Lecture 2: Data Mining Activities

BITS Pilani
Pilani | Dubai | Goa | Hyderabad

Types of Data Mining Activities

At a high level, data mining activities can be categorized into two types:

a. **Description**
b. **Prediction**

Description: Given a large amounts of data, it is not easy to make any interpretation by simple visual observation. A significant processing of this data may help to understand or uncover some patterns within, which in turn may describe the data. This description brings in an insight about the data.

Prediction: Prediction is simpler to observe in the data. With huge volume of events and/or data available, it is possible to predict future occurrence of the event/data or predict value of a variable.

Examples of Prediction:

a. Predicting whether a person is going to buy some particular item
b. Predicting the temperature of a particular region at a particular point-in-time

In summary, prediction is representing what we want to find out with respect to some variable for the future.

Common types of Data Mining Tasks:

More popular and common types of data mining activities are as follows:

	Type	Data Mining Category	Remarks
a.	Classification	Predictive	Use training set to build a model to predict classification
b.	Clustering	Descriptive	
c.	Association Rule Discovery	Descriptive	Creating rules based on observed data
d.	Sequential Patterns Discovery	Descriptive	Creating sequential patterns based on observed data
e.	Regression	Predictive	Prediction of a numerical value based on historical information
f.	Deviation/Anomaly Detection	Predictive	Prediction of unusual happenings or accidental event based on historical information

Classification

A collection of records where the outcome or value of a target attribute that we are looking at is already known. This set of records is known as **training set**. The task here is to find a model to predict the value of **class** attribute as a function of the values of other attributes. The goal of the model is that it should be able to assign a class as accurately as possible even for a new set of records. Usual process is to divide given set of records into training and **test sets**. The training set is used to build the model while test set is used to validate the model or determine the accuracy of the model.

Consider a scenario where a set of customer records are available with their attributes along with the fact that whether they have bought a particular product or not. With this, it is possible to create a model whereby in future this model can be applied on a customer record to predict upfront, if this customer is likely to buy the product or not. Similarly, the same technique can be applied for several other possible events and predict the future occurrence.

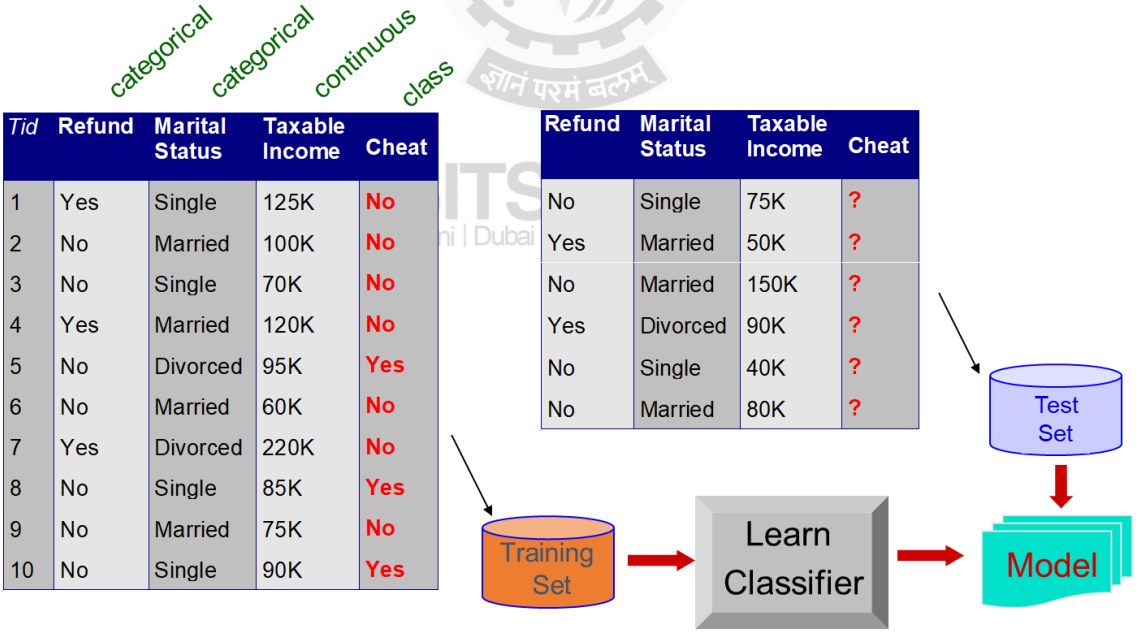
Classification Example:

Consider the following set of tax payers' records where the attributes include marital status, taxable income, whether they have got the refund or not. We need to determine whether they have cheated with respect to their tax returns or not.

