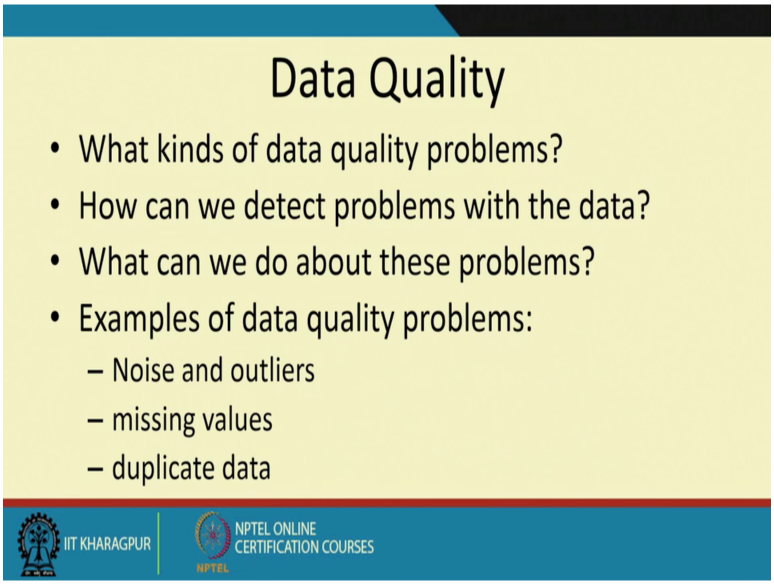
Data Mining

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

Lecture – 03
Data Preprocessing-II

Welcome to the third lecture on data mining. Today we will talk about some of the preprocessing techniques that are necessary before we actually perform the data mining.

(Refer Slide Time: 00:29)



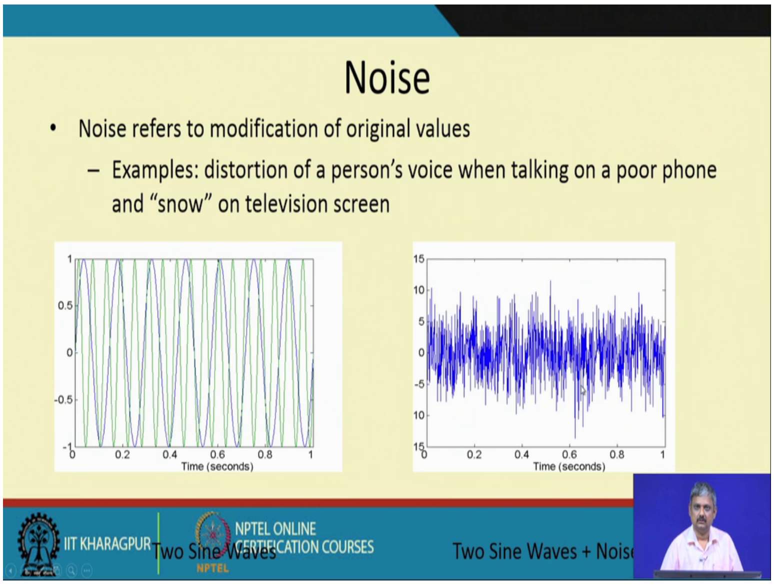
Data Quality

- What kinds of data quality problems?
- How can we detect problems with the data?
- What can we do about these problems?
- Examples of data quality problems:
  - Noise and outliers
  - missing values
  - duplicate data

IIT KHARAGPUR
NPTEL ONLINE CERTIFICATION COURSES
NPTEL

Before doing the preprocessing, we should look at what are the quality, and other issues that are to be considered before performing the preprocessing. So, there are 3 questions that will answer. What kind of quality problems can happen? How do we know that there are problems? And how to solve them? There can be several type of quality problems for example, you might have noise and outliers which is the most common problem. We can have missing values and we can have duplications in the data. A data item have I captured multiple number of times.

(Refer Slide Time: 01:21)



Noise

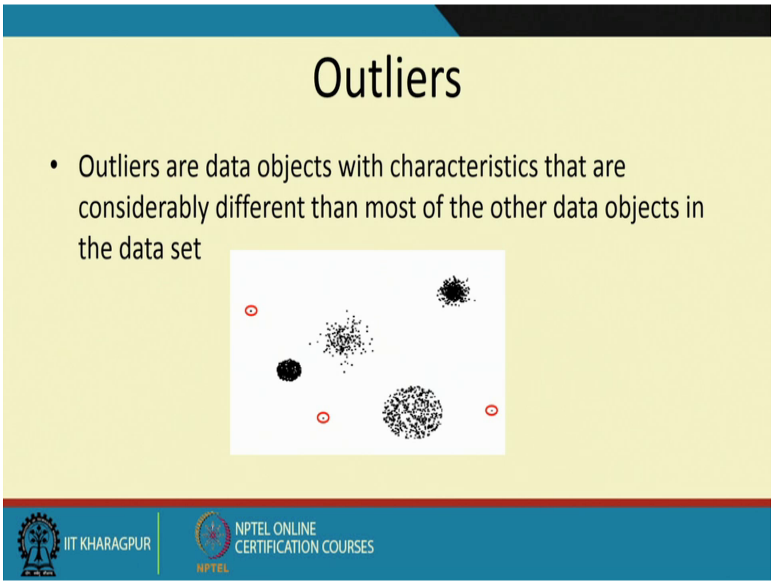
- Noise refers to modification of original values
  - Examples: distortion of a person’s voice when talking on a poor phone and “snow” on television screen

IIT KHARAGPUR NPTEL ONLINE CERTIFICATION COURSES

So, the most common example is a noise.

So, you can see that noise usually is something some kind of distortion on the actual data for example, this is a speech signal and you if we have a noise to it becomes there is a distortion may be due to the telephone or some snow, or something there may be distortion. So, this is the first problem that you will encounter in the data often the data will have some unnecessary component, which is noise we should remove before processing.

(Refer Slide Time: 02:06)



Outliers

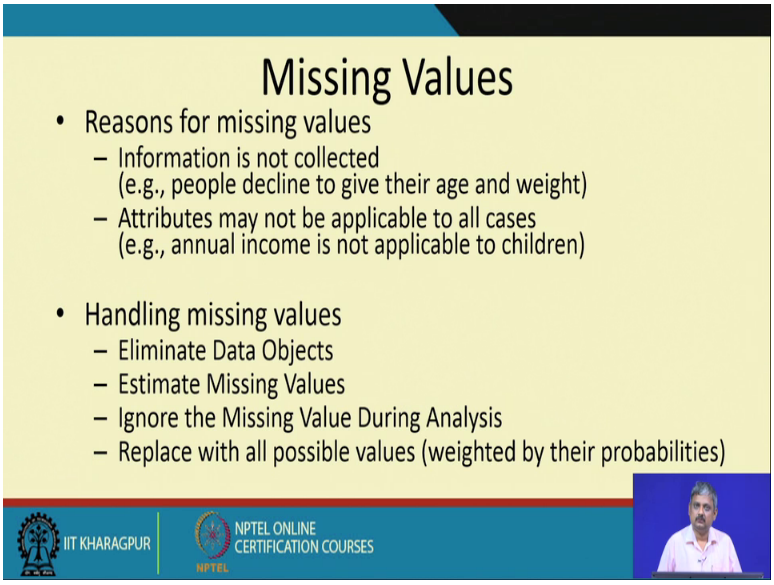
- Outliers are data objects with characteristics that are considerably different than most of the other data objects in the data set

IIT KHARAGPUR NPTEL ONLINE CERTIFICATION COURSES

One thing related to noise is that of outliers. So, these problems so, you can notice that for example, if your data visually looks like these clusters and this red ones are kind of the points which you can sort of perceive not belonging to the other.

