MODULE 1
INTRODUCTION

LESSON 1

Introduction

Keywords: Supervised, Unsupervised, Semi-Supervised, Classification, Clustering

INTRODUCTION

- This course deals with pattern recognition. A **pattern** is either a physical object, for example a *book or a chair* or an abstract notion, like *style of talking*, or *style of writing*. It is also a shared property of a set of objects; for example, *chairs*, *rectangles*, or *blue* colored objects. We illustrate using ellipses and rectangles shown in Figure 1.

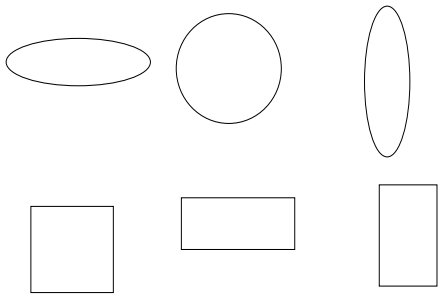

Figure 1: **Ellipses and Rectangles**

Cognition is the act of seeing or perceiving, whereas **recognition** means *as having seen or perceived*. There are three ways of appreciating the activity of **pattern recognition** in a simple manner:

1. **Classification**: Assign a pattern to one of the already known (semantically labelled) classes. For example, consider the two classes of physical objects shown in Figure 1: *ellipses* and *rectangles* where *ellipse* and *rectangle* are class labels. Now the **classification** problem, prominent in pattern recognition, involves:
   (a) Either learn a model or directly use the training data set (collection of labelled patterns) and
   (b) assign a class label to a new pattern (test pattern) or equivalently assign the test pattern to one of the known classes. That is, with respect to objects in Figure 1, given a new object we would like to classify it as either an ellipse or a rectangle.

The Classification problem: We are given a collection of semantically labelled patterns, $\mathcal{X}$, where

$$\mathcal{X} = \{(X_1, C^1), (X_2, C^2), \cdots, (X_n, C^n)\}$$

We need to note the following here:

- The number of classes, K, is fixed and finite. The value of K is known *a priori*. Let the class labels be $C_1,\ C_2,\ \cdots,\ C_K$.
- The set $\mathcal{X}$ is finite and is of size (cardinality) n. Further, X_i represents the i^{th} pattern and C^i is the corresponding semantic class label for $i = 1, \cdots, n$. So, observe that $C^i \in \mathcal{C} = \{C_1,\ C_2,\ \cdots,\ C_K\}$.
- Let X be a test pattern. Then, either we use the training set $\mathcal{X}$ directly or models $M_1,\ M_2,\ \cdots,\ M_K$ learnt from $\mathcal{X}$ to assign a class label, out of $C_1,\ C_2,\ \cdots,\ C_K$, to X. Here, model M_i is learnt from the training patterns drawn from class C_i, for $i = 1,\ 2,\ \cdots,\ K$.

An Example: Let us say that we are given the following collection of *chairs* and *humans* as the training set.

$$\begin{aligned}\mathcal{X} = \{&(X_1, chair), (X_2, chair), (X_3, human), (X_4, human), (X_5, human),\\ &(X_6, chair), (X_7, human), (X_8, chair), (X_9, human), (X_{10}, human)\}\end{aligned}$$

Now the problem is, given a test pattern X, classify X as either *chair* or *human*. In other words, assign one of the two class labels to X.

2. **Clustering**: Assign a pattern to one of the *syntactically labelled classes* or clusters. For example, consider two clusters of patterns, labelled C_1 and C_2. Given a new pattern, assign it to either C_1 or C_2 based on the similarity between the pattern and the collection. Here, the labels are syntactic because we can switch the labels of the two collections without affecting the results. *Clustering* is concerned with grouping of patterns based on similarity. Patterns in a cluster are similar to each other whereas patterns in different clusters are dissimilar.

Clustering Problem: We are given a collection, $\mathcal{X}$, of syntactically labelled patterns, where

$$\mathcal{X} \ = \ \{X_1, \ X_2, \ \cdots, \ X_n\}.$$

Note that the patterns are syntactically labelled using different subscripts. The problem is to partition the set $\mathcal{X}$ into some finite number of blocks or clusters. In other words, we partition $\mathcal{X}$, so that

$$\mathcal{X} \ = \ C_1 \ \cup \ C_2 \ \cup C_3 \ \cdots \ \cup \ C_K$$

where C_i is the i^{th} cluster. Clustering is done so that none of the K clusters is empty and any pair of clusters do not overlap, which means

$$C_i \ \neq \ \emptyset, \ and \ C_i \bigcap \ C_j \ = \emptyset \ for \ i \neq j \ and \ i, j \ \in \ \{1, 2, \ \cdots, \ K\}.$$

An Example of Clustering: Consider a collection of patterns

$$\mathcal{X} \ = \ \{X_1, \ X_2, \ \cdots, \ X_{10}\}.$$

A possible partition, of $\mathcal{X}$ having two clusters is
$C_1 \ = \ \{X_1, X_2, X_4, X_5, X_7, X_8\}$ and $C_2 \ = \ \{X_3, X_6, X_9, X_{10}\}$. Typically, a notion of *similarity* or *matching* is used to partition $\mathcal{X}$. Patterns in C_1 are similar to other patterns in C_1 and patterns in C_2 are similar to other patterns in C_2; a pattern, say X_2, in C_1 is dissimilar to a pattern, say X_9, in C_2. In clustering, it is possible to switch the labels; for example, we have the same partition as above if

$C_1 \ = \ \{X_3, X_6, X_9, X_{10}\}$
$C_2 \ = \ \{X_1, X_2, X_4, X_5, X_7, X_8\}$

3. **Semi-Supervised Classification**: Here, we are given a small collection of semantically labelled patterns and a large collection of syntactically labelled patterns. The problem is to assign a new pattern (test pattern) to one of the classes or equivalently assign a semantic label to the test pattern.

