

Course Name - Recommender Systems
Professor Name - Prof. Mamata Jenamani
Department Name - Industrial and Systems Engineering
Institute Name - Indian Institute of Technology Kharagpur
Week - 02
Lecture - 06

Lecture 06: Data preprocessing

Welcome to the second week of the course on the Recommender System. So, so far we have been talking about the data part of the recommender system literature. So, continuing in this we have seen in last week we have discussed about the data collection methodology, we have discussed about different measurement scales, we also saw how to get certain initial insights about the data considering it in the univariate form as well as you considering their bivariate relationships. So, now we are going ahead and trying to get little bit more understanding about the data before we actually go for statistical and machine learning foundations. So, this is the first step towards this foundation building. So, moving ahead here we will be talking only on different data preprocessing techniques.

So, when we talk about the data preprocessing it is a part of data preparation. So, before you adopt any machine learning algorithms you have to prepare the data so that it is suitable for that particular algorithm. So, data in real world has many problems sometimes this is incomplete. So, of course, the problem in recommender system that you are trying to solve is about how to impute this values of ratings, how to predict this rating.

So, basically those data is missing. So, this is one example of incomplete data with the problem that you are trying to solve in recommender system. Besides this the age the data can be noisy, noisy in the sense if somebody's salary is let us say 10 rupees or age is 222 which are quite impossible. Data can be inconsistent as well. For example, the age of a person is 42 years, but the birthday turns out to be 1997 which means it is fairly impossible.

Sometimes the data need to be formatted for a given software tool or algorithm. For example, if you are talking about let us say decision tree your data needs to be discrete. So, therefore, if in even if you have the data in terms of let us say interval or ratio scale which is continuous quantitative scales you have to now discretize it ok. So, ah so this because of this software tool and a method also we have to do some kind of modification for the data to prepare it for the algorithm. So, these are the major tasks under preprocessing.

Data discretization which is about reducing ah the data in the form of discrete values which is derived from the numerical form. The example that I continue with the example that I gave you if we have age of a person in the age of the customers in the store maybe you can categorize instead of considering the age as a continuous variable we discretize it and maybe we can make the groups such as young adult, young, middle aged, then you can

say middle aged, old, then old and so on. Then the task is data cleaning it is about filling in the missing values, removing taking care of the noisy data, removing outliers, resolving inconsistencies and so on. Next comes data integration. Then the data comes from multiple databases or has multiple dimensions we have to combine them and represent them in a form so that it becomes usable by the algorithm again.

Then comes the data normal transformation here we have to normalized and the data and maybe we have to aggregate the data from multiple sources. And finally, we have data reduction it is about reducing the data in a manner see data consists of a number of rows and a number of columns. Rows basically represent entities and columns represent the features. So, it may so happen that you want to reduce the number of features. So, that is column wise you will be reducing maybe you will be removing some of the unimportant features or you will be representing the data in a latent feature space.

So, that less with less number of features you will be characterizing maximum amount of information. Similarly, row wise also you may reduce and why this row wise reduction is necessary that this process is called sampling. So, out of all the elements that you have all the entities that you have only partially you consider them for your algorithm. Why this has to be done? This has to be done in the initial stage when you are trying to figure out which algorithm is the best. Let us say you are considering 4 algorithms for your work.

How will those algorithms work? You do not to save time probably you would not be using the entire data set. You will be using a representative part which is a subset of this data set that we also you have to reduce the data. So, data reduction though typically is concerned with reducing the number of features sampling can also be termed as data reduction in the sense selecting very less number of representative entities for testing the algorithms. Now comes to discretization of continuous variable. It is about dividing the range of a continuous attribute into intervals.

There are many methods to discretize this and the techniques such as Naive Bayes etcetera require such kind of discretization. Many times this is also useful for generating a summary of the data. Now it is done this discretization. Discretization can be done in many ways, but one of the ways specifically for the numeric data is making the bins. In the sense you have let us say many data points D_1, D_2 and let us say some D_m number of data points.