So, a outlier is something which we know that definitely not come from the same mechanism that the other data has come maybe this outlier has happened due to some sensor error due to some error in notifying some data. Maybe somebody has wrongly in entered a data. For example, if people are noting down height of a person 17, 7-point, 6 point something maybe missed the decimal point, and it has become 611. Or which is way different from the other height values. So, again if we have outliers in the data this may mislead our mining algorithm.

(Refer Slide Time: 03:21)

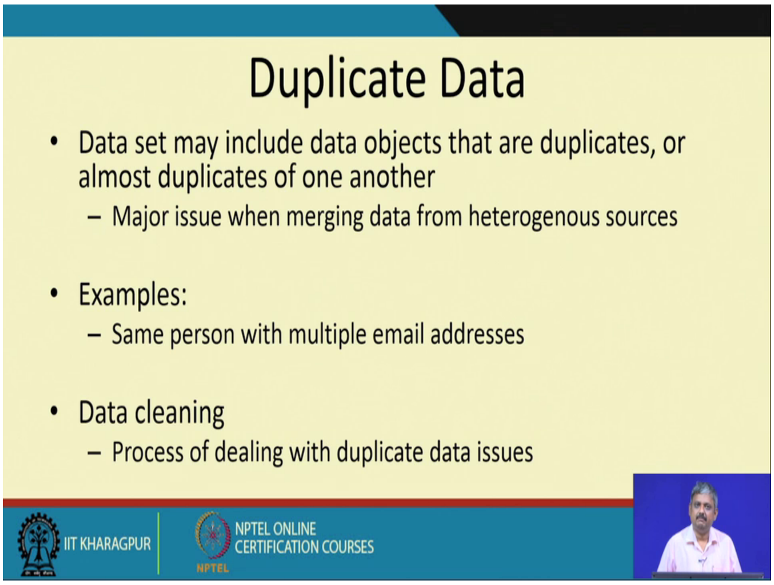


Missing Values

- Reasons for missing values
  - Information is not collected (e.g., people decline to give their age and weight)
  - Attributes may not be applicable to all cases (e.g., annual income is not applicable to children)
- Handling missing values
  - Eliminate Data Objects
  - Estimate Missing Values
  - Ignore the Missing Value During Analysis
  - Replace with all possible values (weighted by their probabilities)

So, we need to remove outliers too. Another common problem is that of missing values. Either people have some information is not collected, maybe it is not available or intentionally some values are not given. For example, people have declined to give their age, and weight. Similarly, annual income for a child below 18 years old is not applicable. So, we have several ways by which we can handle this missing values. We can either eliminate them the straightforward way is to ignore them. We can estimate them, and replace them with the estimated values, weighted by some appropriate probability factors.

(Refer Slide Time: 04:24)



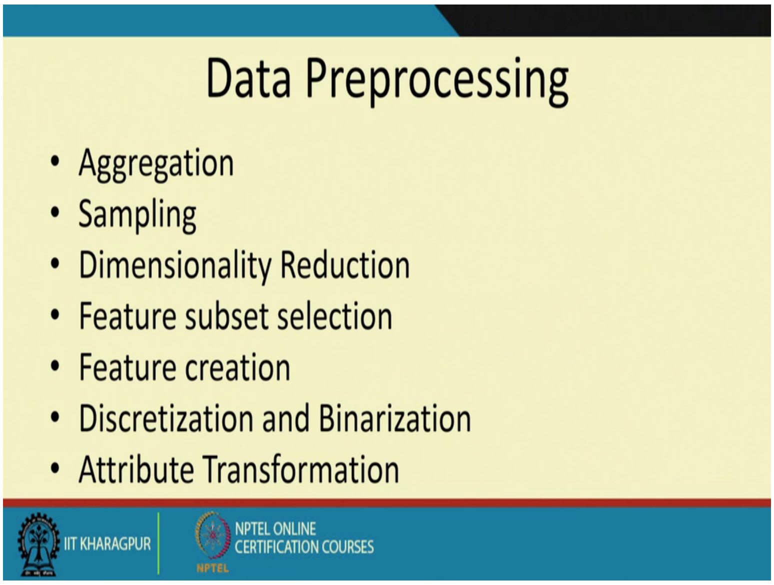
Duplicate Data

- Data set may include data objects that are duplicates, or almost duplicates of one another
  - Major issue when merging data from heterogenous sources
- Examples:
  - Same person with multiple email addresses
- Data cleaning
  - Process of dealing with duplicate data issues

It may also happen that there may be multiple data items representing the same thing. So, for example, this is common when we merge data from heterogeneous sources. So, the same persons record. For example, income tax data and pan card data. So, the same person information has come multiple number of times they are repeated.

So, in the process of data cleaning, we should be very careful to not to include such duplicates. So, that we are not mislead by the outcome. Note that the data mining algorithm gives result it extracts only what is there in the data. So, if your data itself has this kind of errors and duplicates of course, the output of mining algorithm will be erroneous. So, the preprocessing is indeed an very important step. So, this is the third problem that preprocessing algorithm needs to solve.

(Refer Slide Time: 05:38)



Data Preprocessing

- Aggregation
- Sampling
- Dimensionality Reduction
- Feature subset selection
- Feature creation
- Discretization and Binarization
- Attribute Transformation

IIT KHARAGPUR
NPTEL ONLINE CERTIFICATION COURSES
NPTEL

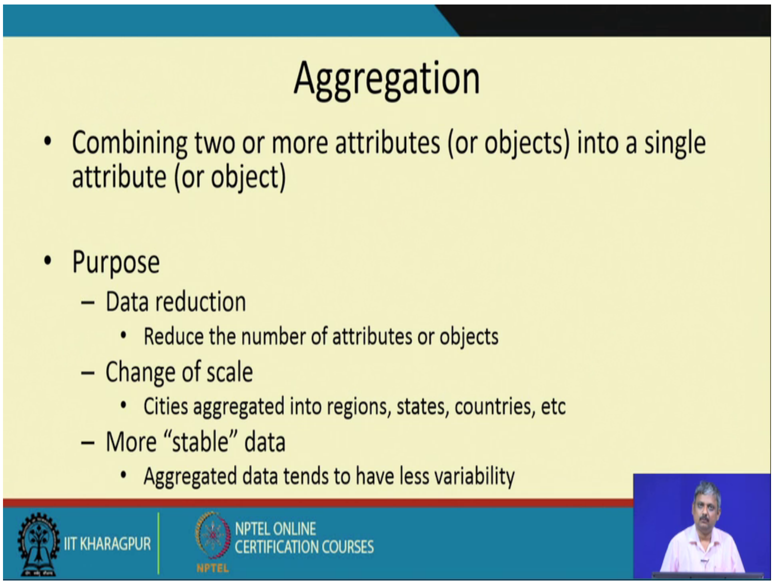
So, what are the different kind of preprocessing algorithm? By the way preprocessing is required not just for improving the data quality, but also sometimes for 2 other purpose. One purpose is after this preprocessing your data may be more valuable, it may be more informative.

It may it may give more information about the event after all this person. Second is it may reduce your computational load. It may throw a unnecessary things and have only the necessary information so that your computational load is reduced. So, here are the preprocessing steps that I usually followed, aggregation, sampling. So, aggregation means sometimes you consider a bunch of data together not a single data a bunch of data together. And the cumulative information of all these data is what is used. The second is sampling here only few representative data is kept rest is thrown away.

The idea is that only the sample is enough only the sample is enough for the processing. So, the next what we have is something called of which the third the 4th and the fifth false is dimensionality reduction. The most common example I would give for example if you go to a doctor with lots of symptoms and lots of measurement, doctor would not look at all of them. Doctor will select few of them. Say we will try to figure out which are the important ones and just pick up few of them. So, this process of picking up only the information that is important is the process of dimensionality reduction. You pick up only the attributes that are important. There are 2 type of dimensionality reduction either is a subset selection another is a feature extraction or creation. The other preprocessing operation is sometimes, you may have to discretize the value, and the direct continuous value may not be useful for example, aids reg these kinds of things you may have to discretize or binarize. Similarly, you may also have to do attribute transformation, you may have to transform the attributes.

