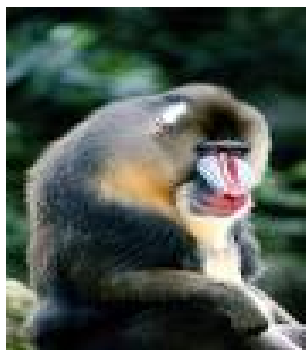


Supervised Learning with Applications

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 $x_i^{(k)}$, $i=1,2, \dots, n_i$, $k=1,2,\dots, C$, 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

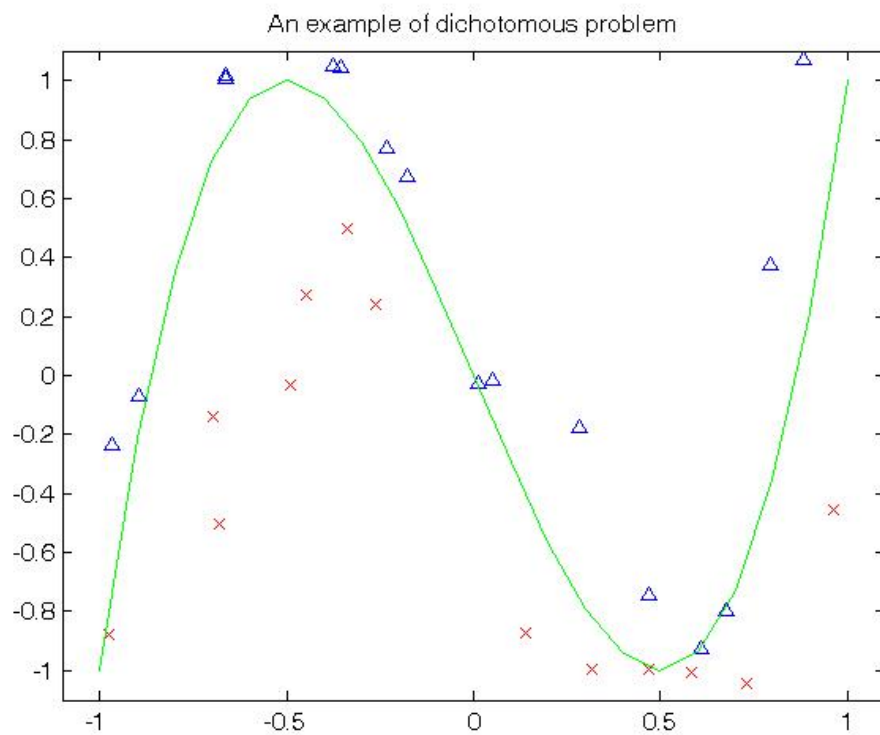
Other quantitative
measurements

2. Decision rules: Classifiers

3. Performance Evaluation

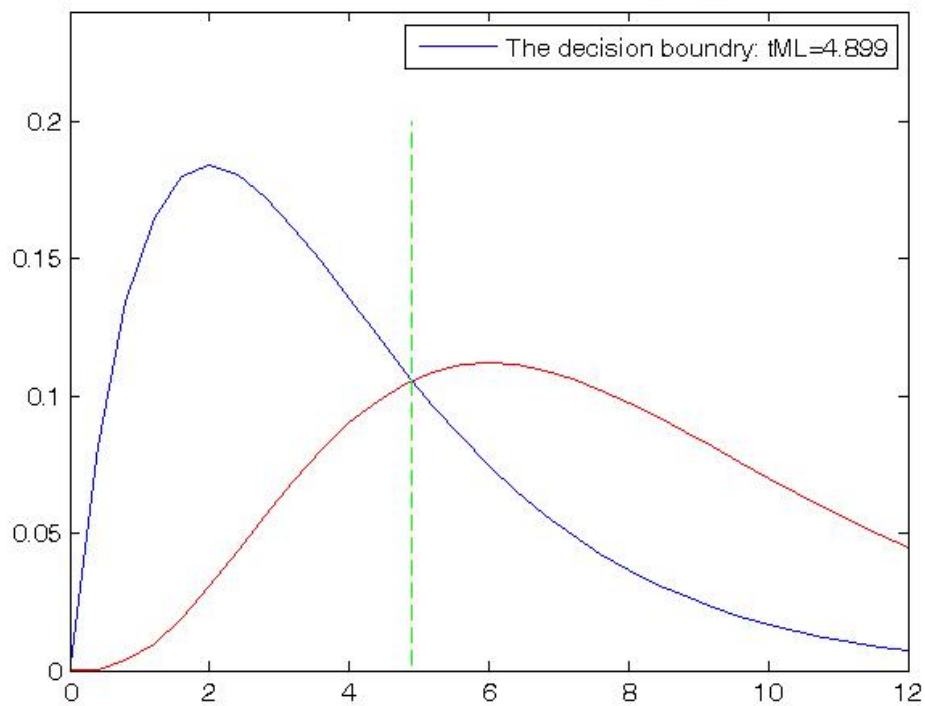
4. Classification

Illustration of Supervised Learning



Two Chi-Square Distributions

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



Introduction to Bayes Decision

- (1) $p(\omega_i)$: a priori probability
- (2) $p(\mathbf{x}|\omega_i)$: class conditional density function
- (3) $p(\omega_i|\mathbf{x})$: a posteriori probability
- (4) $\alpha(\mathbf{x})$: an action (a decision)
- (5) $\lambda(\alpha(\mathbf{x})|\omega_j)$: the loss function

$$p(\text{error}) = \sum_{\mathbf{x}} p(\text{error}|\mathbf{x})p(\mathbf{x}) = \sum_{\mathbf{x}} p(\alpha(\mathbf{x}) \in \omega_i, \mathbf{x} \in \omega_j, i \neq j)p(\mathbf{x})$$

$$R(\alpha(\mathbf{x})|\mathbf{x}) = \sum_{j=1}^C \lambda(\alpha(\mathbf{x})|\omega_j)p(\omega_j|\mathbf{x}): \text{conditional risk for pattern } \mathbf{x}$$

$$\sum_{\mathbf{x}} R(\alpha(\mathbf{x})|\mathbf{x})p(\mathbf{x}): \text{average error of probability (error rate)}$$

- **Bayes Decision Rule**

For each $\mathbf{x}$, find $\alpha(\mathbf{x})$ which minimizes $R(\alpha(\mathbf{x})|\mathbf{x})$

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

Then the Bayes decision rule can be reduced to

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

or

Assign $\mathbf{x}$ to class ω_i if $p(\omega_i|\mathbf{x}) > p(\omega_j|\mathbf{x})$ 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 ω_1 if $p(x|\omega_1) > p(x|\omega_2)$*

The Bayes decision is to *assign x to ω_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 $p(\omega_1) = \frac{1}{7}$, $p(\omega_2) = \frac{6}{7}$ with $r_1 < r_2$.

ML Decision:

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

Bayes Decision:

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

The error probability can be computed by

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

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

♣ $Err_{ML} = 0.2411$ and $Err_{Bayes} = 0.1214$.

Illustration of Bayes vs. ML (1)