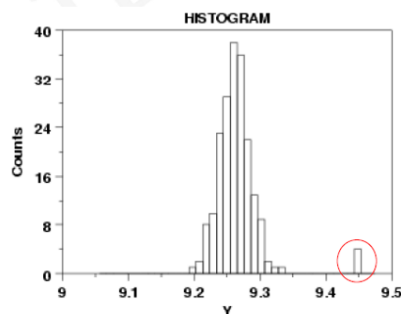
Now what you do you have to this is in continuous scale you have to bring them group them together so that they can put together in one bin. So, this bin can be of two types in equal width bin. In case of equal width bin let us say your total range is between 10 to 100 and you would like to make some 5 bins out of it. So, this entire range which is from 10 to 100 which consists of let us say 90 different possible values I mean the 90 or I mean the between 10 to 100 there can be many values of course, let us say you put them into 5 bins. So, if you put them into 5 bins of equal width and total let us say you can keep something I mean following certain procedure you can let us say starting from 10 to 20 you can put in one bin from 20 to 30 you can put another bin and I mean 20 plus and 30 you can put another bin and so on.

Similarly, 90 to 100 you can put another bin. So, how many bins you have made total 9 bins you have made and in case you have to make 5 bins of course, this range you are supposed to increase for individually. Next comes your equal height bins it may so happen that here there are very few people in this range let us say there are only 2 people here 10 people here and let us say this here again you have some 1 person in between there is another range where you have let us say 100 people. So, height of all these if we consider the frequency of items in each of these it is going to be different. So, it may so happen that you are you are interested to in each group you are interested to consider same number of people.

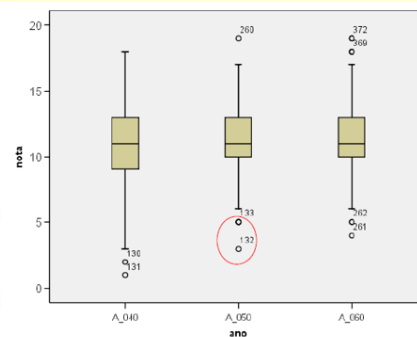
So, in that case probably you do not have to make this interval equal in terms of this range rather in terms of height. So, to make let us say the age groups when we make of equal heights. So, this is frequency and this is let us say your age groups ok. So, it will turn out that to in when you try making them equal height maybe here you have some 10 to 30, here you have 30 to 40, here you have 40 to 60 and so on. So, this range which was same here is now different.

So, you can make equal width binning or equal height binning ok. So, now this is little bit more about equal width binning. So, in case of equal width binning suppose you would like to make a bin number of bins to be n . So, this is the highest value, this is the lowest value and this is you divide. So, in the last example when the highest value was 100 and lowest value was 10 and we wanted to make as 5 bins this is the number of bins and this is the based on this you have to decide your interval.

Outlier Detection in Univariate Data



Compute mean and std. deviation. If the value is three standard deviations away from the mean, it may be considered as an outlier



An observation is an extreme outlier if outside $(Q1-3 \times IQR, Q3+3 \times IQR)$, and declared a mild outlier if it lies outside of the interval $(Q1-1.5 \times IQR, Q3+1.5 \times IQR)$ ($IQR = \text{Inter Quartile Range, } IQR = (Q3 - Q1)$)

So, in case you have to have equal height binning you have to divide the range into n intervals each containing approximately the same number of sample. Generally it is preferred to avoid clumping of the data in a particular bin as I told you. Then comes the next task of preprocessing data cleaning. Missing value of course, we are trying to solve the problem of predicting the ratings which are missing. So, in that case of course, we cannot

ignore ok, but suppose in case of let us say customer details some of the values are missing then what do you do most of the values are available.

So, you can ignore it you can mark another value as a new class for this let us say call it unknown. So, this just becomes another nominal variable. If it is a continuous one you can change it with attribute min and you can sometimes do replace it with certain most probable value depending on some inference that you make. Then you can another issue that comes during data cleaning is to identify the outliers and smooths the noisy data. And we saw that finding the outliers while finding the outliers in case of univariate method how to find the outliers using the box plot etcetera.

