Data Mining & Knowledge Discovery

Data Preprocessing

Md. Biplob Hosen

Lecturer, IIT-JU

Email: biplob.hosen@juniv.edu

Data Preprocessing

- Data preprocessing is a broad area and consists of a number of different strategies and techniques that are interrelated in complex ways.
- **Selecting data objects** and **attributes** for the analysis or for creating/changing the attributes.
- In both cases, the goal is to improve the data mining analysis with respect to time, cost, and quality.
  1. Aggregation
  2. Sampling
  3. Dimensionality Reduction
  4. Feature-subset Selection
  5. Feature Creation
  6. Discretization and Binarization
  7. Variable Transformation

Aggregation

- Process of collecting information from different sources and presenting it in a summarized format.
- Data aggregation plays a vital role in finance, product, operations, and marketing strategies in any business organization.
- Aggregated data is present in the data warehouse that can enable one to solve various issues, which helps solve queries from data sets.
- **Purpose:**
  - Data reduction - reduce the number of attributes or objects.
  - Change of scale - Cities aggregated into regions, states, countries, etc. Days aggregated into weeks, months, or years
  - More “stable” data - aggregated data tends to have less variability.

Aggregation - Procedure

- Data aggregation is needed if a dataset has useless information that can not be used for analysis.
- In data aggregation, the datasets are summarized into significant information, which helps attain desirable outcomes and increases the user experience.
- Data aggregation provides accurate measurements such as sum, average, and count.
- The collected, summarized data helps the business analysts to perform the demographic study of customers and their behavior.
- Aggregated data help in determining significant information about a specific group after they submit their reports.
- With the help of data aggregation, we can also calculate the count of non-numeric data.
- Generally, data aggregation is done for data sets, not for individual data.

Working of Data Aggregators

- Data aggregators refer to a system used in data mining to collect data from various sources, then process the data and extract them into useful information into a draft.
- **Collection of data:** The collection of data means gathering data from different sources. The data can be extracted using the internet of things (IoT), such as : Social media interaction, News headlines, Speech recognition like call centers, Browsing personal data and history of devices.
- **Processing of data:** Once data is collected, the data aggregator determines the atomic data and aggregates it. In the data processing technique, data aggregators use numerous algorithms form the AI or ML techniques, and it also utilizes statistical methodology to process it like the predictive analysis.
- **Presentation of data:** In this step, the gathered information will be summarized, providing a desirable statistical output with accurate data.

Aggregation - Example

- Organizations usually gather information about their online customers and website visitors.
- Here, the data aggregation involves statistics on customers' demographic and behavior matrices such as different age groups of customers and the total number of transactions.
- The marketing team does the data aggregation, which helps them personalize messaging, offers, and more in the user's digital experiences with the brand.
- It also helps the product management team of any organization to know which products generate more revenue and which are not.
- The aggregated data is also used by the financial and company executive, which helps them select how to allocate budget towards marketing or product development strategies.
- It helps to determine the average age of customers buying a specific product, which helps the business management team find the target age group for that specific product. In data aggregation usually prefer to calculate the average age of customers rather than individual customers.
- Calculating the value of voter turnout in a country or state. It is achieved by counting the total number of votes of a candidate in a specific region instead of counting the individual records of the voter.

Data Aggregation with Web Data Integration

- Web Data Integration(WDI) is a time-consuming nature in the data mining field where the data from different websites is aggregated into a single workflow.
- By using WDI, the time taken to aggregate data can be broken down to minutes which increases accuracy and thereby prevent human-made errors.
- By following the use cases provided by varied fields, the company can extract data from other sites to increase efficiency and accuracy.
- The inbuilt quality control in WDI helps in enhancing accuracy.
- It not only aggregates but cleans the data, also prepares it in useful forms for integration or analysis of data.
- If a company wants accuracy in dealing with data, WDI is the inevitable choice.

Sampling

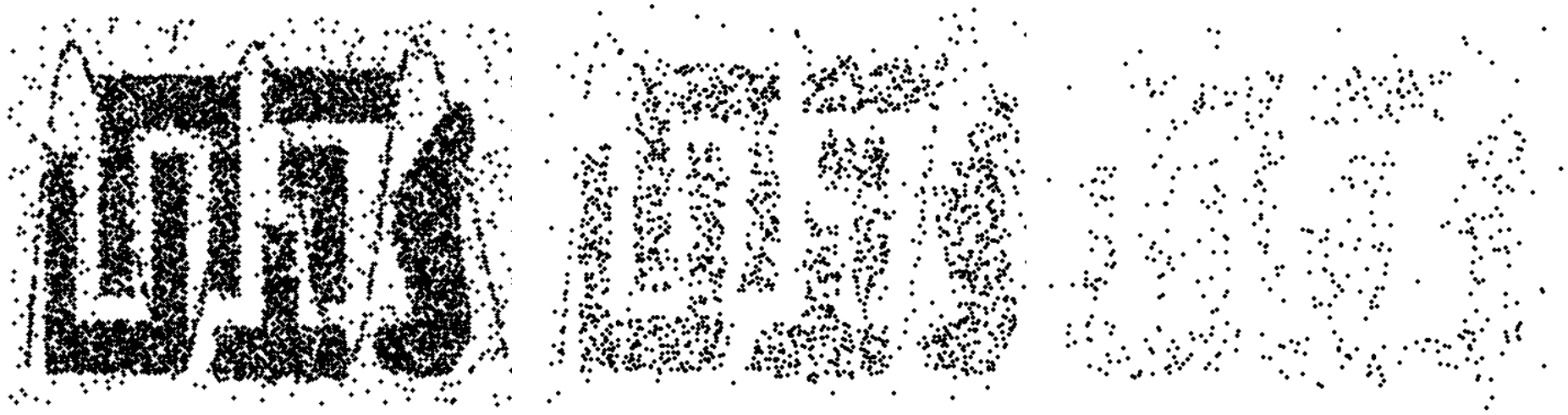
- Sampling is the main technique employed for data reduction.
- It is often used for both the preliminary investigation of the data and the final data analysis.
- Sampling is typically used in data mining because processing the entire set of data of interest is too expensive or time consuming.
- The key principle for effective sampling is the following:
  - Using a sample will work almost as well as using the entire data set, if the sample is representative.
  - A sample is representative if it has approximately the same properties (of interest) as the original set of data.

