Classification

Vladimir Curic
Centre for Image Analysis
Swedish University of Agricultural Sciences
Uppsala University
Outline

- An overview on classification
- Basics of classification
- How to choose appropriate features (descriptors)
- How to perform classification
- Classifiers
Where are we right now?

Figure: Classification in image processing.
What is classification?

- Classification is a process in which individual items (objects/patterns/image regions/pixels) are grouped based on the similarity between the item and the description of the group.

![Diagram of classification process](image-url)
Classification example 1

Figure: Classification of leaves.
Classification example 2

Figure: Histological bone implant images.
Terminology

- Object = pattern = point = sample = vector
- Feature = descriptor = attribute = measurement
- Classifier = decision function (boundary)
- Class
- Cluster
Dataset - example

- By measuring features of many objects we can construct a dataset

<table>
<thead>
<tr>
<th>Object</th>
<th>Perimeter</th>
<th>Area</th>
<th>Label</th>
</tr>
</thead>
<tbody>
<tr>
<td>Apple 1</td>
<td>25</td>
<td>80</td>
<td>1</td>
</tr>
<tr>
<td>Apple 2</td>
<td>30</td>
<td>78</td>
<td>1</td>
</tr>
<tr>
<td>Apple 3</td>
<td>29</td>
<td>81</td>
<td>1</td>
</tr>
<tr>
<td>Pear 1</td>
<td>52</td>
<td>65</td>
<td>2</td>
</tr>
<tr>
<td>Pear 2</td>
<td>48</td>
<td>66</td>
<td>2</td>
</tr>
<tr>
<td>Pear 3</td>
<td>51</td>
<td>63</td>
<td>2</td>
</tr>
</tbody>
</table>
Data set and a representation in the feature space

- Perimeter
- Area
Feature set - Example (Lab 2)

- Colour?
- Area?
- Perimeter?
- ...

[Image of coins]
Features (Descriptors)

- Features are individual measurable properties of a phenomena
- The goal is to choose those features that allow pattern vectors belonging to different classes to occupy compact and disjoint regions
- Discriminating (effective) features
- Independent features
What are good features?

- You might try many, many features, until you find the right ones
- Often, people compute 100s of features, and put them all in a classifier
  - "The classifier will figure out which ones are good"
  - This is wrong!!!
- It is application dependent
Peaking phenomenon (Curse of dimensionality)

- Additional features may actually degrade the performance of a classifier.
- This paradox is called peaking phenomenon (curse of dimensionality).

**Figure:** Curse of dimensionality.
Dimensionality reduction

- Keep the number of features as small as possible
  - Measurements cost
  - Accuracy
- Simplify pattern representation
- The resulting classifier will be faster and will use less memory
- OBS! A reduction in the number of features may lead to a loss in the discriminatory power and thereby lower the accuracy
Feature selection

- Feature selection methods choose features from the original set based on some criteria
- Reduce dimensionality by selecting subset of original features
- Have clear physical meaning
Feature extraction is the process of generating features to be used in the selection and classification tasks. Might not have clear physical meaning.
Supervised vs. Unsupervised classification

- **Supervised**
  - First apply knowledge, then classify
- **Unsupervised**
  - First classify, then apply knowledge

**Figure**: Supervised and unsupervised classification.
Supervised classification

- Train and classify
- Training
  - Find rules and discriminant function that separate patterns in different classes using known examples
- Classification
  - Take a new unknown example and put it into the correct class using discriminant function
Training data

- Training data can be obtained from available training samples
- Classify patterns which are not used during the training stage
- The performance of classifier depends on the number of available training samples as well as the specific values of the samples
- Number of training samples should not be too small!
Training data

- Conditional densities are unknown and they have to be learned from available training data.
- Maybe density is known (for example, Gaussian distribution).
- A common strategy is to replace unknown parameters by their estimated values.
- Less available information - the difficulty of a classification problem increases.
How to choose appropriate training set?
How to choose appropriate training set?

Figure: Original image and training image.
How to choose appropriate training set?