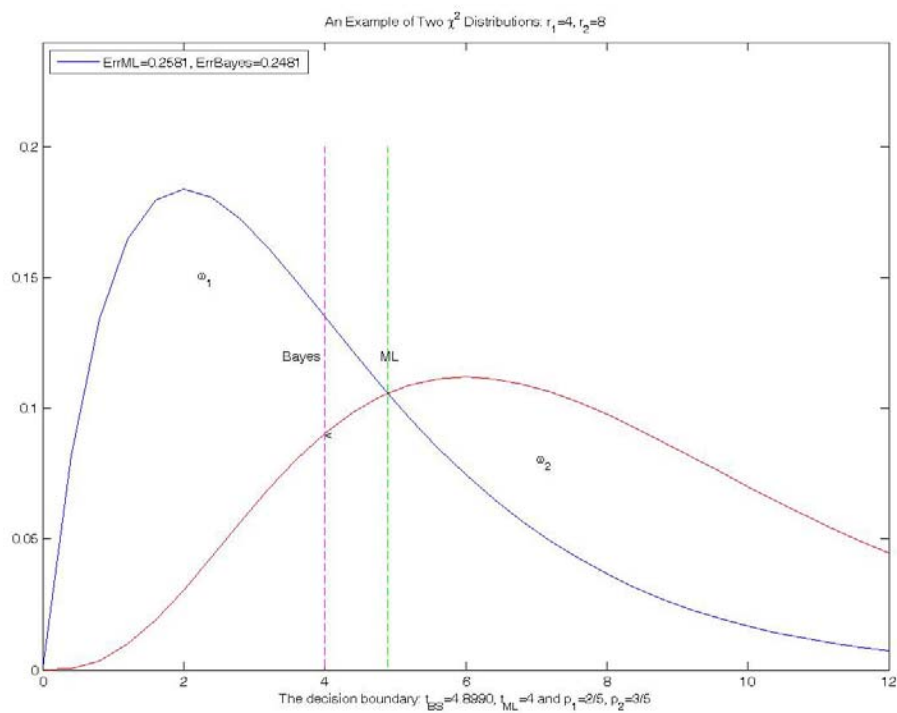
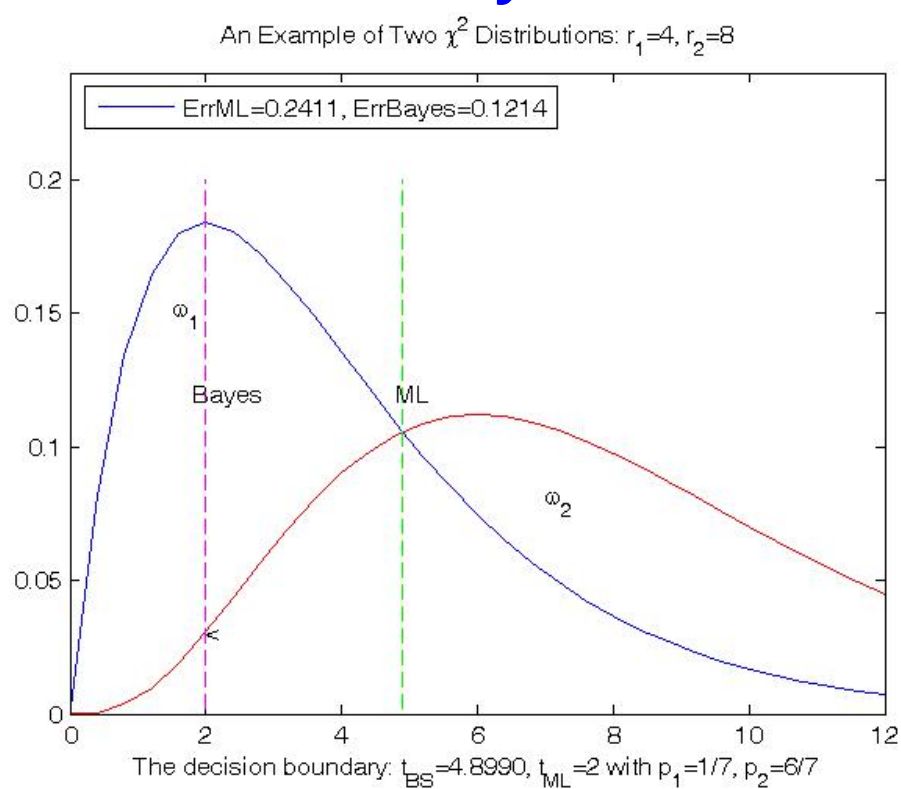


Illustration of Bayes vs. ML (2)