For example, I have to scale it by some factor instead of in 10 to the power 20; I may have to scale to log scale or something. So, that is attribute transformation.

(Refer Slide Time: 09:19)



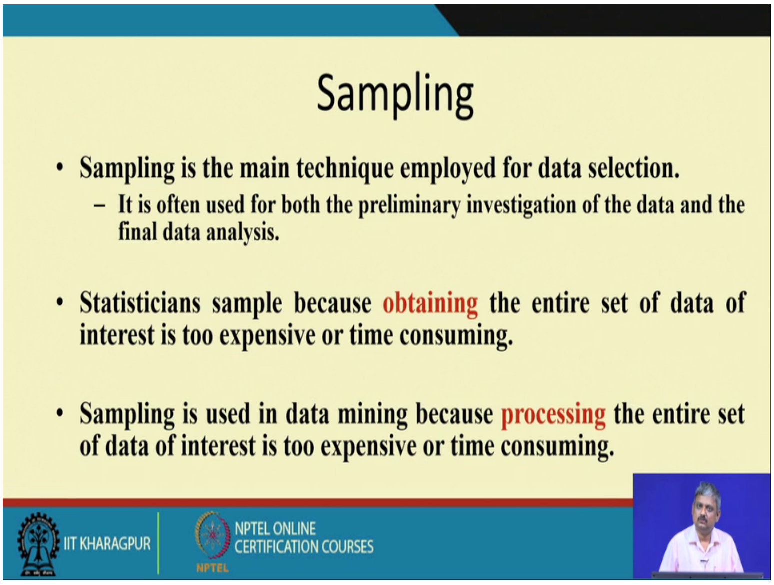
Aggregation

- Combining two or more attributes (or objects) into a single attribute (or object)
- Purpose
  - Data reduction
    - Reduce the number of attributes or objects
  - Change of scale
    - Cities aggregated into regions, states, countries, etc
  - More "stable" data
    - Aggregated data tends to have less variability

IIT KHARAGPUR
NPTEL ONLINE CERTIFICATION COURSES
NPTEL

I tell you some more details of this individual steps, that are important in preprocessing. Aggregation is combining attributes or objects into a single into a single entity. For example, maybe the sales of coca cola or some soft drinks in a entire district. Instead of storing the values in the individual towns in the district, I may add them together and have the entire district sales before I do the mining this reduces the data. This changes the scale, and you have a more kind of average or stable data with just less noise less variability less noise more stable more robust.

(Refer Slide Time: 10:17)



Sampling

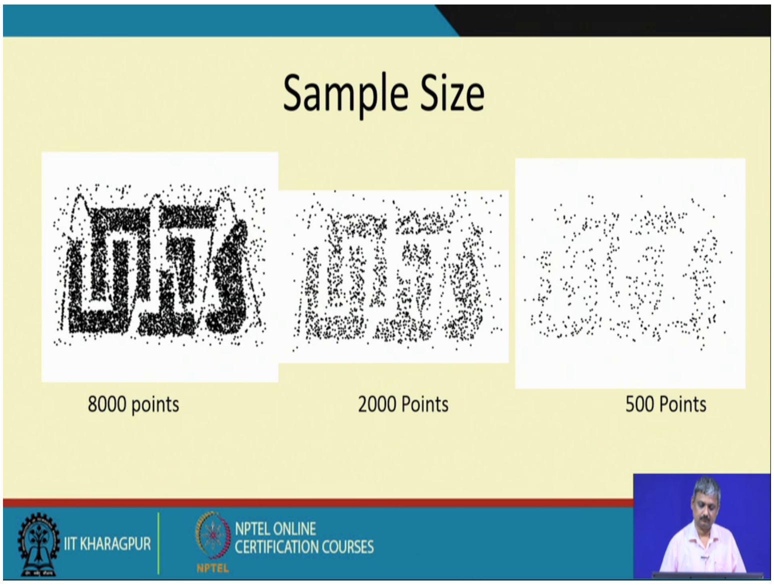
- Sampling is the main technique employed for data selection.
  - It is often used for both the preliminary investigation of the data and the final data analysis.
- Statisticians sample because obtaining the entire set of data of interest is too expensive or time consuming.
- Sampling is used in data mining because processing the entire set of data of interest is too expensive or time consuming.

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

A sort of alternate to this technique is do not add up, you just pick some through (Refer Time: 10:29) is do not add up just sample it.

So, if we take the sales of a soft drinks in a district instead of adding up over all the towns just take few representative towns and take their value, that you would call as a sampling. So, sampling as you know is a very well-known process in statistics. It is very commonly used in statistics, in any sample survey or anything.

(Refer Slide Time: 11:00)



Sample Size

8000 points

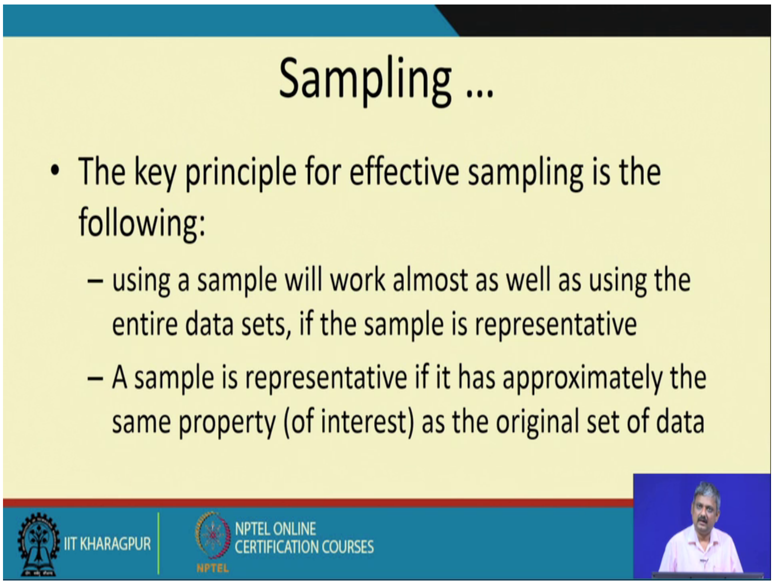
2000 Points

500 Points

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

One important thing to note down in this regard is what your sample size would be. You have to decide the appropriate sample size. So, that you get a fair representative of the data distribution, that should be the goal. So, again that is a challenging problem we will see as we study the algorithms how to decide on.

(Refer Slide Time: 11:28)

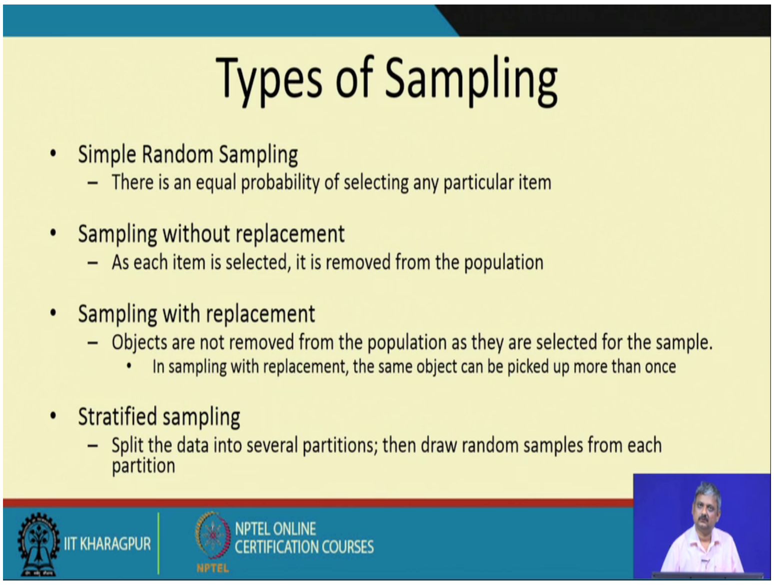


Sampling ...

