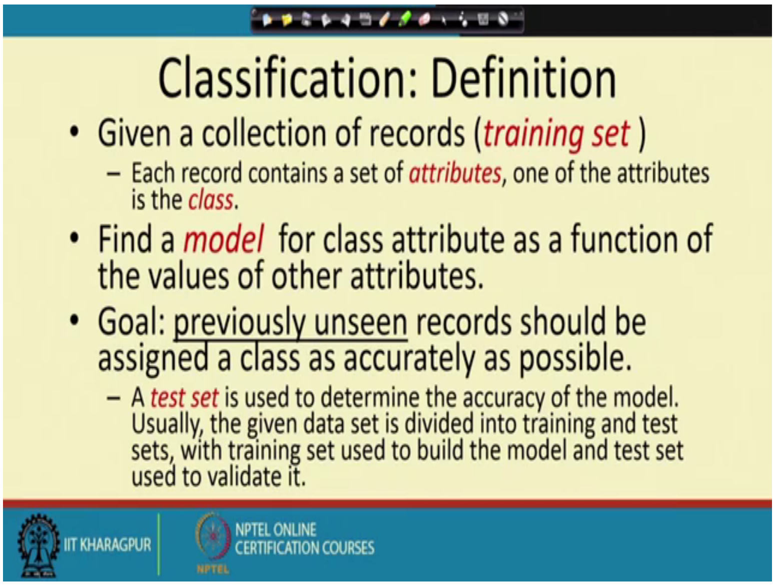
Data Mining
Prof. Pabitra Mitra
Department of Computer Science & Engineering
Indian Institute of Technology, Kharagpur

Lecture – 07
Classification

(Refer Slide Time: 00:35)

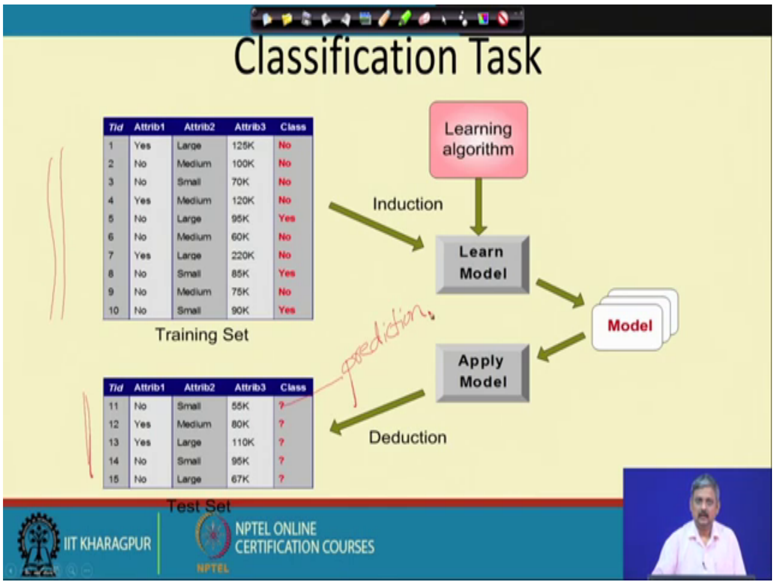


Classification: Definition

- Given a collection of records (*training set*)
  - Each record contains a set of *attributes*, one of the attributes is the *class*.
- Find a *model* for class attribute as a function of the values of other attributes.
- Goal: previously unseen records should be assigned a class as accurately as possible.
  - A *test set* is used to determine the accuracy of the model. Usually, the given data set is divided into training and test sets, with training set used to build the model and test set used to validate it.

Welcome to the lecture on classification algorithms. So, this is the second type of pattern that we will study the first pattern we studied was that of association rules. So, let me define what classification means. The idea is the following it is you are given a set of attributes and a set of problems belonging to a class let me tell you with an example, let me explain you with an example.

(Refer Slide Time: 00:59)



Classification Task

Tid	Attrib1	Attrib2	Attrib3	Class
1	Yes	Large	125K	No
2	No	Medium	100K	No
3	No	Small	70K	No
4	Yes	Medium	120K	No
5	No	Large	95K	Yes
6	No	Medium	60K	No
7	Yes	Large	220K	No
8	No	Small	85K	Yes
9	No	Medium	75K	No
10	No	Small	90K	Yes

Training Set

Tid	Attrib1	Attrib2	Attrib3	Class
11	No	Small	55K	?
12	Yes	Medium	80K	?
13	Yes	Large	110K	?
14	No	Small	95K	?
15	No	Large	67K	?