Similarly, you can use other methods as well. And once you correct this you have to look for inconsistent data that I saw one example earlier somebody's birthday was ah was 24 birthday was in in sorry birth date was in 1997 and age was 24. So, similarly redundancy you have that is caused by data integrity can also be the data integration can has to be avoided as well. So, this is what I told you, you avoid the outliers using this box plots and extreme outliers are thing which are beyond 3 times the away from the first quartile and 3 times away ah, 3 times IQR value away from the first quartile or beyond third quartile 3 times IQR value. But you also have certain mild outliers which can be in this range ok.

So, using this formula you can this figure is also not necessary using this formula you can remove the outliers. You can make certain observation of the outlier if you look the univariate distribution and find out something which is quite away ok that can be on some outlier. There are many statistical methods for detecting outlier in multivariate data. So, in even if we are considering univariate outlier sometimes this can be misleading. For example, if somebody's age is 10 or below he is likely to watch kid movies.

Now if only few such people are there you cannot call them as outliers this is a specific group and their viewing pattern is associated with their age. So, therefore, multiple variables together can give you this detail ok. So, one such multivariate outlier detection method is using something called Mahalanobis distance, Mahalanobis distance ok. So, higher this distance you say more probable is that the point is an outlier. Here in this formula this is your covariance matrix, this is the average values of the data and x_i is the data element.

- **Statistical Methods**

- Mahalanobis Distance
- Outliers: Multivariate data points with large distances

$$M_i = \left(\sum_{i=1}^n (\mathbf{x}_i - \bar{\mathbf{x}}_n)^T \mathbf{V}_n^{-1} (\mathbf{x}_i - \bar{\mathbf{x}}_n) \right)^{1/2} \quad \mathbf{V}_n = \frac{1}{n-1} \sum_{i=1}^n (\mathbf{x}_i - \bar{\mathbf{x}}_n) (\mathbf{x}_i - \bar{\mathbf{x}}_n)^T$$

When I say data element, data element again will have a number of attributes ok. So, taking all the attributes together you make this covariance matrix ok. So, when you make this covariance matrix n is the number of attributes and over all the attributes you make this. There are many data mining methods using certain distance based measure if you find an observation is defined as a distance based on outlier if it is at least a fraction β if at least β at least a fraction β of this observation in the data set are further than r from it. So, some such heuristics you can make this is basically heuristics based method.

There are many one class classifiers in which you try making one class consisting of the data whatever is outside that class you consider them as outlier. Say similarly you have clustering based methods in clustering typically a small side cluster you may consider as an outlier, but it may be a misleading situation you have to be actually going through further detail to explore why such a thing has happened, but such small clusters sometimes are considered outlier compared to the larger clusters. But the example that I was telling if there are very few movie viewers who are watching let us say kids movies then they cannot be considered as outliers they are supposed to be different class of viewers altogether their number may be small, but putting them together in the analysis along with adult viewers may lead to misleading results. So, therefore, you may consider them as outlier when you cluster the data, but that cluster needs special attention may be you are supposed to deal with that cluster with in a different manner and run a separate algorithm for that. Now next task is handling missing values.

So, when it comes to handling missing values this missing values can be more in one record see after all we are considering the data where we have the entities here and attributes here ok. So, let us say we have one entity in which most of the values are missing only two values are there then we may drop this entity completely. Similarly, there is one attribute column where nobody has given any data. So, that attribute you may think of dropping ok, but usually when you ignore the records or ignore the columns completely because very less number of values are present you may also be missing some information ok. So, therefore, while taking these decisions also you have to be little careful.

So, data description actually comes to rescue when such kind of situations arise. So, you can also fill on this missing values manually if necessary, but that is a tedious task. So, you have to have some kind of algorithms for this purpose ok. So, how do you handle missing values? You can use a global constant to fill in the missing value use the attribute mean to fill in the missing value. It use the attribute mean for all sample belonging to same class to fill in the missing value same class in the cells suppose you have clustered it you know that for average age of all the young when you categorize the data into young young adult and so on use the most probable value in the missing place.