Matlab Code

- `r1=4; r2=8; p1=1/7; p2=1-p1;`
- `X1=0:0.4:12; X2=0:0.4:12;`
- `Y1=X1.*exp(-X1/2)/4;`
- `Y2=(X2.^3).*exp(-X2/2)/96;`
- `tML=sqrt(24);`
- `tBS=2;`
- `x=tML; y=x*exp(-x/2);`
- `z=tBS; w=(z^3)*exp(-z/2)/96;`
- `X3=zeros(11)+tML; Y3=0:0.02:0.2;`
- `X5=zeros(11)+tBS; Y5=0:0.02:0.2;`
- `t=tML;`
- `ErrML=p1*exp(-t/2)*(2*t+4)/4+p2*(1.0-exp(-t/2)*(2*t^3+12*t^2+48*t+96)/96);`
- `t=tBS;`
- `ErrBS=p1*exp(-t/2)*(2*t+4)/4+p2*(1.0-exp(-t/2)*(2*t^3+12*t^2+48*t+96)/96);`
- `plot(X1,Y1,'b-',X2,Y2,'r-',X3,Y3,'g--',X5,Y5,'m--');`
- `axis([0,12,0,0.24]);`
- `legend('The decision boundry: tML=4.8990, tBS=2','Location','NorthEast');`
- `legend('ErrML=0.2411, ErrBayes=0.1214','Location','NorthWest');`
- `text(1.5,0.15,'omega_1'); text(7.0,0.08,'omega_2');`
- `text(x,y,'<'); text(4.8,0.12,'ML');`
- `text(z,w,'<'); text(1.8,0.12,'Bayes');`
- `title('An Example of Two \chi^2 Distributions: r_1=4, r_2=8');`
- `xlabel('The decision boundary: t_{BS}=4.8990, t_{ML}=2 with p_1=1/7, p_2=6/7');`

Image Segmentation

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

Y :

199 201 101 103

213 198 105 130

228 109 119 160

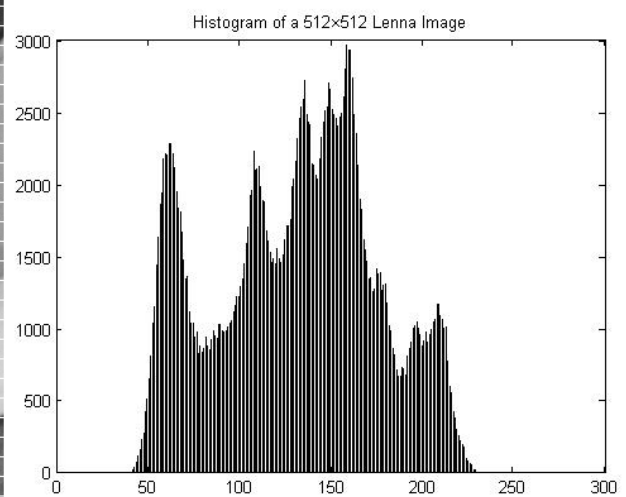
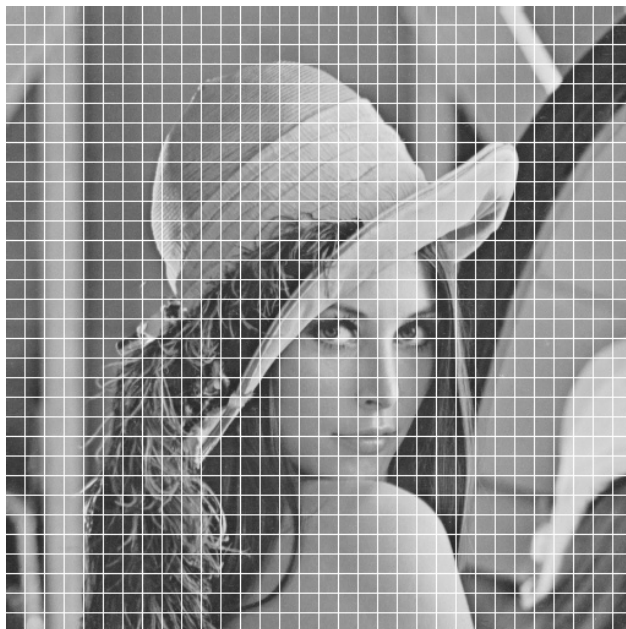
X :