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.

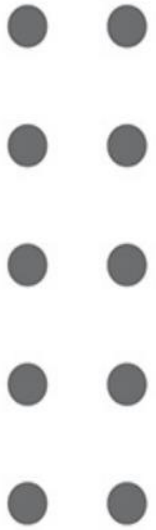
Sample Size

- Once a sampling technique has been selected, it is still necessary to choose the sample size.
- Larger sample sizes increase the probability that a sample will be representative, but they also eliminate much of the advantage of sampling.
- Conversely, with smaller sample sizes, patterns can be missed or erroneous patterns can be detected.

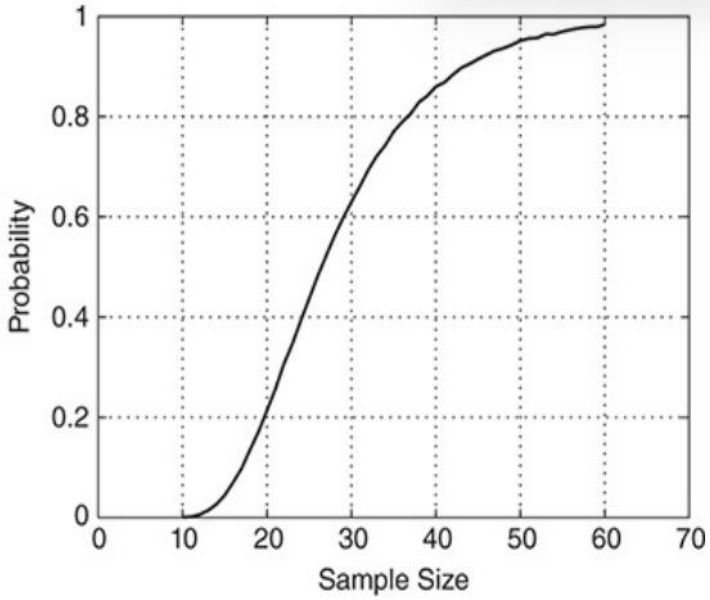


Sample Size

- To illustrate that determining the proper sample size requires a methodical approach, consider the following task:
- **Given a set of data consisting of a small number of almost equalsized groups, find at least one representative point for each of the groups.**
- Assume that the objects in each group are highly similar to each other, but not very similar to objects in different groups.



(a) Ten groups of points.



(b) Probability a sample contains points from each of 10 groups.

Figure 2.10.
Finding representative points from 10 groups.

- Figure 2.10(a) shows an idealized set of clusters (groups) from which these points might be drawn.
- Figure 2.10(b) shows the probability of getting one object from each of the 10 groups as the sample size runs from 10 to 60.
- Interestingly, with a sample size of 20, there is little chance (20%) of getting a sample that includes all 10 clusters. Even with a sample size of 30, there is still a moderate chance (almost 40%) of getting a sample that doesn't contain objects from all 10 clusters.

Progressive Sampling

- The proper sample size can be difficult to determine, so adaptive or progressive sampling schemes are sometimes used.
- These approaches start with a small sample, and then increase the sample size until a sample of sufficient size has been obtained.
- While this technique eliminates the need to determine the correct sample size initially, it requires that there be a way to evaluate the sample to judge if it is large enough.
- Suppose, for instance, that progressive sampling is used to learn a predictive model.
- Although the accuracy of predictive models increases as the sample size increases, at some point the increase in accuracy levels off.
- We want to stop increasing the sample size at this leveling-off point.
- By keeping track of the change in accuracy of the model as we take progressively larger samples, and by taking other samples close to the size of the current one, we can get an estimate of how close we are to this leveling-off point, and thus, stop sampling.

Curse of Dimensionality

- Many types of data analysis become significantly harder as the dimensionality of the data increases.
- Specifically, as dimensionality increases, the data becomes increasingly sparse in the space that it occupies.
- Thus, the data objects we observe are quite possibly not a representative sample of all possible objects.
- For classification, this can mean that there are not enough data objects to allow the creation of a model that reliably assigns a class to all possible objects.
- For clustering, the differences in density and in the distances between points, which are critical for clustering, become less meaningful
- As a result, many clustering and classification algorithms (and other data analysis algorithms) have trouble with high-dimensional data leading to reduced classification accuracy and poor quality clusters.