So, when you use the most probable value in the missing place you can use certain information in inference based mechanisms like using some kind of Bayesian formula or decision tree identifying the relation as we go ahead we will be knowing more about this Bayesian methods decision tree etcetera at least if not in depth we will have overview of all

category of machine learning algorithm that are typically used. So, some such algorithms can also be used to handle the missing data problems. Then next is your data integration when it comes to data integration again once again data has columns and data has rows schema basically represents the columns. So, when you get data from multiple sources it may so happen there are common columns. So, you have to be really careful while merging you should not be replicating the same column again and again.

So, similarly there can be when you collect the data and try integrating there can be row replication of row itself. So, again you have that has to be taken care of ok. So, while detecting this resolving this data value conflicts it may happen because some of the real world entities and attribute values they derive from different sources may be different. For example, in some source the data is represented in metric unit and in some in British unit. So, in that sense same data can have different form then data transformation data transformation again is about smoothing the data and removing the noise.

- min-max normalization

$$v' = \frac{v - \min_A}{\max_A - \min_A} (\text{new_max}_A - \text{new_min}_A) + \text{new_min}_A$$

- z-score normalization (standardization)

$$v' = \frac{v - \text{mean}_A}{\text{stand_dev}_A}$$

- normalization by decimal scaling

$$v' = \frac{v}{10^j}$$

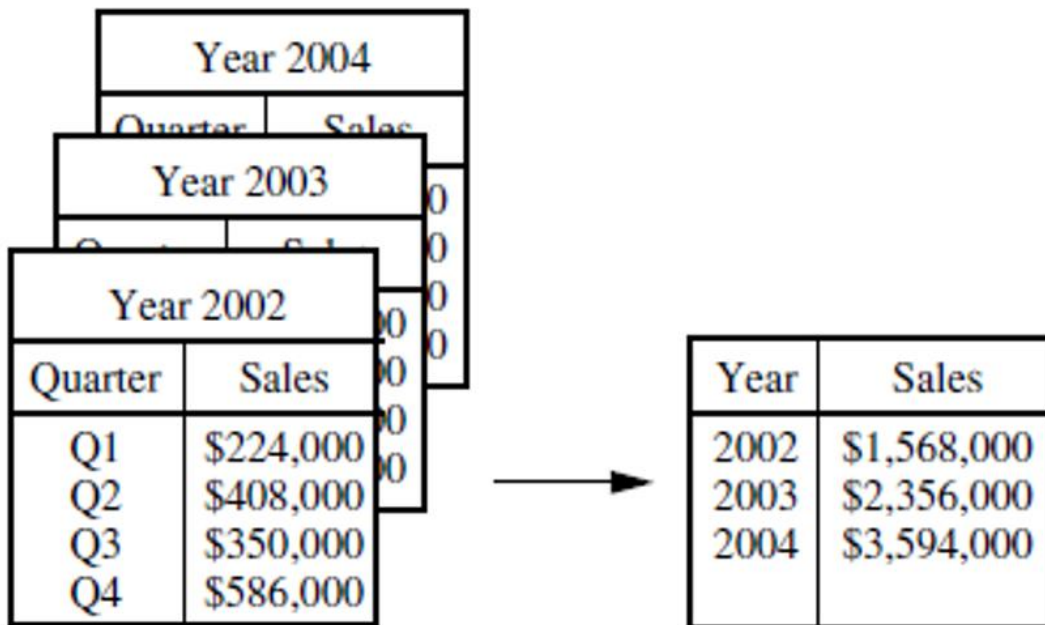
Where j is the smallest integer such that $\text{Max}(|v'|) < 1$

So, this can be done by binning regression method and clustering or you can aggregate the value which is about summarizing from the data cubes. You can make generalization using concept hierarchy you can have attribute or feature construction by combining more than one attribute or you can normalize. So, let us look at how do we normalize this data. So, there can be various ways in which you normalize and one why this normalization is necessary because many times if the data attribute in at one attribute has very small value let us say in terms of fractions let us say between 0 to 1 and other one has the value in terms of millions then the algorithm that to which you are feeding this data may end up making some kind of numerical errors which is undesirable. So, therefore, probably you would like to bring the data almost into the same range let us say from 0 to 1.