- The key principle for effective sampling is the following:
  - using a sample will work almost as well as using the entire data sets, if the sample is representative
  - A sample is representative if it has approximately the same property (of interest) as the original set of data

This representative sample the idea is it should work as well as the entire data.

(Refer Slide Time: 11:38)

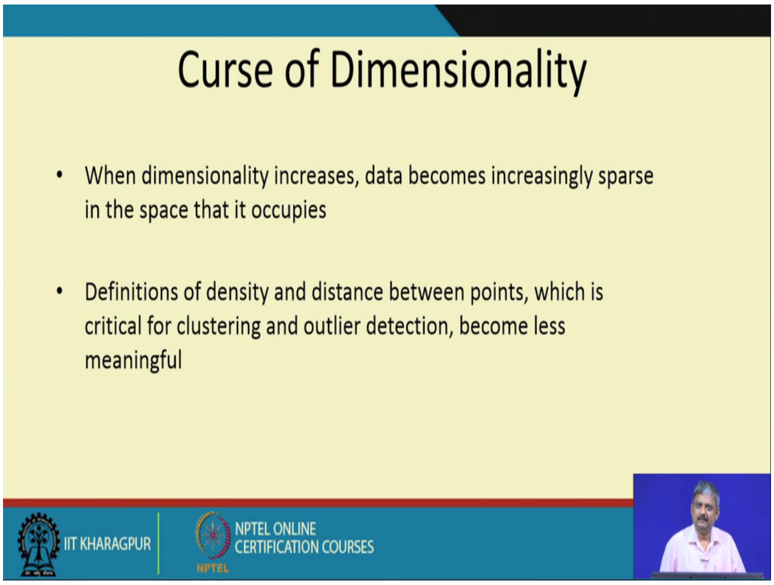


Types of Sampling

- Simple Random Sampling
  - There is an equal probability of selecting any particular item
- Sampling without replacement
  - As each item is selected, it is removed from the population
- Sampling with replacement
  - Objects are not removed from the population as they are selected for the sample.
    - In sampling with replacement, the same object can be picked up more than once
- Stratified sampling
  - Split the data into several partitions; then draw random samples from each partition

So, some of the common sampling types are simple random, without replacement stratified; that means, you take equal number of samples from each class each group, others one you can clearly understand that sampling without replacement than without each of these techniques have their own advantages and disadvantages. They so, there are many other type of sampling also bucket sampling, which is suitable for a large volume of data and many other techniques are there, but depending on the problem some sampling algorithm is suitable. Which again when we discuss the specific algorithms we will decide which is the best algorithm.

(Refer Slide Time: 12:36)



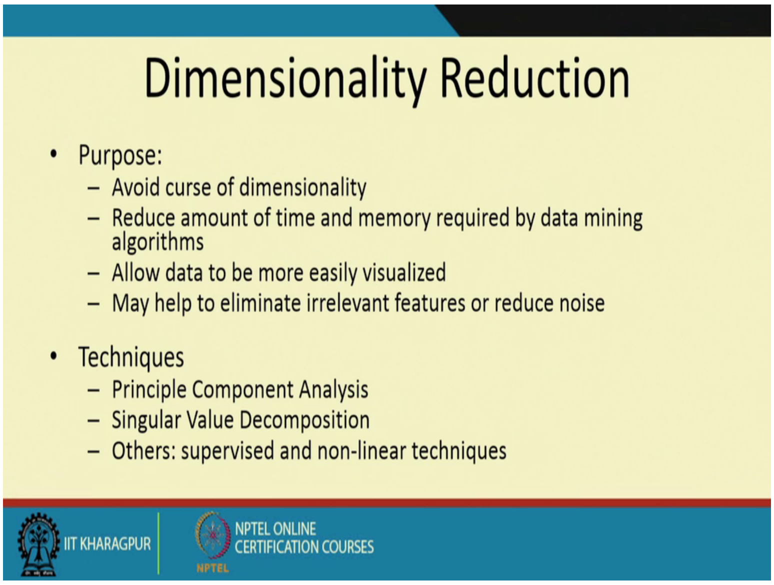
Curse of Dimensionality

- When dimensionality increases, data becomes increasingly sparse in the space that it occupies
- Definitions of density and distance between points, which is critical for clustering and outlier detection, become less meaningful

Next, we have the dimensionality reduction. The problem is higher the dimension of the data. The more number of attributes, the difficult it is to model both in terms of time and in the complexity of the models.

So, it is wise to reduce the dimensionality as much as possible.

(Refer Slide Time: 13:02)



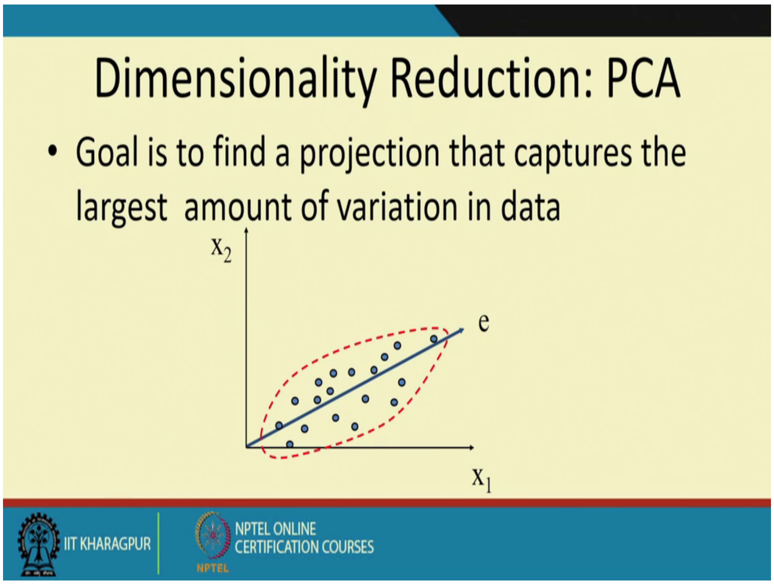
Dimensionality Reduction

- Purpose:
  - Avoid curse of dimensionality
  - Reduce amount of time and memory required by data mining algorithms
  - Allow data to be more easily visualized
  - May help to eliminate irrelevant features or reduce noise
- Techniques
  - Principle Component Analysis
  - Singular Value Decomposition
  - Others: supervised and non-linear techniques

IIT KHARAGPUR NPTEL ONLINE CERTIFICATION COURSES

So, reduce the memory and time requirement, easier visualization eliminating irrelevant attributes. There are 2 types of techniques supervised and unsupervised, unsupervised which tells you that the supervised somebody tells that this is good unsupervised nobody tells. And one of the common technique here is the famous principal component analysis, which is the most widely adopted technique.

(Refer Slide Time: 13:40)



Dimensionality Reduction: PCA

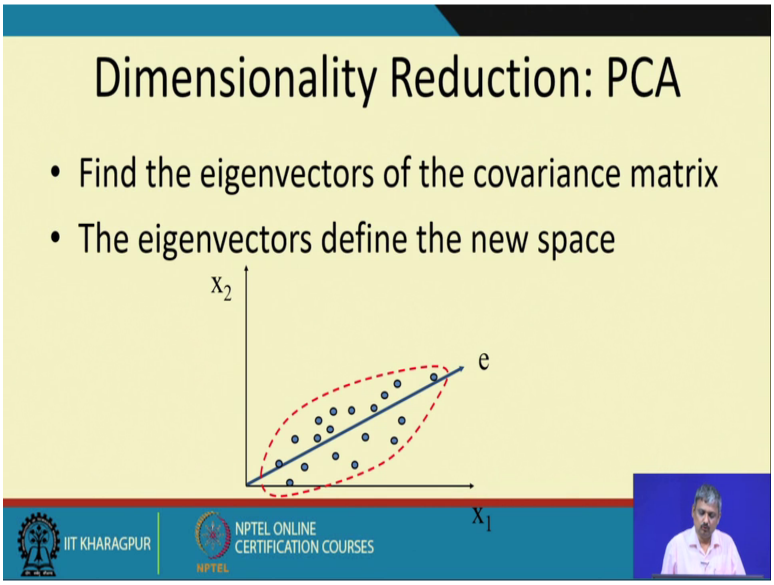
- Goal is to find a projection that captures the largest amount of variation in data

IIT KHARAGPUR NPTEL ONLINE CERTIFICATION COURSES

So, what is done in principal component analysis is can be very easily visualized using this data? So, each of these blue dots are say for example, data points with 2 measurements X 1 and X 2. Suppose I ask you that inst, I would not allow you 2 measurements, a 2 coordinates X 1 and X 2. Give me a single coordinate which may be tilted not same as X 1 and X 2 along which if I project my data, it would retain the maximum information of the data.