Dimensionality Reduction

- Consider a set of documents, where each document is represented by a vector whose components are the frequencies with which each word occurs in the document. In such cases, there are typically thousands or tens of thousands of attributes (components), one for each word in the vocabulary.
- As another example, consider a set of time series consisting of the daily closing price of various stocks over a period of 30 years. In this case, the attributes, which are the prices on specific days, again number in the thousands.
- **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:**
  - Principal Components Analysis (PCA).
  - Singular Value Decomposition.
  - Others: supervised and non-linear techniques.

Principal Component Analysis

- Principal Component Analysis is an unsupervised learning algorithm that is used for the dimensionality reduction in machine learning.
- It is a statistical process that converts the observations of correlated features into a set of linearly uncorrelated features with the help of orthogonal transformation.
- These new transformed features are called the Principal Components.
- It is a technique to draw strong patterns from the given dataset by reducing the variances.
- PCA works by considering the variance of each attribute because the high attribute shows the good split between the classes, and hence it reduces the dimensionality.
- Some real-world applications of PCA are image processing, movie recommendation system, optimizing the power allocation in various communication channels.
- It is a feature extraction technique, so it contains the important variables and drops the least important variable.

Steps for PCA algorithm

1. **Getting the dataset:**
- Firstly, we need to take the input dataset and divide it into two subparts X and Y, where X is the training set, and Y is the validation set.
2. **Representing data into a structure:**
- Now we will represent our dataset into a structure. Such as we will represent the two-dimensional matrix of independent variable X. Here each row corresponds to the data items, and the column corresponds to the Features. The number of columns is the dimensions of the dataset.
3. **Standardizing the data:**
- In this step, we will standardize our dataset. Such as in a particular column, the features with high variance are more important compared to the features with lower variance.
- If the importance of features is independent of the variance of the feature, then we will divide each data item in a column with the standard deviation of the column. Here we will name the matrix as Z.
4. **Calculating the Covariance of Z:**
- To calculate the covariance of Z, we will take the matrix Z, and will transpose it. After transpose, we will multiply it by Z. The output matrix will be the Covariance matrix of Z.

Steps for PCA algorithm

5. **Calculating the Eigen Values and Eigen Vectors:**

- Now we need to calculate the eigenvalues and eigenvectors for the resultant covariance matrix Z. Eigenvectors or the covariance matrix are the directions of the axes with high information. And the coefficients of these eigenvectors are defined as the eigenvalues.

6. **Sorting the Eigen Vectors:**

- In this step, we will take all the eigenvalues and will sort them in decreasing order, which means from largest to smallest. And simultaneously sort the eigenvectors accordingly in matrix P of eigenvalues. The resultant matrix will be named as P*.

7. **Calculating the new features Or Principal Components:**

- Here we will calculate the new features. To do this, we will multiply the P* matrix to the Z. In the resultant matrix Z*, each observation is the linear combination of original features. Each column of the Z* matrix is independent of each other.
- Remove less or unimportant features from the new dataset.
- The new feature set has occurred, so we will decide here what to keep and what to remove. It means, we will only keep the relevant or important features in the new dataset, and unimportant features will be removed out.

Feature Subset Selection

- Another way to reduce the dimensionality is to use only a subset of the features.
- While it might seem that such an approach would lose information, this is not the case if **redundant** and **irrelevant** features are present.
- **Redundant features** duplicate much or all of the information contained in one or more other attributes.
- For example, the purchase price of a product and the amount of sales tax paid contain much of the same information.
- **Irrelevant features** contain almost no useful information for the data mining task at hand.
- For instance, students' ID numbers are irrelevant to the task of predicting students' grade point averages.
- Redundant and irrelevant features can reduce classification accuracy and the quality of the clusters that are found.

