Outline

- Patterns and Classes
- Decision-Theoretic Methods
- Minimum Distance Classifier
- Matching by Correlations
- Bayes Classifiers
- Assignments
Patterns & Classes

- Patterns are usually vectors comprising qualitative or structural descriptors. Say there are \( n \) number of patterns (often referred to as features)

\[
x = [x_1, x_2, x_3, \cdots, x_n]^\top
\]

- Class is a family of patterns that share common properties. Say there are \( W \) number of classes

\[
\omega_1, \omega_2, \omega_3, \cdots, \omega_W
\]

- Classical works is known as Fisher discriminant analysis (1963)
Example

\[ x_1 = \text{petal length} \]
\[ x_2 = \text{petal width} \]

2-Patterns

\[ \omega_1 = \text{Iris virginica} \]
\[ \omega_2 = \text{Iris versicolor} \]
\[ \omega_3 = \text{Iris setosa} \]

3-Classes
Choice of Features

- Degree of class separability depends strongly on the choice of descriptors selected for an application.

- Patterns can be noisy. See noisy signature in the figure. Other quantitative features can be statistical moments, Fourier descriptors, etc.

- Features with structural relations include the size and positions of minutiae that describes the ridge properties of fingerprints called primitive components such as abrupt endings, branching, merging and disconnected segments.

\[
x_1 = r(\theta_1) \\
x_2 = r(\theta_2) \\
\ldots \\
x_n = r(\theta_n)
\]
Choice of Features

- String of primitive components (say a and b) associated with a boundary shape often adequately describes structural patterns from head-to-tail.

\[ \cdots \text{abababab} \cdots \implies \text{patterns} \]

- Hierarchical ordering of patterns often produces tree structure of the problem of classification.
Choice of Features

Satellite image

Tree description of image
Feature Selection

- Noisy features are refined using the criterion of maximum relevancy and minimum redundancy of the features among the classes.

- An example of selection of features use Fisher criterion that uses the maximum of the ratio between inter-class and intra-class distances of the features to select only the discriminative set of features.

- Refinement of features not only improves the classification accuracy but also reduces the computational complexity.
Decision-Theoretic Recognition

- Let the decision functions be \( d_1(x), d_2(x), \ldots, d_w(x) \).

- A pattern \( x \) belongs to a class \( \omega_i \) if
  \[
  d_i(x) > d_j(x) \quad j = 1, 2, \ldots, W \quad i \neq j
  \]

- Ties are resolved arbitrarily.

- Decision boundary of classes \( \omega_i \) and \( \omega_j \) is
  \[
  d_{ij}(x) \implies d_i(x) - d_j(x) = 0
  \]

  such that
  \[
  d_{ij}(x) > 0 \implies \text{pattern class } \omega_i
  \]
  \[
  d_{ij}(x) < 0 \implies \text{pattern class } \omega_j
  \]
Minimum Distance Classifier

- Use Euclidean distance of feature vectors to determine a class.

- Let \( N_j \) is the number of pattern vectors of class \( \omega_j \). Then the mean of pattern vector is

\[
\mathbf{m}_j = \frac{1}{N_j} \sum_{x \in \omega_j} \mathbf{x}_j \quad j = 1, 2, \ldots, W
\]

- The distance of a given pattern vector \( \mathbf{x} \) from the mean vector is

\[
D_j(\mathbf{x}) = \left\| \mathbf{x} - \mathbf{m}_j \right\| = \left( \mathbf{x}^\top \mathbf{x} - \mathbf{x}^\top \mathbf{m}_j - \mathbf{x} \mathbf{m}_j^\top + \mathbf{m}_j^\top \mathbf{m}_j \right)^{1/2}
\]
Minimum Distance Classifier

- Since the term $x^T \mathbf{x}$ is independent of class label $j$ the decision function is

$$d_j(x) = x^T \mathbf{m}_j - \frac{1}{2} \mathbf{m}_j^T \mathbf{m}_j \quad j = 1, 2, \ldots, W$$

- The decision boundary is

$$d_{ij}(x) = d_i(x) - d_j(x)$$

$$= x^T (\mathbf{m}_i - \mathbf{m}_j) - \frac{1}{2} (\mathbf{m}_i - \mathbf{m}_j)^T (\mathbf{m}_i + \mathbf{m}_j) = 0$$

- The decision surface is perpendicular to the bisector of the line segment joining $\mathbf{m}_i$ and $\mathbf{m}_j$

$n = 2 \Rightarrow$ line bisector

$n = 3 \Rightarrow$ plane bisector

$n > 3 \Rightarrow$ hyperplane bisector
Minimum Distance Classifier

Example

$\omega_1 \Rightarrow \text{Class Iris versicolor}$
$\omega_2 \Rightarrow \text{Class Iris setosa}$