Figure: Original image and training image.
Object-wise and pixel-wise classification

- **Object-wise classification**
  - Uses shape information to describe patterns
  - Size, mean intensity, mean color, etc.

- **Pixel-wise classification**
  - Uses information from individual pixels
  - Intensity, color, texture, spectral information
Object-wise classification

- Lab 2

No. of objects

50 10 kr 70 1 kr 5 kr
50 öre (new) 50 öre (old)

Centre for Image Analysis
Swedish University of Agricultural Sciences
Uppsala University
Object-wise classification

- Segment the image into regions and label them
- Extract features for each pattern
- Train classifier on examples with known class to find discriminant function in the feature space
- For new examples decide their class using the discriminant function
Pixel-wise classification

- 256*256 patterns (pixels)
- 3 features (red, green and blue color)
- Four classes (stamen, leaf, stalk and background)
Pixel-wise classification

- The pattern is a pixel in a non-segmented image
- Extract (calculate) features for each pixel e.g., color, gray-level representation of texture
- Train classifier
- New samples are classified with a classifier
Classifiers

- Once a feature selection finds a proper representation, a classifier can be designed using a number of possible approaches.
- The performance of classifier depends on the interrelationship between sample size, number of features and classifier complexity.
- The choice of a classifier is a difficult problem!!!
- It is often based on which classifier(s) happen to be available or best known to the user.
A discriminant function for a class is a function that will yield larger values than functions for other classes if the pattern belongs to the class

\[ d_i(x) > d_j(x) \quad j = 1, 2, ..., N; \quad j \neq i \]

For \( N \) pattern classes, we have \( N \) discriminant functions.

The decision boundary between class \( i \) and class \( j \) is

\[ d_j(x) - d_i(x) = 0 \]
Decision boundary
Decision boundary

Figure: Decision boundary.
Linear and quadratic classifier

Figure: Linear (left) and quadratic (right) classifier.
Thresholding - simple classification

- Classify image into foreground and background

Figure: Original image and thresholded image.
Bayes classifiers

- Based on a priori knowledge of class probability
- Cost of errors
- Minimum distance classifier
- Maximum likelihood classifier
Minimum distance classifier

- Training is done using the objects (pixels) of known class
- Each class is represented by its mean vector $\mathbf{m}_j$
- Mean of the feature vectors for the object within the class is calculated
- New objects are classified by finding the closest mean vector
- Determine closeness

$$D_j(x) = \|x - \mathbf{m}_j\|, j = 1, 2, ..., N$$
Minimum distance classifier
Limitations of Minimum distance classifier

- When this classifier should be used?
- Optimal performance when distributions of the classes form spherical shape
Minimum distance classifier - equivalence

- Use $\mathbf{x} = (\mathbf{x}^T \mathbf{x})^{\frac{1}{2}}$ to show the equivalence
- Minimal: $D_j(\mathbf{x}) = \|\mathbf{x} - \mathbf{m}_j\|, j = 1, 2, \ldots, N$
- Maximal $d_j(\mathbf{x}) = \mathbf{x}^T \mathbf{m}_j - \frac{1}{2} \mathbf{m}_j^T \mathbf{m}_j, j = 1, 2, \ldots, N$
Bayes’ classifiers

- Bayes’ formula

\[ p(A|B) = \frac{p(A)p(B|A)}{p(B)} \]

- Classify according to the largest probability (taking variance and covariance into consideration)

- If not other specified, assume that distribution within each class is Gaussian

- The distribution in each class can be described by a mean vector and covariance matrix
The optimal rule for minimizing the risk can be stated as follows:

Assign input pattern $\mathbf{x}$ to class $\omega_i$ for which the conditional risk

$$r_i(\mathbf{x}) = \frac{1}{p(\mathbf{x})} \sum_{j=1}^{N} L(\omega_i, \omega_j) \cdot p(\omega_j | \mathbf{x}) P(\omega_k)$$

is smallest

- $L(\omega_i, \omega_j) = L_{ij}$ is the loss when putting $\mathbf{x}$ in $\omega_j$ when it belongs to $\omega_i$
- $p(\omega_j | \mathbf{x})$ is the probability that pattern $\mathbf{x}$ comes from class $\omega_i$
- $P(\omega_k)$ is the probability of class $\omega_i$ occurring

