CS-F415: Data Mining

10 Preprocessing Of The Data

Dr. Kamlesh Tiwari
Associate Professor, Department of CSIS,
BITS Pilani, Pilani Campus, Rajasthan-333031 INDIA

Sept 01, 2024 M/T/W - 9/8/9 am 6108 @ BITS-Pilani [Jan-May 2024]

http://ktiwari.in/dm

Data Preprocessing

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

After perprocessing data will be more informative and will reduce computational load.

Aggregation

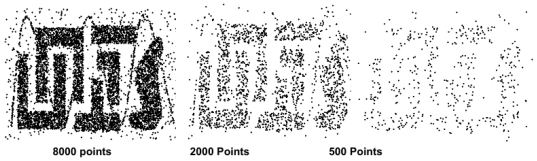
Combining two or more attributes (or objects) into a single attribute (or object)

- Reduces the number of attributes or objects
- Cities aggregated into regions, states, countries, etc.
- Days aggregated into weeks, months, or years
- More stable data - aggregated data tends to have less variability

Sampling

Sampling is the main technique employed for data reduction.

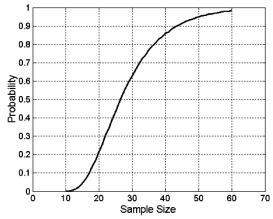
- Used for preliminary investigation and the final analysis.
- Because processing the entire set of data of interest is too expensive or time consuming.
- Using a sample will work almost as well as using the entire data set, if the sample is representative
- A sample is representative if it has approximately the same properties (of interest) as the original set of data



Types of Sampling

- **Simple Random Sampling** - Same selection probability
  - Sampling without replacement: selected item is removed
  - Sampling with replacement: NOT removed - item can appear again
- **Stratified sampling**
  Split the data into several partitions; then draw random samples from each partition
  Ensure that each subgroup is adequately represented

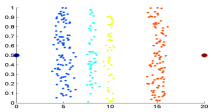
Sample size to get at least one object from each of 10 equal-sized groups.



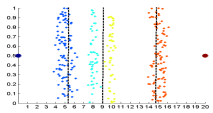
Discretization

Process of converting a continuous attribute into an ordinal attribute

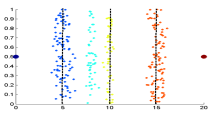
Data



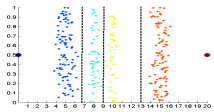
Equal frequency



Equal interval width



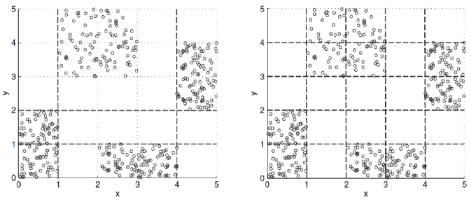
K-means



- Potentially ∞ values are mapped to small number of categories
- Discretization is useful for unsupervised and supervised both

Discretization in Supervised Settings

- 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

Binarization: One-Hot Encoding

- Binarization maps a continuous or categorical attribute into one or more binary variables.

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

Attribute Transformation

- Attribute transform is a function that maps the entire set of values of a given attribute to a new set of replacement values. Simple functions: x^k, $\log(x)$, e^x, $|x|$
- Normalization refers to various techniques to adjust to differences among attributes in terms of frequency of occurrence, mean, variance, range. Take out unwanted, common signal, e.g., seasonality
- Z Score: In statistics, standardization refers to subtracting off the means and dividing by the standard deviation

Thank You!

Thank you very much for your attention!

Queries ?