The decision functions are

$$d_1(x) = x^T m_1 - \frac{1}{2} m_1^T m_1$$
$$= 4.3x_1 + 1.3x_2 - 10.1$$

$$d_2(x) = x^T m_2 - \frac{1}{2} m_2^T m_2$$
$$= 1.5x_1 + 0.3x_2 - 1.17$$

The decision boundary is

$$d_{12}(x) = d_1(x) - d_2(x)$$
$$= 2.8x_1 + x_2 - 8.9 = 0$$

$m_1 = \begin{bmatrix} 4.3 \\ 1.3 \end{bmatrix}$ and $m_2 = \begin{bmatrix} 1.5 \\ 0.3 \end{bmatrix}$
Minimum Distance Classifier

- Decision boundary of two classes is a line

\[ d_{12}(x) > 0 \Rightarrow \text{pattern class } 1 \]
\[ d_{12}(x) < 0 \Rightarrow \text{pattern class } 2 \]
Minimum Distance Classifier

- Minimum distance classifier is computationally very fast

- The classifier shows optimum performance if the distribution of patterns for each class about its mean is in the form of a spherical hyper-cloud in n-dimensional space

- Example of large mean separation and small class spread happens in designing E-13B font character set used by the American Banker’s Association.

- The 14 characters were designed on 9x7 grid to facilitate reading. The characters are printed using magnetic materials. Scanning produces 1D electrical waveform which are distinct.
Minimum Distance Classifier

- E-13B font character used by American Banker’s Association.
- Area of waveform remains approximately constant and zero derivative near the middle of the character.
- The mean vectors are \( m_j \), \( j = 1, 2, \ldots, 14 \).
- High classification speed is obtained using the minimum distance classifier.


Matching by Correlation

- Let the template for recognition is $w(x, y)$
- The correlation coefficients

$$
\gamma(x, y) = \frac{\sum \sum [w(s, t) - \overline{w}] \sum \sum [f(x + s, y + t) - \overline{f}(x + s, y + t)]}{\left[ \sum \sum [w(s, t) - \overline{w}]^2 \sum \sum [f(x + s, y + t) - \overline{f}(x + s, y + t)]^2 \right]^{1/2}}
$$

- The correlation coefficient can be measured using Fourier transform efficiently

$$
\gamma(x, y) = f(x, y) \otimes w(x, y) = F^*(u, v) W(u, v)
$$

- The maximum correlation $(-1 \leq \gamma \leq 1)$ gives the positional information of the template

© Dr. S. M. Mahbubur Rahman
Matching by Correlation

(a) Satellite image of hurricane
(b) Template of eye of storm
(c) Correlation coefficients
(d) Position of eye estimated
Bayes Classifier

- Let the probability that a pattern in test \( x \) comes from the class \( \omega_i \) is \( p(\omega_i | x) \).

- Let \( L_{ij} \) denote the loss when the classifier decides \( x \) came from the class \( \omega_j \), when its actual class is \( \omega_i \).

- Since the pattern \( x \) may belong to anyone of the \( W \) classes, the average loss incurred in assigning \( x \) to the class \( \omega_j \) is

\[
r_j(x) = \sum_{k=1}^{W} L_{kj} p(\omega_k | x)
\]

- Using Bayes formula the conditional average risk or loss can be written as

\[
r_j(x) = \frac{1}{p(x)} \sum_{k=1}^{W} L_{kj} p(x | \omega_k) p(\omega_k)
\]
Bayes Classifier

- Here $p(x | \omega_k)$ ➔ PDF of patterns in class $\omega_k$
- $P(\omega_k) ➔$ Probability of occurrence of class $\omega_k$

- Since the term $p(x)$ is positive and common to all $r_j(x)$, when $j = 1, 2, \cdots, W$ it can be dropped for estimating smallest or largest value of the loss. Hence, the average loss function reduces to

$$r_j(x) = \sum_{k=1}^{W} L_{kj} p(x | \omega_k) P(\omega_k)$$

- The classifier that minimizes the total average risk is called the Bayes classifier
Bayes Classifier

- The classifier assigns an unknown pattern \( \mathbf{x} \) in the class \( \omega_i \) only if \( r_i(\mathbf{x}) < r_j(\mathbf{x}) \) for \( j = 1, 2, \ldots, W ; i \neq j \), i.e.,

\[
\sum_{k=1}^{W} L_{ki} p(\mathbf{x} | \omega_k) P(\omega_k) < \sum_{q=1}^{W} L_{qj} p(\mathbf{x} | \omega_q) P(\omega_q) \quad \forall j \ i \neq j
\]