So, there are various ways you can do it you can do minimax normalization in which following this formula in which you will be putting the value you will be bringing the value to certain maximum and minimum defined by you. There can be certain z score

normalization in which you will be following the you will be you will assume that they it follows the normal distribution and from the data you will be subtracting the mean and divide by standard deviation. Then you can also normalize the data by decimal scaling. So, as I was telling let us say some value is in fractions let us say 0.0001 or something another value is in 10 to the power something.

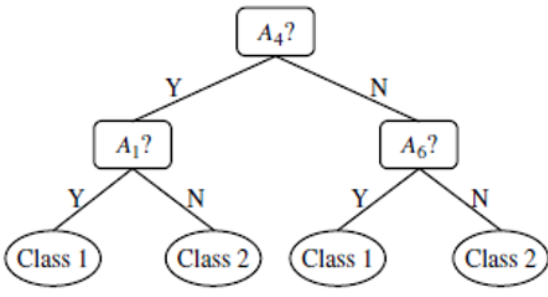
Data cube aggregation



Then if you divide the higher value with this then probably you will be bringing it to the same scale. This data reduction can happen by aggregating the values in the data cube by selecting a subset of the attributes by reducing the numericity and reducing the dimensionality. As we move ahead we are going to look at all this. So, this is one example of data cube aggregation. In case of data cube aggregation as you see here in this data cube you have quarter wise cells in each year quarter wise cells in each year first year, second year, third year now we have to combine them together.

So, instead of representing it in quarter level we will be representing it in terms of years. So, how do we come to years? We will be combining this data together and represent in terms of year combining this data together add this add this together and represent in terms of year. So, this is how we reduce the data by aggregating from the data cube. You can also reduce the data through certain attribute subset selection procedure. This is just one example of attribute subset selection procedure.

Attribute subset selection

Forward selection	Backward elimination	Decision tree induction
Initial attribute set: $\{A_1, A_2, A_3, A_4, A_5, A_6\}$ Initial reduced set: $\{\}$ $\Rightarrow \{A_1\}$ $\Rightarrow \{A_1, A_4\}$ \Rightarrow Reduced attribute set: $\{A_1, A_4, A_6\}$	Initial attribute set: $\{A_1, A_2, A_3, A_4, A_5, A_6\}$ $\Rightarrow \{A_1, A_3, A_4, A_5, A_6\}$ $\Rightarrow \{A_1, A_4, A_5, A_6\}$ \Rightarrow Reduced attribute set: $\{A_1, A_4, A_6\}$	Initial attribute set: $\{A_1, A_2, A_3, A_4, A_5, A_6\}$  \Rightarrow Reduced attribute set: $\{A_1, A_4, A_6\}$