2 2 1 1

2 2 1 1

2 1 1 2

Image Lenna and Its Histogram



Application to Image Segmentation

◇ $X|\omega_1 \sim N(\mu_1, \sigma_1^2)$ and $X|\omega_2 \sim N(\mu_2, \sigma_2^2)$

$$p(x|\omega_i) = \frac{1}{\sqrt{2\pi}\sigma_i} \exp[-(x - \mu_i)^2/2\sigma_i^2], \quad i = 1, 2$$

The Maximum Likelihood (ML) decision is to *assign* x to ω_1 if $p(x|\omega_1) > p(x|\omega_2)$

The Bayes decision is to *assign* x to ω_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

$$\begin{aligned} Err &= p(\omega_1) \int_T^\infty \frac{1}{\sqrt{2\pi}\sigma_1} \exp[-(x - \mu_1)^2/(2\sigma_1^2)] dx \\ &+ p(\omega_2) \int_0^T \frac{1}{\sqrt{2\pi}\sigma_2} \exp[-(x - \mu_2)^2/(2\sigma_2^2)] dx \end{aligned}$$

□ 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

$$\sigma_B^2 = p(\omega_1)p(\omega_2)(\mu_1 - \mu_2)^2$$

$$Fisher = \frac{(\mu_1 - \mu_2)^2}{\sigma_1^2 + \sigma_2^2}$$

ICM Segmentation Algorithm

1. Given an image Y , initialize a labeling X
2. For $t=1:mxn$
 $X(t) \leftarrow g_0$ if
 $\Pr(X(t)=g_0|X_{N(t)}, Y) > \Pr(X(t)=g|X_{N(t)}, Y)$ 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

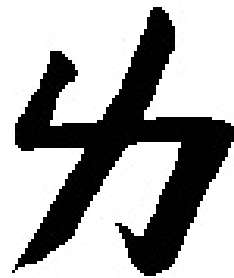
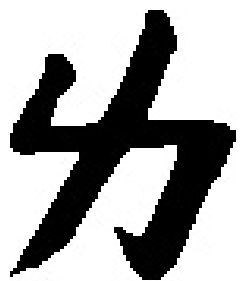
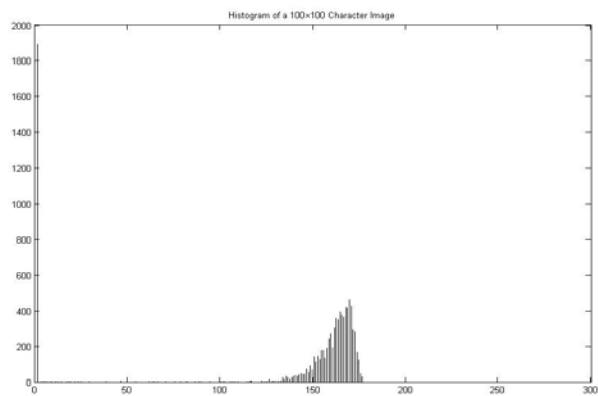
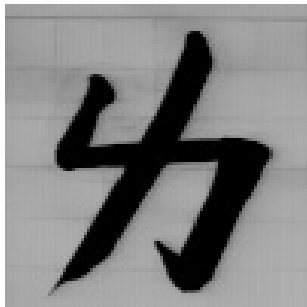