Refund	Marital Status	Taxable Income	Cheat
No	Single	75K	?
Yes	Married	50K	?
No	Married	150K	?
Yes	Divorced	90K	?
No	Single	40K	?
No	Married	80K	?

In order to build a model, we need a training set which is nothing but the past data where the tax department has done the investigation and found out who have cheated and who have not. Attributes such as the marital status, taxable income, refund etc can be used to build and train a model since the outcome (whether that person has cheated or not) is already known for these records.

A classifier is used as shown below to learn these attributes and create a model which will be able to predict or classify new records (test set) for which the outcome is not known.



Note that new records will not be identical to training set. However, we can always make sense based on the values. For example, a person with income of $125,000 or $100,000 has not cheated and a person with an income in some other range such as $30,000 has cheated. Therefore, what is the likelihood of a person cheating if he or she has an income of

$110,000? Hence the function of the classifier is to understand the training data and create a model which can be based on all the attributes.

Note that the above example is an extremely simple case where there are only three attributes and only ten records. However in reality, there will be large dimensional problems which are common in data mining domain. So, here they need to have a combination of all the dimensions and look at what is the likely outcome.

Classification Application 1: Direct Marketing

Many organizations spend a lot of money in promoting their product to a large set of customers. However, if they identify or classify the customers based on their past purchasing habits, then it is quite likely that they will need to promote their product with far lesser number of people and achieve a high hit rate. This reduces the promotional expenditure significantly.

Goal: Reduce the cost of mailing by targeting a set of customers likely to buy a new cell-phone product

Approach:

Training Set: Use data for a similar product introduced earlier.

Class attribute: From the past data, we know who bought / did not buy the product

Other attributes: Collect various demographic, lifestyle, company-interaction related information, type of business, their locality, annual income etc.

Classifier: Use all the information collected as input attributes to learn a classifier model

Classification Application 2: Fraud Detection

Growing volume of online transactions and possible fraudulent transactions has necessitated Fraud detection. The repository of fraudulent transactions is available with the credit or debit card issuers or those companies which are dealing with such transactions. It is therefore possible to create a model to detect a fraudulent transaction based on the attributes. For example, if a particular type of transaction happens for a particular customer, at a particular location, for a particular value or many such conditions can possibly signal a fraudulent transaction. The attributes involved can be numerous, however, this will help detect such transactions quickly and prevent further loss. Without such fraud detection mechanism, a lot of time would be lost in victim's complaint, investigation and finally catching the person involved. Whereas, by creating a model which can predict or identify a transaction or classify a transaction as a potential fraud, we can prevent the fraud easily.

Goal: Prevent fraudulent cases in credit card transactions

Approach:

Training Set: Use credit card transactions and the information on its account holder as attributes such as when does a customer buy, what does he/she buy, how often does he/she pay on time, etc.

Class attribute: Label the past transactions as fraud or fair transactions. This forms the class attribute.

Classifier: Learn a model for the class of the transactions.

Detect Fraud: Use this model to detect fraud by observing credit card transactions on an account.

Classification Application 3: Customer Attrition/Churn

Businesses always strive to retain their customers and hence are concerned about customer attrition. This is a common concern due to the high cost of customer acquisition.

Let us take a case of a telecom company which is trying to find out whether a customer is going to stay or is he/she going to move out to another competitor. It is possible for them to take some action if they can predict the attrition. Similarly, customer attrition is common to all the businesses and will help if predicted upfront.

Goal: To predict whether a customer is likely to be lost to a competitor

Approach:

Training Set: Use detailed record of transactions with each of the past and present customers to find attributes. You may consider attributes related to calls by the customer such as frequency of customer call, location of the call, time-of-the day and details such as financial status, marital status etc.

Class attribute: Label the customers as loyal or disloyal. This forms the class attribute.

Classifier: Find a model for loyalty.

Classification Application 4: Sky Survey Cataloguing

Data mining can be applied to astronomy as well. Telescopes are used to observe the sky and they constantly send images. However, it is almost impossible to identify celestial objects manually considering the size of the data and types of objects that exist in this universe. Therefore applying classification techniques stand a higher chance of detecting celestial bodies.