Feature Subset Selection

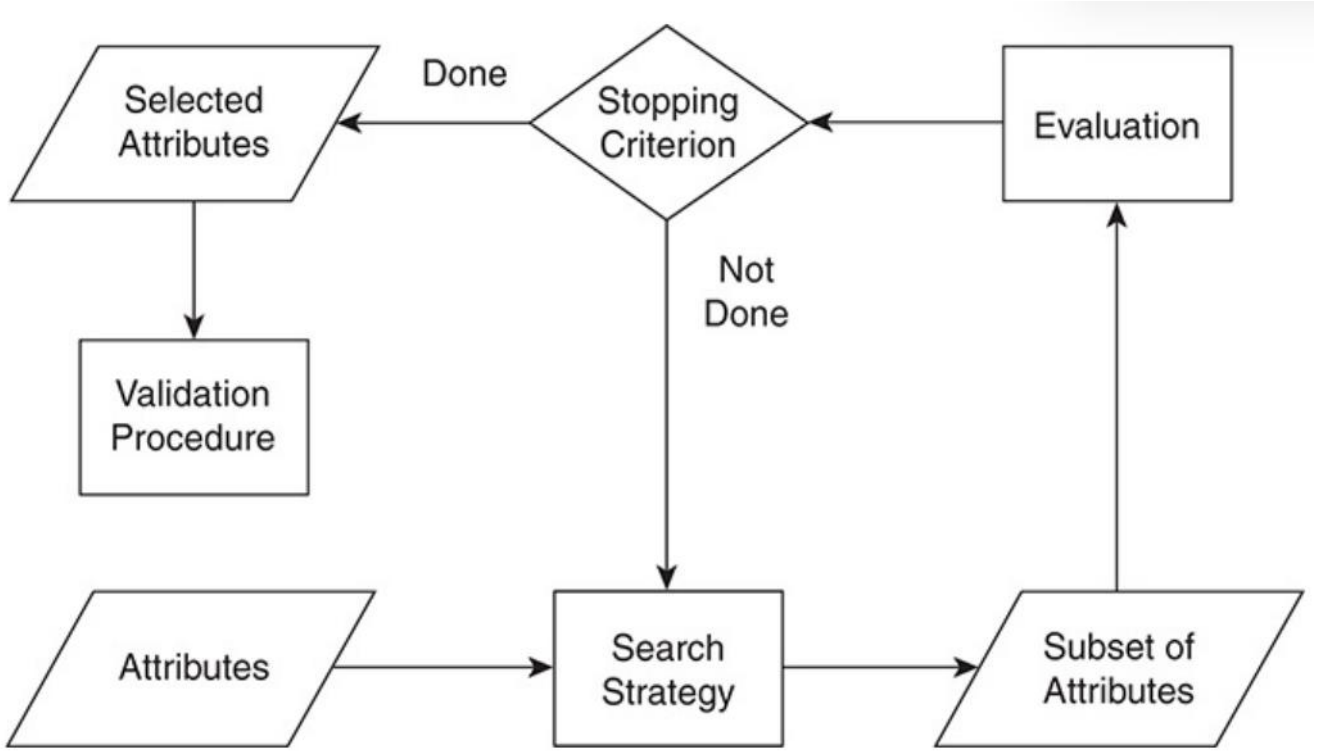
- While some irrelevant and redundant attributes can be eliminated immediately by using common sense or domain knowledge, selecting the best subset of features frequently requires a systematic approach.
- The ideal approach to feature selection is to try all possible subsets of features as input to the data mining algorithm of interest, and then take the subset that produces the best results.
- This method has the advantage of reflecting the objective and bias of the data mining algorithm that will eventually be used.
- Unfortunately, since the number of subsets involving n attributes is 2^n, such an approach is impractical in most situations and alternative strategies are needed.
- There are three standard approaches to feature selection: **embedded, filter,** and **wrapper**.

Feature Subset Selection

1. **Embedded approaches:**
- Feature selection occurs naturally as part of the data mining algorithm.
- Specifically, during the operation of the data mining algorithm, the algorithm itself decides which attributes to use and which to ignore.
- Algorithms for building decision tree classifiers operate in this manner.
2. **Filter approaches:**
- Features are selected before the data mining algorithm is run, using some approach that is independent of the data mining task.
- For example, we might select sets of attributes whose pairwise correlation is as low as possible so that the attributes are non-redundant.
3. **Wrapper approaches:**
- These methods use the target data mining algorithm as a black box to find the best subset of attributes, in a way similar to that of the ideal algorithm described above, but typically without enumerating all possible subsets.

Feature Subset Selection - Architecture

- The feature selection process is viewed as consisting of four parts: a measure for evaluating a subset, a search strategy that controls the generation of a new subset of features, a stopping criterion, and a validation procedure.
- Filter methods and wrapper methods differ only in the way in which they evaluate a subset of features.
- For a wrapper method, subset evaluation uses the target data mining algorithm, while for a filter approach, the evaluation technique is distinct from the target data mining algorithm.



Feature Weighting

- Feature weighting is an alternative to keeping or eliminating features.
- More important features are assigned a higher weight, while less important features are given a lower weight.
- These weights are sometimes assigned based on domain knowledge about the relative importance of features.
- Alternatively, they can sometimes be determined automatically.
- For example, some classification schemes, such as support vector machines, produce classification models in which each feature is given a weight.
- Features with larger weights play a more important role in the model.
- The normalization of objects that takes place when computing the cosine similarity can also be regarded as a type of feature weighting.

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:

1. Feature extraction:
   - Example: extracting edges from images.
2. Feature construction:
   - Example: dividing mass by volume to get density.
3. Mapping data to new space:
   - Example: Fourier and wavelet analysis.

Discretization and Binarization

- Some data mining algorithms, especially certain classification algorithms, require that the data be in the form of categorical attributes.
- Algorithms that find association patterns require that the data be in the form of binary attributes.
- Thus, it is often necessary to transform a continuous attribute into a categorical attribute (discretization), and both continuous and discrete attributes may need to be transformed into one or more binary attributes (binarization).
- Additionally, if a categorical attribute has a large number of values (categories), or some values occur infrequently, then it can be beneficial for certain data mining tasks to reduce the number of categories by combining some of the values.
- As with feature selection, the best discretization or binarization approach is the one that “produces the best result for the data mining algorithm that will be used to analyze the data.”