So, for you can see that if your data looks like this, and e is a direction, e is a direction e is a tilted direction somewhere in between X 1 and X 2, on this line e if I project my blue points, I sort of project them onto this single dimension instead of 2 X 1 and X 2; the data sort of retains the characteristics of the original data it returns the it looks as if the same the distribution the spread all these are more or less preserved. So, this direction e is called the principal component of this data, is called the principal component of this data.

(Refer Slide Time: 15:15)



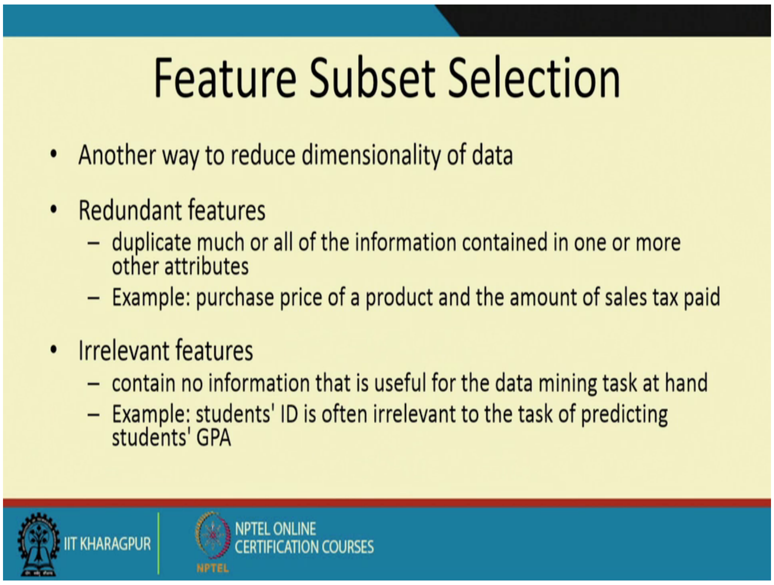
Dimensionality Reduction: PCA

- Find the eigenvectors of the covariance matrix
- The eigenvectors define the new space

How to find that? So, what we do is that find the covariance matrix of the data what is the covariance matrix we find out the variation of the data along each X 1 and X 2 and X 1 with respect to X 2 x 2 with respect to x 1?

So, how much the values of X 2 of different point deviates from the mean value similarly X 1 deviates from the mean value, similarly X 2 deviates some mean of X 1 and so on. So, you get a 2 by 2 matrix in this case it turns out that that principle component direction e is nothing but the direction of the principal eigenvector of this covariance matrix the direction, where the principal eigenvector points 2 principal eigenvector means the eigenvector corresponding to the largest eigen value.

(Refer Slide Time: 16:28)



Feature Subset Selection

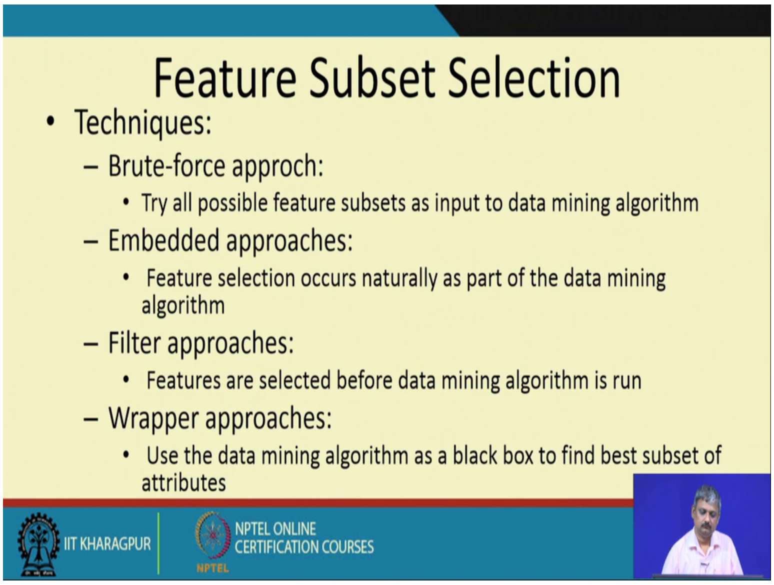
- Another way to reduce dimensionality of data
- Redundant features
  - duplicate much or all of the information contained in one or more other attributes
  - Example: purchase price of a product and the amount of sales tax paid
- Irrelevant features
  - contain no information that is useful for the data mining task at hand
  - Example: students' ID is often irrelevant to the task of predicting students' GPA

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

So, in this PCA or principal component analysis what we do is that we select all together a new direction e, but suppose I ask you in between X 1 and X 2 which one should I choose. So, that I get the best different I cannot just both I choose have to choose only 1.

So, this problem is known as the feature subset selection problem, it need not be one it may be say 2 out of 5. So, this problem is known as the feature selection problem, this is also an important preprocessing step. This is usually performed before actual processing is carried on.

(Refer Slide Time: 17:17)

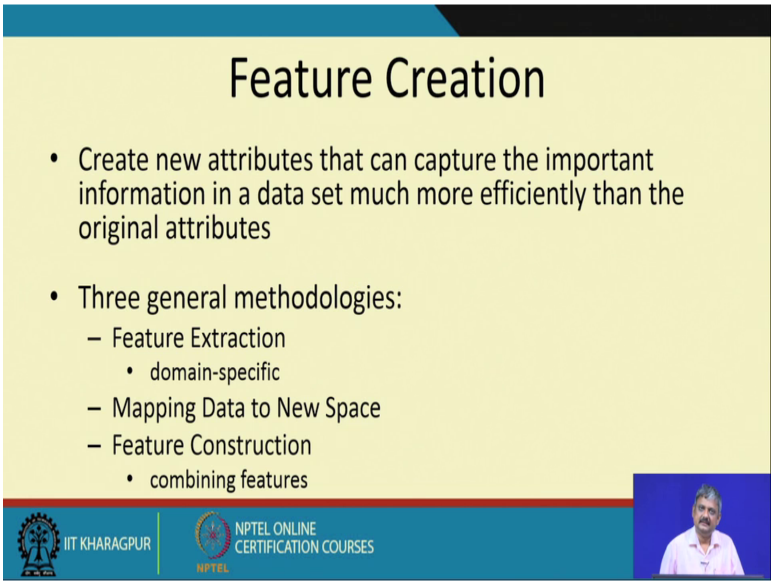


Feature Subset Selection

- Techniques:
  - Brute-force approch:
    - Try all possible feature subsets as input to data mining algorithm
  - Embedded approaches:
    - Feature selection occurs naturally as part of the data mining algorithm
  - Filter approaches:
    - Features are selected before data mining algorithm is run
  - Wrapper approaches:
    - Use the data mining algorithm as a black box to find best subset of attributes

These are the approaches so, what you basically do is that try out all combination of features, find out which one is best select that. How do I determine which is best? By seeing some criteria which are evaluation criteria. So, there are many ways like embedded approach, filter approach, wrapper approach, which would which would find out the best subset.