- The loss function is zero and unity for correct and incorrect decisions, respectively. Hence,

\[
L_{ij} = 1 - \delta_{ij}
\]

- The average loss function becomes

\[
r_j(\mathbf{x}) = \sum_{k=1}^{W} (1 - \delta_{kj}) p(\mathbf{x} | \omega_k) P(\omega_k)
= p(\mathbf{x}) - p(\mathbf{x} | \omega_j) P(\omega_j)
\]
Bayes Classifier

- The Bayes classifier assigns the unknown pattern $\mathbf{x}$ in the class $\omega_i$ only if

$$p(\mathbf{x}) - p(\mathbf{x} | \omega_i)P(\omega_i) < p(\mathbf{x}) - p(\mathbf{x} | \omega_j)P(\omega_j)$$

for $j = 1, 2, \ldots, W; i \neq j$

- In other words, the classifier considers

$$p(\mathbf{x} | \omega_i)P(\omega_i) > p(\mathbf{x} | \omega_j)P(\omega_j)$$

for $j = 1, 2, \ldots, W; i \neq j$

- In conclusion, the decision function of Bayes classifier under the 0-1 loss function is a conditional density given by

$$d_j(\mathbf{x}) = p(\mathbf{x} | \omega_j)P(\omega_j)$$

for $j = 1, 2, \ldots, W$
Bayes Classifier

- The Bayes classifier needs PDF of patterns in each of the classes $p(x|\omega_j)$ and the probability of occurrence of each class $P(\omega_j)$.

- The PMFs are usually inferred from previous knowledge or training set of the problem. For example, if all classes are equally likely then

$$P(\omega_j) = \frac{1}{W}$$

- The PDFs $p(x|\omega_j)$ need multivariate analysis since the dimension of $x$ is $n$. In practice, analytical solutions are obtained assuming parametric PDFs for patterns in a class.
Naïve Bayes Classifier

- The Bayes classifier assuming Gaussian PDF of patterns in each of the classes is known as Naïve Bayes classifier. Example for two equi-probable patterns has a decision boundary at \( X_0 \).

- The decision function is

\[
d_j(x) = p(x | \omega_j)p(\omega_j) = \frac{1}{\sqrt{2\pi\sigma_j^2}} e^{-\frac{(x-m_j)^2}{2\sigma_j^2}} P(\omega_j)
\]

- For \( j = 1,2 \)

- On the decision boundary

\[
d_1(x_0) = d_2(x_0)
\]

- If classes are equi-probable

\[
P(\omega_1) = P(\omega_2)
\]

- Then, \( X_0 \) is s.t.

\[
p(x_0 | \omega_1) = p(x_0 | \omega_2)
\]
Naïve Bayes Classifier

- The Naïve Bayes classifier assuming multivariate Gaussian PDF of patterns with dimension $n$ can be obtained.

- The PDF for $j$-th pattern class is

$$p(x \mid \omega_j) = \frac{1}{\sqrt{(2\pi)^n |C_j|}} e^{-\frac{1}{2}(x-m_j)^T C_j^{-1}(x-m_j)}$$

where

Mean vector

$$m_j = E_j \{x\} \approx \frac{1}{N_j} \sum_{x \in \omega_j} x$$

Covariance-Matrix

$$C_j = E_j \{(x-m_j)(x-m_j)^T\} \approx \frac{1}{N_j} \sum_{x \in \omega_j} (xx^T - m_j m_j^T)$$

- The off-diagonal elements $c_{jk}$ are zero when $x_j$ and $x_k$ are uncorrelated.
Naïve Bayes Classifier

- Using $\ln$ as monotonic function the decision function of Bayes classifier becomes

$$d_j(x) = \ln(p(x | \omega_j)P(\omega_j))$$

$$= \ln P(\omega_j) + \ln p(x | \omega_j)$$

- Considering multivariate Gaussian as prior for pattern classes

$$d_j(x) = \ln P(\omega_j) + \ln p(x | \omega_j)$$

$$= \ln P(\omega_j) - \frac{n}{2} \ln(2\pi) - \frac{1}{2} \ln |C_j| - \frac{1}{2} [(x - m_j)^T C_j^{-1}(x - m_j)]$$

- Decision function is hyperquadratic
Naïve Bayes Classifier

- If the covariance matrix is equal for all classes, i.e., \( C_j = C \) for \( j = 1, 2, \ldots, W \)

- The linear decision function in terms of hyper-plane is obtained by eliminating all terms except with terms having \( j \)

\[
d_j(x) = \ln P(\omega_j) + x^T C^{-1} m_j - \frac{1}{2} m_j^T C^{-1} m_j
\]

for \( j = 1, 2, \ldots, W \)

- If the covariance matrix \( C = I \) and classes have \( P(\omega_j) = \frac{1}{W} \)

\[
d_j(x) = x^T m_j - \frac{1}{2} m_j^T m_j
\]

for \( j = 1, 2, \ldots, W \)

- Bayes classifier becomes minimum distance classifier (hyper-spheres) when (i) pattern classes are Gaussian, (ii) covariance matrix is identity matrix and (iii) all classes are equally likely
Naïve Bayes Classifier

- Examples of 2 pattern classes each having 3 dimensions of patterns.