Binarization

- A simple technique to binarize a categorical attribute is the following: If there are m categorical values, then uniquely assign each original value to an integer in the interval If the attribute is ordinal, then order must be maintained by the assignment.
- Next, convert each of these m integers to a binary number. Since binary digits are required to represent these integers, represent these binary numbers using n binary attributes.
- To illustrate, a categorical variable with 5 values {awful, poor, OK, good, great} would require three binary variables X1, X2, X3.

Table 2.6. Conversion of a categorical attribute to five asymmetric binary attributes.

Categorical Value	Integer Value	x_1	x_2	x_3	x_4	x_5
awful	0	1	0	0	0	0
poor	1	0	1	0	0	0
OK	2	0	0	1	0	0
good	3	0	0	0	1	0
great	4	0	0	0	0	1

Binarization

- A simple technique to binarize a categorical attribute is the following: If there are m categorical values, then uniquely assign each original value to an integer in the interval If the attribute is ordinal, then order must be maintained by the assignment.
- Next, convert each of these m integers to a binary number. Since binary digits are required to represent these integers, represent these binary numbers using n binary attributes.
- To illustrate, a categorical variable with 5 values {awful, poor, OK, good, great} would require three binary variables X1, X2, X3.

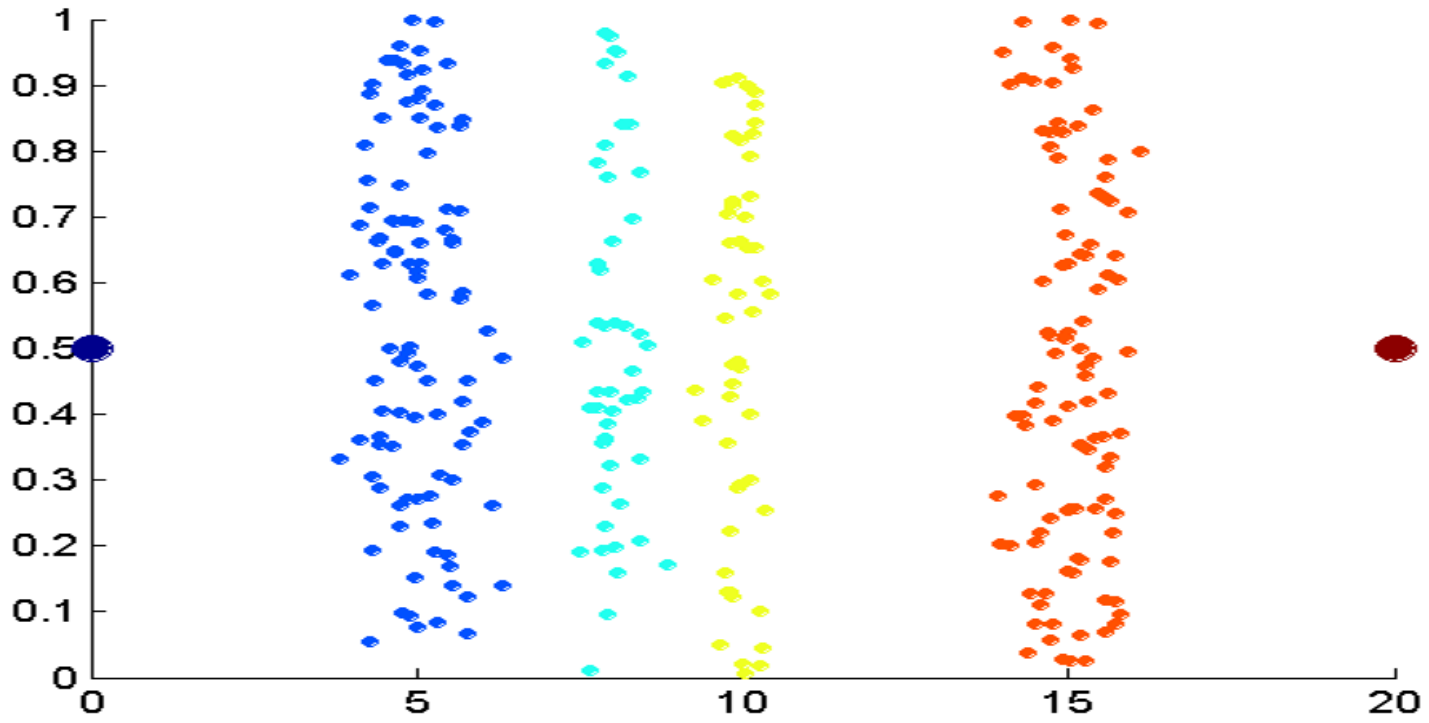
Table 2.6. Conversion of a categorical attribute to five asymmetric binary attributes.

