y_5,'m--'); \)
- \( \text{axis([0,12,0,0.24]);} \)
- \( \text{legend('The decision boundry: } t_{ML}=4.8990, \ t_{BS}=2'; 'Location','NorthEast');} \)
- \( \text{legend('Err}_{ML}=0.2411, \ Err}_{Bayes}=0.1214', 'Location','NorthWest') \)
- \( \text{text(1.5,0.15,'omega_1'); text(7.0,0.08,'omega_2');} \)
- \( \text{text(x,y,'<'); text(4.8,0.12,'ML');} \)
- \( \text{text(z,w,'<'); text(1.8,0.12,'Bayes');} \)
- \( \text{title('An Example of Two } \chi^2 \text{ Distributions: } r_1=4, \ r_2=8') \)
- \( \text{xlabel('The decision boundary: } t_{(BS)}=4.8990, \ t_{(ML)}=2 \text{ with } p_1=1/7, \ p_2=6/7') \)
Image Segmentation

Let $Y$ be an $mxn$ intensity image having 256 gray levels, the problem is to label each site or pixel in an $mxn$ lattice $X$ to optimize the prescribed criterion.

$Y$:                              $X$:

199 201 101 103 2 2 1 1
213 198 105 130 2 2 1 1
228 109 119 160 2 1 1 2
Image Lenna and Its Histogram
Application to Image Segmentation

\[ X_{\omega_1} \sim N(\mu_1, \sigma_1^2) \text{ and } X_{\omega_2} \sim N(\mu_2, \sigma_2^2) \]

\[ p(x_{\omega_1}) = \frac{1}{\sqrt{2\pi}\sigma_1} \exp\left(-\frac{(x - \mu_1)^2}{2\sigma_1^2}\right), \quad i = 1, 2 \]

The Maximum Likelihood (ML) decision is to assign \( x \) to \( \omega_1 \) if \( p(x_{\omega_1}) > p(x_{\omega_2}) \)

The Bayes decision is to assign \( x \) to \( \omega_1 \) if \( p(x_{\omega_1})p(\omega_1) > p(x_{\omega_2})p(\omega_2) \)

Note that ML is the special case by assuming \( p(\omega_1) = p(\omega_2) = \frac{1}{2} \) which need not be true in practical applications.

The error probability can be computed by

\[
Err = p(\omega_1) \int_0^T \frac{1}{\sqrt{2\pi}\sigma_1} \exp\left[-\frac{(x - \mu_1)^2}{2\sigma_1^2}\right]dx \\
+ p(\omega_2) \int_T^\infty \frac{1}{\sqrt{2\pi}\sigma_2} \exp\left[-\frac{(x - \mu_2)^2}{2\sigma_2^2}\right]dx
\]

\( \square \) As soon as \( T \) is chosen, Parameters \( p(\omega_i), \mu_i, \sigma_i^2, \ i = 1, 2 \) could be estimated, so is \( Err \).

• Otsu (1979) and Fisher (1936) chose \( T \) to maximize the following criterion, respectively

\[ \sigma_T^2 = p(\omega_1)p(\omega_2)(\mu_1 - \mu_2)^2 \]

\[ Fisher = \frac{(\mu_1 - \mu_2)^2}{\frac{1}{p(\omega_1)} + \frac{1}{p(\omega_2)}} \]
ICM Segmentation Algorithm

1. Given an image \( Y \), initialize a labeling \( X \)
2. For \( t=1:mn \)
   \[ X(t) \leftarrow g_0 \text{ if } \Pr(X(t)=g_0|X_{N(t)},Y) > \Pr(X(t)=g|X_{N(t)},Y) \text{ for } g,g_0 \]
3. Repeat step 2 until “convergence” (6 runs)
4. \( X \) is the required labeling