(Refer Slide Time: 17:56)



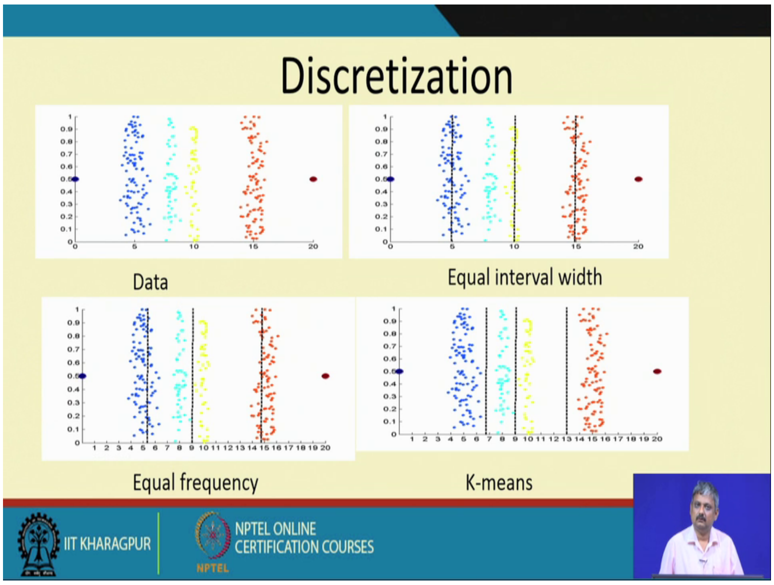
Feature Creation

- Create new attributes that can capture the important information in a data set much more efficiently than the original attributes
- Three general methodologies:
  - Feature Extraction
    - domain-specific
  - Mapping Data to New Space
  - Feature Construction
    - combining features

Similarly, there are methods in which you combine these attributes or features into a new space called a feature creation into a new domain specific place.

For example, in physics we do like say force into time is work. You combine some values to create more meaningful information.

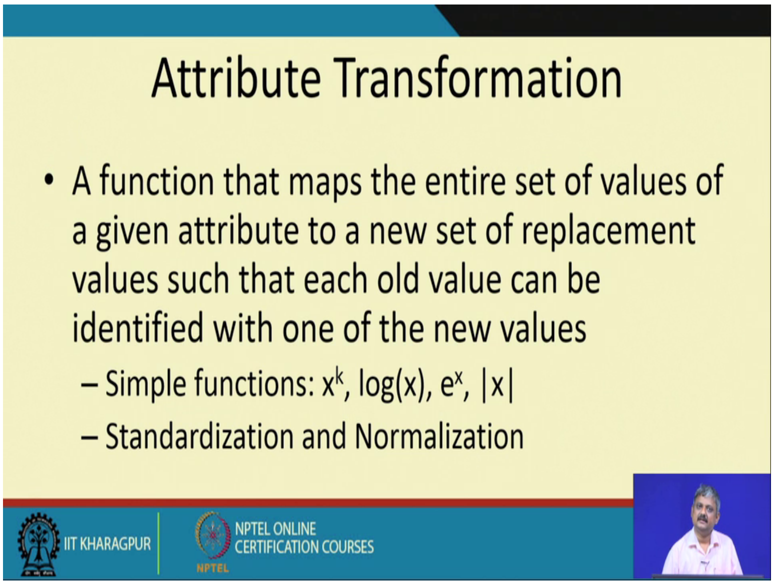
(Refer Slide Time: 18:30)



Discretization

So, the next preprocessing task is that of discretization. What it means? It means that suppose I have continuous data. Say, these different colored points. Instead of this continuous value can I represent them by some intervals or discrete bins. Say age of a person you might have seen in customer service they give 10 to 20 20 to 30 and so on. So, the question is what is this right partition. So, equal frequency or equal interval or some clustering which is k means we will come to it later. What is the right technique? So, there are different discretization techniques that one might follow.

(Refer Slide Time: 19:34)



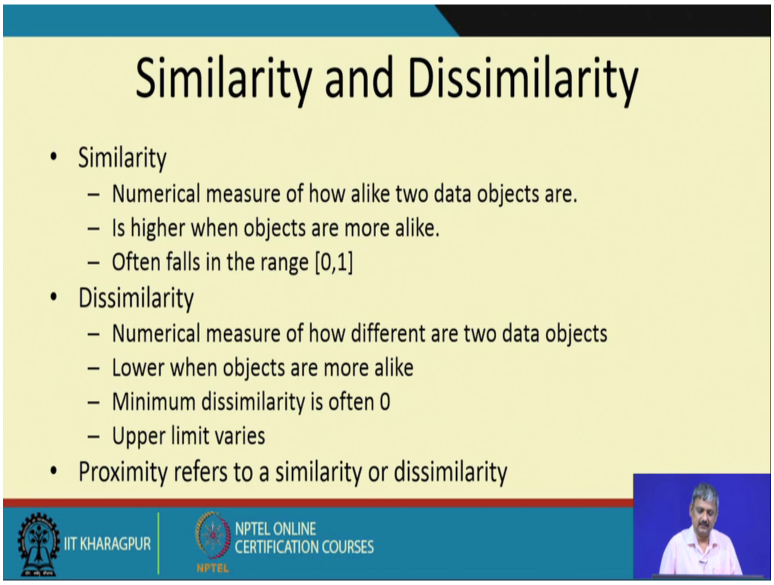
Attribute Transformation

- A function that maps the entire set of values of a given attribute to a new set of replacement values such that each old value can be identified with one of the new values
  - Simple functions: x^k, log(x), e^x, |x|
  - Standardization and Normalization

IIT KHARAGPUR
NPTEL ONLINE CERTIFICATION COURSES
NPTEL

So, these I also talked about either normalization, standardization applying some function to do the attribute transformation. Again, there are no general techniques of doing this. Many of these are domain dependent. You have to look at the data it is like a art. You as you gain experience in data mining you know what is the transformation, you should do what kind of normalization you should do. So, the purpose is if you do this you have more useful information.

(Refer Slide Time: 20:22)

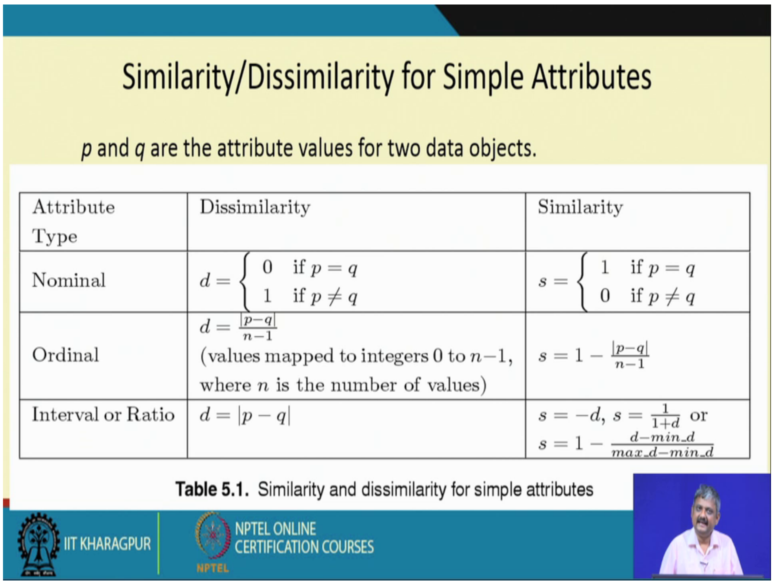


Similarity and Dissimilarity

- Similarity
  - Numerical measure of how alike two data objects are.
  - Is higher when objects are more alike.
  - Often falls in the range [0,1]
- Dissimilarity
  - Numerical measure of how different are two data objects
  - Lower when objects are more alike
  - Minimum dissimilarity is often 0
  - Upper limit varies
- Proximity refers to a similarity or dissimilarity

So, after these preprocessing, that is another quick thing before we need doing before doing any data. Mining algorithm is a notion of similarity or dissimilarity between data items. Some numerical measure how different or similar are 2 data items.