Test Set

Learning algorithm

Induction

Learn Model

Model

Apply Model

Deduction

prediction

IIT KHARAGPUR

NPTEL ONLINE CERTIFICATION COURSES

NPTEL

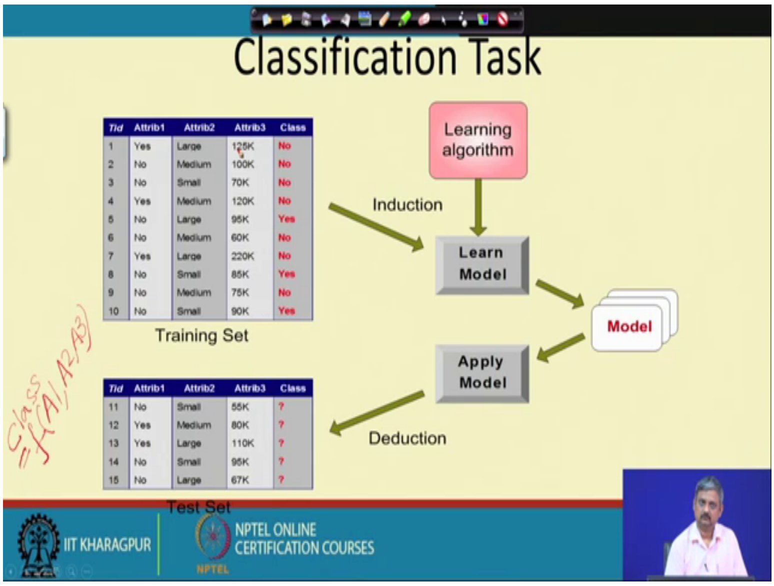
So, I will show. Suppose I have a table like this, I will come back to the details of this. Let me you just focus on the table. Look at this table. So, you see that this is a table consisting of some attributes attribute one, attribute two, attribute three, whose values are like yes, no, large, medium, small, 125 k, 100 k some values. And the final column marked in red is something called a class. So, each of the rows are object. So, you can imagine sort of you can think of them suppose it is like a each row is a person - record of a person, who attribute one is let say he is employed or unemployed yes or no. Attribute 2 is maybe he stays in a small home or a large home or a medium home. Attribute 3 is may be his annual income in rupees 100 k or 125 is annual income.

And then maybe I want to I know that this each of this person each of this rows belong to one of two categories, either you give them loan if they apply for a loan, they will repay their loan or you should not give them loan, yes or no. I should give them loan or I should not give them loan. So, what I will do is that I will have something called a experience or a training set where I will have ten such persons this attribute details and the class to be loan to be given or not, known. So, I have this ten persons for which it is known and my job is for five more persons whose attribute values I know these employed no, house small, income 55 k for them whether they are actually given loan or not that I do not know I want to predict whether they should be given loan or not. So, these persons I know, they are given loan, they are not given loan, they are given loan they are not given loan and this five persons, I do not know whether given loan or not.

So, what I want to do is to kind of predict what the values of this class attribute should be for each of this five points. So, this I call my test set or unknown examples, this I call my unknown examples. So, for these five examples, I should predict. So, this is my prediction problem. So, the value to be predicted can be a group that is single person falls into; in this case only two groups, loan given or loan not given. In general, you can have more group, so these groups are called classes of that person classes or categories of that person. So, basically what I am trying to do is that knowing the class or category of say 10 training set persons, I want to predict the class of some new persons testing persons for which class is currently unknown, but I want to predict them from this history from this previous examples. So, this problem is known as the classification problem.

So, let me write them down in plain English. Given a training set, which is a collections of rows or records, each consisting of some attribute, one of which is a class. Find a function in this case that is the pattern which will take as input this attributes and predict the class. Take as input this attribute and predict the class.

(Refer Slide Time: 06:01)



Classification Task

Tid	Attrib1	Attrib2	Attrib3	Class
1	Yes	Large	125K	No
2	No	Medium	100K	No
3	No	Small	70K	No
4	Yes	Medium	120K	No
5	No	Large	95K	Yes
6	No	Medium	60K	No
7	Yes	Large	220K	No
8	No	Small	85K	Yes
9	No	Medium	75K	No
10	No	Small	90K	Yes

Training Set

Tid	Attrib1	Attrib2	Attrib3	Class
11	No	Small	55K	?
12	Yes	Medium	80K	?
13	Yes	Large	110K	?
14	No	Small	95K	?
15	No	Large	67K	?