- Given that

  \[ P(\omega_1) = P(\omega_2) = \frac{1}{2} \]

  \[ m_1 = \frac{1}{4} \begin{bmatrix} 3 \\ 1 \\ 1 \end{bmatrix} \quad m_1 = \frac{1}{4} \begin{bmatrix} 1 \\ 3 \\ 3 \end{bmatrix} \]

  \[ c_1 = c_2 = \frac{1}{16} \begin{bmatrix} 3 & 1 & 1 \\ 1 & 3 & -1 \\ 1 & -1 & 3 \end{bmatrix} \]

- The decision function

  \[ d_j(x) = x^T c^{-1} m_j - \frac{1}{2} m_j^T c^{-1} m_j \quad \text{for} \quad j = 1, 2 \]

- The decision surface

  \[ d_1(x) - d_2(x) = 8x_1 - 8x_2 - 8x_3 + 4 \]
Naïve Bayes Classifier

- Examples of 3 pattern classes each having 4 dimensions of patterns in classifying objects of multispectral images.

- The patterns are
  
  \[ X_1 = \text{Visible blue (Band-1)} \]
  
  \[ X_2 = \text{Visible green (Band-2)} \]
  
  \[ X_3 = \text{Visible red (Band-3)} \]
  
  \[ X_4 = \text{Infra red (Band-4)} \]

- The classes are
  
  class #1 = Water body
  
  class #2 = Urban development
  
  class #3 = Vegetation
Bayes Classifier

(a) (b) (c)
(d) (e) (f)
(g) (h) (i)

(a)-(d) Four spectral images, (e) training regions for 3-type of objects (f) classification results of training sets (observe errors shown in black dots) (g) water body (h) urban area (i) vegetation for whole image
Bayes Classifier

- Confusion matrix for the results of classification
- Condition is $P(\omega_j) = \frac{1}{3}$ for $j = 1, 2, 3$

### Training Patterns

<table>
<thead>
<tr>
<th>Class</th>
<th>No. of Samples</th>
<th>Classified into Class</th>
<th>% Correct</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>1</td>
<td>484</td>
<td>482</td>
<td>2</td>
</tr>
<tr>
<td>2</td>
<td>933</td>
<td>0</td>
<td>885</td>
</tr>
<tr>
<td>3</td>
<td>483</td>
<td>0</td>
<td>19</td>
</tr>
</tbody>
</table>

### Independent Patterns

<table>
<thead>
<tr>
<th>Class</th>
<th>No. of Samples</th>
<th>Classified into Class</th>
<th>% Correct</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>1</td>
<td>483</td>
<td>478</td>
<td>3</td>
</tr>
<tr>
<td>2</td>
<td>932</td>
<td>0</td>
<td>880</td>
</tr>
<tr>
<td>3</td>
<td>482</td>
<td>0</td>
<td>16</td>
</tr>
</tbody>
</table>
Assignment

Problem #1

Show how the minimum distance classifier discussed in connection with theory could be implemented by using $W$ resistor banks ($W$ is the number of classes), a summing junction at each bank (for summing currents), and a maximum selector capable of selecting the maximum of $W$ inputs, where the inputs are currents.

Problem #2

The following pattern classes have Gaussian probability density functions: $\omega_1: \{(0, 0)^T, (4, 0)^T, (4, 4)^T, (0, 4)^T\}$ and $\omega_2: \{(5, 5)^T, (7, 5)^T, (7, 7)^T, (5, 7)^T\}$.

(a) Assume that $P(\omega_1) = P(\omega_2) = \frac{1}{4}$ and obtain the equation of the Bayes decision boundary between these two classes.

(b) Sketch the boundary.
Problem #3

Repeat Problem #2, but use the following pattern classes: \( \omega_1: \{(-1, 0)^T, (0, -1)^T, (1, 0)^T, (0, 1)^T \} \) and \( \omega_2: \{(-4, 0)^T, (0, -4)^T, (4, 0)^T, (0, 4)^T \} \). Observe that these classes are not linearly separable.