Goal: To predict class (star or galaxy) of sky objects, especially visually faint ones based on the telescopic survey images (from Palomar Observatory). Given data: About 3000 images with 23,040 X 23,040 pixels per image.

Approach:

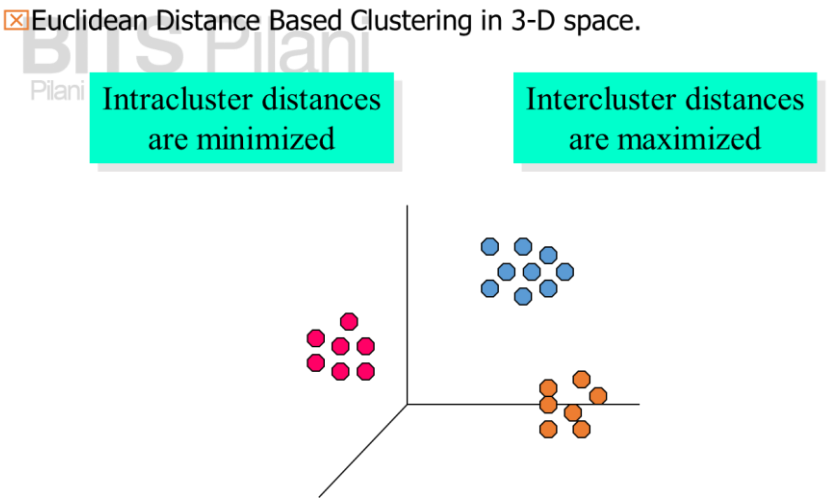
- Segment the image.
- Measure image attributes (features) – 40 of them per object.
- Model the class based on these features.
- Success story: 16 new high red-shift quasars were found. These are some of the farthest objects difficult to find.

Clustering

Clustering is concerned with separating out data points or records based on relationship among them. This separation or clustering helps businesses in various ways and finds many applications.

In a given set of data points with certain similarity measure among them, find clusters such that data points in one cluster are more similar to one another and data points in separate clusters are less similar to one another. This means to find out what is similar and what is dis-similar among the given data points. In general, similarity measures can be diverse. One simple measure that can be used is Euclidian distance, if the attribute values are continuous and have a mechanism to normalize the values. But there can be several other measures using which one can find the distance and identify whether some points are representing similar objects or distinct objects.

Illustration of Clustering:



The data points here are located in three-dimensional space. However, considering the proximity, which, in this case, is simple physical proximity, one can see that there are three clusters: one, two and three; coloured distinctly. Objective behind this clustering is that, within a cluster, all the objects are very close to each other or distances are minimal and between the clusters, the distances are higher. Clustering in this way gives a better picture of the differentiation among the data points or items that we are trying to understand.

Assume that these items are the records of some particular type of material and the objective is to cluster them as either good or bad quality. With some clear attributes and clear mechanism to differentiate, clustering technique will help to differentiate them distinctively.

Clustering Application 1: Market Segmentation

There are many corporations which are dealing with a large number of customers, but all customers are not suitable for all the products. Therefore, it is a good idea to segment the customers or the potential customers so that they can target the right product to the right segment of the customers. This helps to reduce the cost of the advertisement.

Consider an automobile manufacturer of who manufactures both high-end and lower-end cars. If they segment the customers or the potential customers, it is possible to know whom they should be targeting which product.

Goal: Subdivide a market into distinct subsets of customers where any subset may conceivably be selected as a market segment to be reached with a distinct marketing mix.

Approach:

Collect different attributes of customers based on their geographical and lifestyle related information.

- Find clusters of similar customers.
- Measure the clustering quality by observing buying patterns of customer in same cluster Vs those from different clusters.

Clustering Application 2: Document Clustering

There are billions of documents which are available in the public domain and the application is to find groups of documents that are similar to one another. Set a scope by identifying the documents which are related to some particular topic. This can be achieved by identifying the documents which are similar and which are dissimilar. List the words that are used within the document or the frequently occurring terms and the synonyms, etc. to cluster some set of documents related to the scope set initially.

If we are able to cluster the documents, say related to biology or civil engineering, etc, then the task of further analysing the documents is significantly simplified.

Goal: To find groups of documents which are similar to each other based on the important terms appearing in them.

Approach:

- To identify frequently occurring terms in each document. Form a similarity measure based on the frequencies of different terms. Use it to cluster.
- Gain: Information Retrieval can utilize the clusters to relate a new document or search term to cluster the documents.