Image Segmentation: ICM vs. Otsu

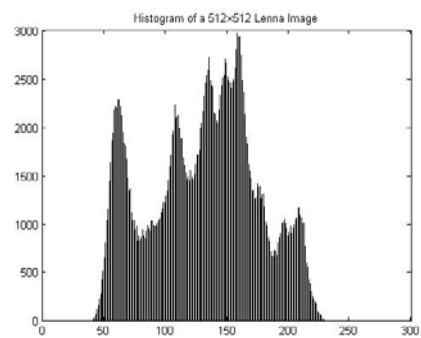
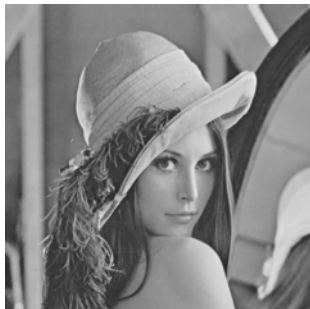
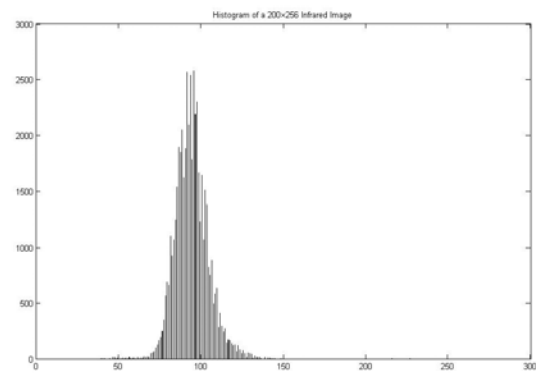
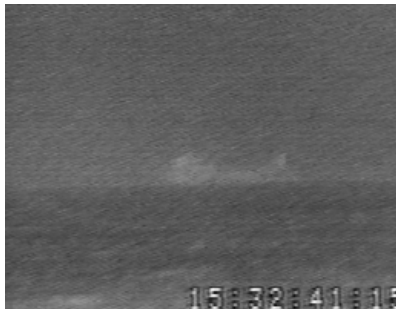
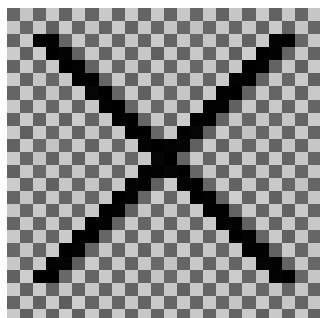
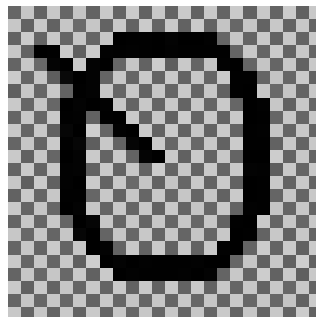
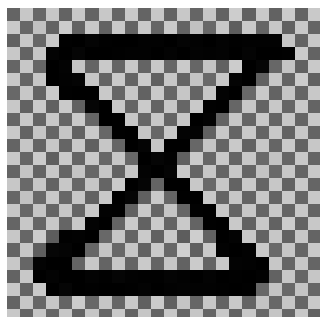


Image Segmentation: ICM vs. Otsu



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.	0
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

Product **C**omponent **A**nalysis (**PCA**) is
*a multivariate statistical technique that is
often useful in reducing dimensionality of a
collection of unstructured random variables
for analysis and interpretation.*

Problem Statement

- Let $\mathbf{X}$ be a D -dimensional random vector with covariance matrix C , the problem is to consecutively find the unit vectors $\mathbf{a}_1, \mathbf{a}_2, \dots, \mathbf{a}_D$ such that $y_i = \mathbf{x}^t \mathbf{a}_i$ with $Y_i = \mathbf{X}^t \mathbf{a}_i$ 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 $k=3, 4, \dots, D$ 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 $x_1, x_2, \dots, x_n$.

1. Compute mean vector $\mathbf{m} = \frac{1}{n} \sum_{i=1}^n x_i$
2. Compute covariance matrix $\mathbf{C} = \frac{1}{n} \sum_{i=1}^n (x_i - \mathbf{m})(x_i - \mathbf{m})^t$
3. Compute eigenvalue/eigenvector pairs (λ_i, u_i) of $\mathbf{C}$, where $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_D$
4. Compute the first d principal components $y_i^{(j)} = x_i^t u_j$, for each x_i , $i=1,2,\dots,n$, along the direction u_j , $j=1,2,\dots,d$

Illustration of PCA (1)