Test Set

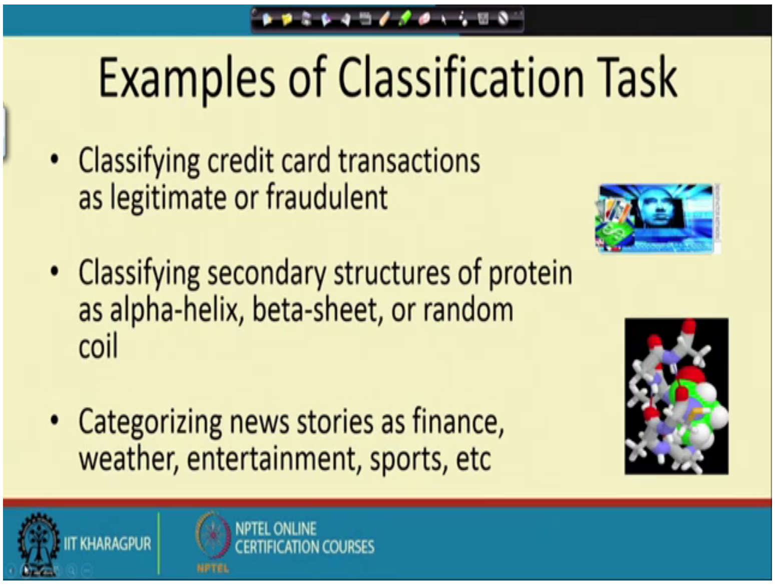
Class = f(A1, A2, A3)

So, in this case for examples I can what I want to do is that, if I want to write class as a function of this attribute 1, attribute 2, attribute 3 as a function of this. So, this function will take on this three values yes, large, 125 k, and return - yes or no. What are the parameters of this function that I will have to learn from these 10 examples. So, let me coming back to the definition. W have to find a function or a model from the previous records which would classify the new records the future records and assign a class to them based on that attribute values. This problem is known as the classification problem.

You can actually see that this is a very common problem that human being solve. How they solve it? Suppose, you have gone to a doctor and every patient is categorized by some attribute say has high temperature or not, has cough or not, has fewer or not; and based on this symptoms and attributes the doctor classifies the patient into a disease say influenza or flu or malaria puts it into a class. Predicts the class for a how does the doctor does it, she has looked at various influenza and malaria patients in the past; she has trained over them. And from this previous experience she has spitted a model and for a new patient comes, using this model she tries to predict the class of that new person.

Similarly, let say suppose you are crossing a road and you are seen a car or a person or a vehicle. So, a kid saw the parent tells that this is a car, this is a truck, this is a cow. So, this is a training that the kid is going through, this is called super wise learning this classification is also sometimes called super wise learning. And when the kid sees a new person the kid can knew say car or a bus or a human or a cow, the kid can classify it. So, this is what is the classification problem. So, if I draw it pictorially there are two steps in classification, one is from the training set you learn a model and then apply the model on a new problem.

(Refer Slide Time: 09:28)

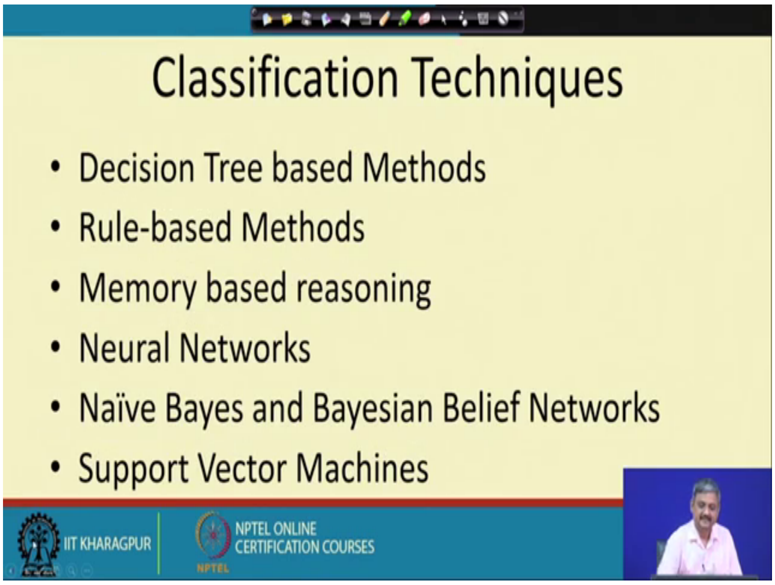


Examples of Classification Task

- Classifying credit card transactions as legitimate or fraudulent
- Classifying secondary structures of protein as alpha-helix, beta-sheet, or random coil
- Categorizing news stories as finance, weather, entertainment, sports, etc

Here are some examples say classifying credit card, classifying proteins into one of this groups, categorizing new stories as finance or which category type of new story it is, classifying email as spam or non spam, all these are examples of classification task. Commonly in your practice, we will encounter many such tasks.

(Refer Slide Time: 10:17)



Classification Techniques

- Decision Tree based Methods
- Rule-based Methods
- Memory based reasoning
- Neural Networks
- Naïve Bayes and Bayesian Belief Networks
- Support Vector Machines

Now, depending on this what this function or model you take you can have different type of classifiers, decision tree methods, rule based methods, memory reasoning neural networks. So, in our course, we will gradually go through them. So, today I end my lecture here.

Thank you. In the next class, we will go into details of this individual techniques starting with the decision tree. I hope you have understood the fundamental problem of classification. Next, we will study different techniques to perform the classification.

Thank you.