Clustering Application 3: Clustering S&P Stock Data

Stock market has many industries like Information Technology, Chemical, Metals and so on. The behaviour of each stock is dependent on the industry segment to which it belongs. One can take past history data and observe the movements and accordingly create a clustering. This clustering may provide insight into the business and understanding the relation between the various stocks.

This application helps portfolios managers to create de-risked portfolios. They can create various clusters with a diverse behaviour, which help to choose the stocks from multiple clusters and create a more stable portfolio.

- Observe stock movement everyday
- Clustering points: Stock (Up/Down)
- Similarity Measure: Two points are more similar if the events described by them frequently happen together on the same day. Association rules are used to quantify a similarity measure.

	Discovered Clusters	***Industry Group***
1	Applied-Matl-DOWN,Bay-Network-Down,3-COM-DOWN, Cabletron-Sys-DOWN,CISCO-DOWN,HP-DOWN, DSC-Comm-DOWN,INTEL-DOWN,LSI-Logic-DOWN, Micron-Tech-DOWN,Texas-Inst-Down,Tellabs-Inc-Down, Natl-Semiconduct-DOWN,Oracl-DOWN,SGI-DOWN, Sun-DOWN	Technology1-DOWN
2	Apple-Comp-DOWN,Autodesk-DOWN,DEC-DOWN, ADV-Micro-Device-DOWN,Andrew-Corp-DOWN, Computer-Assoc-DOWN,Circuit-City-DOWN, Compaq-DOWN, EMC-Corp-DOWN, Gen-Inst-DOWN, Motorola-DOWN,Microsoft-DOWN,Scientific-Atl-DOWN	Technology2-DOWN
3	Fannie-Mae-DOWN,Fed-Home-Loan-DOWN, MBNA-Corp-DOWN,Morgan-Stanley-DOWN	Financial-DOWN
4	Baker-Hughes-UP,Dresser-Inds-UP,Halliburton-HLD-UP, Louisiana-Land-UP,Phillips-Petro-UP,Unocal-UP, Schlumberger-UP	Oil-UP

Association Rule Discovery: Definition

A common data mining task is Association Rule discovery which draws association among the items that are occurring together. This is also sometimes called as a market basket analysis because it is widely used to understand the purchasing habits of the people. For example, when someone buys bread, he/she is likely to buy butter as well.

Consider the following table which shows the transactions at a point of sale. Observing these sales, one may be able to identify sale of certain items occur together. For example, in this case, when a customer is buying milk, he/she is likely to buy coke as well. Knowing such association rules can be of interest to many people within an organization such as a large grocery chain where there may be billions of transactions. So if these rules are discovered, then it will become easier for them to organize their business and the inventory in a better way.

TID	Items
1	Bread, Coke, Milk
2	Beer, Bread
3	Beer, Coke, Diaper, Milk
4	Beer, Bread, Diaper, Milk
5	Coke, Diaper, Milk

Rules Discovered:
{Milk} --> {Coke}

Association Rule Discovery: Application 1: Marketing and Sales Promotion

Consider that a rule discovered is {Bagels,...} → {Potato Chips}

The rule discovered says that sale of Potato chips is consequent of sale of bagel, which is the antecedent. Discovering such rules may be challenging; however, it means that, a consequent is helping to determine what should be done to boost its sales. Therefore the product which is on the consequent side may be promoted by doing something with respect to the products in the antecedent.

Potato chips as consequent → can be used to determine what should be done to boost its sales.

In addition, it is possible to know what products will be affected if the antecedent is out of inventory.

Bagels in the antecedent → can be used to see which products will be affected if the store discontinues selling bagels.

Bagels in antecedent & Potato Chips in consequent → can be used to see what products should be sold with bagels to promote sale of potato chips

Note that association rules may extend even to groups of items which help businesses to stay organized.

Association Rule Discovery: Application 2: Inventory Management

Consider the case of a maintenance company concerned with the maintenance of the consumer goods. It is likely that a technician will be going around the town, repairing various complaints. We need to determine what components the technician should carry with him. The problem here is to identify the association among the types of defects that are received, what is the type of the repair work that needs to be performed. With that, it is easier to create the inventory of the components that the technician should be carrying while going to the customer sites for repair work. The goal and solution approach is given below.

Goal: A consumer appliance repair company wants to anticipate the nature of repairs on its consumer products and keep the service vehicle equipped with right parts to reduce on the number of visits to consumer households.