Semi-Supervised Classification Problem: We are given a collection, $\mathcal{X}$, given by

$$\mathcal{X} = \{(X_1, C^1), \cdots (X_l, C^l), X_{l+1}, \cdots X_{l+u}\}$$

where l patterns are semantically labelled and u patterns are syntactically labelled. The problem is to build models $M_1, M_2, \cdots M_K$ corresponding to classes $C_1, C_2, \cdots, C_K$ respectively. Now given a new pattern, X, classify it to one of the K classes using the models built.

An Example: Given a set, $\mathcal{X}$, of patterns given by

$$\mathcal{X} = \{(X_1, human), (X_2, chair), X_3, X_4, X_5, X_6, X_7\}$$

the problem is to assign a class label of *chair* or *human* to a new pattern (test pattern) X.

The popularity of pattern recognition (PR) may be attributed to its application potential; there are several important applications. For example,

- **document recognition:** there are a variety of applications including classification and clustering of
  - email messages and web documents; one requirement is to recognize whether a *mail is spam* or not.
  - fingerprints, face images, and speech signals which form an important variety of documents used in *biometrics*.
  - *health records* which may include x-ray images, ultrasound images, ECG charts and reports on various tests, diagnosis, and prescriptions of medicines.
  - *legal records* including judgments delivered, petitions and appeals made.
- **remote sensed data analysis:** for example, images obtained using satellite or aerial survey are analysed to discriminate healthy crops from deceased crops.
- **bioinformatics:** Here, classification and clustering of DNA and protein sequences is an important activity.

- **semantic computing:** Knowledge in different forms is used in clustering and classification to facilitate natural language understanding, software engineering, and information retrieval.
- There are plenty of other areas like *agriculture, education, and economics* where pattern recognition tools are routinely used.

Abstractions

In machine recognition of patterns, we need to process patterns so that their representations can be stored on the machine. Not only the pattern representations, but also the classes and clusters need to be represented appropriately. In pattern recognition, inputs are abstractions and the outputs also are abstractions.

- As a consequence, we do not need to deal with all the specific details of the individual patterns.
- It is meaningful to summarize the data appropriately or look for an apt abstraction of the data.
- Such an abstraction is friendlier to both the human and the machine.
- For the human it is easy for comprehension and for the machine it reduces the computational burden in the form time and space required for processing.
- Generating an abstraction from examples is a well-known paradigm in machine learning.
- Specifically, learning from examples or supervised learning and learning from observations or clustering are the two important machine learning paradigms that are useful here.
- In artificial intelligence, the machine learning activity is enriched with the help of domain knowledge; abstractions in the form of rule-based systems are popular in this context.
- In addition data mining tools are useful when the set of training patterns is large.

- So, naturally pattern recognition overlaps with machine learning, artificial intelligence and data mining.

Two popular paradigms for pattern recognition are:

- **statistical pattern recognition:** In this case, **vector-spaces** are used to represent patterns and collections of patterns. Vector-space representations are popular in *information retrieval*, *data mining*, and *statistical machine learning*. Abstractions like vectors, graphs, rules or probability distributions are used to represent clusters and classes.

- **syntactic pattern recognition:** In this case, patterns are viewed as sentences in a formal language like mathematical logic. So, it is useful in describing classes and clusters of well-structured patterns. This paradigm is popular as *linguistic* or *structural pattern recognition*.

- Readers interested in some of these applications may refer to popular journals such as Pattern Recognition (www.elsevier.com/locate/pr) and IEEE Transactions on Pattern Analysis and Machine Intelligence (www.computer.org/tpami) for details. Similarly, for specific application areas like bioinformatics refer to Bioinformatics (http://bioinformatics.oxfordjournals.org/) and for semantic computing refer to International Journal of Semantic Computing (www.worldscinet.com/ijsc/). An excellent introduction to syntactic pattern Recognition is providede by *Syntactic Pattern Recognition: An Introduction* by RC Gonzalez and MG Thomason, Addision-Wesley, 1978.

Assignment

Solve the following problems:

1. Consider the data, of four adults indicating their health status, shown in the following table. Devise a simple classifier that can properly classify all the four patterns. How is the fifth adult having a weight of 65 KGs classified using the classifier?

Weight of Adults in KGs	Class label
50	Unhelathy
60	Healthy
70	Healthy
80	Unhealthy

2. Consider the data items bought in a supermarket. The features include cost of the item, size of the item, colour of the object and the class label. The data is shown in the following table. Which feature would you like to use for classification? Why?

item no	cost in Rs.	volume in cm^3	colour	Class label
1	10	6	blue	inexpensive
2	15	6	blue	inexpensive
3	25	6	blue	inexpensive
4	150	1000	red	expensive
5	215	100	red	expensive
6	178	120	red	expensive