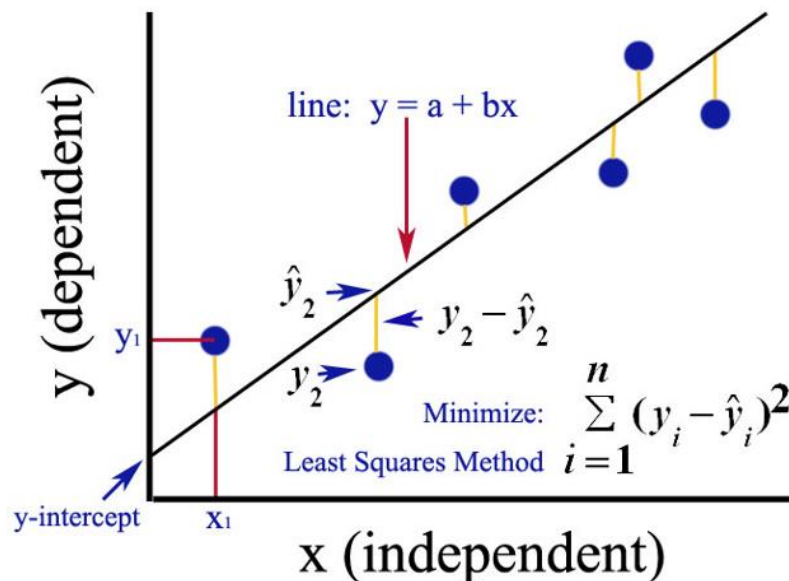
There can be many other ways forward selection in which you will be using one attribute one attribute at a time to run the model then add second attribute then add more number of attributes you keep on adding till you find there is no improvement in result by adding additional attributes. Similarly, you start your process from taking all the attributes together and you keep eliminating till you end up getting the minimum number of attributes which give you good result. You can use decision tree induction as well suppose the from your data you find out while making a decision tree you find out that in the nodes only A_1 A_2 A_3 A_4 and A_6 are existing rest A_3 A_5 and A_6 are becoming irrelevant when I we tried making this decision tree for classification process. So, therefore, we can keep these 3 attributes as important and drop remaining attributes. For numeracy reduction we can keep the attributes as important.

For reduction again there are various methods it can be either parametric or it can be nonparametric. By parametric we mean we mean to say we assume the data fit certain distribution and parameter of that distribution we use in some cells for reducing the numeracy. In case of nonparametric method no such distribution or additional adoption assumption no such distributional assumption is made. So, histogram clustering sampling etcetera can be used for this purpose. Now discretization and concept hierarchy generation when we talk about this we look at the raw data and try it aggregating it at a higher conceptual level.

So, this is one example of a parametric method specifically here regression is used. So, what can be done let us say these are your data points, but your regression line provide you certain average behavior. So, instead of representing this presenting this data at this point

corresponding point here we can represent for this data we can have this point for this data we can have this point. So, all this basically will be representing on the regression line that is how the variability will among the data points will actually decrease. In case of nonparametric method as I told you we can have let us say some binning approach in which let us say 1 to 10 this is 11 to 20 this is 21 to 30.

Parametric Methods -Regression

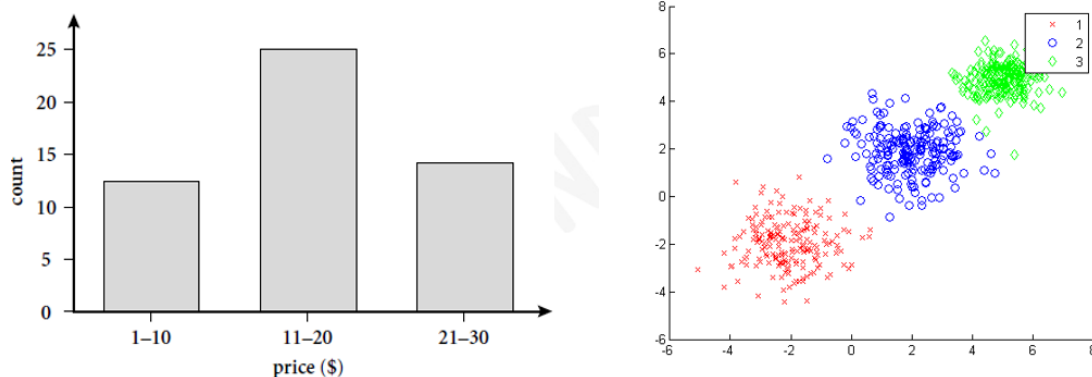


So, this one is 1 to 10 can be represented as one group let us say group 1 this will be group 2 this will be group 3 and so on. Similarly when we cluster we can have distinct clusters let us say this can we talk it at group 1 this is group 2 this is group 3. So, this is how so many data points will be represented by only 3 values. So, all these data points with here let us say frequency is around 13 or so. So, all these 13 elements will be represented by 1 here all the 25 elements will be represented by 2 and so on.