Approach: Process the data on tools and parts required in previous repairs at different customer sites and discover the co-occurrence patterns.

Sequential Patterns Discovery: Definition

There are many events that happen in patterns. If some event has occurred, it is quite likely that after a certain period of time, another event will occur. For example, if a student completes a basic course, it is likely that he/she will be planning to take up a related advanced course. Similarly, if a person buys a particular product, he/she is likely to buy certain accessories that work with the product. There may be several such sequences that could be observed around. Such sequences will help the businesses in organizing themselves better.

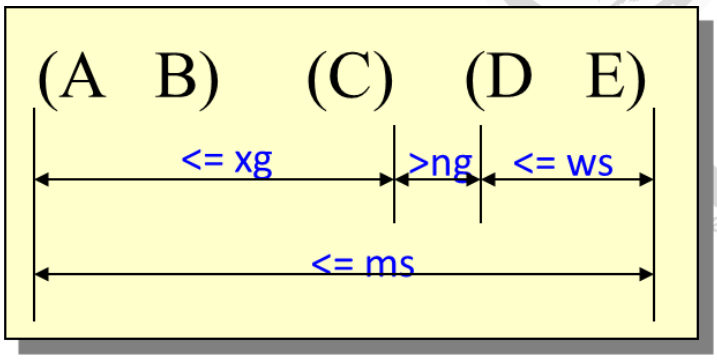
Given a set of objects (each object associated with its own timeline of events) find the rules that predict strong sequential dependencies among different events:

(A B) (C) ⟶ (D E)

It can be noted that the patterns include associated timelines or the timing constraints between the occurrences, such as the maximum gap, the minimum gap, window size of the events, span within which the events are likely to happen, etc. Therefore, once the sequential pattern is arrived at, next step is to identify each one of these constraints that will provide a better understanding of the events.

Event occurrences in the pattern are governed by timing constraints as shown below:

xg → maxgap, ng → mingap, ws→ window size, ms→ maxspan



Sequential Patterns Discovery: Examples

Consider a problem in telecommunications domain. There will be an inverter problem when an excessive line current is drawn. There is likelihood that there will be a fire soon enough within the zone. Therefore, if we know the sequence that whenever there is an inverter problem and there is an excessive line current, a fire alarm is likely to be created based on past experience. This helps to take some action early enough to prevent any damage.

Telecommunications Alarm Logs:

(Inverter_Problem Excessive_Line_Current)
(Rectifier_Alarm) → (Fire_Alarm)

Consider another example of point of sale transaction sequences given below. A person is likely to buy a particular product following some other product. This gives the business insight into the customer behaviour and they can be better prepared.

Point of Sale transaction sequences:

- **Computer Book Store:**
  (Intro_To_Visual_C) (C++_Primer) → (Perl_for_Dummies, Tcl_Tk)
- **Aesthetic Apparel Store:**
  (Shoes) (Racket, Racketball) → (Sports_Jacket)

Regression:

Regression tries to predict continuous valued variables based on the values of the some other points of the data. For example, based on the temperature variation over a period of time, it should be possible to predict the temperature at another point of time. It involves creating a mathematical model that represents the various points. This model can be either linear or nonlinear, into which the existing points fit. This model will help in predicting an element at another point of time or a point in another coordinate. Note that regression is studied in depth in statistics and neural network fields.

Regression: Examples

Predicting Sales of new product based on advertising expenditure:
In order to achieve this, past data needed are various levels of advertising expenditures and their impact on various levels of the sales. Based on these, one can predict new data points as required.

Predicting wind velocities:
Wind velocities are dependent on parameters such as temperature, humidity, air pressure etc. Hence it is possible to predict wind velocities as a function of temperature, humidity, air pressure etc. A regression model may be linear or curve in case it is two dimensional. In case of three dimensions the model will be a plane and when there are more than three dimensions, it would be hyper plane.

Prediction of stock market indices:
This involves predicting values for the future based on time series prediction.

Deviation/Anomaly Detection:

Deviation or anomaly detection involves looking at the significant deviations from the normal behaviour and thereby trying to identify undesirable events that might have occurred. For example, a credit card fraud based on a transaction which is significantly different from the normal transactions. Similarly, a sudden increase in network traffic in an unreasonable way and not as per the normal pattern may signal that the network is under attack or there is some form of intrusion. Therefore, if a model is created based on existing normal behaviour, then it will be easier to predict any anomalous behaviour.