Categorical Value	Integer Value	x_1	x_2	x_3	x_4	x_5
awful	0	1	0	0	0	0
poor	1	0	1	0	0	0
OK	2	0	0	1	0	0
good	3	0	0	0	1	0
great	4	0	0	0	0	1

Discretization

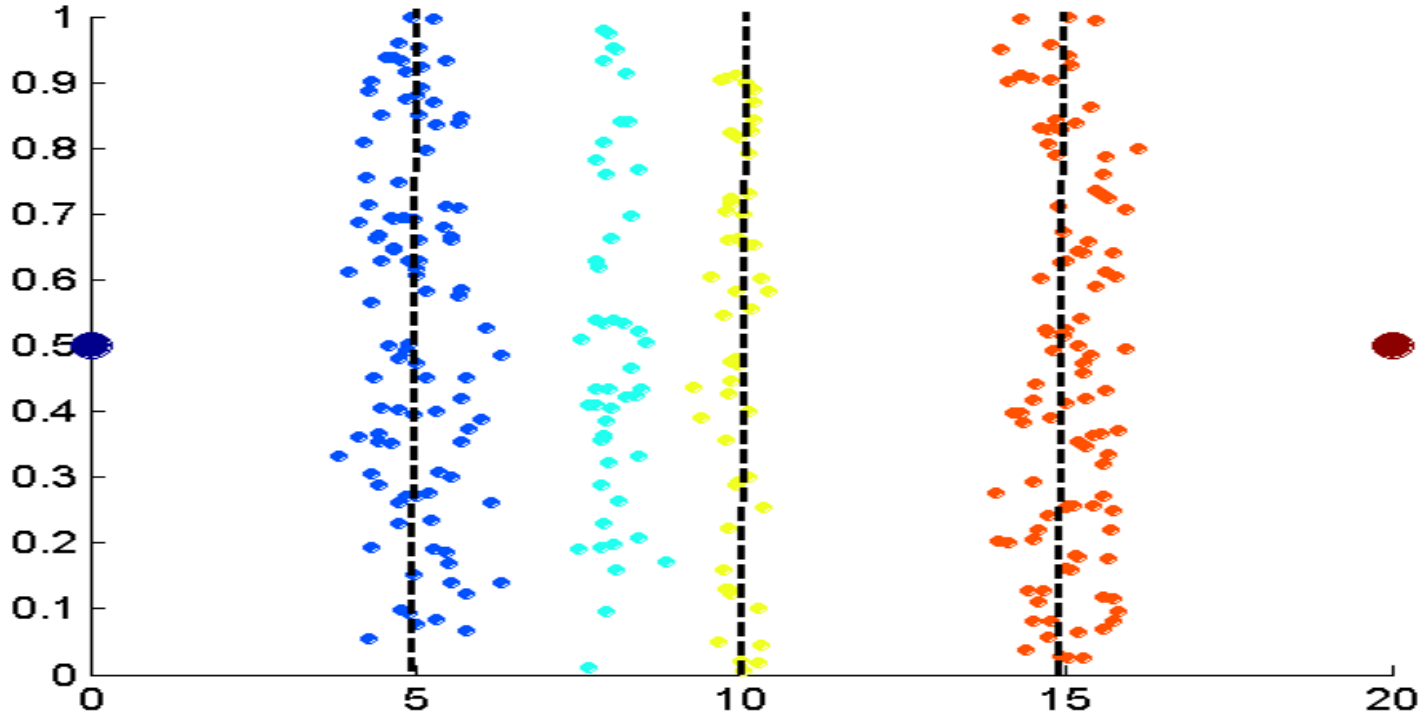
- Discretization is the process of converting a continuous attribute into an ordinal attribute.
- A potentially infinite number of values are mapped into a small number of categories.
- Transformation of a continuous attribute to a categorical attribute involves two subtasks: 1) deciding how many categories, n , to have and 2) determining how to map the values of the continuous attribute to these categories.
- In the first step, after the values of the continuous attribute are sorted, they are then divided into n intervals by specifying split points.
- In the second, rather trivial step, all the values in one interval are mapped to the same categorical value.
- The result can be represented either as a set of intervals {(x0, x1], (x1,x2], …, (x_{n-1}, x_{n-1}]}.