So, this is how also you reduce numerosity and this is again generalizing certain value to a higher conceptual level. For example, suppose in a non numeric nominal scale you have something represented laptops represented as IBM or something else desktops with Dell and other companies office software with Microsoft and other thing. So, therefore, these are a little more generalized in hierarchy concept hierarchy. So, in this concept hierarchy let us say mouse from various agencies can be put together and termed as mouse. Similarly, wrist pad mouse etcetera and all other computer accessories can put together if you want to further reduce the numerosity you can have this thing.

So, at this level you have 4 different categories 1, 2, 3, and 4 to represent 4 categories. At this level if you go you have 1, 2, 3, 4, 5, 6, 7, 8 categories. So, in the data when you replace laptop you replace it 1, desktop with 2, office with 3, antivirus with 4 and so on. This is the case of when you have a non numeric data which is basically nominal in nominal scale. In a continuous scale also you can reduce in the similar manner.

Non-parametric methods

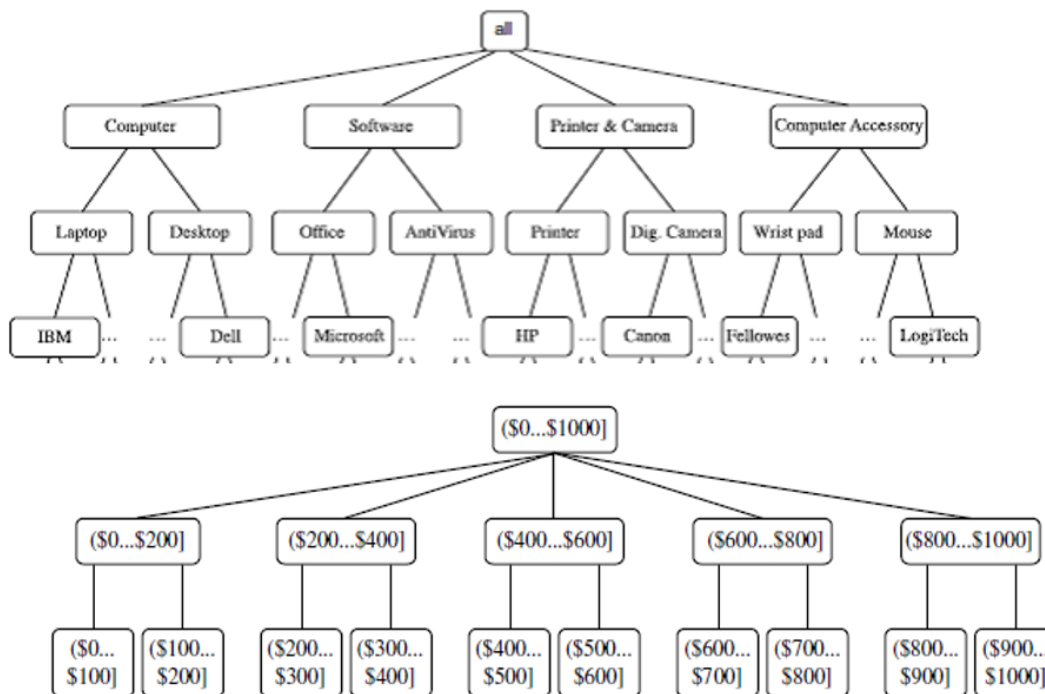


For example, here you have taken your data and discretized it and your range here is 1 to 100, 100 to 200 and so on. So, this 1 to 100 if you want to, here how many numbers you will be using for this is group 1, group 2, group 3, group 4. So, you will have total 8 groups, but if you want to reduce this numerosity further both these together you can call as group 1, this is group 2. So, you will have 4 groups together and if you want to further reduce probably only 1 group and probably you would not be doing because otherwise everywhere this will be 1 and this the variable will become not so useful.