(Refer Slide Time: 20:45)



Similarity/Dissimilarity for Simple Attributes

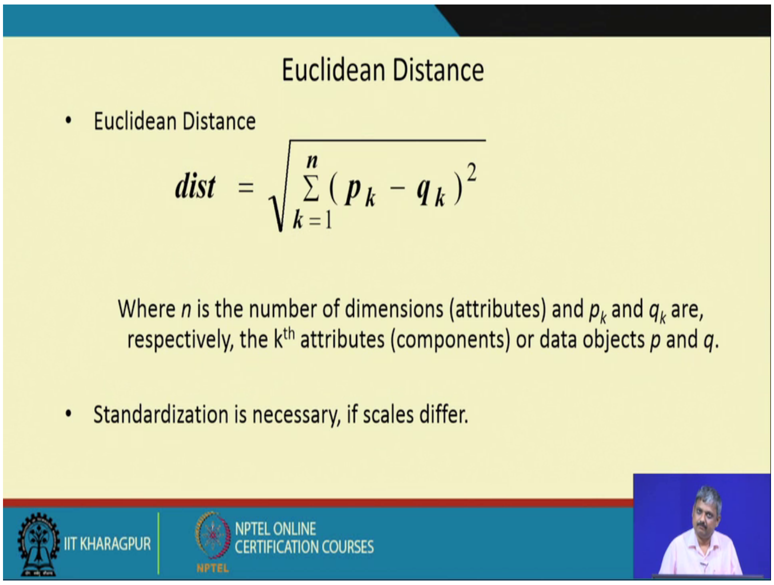
p and *q* are the attribute values for two data objects.

Attribute Type	Dissimilarity	Similarity
Nominal	$d = \begin{cases} 0 & \text{if } p = q \\ 1 & \text{if } p \neq q \end{cases}$	$s = \begin{cases} 1 & \text{if } p = q \\ 0 & \text{if } p \neq q \end{cases}$
Ordinal	$d = \frac{\lvert p-q \rvert}{n-1}$ (values mapped to integers 0 to $n-1$, where n is the number of values)	$s = 1 - \frac{\lvert p-q \rvert}{n-1}$
Interval or Ratio	$d = \lvert p - q \rvert$	$s = -d$, $s = \frac{1}{1+d}$ or $s = 1 - \frac{d - min\_d}{max\_d - min\_d}$

Table 5.1. Similarity and dissimilarity for simple attributes

So, for the case of continuous or interval or these kind of attributes continuous values, you can use the distance some Euclidean or other distance as the dissimilarity, but for other type of attributes that you studied, you should have alternate definitions. So, here are some common definition say nominal say mode a binary, ordinal some mapping value, interval, some difference. So, these are some functional forms which are commonly used you can have your own form too.

(Refer Slide Time: 21:33)



Euclidean Distance

- Euclidean Distance

$$dist = \sqrt{\sum_{k=1}^{n} (p_k - q_k)^2}$$

Where n is the number of dimensions (attributes) and p_k and q_k are, respectively, the k^th attributes (components) or data objects p and q.

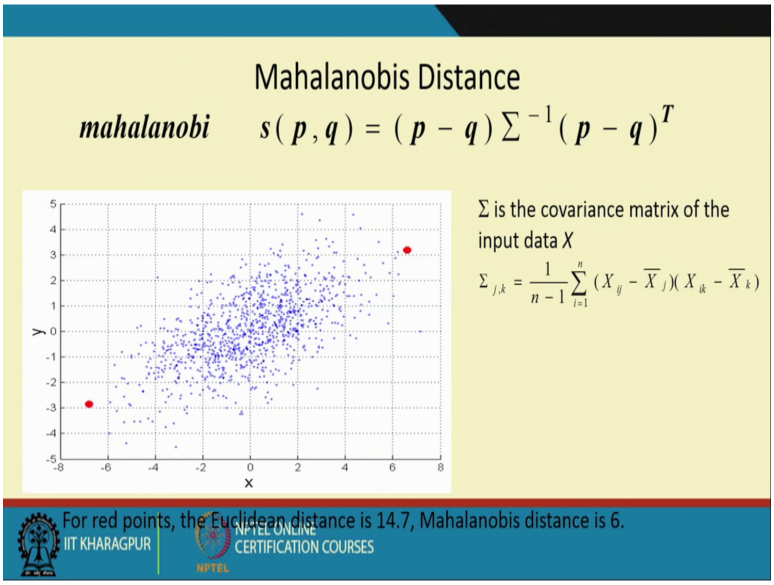
- Standardization is necessary, if scales differ.

IIT KHARAGPUR
NPTEL ONLINE CERTIFICATION COURSES
NPTEL

And if it is continuous we can define distance for similarity. The thing is that all these similarity values and all these things we will need them later when we develop our algorithm. So, in today’s lecture what I am doing I am giving a broad idea, not details of any method. I am giving a broad idea about the steps and measures you should give all these are preparation before you actually do your data mining.

So, here is the common distance function the Euclidean distance, it is difference between every dimension k p and q values squared them up add them up take the square root.

(Refer Slide Time: 22:18)



Mahalanobis Distance

$$mahalanobi \quad s(p,q) = (p-q)\Sigma^{-1}(p-q)^T$$

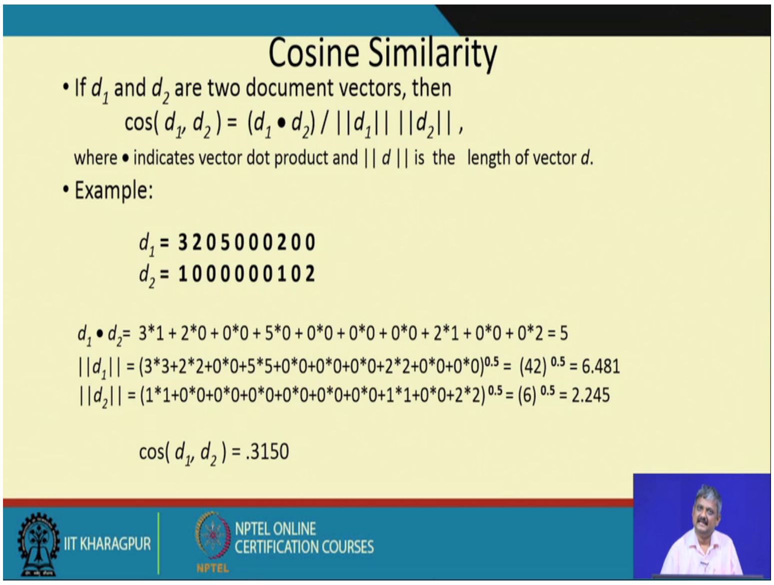
Σ is the covariance matrix of the input data X

$$\Sigma_{j,k} = \frac{1}{n-1}\sum_{i=1}^{n}(X_{ij} - \overline{X}_j)(X_{ik} - \overline{X}_k)$$

For red points, the Euclidean distance is 14.7, Mahalanobis distance is 6.

You can have other distances. So, this is same as the Euclidean, but with weight edge factor some attributes sigma, some attributes are given more weight as than the others the attributes, which are spread apart are given more weight edge cosine similarity.

(Refer Slide Time: 22:36)



Cosine Similarity

- If d_1 and d_2 are two document vectors, then

  $\cos(d_1, d_2) = (d_1 \bullet d_2) / ||d_1||\ ||d_2||$,

  where • indicates vector dot product and || *d* || is the length of vector *d*.

