CS432 – Data Mining – Assignment 1

Report

Q1 – Folder Task1 (exec file: run.m)

(a): script3.m Line 18
(b): Top5: script3.m Line 44, Boxplots: script4.m
(c)/(d)/(e): histograms/scatterplots/trend: script5.m
(f): Total 340 teams. Top five being: Chennai, Mumbai, Lancashire, Hampshire, Warwichshire in the order as is. Better performing teams had also played one of the most number of matches. Chennai super kings won more times when it played against Mumbai Indians.

Q2 – Folder Task2

(a): Basic Stats generated by RapidMiner "Statistics" showing Average, Min, Max, Missing Values.
Top 5 attributes had 1675 missing values out of 1994 total rows == 84%
(b): 2_replace_missing.csv
(c): 3_normalize_all.csv

Screenshot provided.

Q3 – Folder Task3 (exec file: run.m)

(a): script1.m: Line 13 to Line 19
(b): script1.m: Line 21 to Line 298 covering all scatter plots AUC crap
(c): script1.m: Line 303: Chi Value = 701.2505
(d): Alcohol has the most correlation with Quality. More alcohol in wine generally meant better quality, Volatile Acidity has most correlation with Red Wine. A very high Chi Square Value so null hypothesis was rejected.

Q4 – Folder Task4 (exec file: script_1.m)

(a): script_1.m: Line 7 – Center this dataset
(b): script_1.m: Line 12 – standardize
(c): script_1.m: Line 17 – PCA
(d): script_1.m: Line 32 – NNMF
(e): script_1.m: Line 43 – Factor analysis
(f): All these methods helped in reducing data while losing minimum information with best understanding/correlation between attributes.

- Centering Data simplifies notation and computations.
- Z-Score Standardization (a) allows us to calculate the probability of a score occurring within our normal distribution and (b) enables us to compare two scores that are from different normal distributions.
- Principal Component Analysis is a linear subspace projection technique used to down-sample high-dimensional datasets (and minimize the re-projection error). It is calculated by solving an algebraic eigenvalue problem: finding the eigenvectors (PC's) of the covariance matrix of your original dataset. The resulting PC corresponding to the largest eigenvalue is the line of highest variance. The PC corresponding to the second largest eigenvalue is the direction of second highest variance (and is orthogonal to all previous PC's).
- NNMF is a group of algorithms in multivariate analysis and linear algebra where a matrix V is factorized into (usually) two matrices W and H, with the property that all three matrices have no negative elements. This non-negativity makes the resulting matrices easier to inspect. Since the problem is not exactly solvable in general, it is commonly approximated numerically.
- Factor analysis is best related to linear regression as a system of linear equations with a missing right hand side (independent variable).

Source: Wikipedia, Quora etc

Q5 – Folder Task5 - Bonus

Tool Used: OpenRefine

a. Basic stats generated using Facet tool in open refine. Entries sorted by count to get the highest no. of missing values for each columns. Top five were col 102,103,104,105,106.
b. Missing values were replaced by the median of the attribute. Mean was not used since more than 75% of the values were missing for Top 5 columns.
c. Filter tool can be used to find all cells with "?" and their values replaced.
d. Cluster and Edit feature used to remove groups of the same data. None was found since data was numeric.
e. Sort rows by a specific attribute if you need Top 5 or Top 10 values from dataset etc.
f. Screenshots provided