Sampling can be another procedure. Now, here when we talk about the sampling we consider it in two sense for reducing the data size of the data specifically when we are at the stage where we are trying to figure out what is a good algorithm probably we will try to reduce the number of entities. So, when you reduce the number of entities we can sample. So, when you sample there are various procedures for sampling here this is one example of a random sampling with and without replacement. In case of look at this here 5 I mean that this transaction 5 transaction 1 8 and 6 each occur only once.

So, when you sample. So, which means once they are sampled they need not be considered again once they come to the sample set they need not be considered again and here T 4 appears twice. So, which means once it is sampled, but it was put back. So, it was with

replacement this was without replacement. So, similarly there can be something called stratified sampling in case of stratified let us say this is stratified according to age.

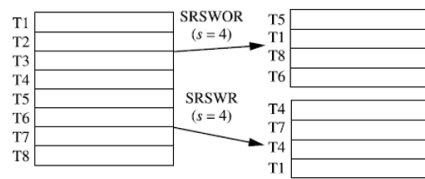


So, this is the first strata, second strata, third strata. So, out of that you will be choosing let us say from youth you have chosen 2 people from middle age you have chosen 4 from senior citizen you have chosen 1 and so on. You clustered this in certain manner and from each cluster represent each cluster why they are clustered because they belong to one cluster because they have certain common characteristics. So, from each cluster you choose few ok. Similarly, this process can also be used to reduce the numeracy numeracy in the sense let us say cluster 1 all the elements will have certain value, but let us say all the youth will have certain age, but ultimately they will be represented by youth only ok.

Now, comes to the dimensionality reduction. Now coming to this dimensionality reduction there are many ways in which you can do it. First one is your principal component analysis in which instead of using all the features you combine the features take a linear combination of the features and present only top few features which explains maximum variability in the data. And those features which you create by linear combination of other features we can call them as some kind of latent features which we discovered from the original data. Similarly you can have matrix decomposition approaches like singular value decomposition where again you can discover latent features and only few latent features you take I mean it depends on the rank of the matrix. So, you make a lower rank approximation of the matrix and then comes your signal processing techniques like that of let us say your input is let us say some song.

Sampling

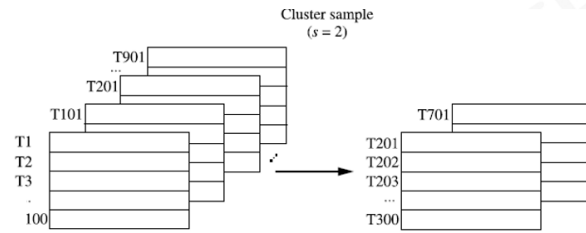
Simple random sample with/without replacement



Stratified sample
(according to age)

T38	youth	T38	youth
T256	youth	T391	youth
T307	youth	T117	middle_aged
T391	youth	T138	middle_aged
T96	middle_aged	T290	middle_aged
T117	middle_aged	T326	middle_aged
T138	middle_aged	T387	middle_aged
T263	middle_aged	T69	senior
T290	middle_aged	T284	senior
T308	middle_aged		
T326	middle_aged		
T387	middle_aged		
T69	senior		
T284	senior		

A **population** is the set of all items that possess a certain characteristic of interest.
 A **sample** is a subset of a population.



So, you can consider the Fourier transform discrete Fourier transform and take few top few components ok. So, with this we finish this one these are my references again and most of these that I covered here are from this Han and Kamber data mining book. And to summarize data processing is an important part for improving the stability of the algorithms data discretization cleaning integration transformation and reduction are the broad category of preprocessing. Discretization of the continuous variable can be done by dividing the range of the continuous variable into intervals using binning. The task on the data cleaning are missing value estimation, outlier removal etcetera.

Data integration deals with schema integration and detecting and resolving the data value conflicts. Data transformation involves the tasks such as smoothing, aggregation, generalization, attribute or feature construction and normalization. Data reduction involves the tasks such as data cube aggregation, attribute subset selection, dimensionality reduction and numericity reduction. Thank you.