lec5-Exploration and cleaning

Data exploration and data cleaning

- ❖Data cleaning
- ❖Data exploration
- ❖Clustering
  - ➢Why clustering?
  - ➢What is clustering?
  - ➢How to do clustering?
- ❖Similarity and dissimilarity measurement

Data quality problems

Noise

modification of original values

→ Denoise data

Outliers

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

→ Remove outliers

Missing values

Reasons:
- Information is not collected
- Attributes may not be applicable to all cases

→ Handling missing data
- Eliminate Data Objects
- Estimate Missing Values
- Ignore the Missing Value During Analysis
- Replace with all possible values (weighted by their probabilities)

Duplicate data

Major issue when merging data from heterogeneous sources

→ Remove duplicates

Unnormalized data

not on the similar level of measurement

→ Data normalization

- Min-max normalization

$$v' = \frac{v - v^{min}}{v^{max} - v^{min}}$$

- Z-score normalization

$$v' = \frac{v - Mean(v)}{\mathrm{Std}(v)}$$

Assumption: Gaussian distribution

Unnormalized Non-Gaussian Data (eg. Gender)

→ Categorical data encoding

one-hot encoding

→ Data cleaning

Data exploration

Summary statistics

summarize properties of the data

- Measures of location: mean and median

mean is very sensitive to outliers

$$mean(x) = \frac{1}{m}\sum_{i=1}^{m} x_i$$

$$median(x) = \begin{cases} x_{(r+1)} & \text{if } m \text{ is odd, i.e., } m = 2r+1 \\ \frac{1}{2}\left(x_{(r)} + x_{(r+1)}\right) & \text{if } m \text{ is even, i.e., } m = 2r \end{cases}$$

- Measure of spread: range and variance
  - Range: max - min
  - Sensitive to outliers

other measures
- ➢**Median absolute deviation (MAD)**
  - $median([x_1 - mean(x)], \ldots, [x_m - mean(x)])$
- ➢**Interquartile range**
  - $x_{75\%} - x_{25\%}$

- Percentiles

p-th percentile is a value of X such that p% of the observed values of X are ***less than*** Xp

- Frequency and mode
  - frequency: the percentage of time the value occurs
  - mode: most frequent attribute value

Exploratory visualization

conversion of data into a visual or tabular format

- Histograms

shows the distribution of values of a single variable

- Two-dimensional histograms

Show the joint distribution of the values of two attributes

- Box Plots

distribution of data

Clustering

Better organization
Faster searching

→ Process of clustering
1. Data to be clustered
2. Similarity measurement
3. Clustering algorithm (the executive procedure)

Similarity and dissimilarity

- Similarity: Numerical measure of how ***alike*** two data objects are
- Dissimilarity (distance): Numerical measure of how ***different*** are two data objects

→ Cosine similarity

$$cos(d_1, d_2) = \frac{d_1 \cdot d_2}{(|d_1| * |d_2|)}$$

→ Correlation

measures the ***linear relationship*** between objects

$$\rho_{X,Y} = \mathrm{corr}(X,Y) = \frac{\mathrm{cov}(X,Y)}{\sigma_X \sigma_Y} = \frac{\mathrm{E}[(X-\mu_X)(Y-\mu_Y)]}{\sigma_X \sigma_Y}$$

→ Euclidean distance

$$Ed(\boldsymbol{p}, \boldsymbol{q}) = \sqrt{\sum_{i=1}^{m} (p_i - q_i)^2}$$

→ Minkowski distance

generalization of Euclidean Distance

$$dist(\boldsymbol{p}, \boldsymbol{q}) = \left(\sum_{i=1}^{m} |p_i - q_i|^r\right)^{\frac{1}{r}}$$

Extra details (not included in lecture)

Clustering VS Classification

Parameter	CLASSIFICATION	CLUSTERING
Type	used for supervised learning	used for unsupervised learning
Basic	process of classifying the input instances based on their corresponding class labels	grouping the instances based on their similarity without the help of class labels
Need	it has labels so there is need of training and testing dataset for verifying the model created	there is no need of training and testing dataset
Complexity	more complex as compared to clustering	less complex as compared to classification
Example Algorithms	Logistic regression, Naive Bayes classifier, Support vector machines, etc.	k-means clustering algorithm, Fuzzy c-means clustering algorithm, Gaussian (EM) clustering algorithm, etc.

- **Clustering**

identifies **similarities** between objects, which it groups according to those characteristics in common and which differentiate them from other groups of objects

- **Classification**

uses **predefined classes** in which objects are assigned