Unsupervised Discretization



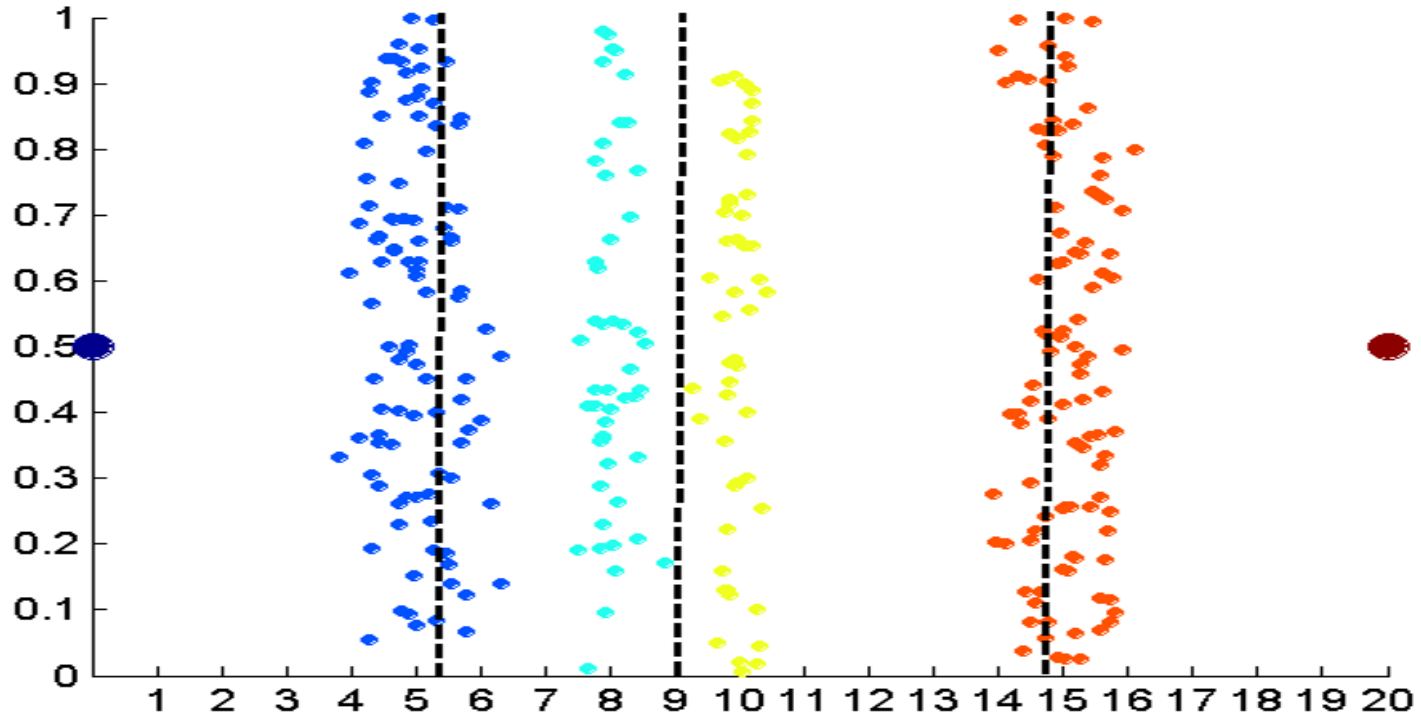
Data consists of four groups of points and two outliers. Data is one-dimensional, but a random y component is added to reduce overlap.

Unsupervised Discretization



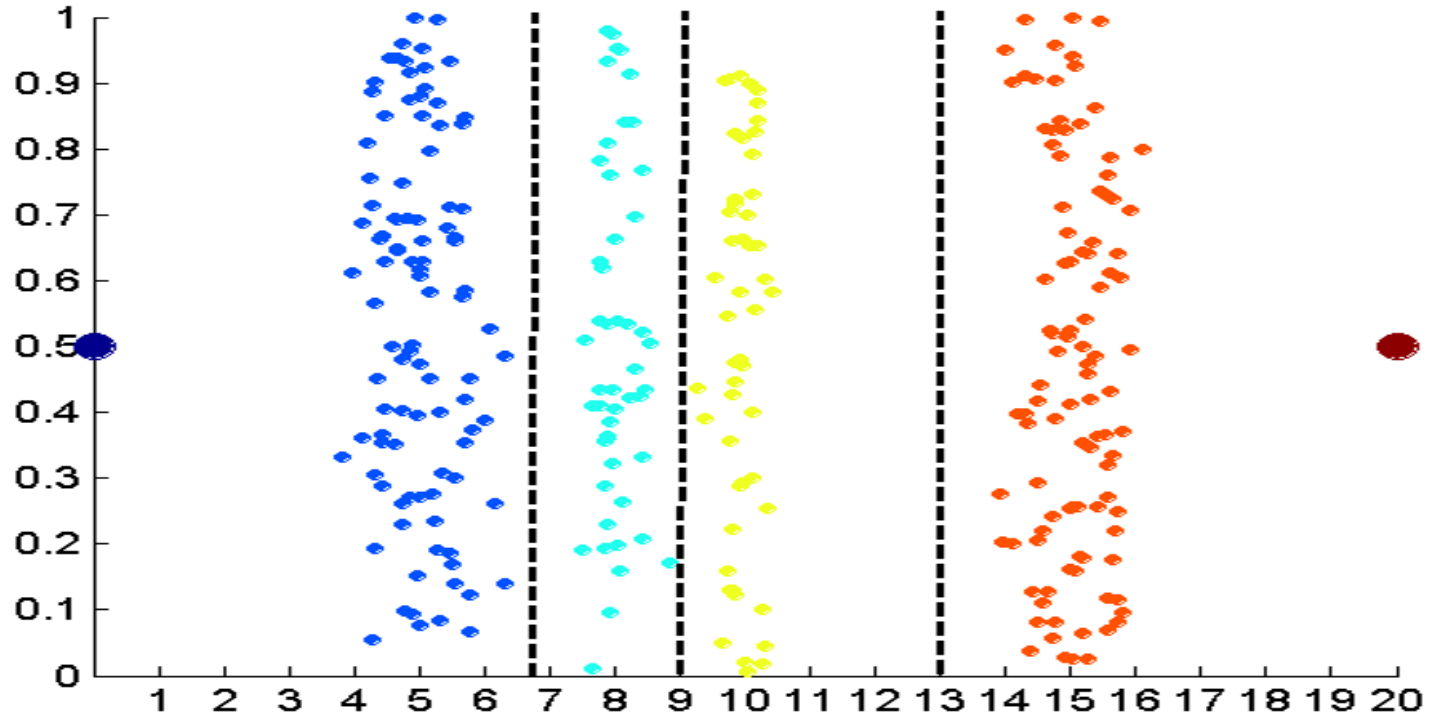
Equal interval width approach used to obtain 4 values.