Chaur-Chin Chen and Richard C. Dubes

*Environmental Studies and ICM Segmentation Algorithm*,
Journal of Information Science and Engineering,
Vol. 6, 325-337, 1990.
Image Segmentation: ICM vs. Otsu
Image Segmentation: ICM vs. Otsu
Image Segmentation: ICM vs. Otsu
Munson’s Handprinted Characters

f1  f2  f3  f4  f5  f6  f7  f8  L
11. 3. 2. 4. 10. 3. 2. 6. 8
4. 8. 2. 3. 4. 10. 3. 7. O
11. 2. 10. 3. 11. 4. 11. 3. X
Features of Characters 8, O, X

7. 13. 5. 5. 6. 13. 2. 3. 8
5. 13. 6. 4. 6. 13. 3. 13. 8
9. 10. 6. 6. 8. 10. 2. 3. 8
7. 7. 6. 6. 8. 7. 2. 3. 8
8. 7. 6. 6. 8. 7. 2. 0. 8

7. 7. 5. 6. 3. 3. 4. 6. O
7. 6. 7. 6. 3. 4. 6. 5. O
6. 6. 5. 5. 4. 3. 4. 5. O
8. 8. 7. 6. 5. 7. 5. 5. O
6. 8. 7. 5. 5. 6. 2. 2. O

10. 7. 6. 6. 9. 9. 7. 10. X
10. 4. 4. 4. 8. 8. 3. 10. X
10. 7. 4. 4. 9. 9. 3. 9. X
7. 7. 6. 5. 10. 10. 10. 8. X
6. 10. 6. 10. 8. 8. 13. 4. X
Feature Extraction by Principal Component Analysis

Principal Component Analysis (PCA) is a multivariate statistical technique that is often useful in reducing dimensionality of a collection of unstructured random variables for analysis and interpretation.
Let $X$ be a $D$-dimensional random vector with covariance matrix $C$, the problem is to consecutively find the unit vectors $a_1, a_2, \ldots, a_D$ such that $y_i=x^t a_i$ with $Y_i=X^t a_i$ that satisfies

1. $\text{var}(Y_1)$ is the maximum
2. $\text{var}(Y_2)$ is the maximum subject to $\text{cov}(Y_2, Y_1)=0$
3. $\text{var}(Y_k)$ is the maximum subject to $\text{cov}(Y_k, Y_i)=0$, where $k=3, 4, \ldots, D$ and $i<k$.

- $Y_i$ is called the $i$-th principal component
- Feature extraction by PCA is called PCP
Methodology of Practical PCA

Given observations $x_1, x_2, \ldots, x_n$.

1. Compute mean vector $\mathbf{m} = \frac{1}{n} \sum_{i=1}^{n} x_i$
2. Compute covariance matrix $\mathbf{C} = \frac{1}{n} \sum_{i=1}^{n} (x_i - \mathbf{m})(x_i - \mathbf{m})'$
3. Compute eigenvalue/eigenvector pairs $(\lambda_i, \mathbf{u}_i)$ of $\mathbf{C}$, where $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_D$
4. Compute the first $d$ principal components $y_i^{(j)} = x_i^t \mathbf{u}_j$, for each $x_i$, $i=1,2,\ldots,n$, along the direction $\mathbf{u}_j$, $j=1,2,\ldots,d$
Illustration of PCA (1)
Illustration of PCA (2)

The 3rd and 4th principal components of 80x Data
Texture Discrimination

- D04: Pressed Cork
- D23: Beach Pebbles
- D77: Cotton Canvas
Shape Discrimination

• Canada Map
• Italy Map
• Taiwan Map
Other Applications

- Bioinformatics
- Cheminformatics
  - Quantitative structure-activity relationship
- Handwriting recognition
- Information retrieval
- Object recognition in computer vision
- Optical character recognition
- Spam detection
- Speech recognition
- Biosignal discrimination (normal vs. abnormal)
Main Steps

• Feature Selection/Extraction
• Decision Rules
• Recognition Rates
Microarray Image Data Analysis
Microarray Image Data Analysis

Each gene expression is a feature which is measured as average spot brightness

Top: Tumor Tissues
Bottom: Normal Tissues
Thank You For Your Attention

Questions and Comments