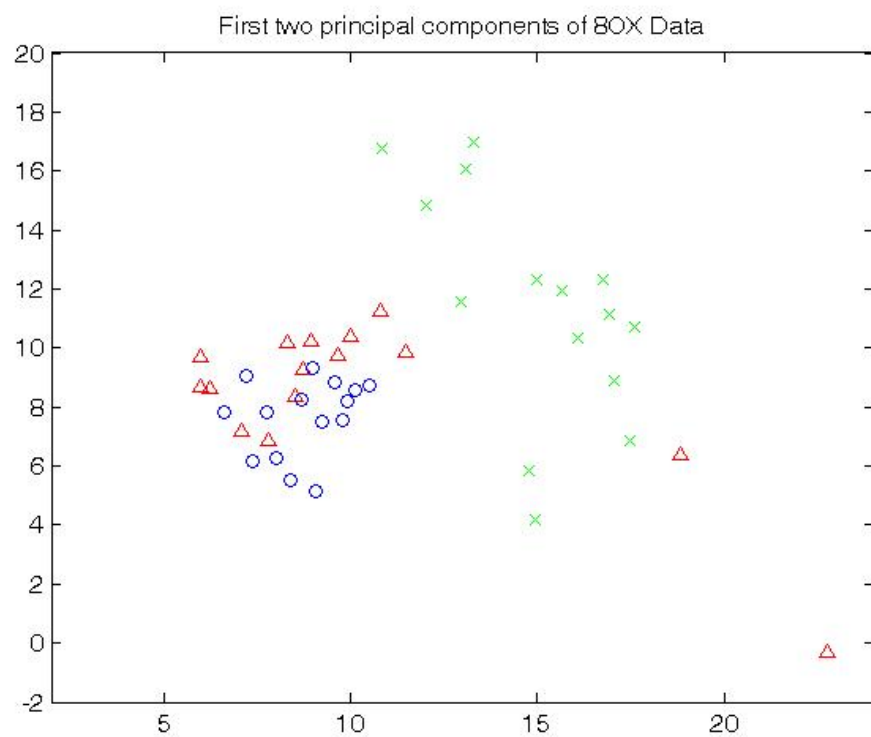
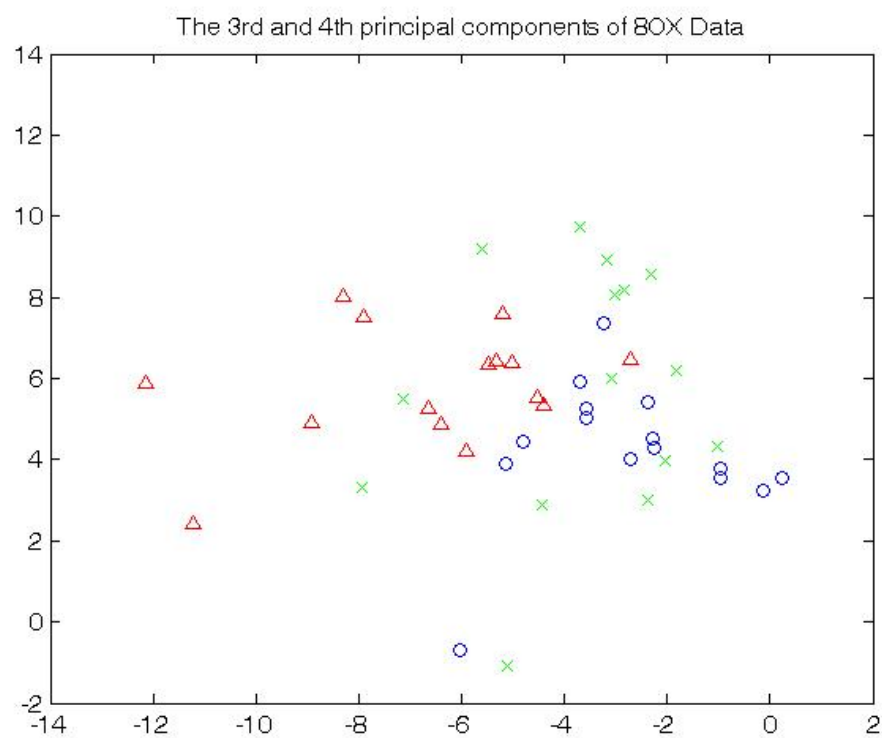
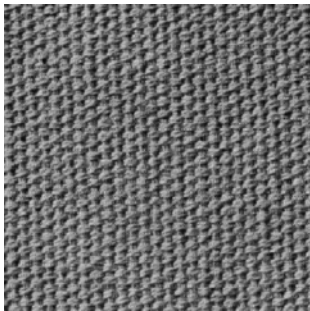
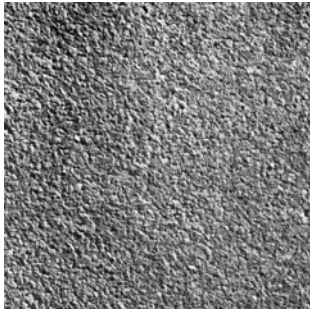