Unsupervised Discretization



Equal frequency approach used to obtain 4 values.

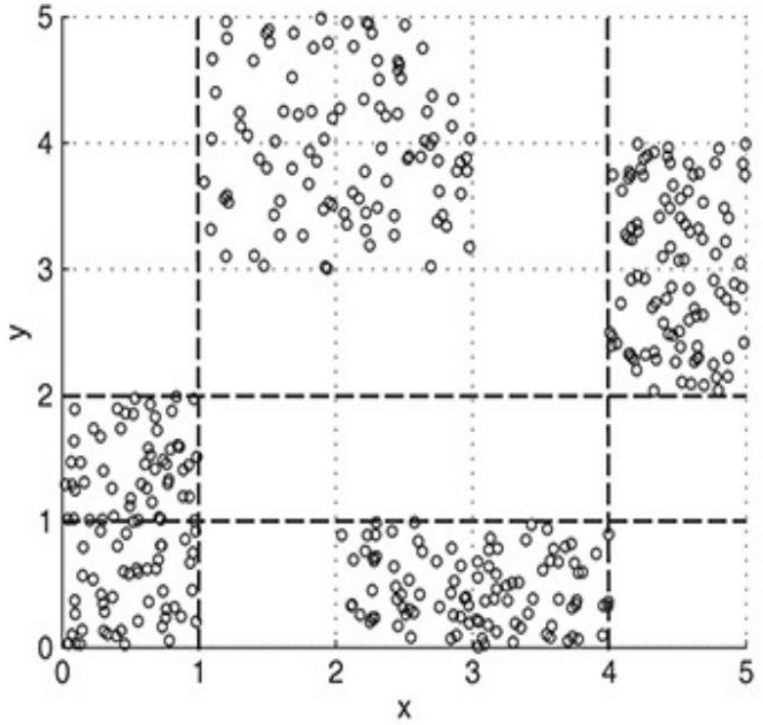
Unsupervised Discretization



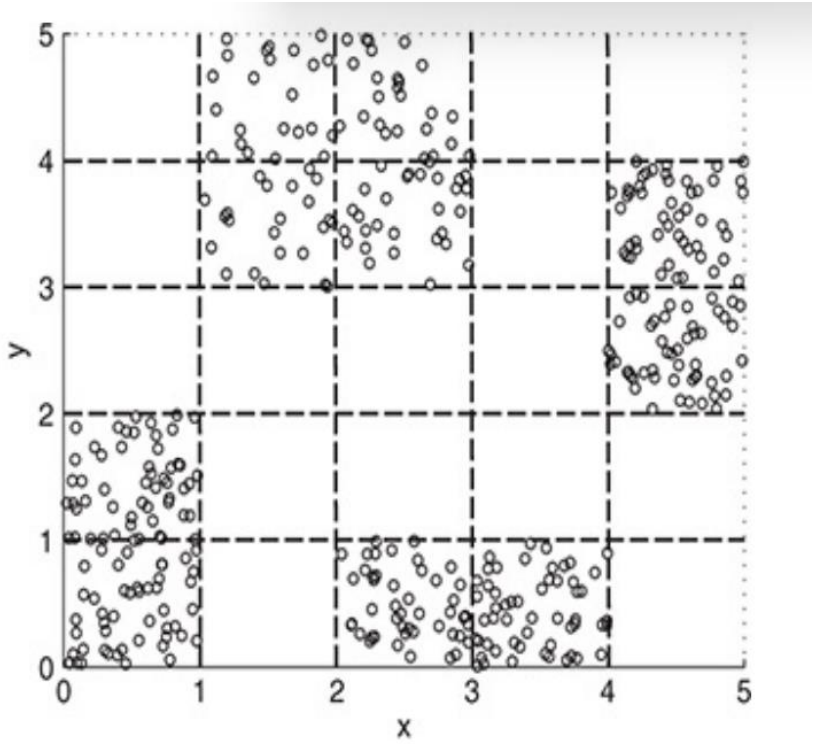
K-means approach to obtain 4 values.

Supervised Discretization

- Many classification algorithms work best if both the independent and dependent variables have only a few values
- We give an illustration of the usefulness of discretization using the following example.



(a) Three intervals



(b) Five intervals

Variable Transformation

- An attribute transform is 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|.

- Consider the variable of interest is the several data bytes in a session and the several bytes range from 1 to 1 billion. This is a huge range, and it may be advantageous to compress it by using a log10 transformation. In this case, sessions that transferred 108 and 109 bytes would be more similar to each other than sessions that transferred 10 and 1000 bytes (9 - 8 = 1 versus 3 - 1 = 2).

Normalization:

- Another common type of variable transformation is the standardization or normalization of a variable.
- The objective of standardization or normalization is to create a whole group of values that have a specific property.
- A common instance is that of "standardizing a variable" in statistics. If x’ is the mean (average) of the attribute values and sx, is their standard deviation, then the transformation x’ = (x –x’)/ sx) creates a new variable that has a mean of 0 and a standard deviation of 1.

Thank You!