- Example:

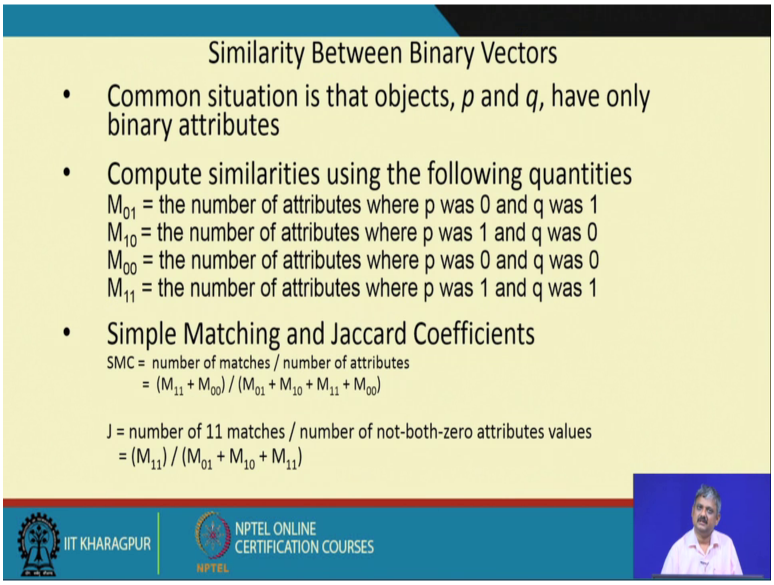
  $d_1 = 3\,2\,0\,5\,0\,0\,0\,2\,0\,0$
  $d_2 = 1\,0\,0\,0\,0\,0\,0\,1\,0\,2$

  $d_1 \bullet d_2 = 3*1 + 2*0 + 0*0 + 5*0 + 0*0 + 0*0 + 0*0 + 2*1 + 0*0 + 0*2 = 5$
  $||d_1|| = (3*3+2*2+0*0+5*5+0*0+0*0+0*0+2*2+0*0+0*0)^{0.5} = (42)^{0.5} = 6.481$
  $||d_2|| = (1*1+0*0+0*0+0*0+0*0+0*0+0*0+1*1+0*0+2*2)^{0.5} = (6)^{0.5} = 2.245$

  $\cos(d_1, d_2) = .3150$

As if these attributes are vectors cosine similarity is nothing but the cos of the angle between them the dot product. So, this you can use for anything a similar thing is something called as a Jaccard coefficient.

(Refer Slide Time: 23:00)



Similarity Between Binary Vectors

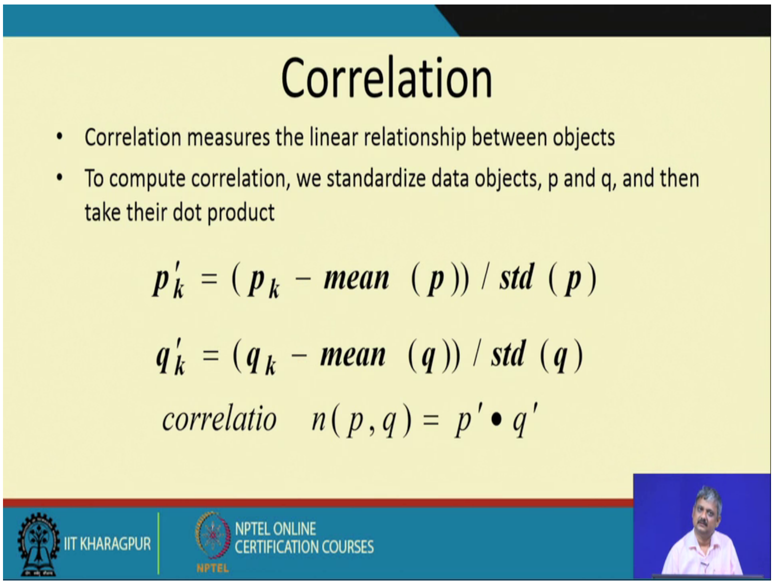
- Common situation is that objects, *p* and *q*, have only binary attributes
- Compute similarities using the following quantities
  M_{01} = the number of attributes where p was 0 and q was 1
  M_{10} = the number of attributes where p was 1 and q was 0
  M_{00} = the number of attributes where p was 0 and q was 0
  M_{11} = the number of attributes where p was 1 and q was 1
- Simple Matching and Jaccard Coefficients
  SMC = number of matches / number of attributes
  $= (M_{11} + M_{00}) / (M_{01} + M_{10} + M_{11} + M_{00})$

  J = number of 11 matches / number of not-both-zero attributes values
  $= (M_{11}) / (M_{01} + M_{10} + M_{11})$

If these values are binary, you can treat them as sets, and you can say number of matches divided by number of number of total number of attributes number of attributes.

Which are common this can be shown that this is equivalent to a cosine similarity if we consider binary vectors. Basically, number of attributes where they are one in both, with the vectors.

(Refer Slide Time: 23:31)



Correlation

- Correlation measures the linear relationship between objects
- To compute correlation, we standardize data objects, p and q, and then take their dot product

$$p'_k = (p_k - mean\ (p)) / std\ (p)$$

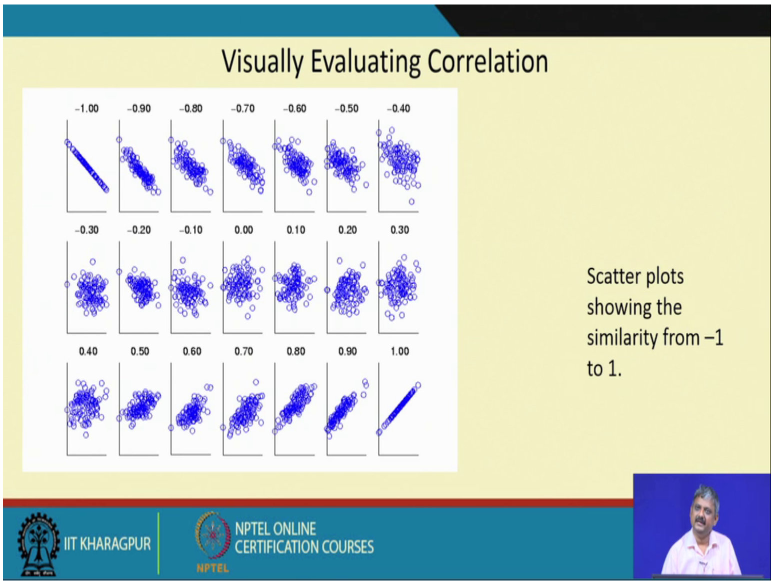
$$q'_k = (q_k - mean\ (q)) / std\ (q)$$

$$correlatio\quad n(p,q) = p' \bullet q'$$

IIT KHARAGPUR
NPTEL ONLINE CERTIFICATION COURSES
NPTEL

Correlation so, it is the mean subtracted divided by the standard deviation, and the product dot product of these 2 that would give you the correlation.

(Refer Slide Time: 23:52)



Visually Evaluating Correlation

Scatter plots showing the similarity from −1 to 1.

So, this is the meaning of correlation, if they are correlated; that means, they fall on a line and so on. Negative correlation means negative line. So, this is all these are the preprocessing steps that you would like to undergo before we actually do the data mining. So, as a summary what you should learn from this lecture is that, what are the quality measures?

What are the sources of poor quality in data? Like noise outlier and such things, what are the preprocessing steps that helps you improve data quality, reduce time produce better relevant model? These steps are sampling aggregation dimensionality reduction feature selection discretization noise removal. And then finally, you need some kind of similarity measure between the data, which tells you that which to how similar 2 data items are this is also a kind of preprocessing step, you need to a priori fix this before you do the algorithm. After all these steps are ready you are kind of ready for the mining algorithm is to run. There are different kinds of mining algorithm that I had already mentioned in the introduction like association rules classification clustering. So, what we will do in the next class.

We will assume that all these preprocessing is done, and study these algorithms. We will start with the association rule method in the next class. And find out how you can apply them what is the algorithm what is the output. So, this is for the preprocessing and distance measured this lecture. This closes the initial part of your course the introduction and the preprocessing. Next, we will move on to a algorithm.

Thank you.