Illustration of PCA (2)



Texture Discrimination



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

Shape Discrimination



- Canada Map
- Italy Map
- Taiwan Map

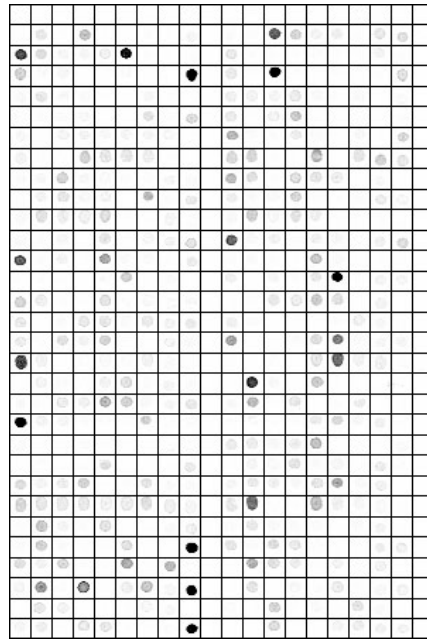
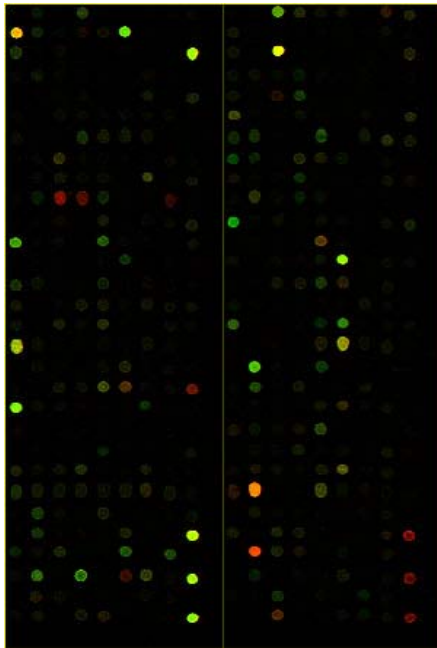
Other Applications

- 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)

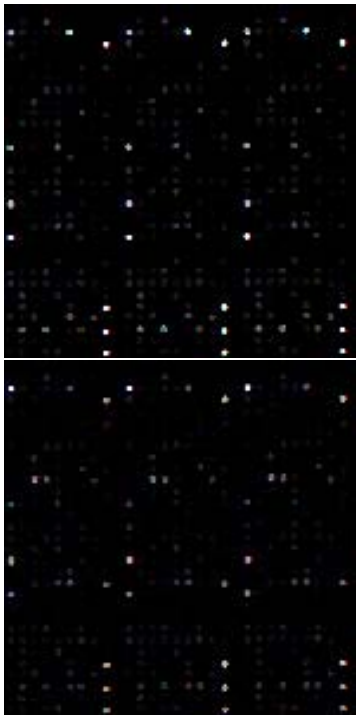
Main Steps

- 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