Assign input pattern $\mathbf{x}$ to class $\omega_k$ if the conditional risk $r_k$ is the smallest
Simplification 1: 0-1 Loss function

- Medicine: false negatives and false positives
- More costly are false negatives
- Often assume that all errors have equal cost

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

- \( \delta_{ij} \) Dirac function
- Discriminant function (decision boundary)

\[ d_j(x) = p(x|\omega_j)P(\omega_j) \]
Simplification 2: Assume Gaussian Probability Density Function

- $p(x|\omega_j)$ is a Gaussian for all classes
- Covariance matrix $C_j = E((x - m_j)(x - m_j)^T)$
- Discriminant function
  
  $$d_j(x) = P(\omega_j) - \frac{1}{2} \ln |C_j| - \frac{1}{2}((x - m_j)^T C_j^{-1}(x - m_j))$$
Simplification 3: Assume Equal Probability

- $P(\omega_j)$ is equal for all classes
- Discriminant function
  \[ d_j(x) = -\frac{1}{2} \ln |C_j| - \frac{1}{2} ((x - m_j)^T C_j^{-1} (x - m_j)) \]
- Maximum Likelihood classifier
  - We have 0-1 loss function
  - The pattern classes are Gaussian
  - All classes are equally likely to occur
Simplification 4: Assume Same Covariance

- All classes have the same covariance matrix
- Discriminant function
  \[ d_j(x) = ((x - m_j)^T C^{-1}(x - m_j)) \]
- Linear discriminant function
- Hyperplanes are decision boundaries
- This discriminant function is also called Mahalanobis distance
Simplification 5: Assume Independent Patterns and Same Variance

- All classes have the same variance
- No covariance
- $C$ is identity matrix
- Discriminant function
  \[ d_j(x) = (x - m_j)^T (x - m_j) \]
- This is the Minimum distance classifier
Unsupervised classification

- Difficult, expensive or even impossible to rely label training sample with its true category (It is not possible to obtain ground truth)
- Patterns within a cluster are more similar to each other than are patterns belonging to different clusters
Unsupervised classification

- Difficult
- How to determine the number of clusters $K$?
- The number of clusters often depends on resolution (fine vs. coarse)
**K-means**

- **Step 1.** Select an initial partition with $K$ clusters. Repeat steps 2 through 4 until the cluster membership stabilizes.
- **Step 2.** Generate a new partition by assigning each pattern to its closest cluster center.
- **Step 3.** Compute new cluster centres as the centroids of the clusters.
- **Step 4.** Repeat steps 2 and 3 until an optimum value of the criterion function is found (the cluster are stabilized).
$K$-means
$K$—means
$K$-means - disadvantages

- Different initialization can result in different final clusters
- Fixed number of clusters can make it difficult to predict what $K$ should be
- It is helpful to rerun the classification using the same as well as different $K$ values, to compare the achieved results
- 10 different initializations for 2D data
- For $N$-dimensional data 10 different initializations is often not enough!
Problem

The following eight points $A_1(2, 10)$, $A_2(2, 5)$, $A_3(8, 4)$, $A_4(5, 8)$, $A_5(7, 5)$, $A_6(6, 4)$, $A_7(1, 2)$, $A_8(4, 9)$ should be classify into three clusters using $K$–means clustering. Initial cluster centres are: $A_1(2, 10)$, $A_4(5, 8)$ and $A_7(1, 2)$. Find the three cluster centres after the first iteration. The distance function between two points $A(x_a, y_a)$ and $B = (x_b, y_b)$ is defined as $d(A, B) = |x_a - x_b| + |y_a - y_b|$. 
Summary and conclusion

- Classification is needed in image processing
- It is highly application dependent
- Reading material: Chapters 12-12.2.3
- Problems: 12.1, 12.2, 12.7 and one problem given during the lecture