CS685: Data Mining

Assignment 1 (100 marks)

Due on: 8th September, 2018, 10:00pm

Please download the data, `datagov.zip` from the course website and/or canvas.

The data is about the various states and union territories of India from `http://data.gov.in` and is of 2011. The categories are demography, education, and economy.

Assume that there are six (6) classes of India, according to regional divisions. The divisions are given in the file `regions.csv`.

You can use any software in Linux as long as it is free.

Q1 (15 marks)

Handle the missing values using average of other members in the same class.

Q2 (20 marks)

Find top-5 states/union territories that are representative of India (i.e., closest to the vector corresponding to India).

Do this exercise with and without normalizing the data.

Q3 (10 marks)

To handle redundancy under each category (demography, education, economy), find the correlations among the attributes within a category and across categories.

What can you conclude?

Q4 (25 marks)

Find the two (2) most important attributes in each category and across categories using the multi-class Relief algorithm.

Q5 (15 marks)

Plot all the points (each point represents a state/union territory) in a 2D plot from the output of Q4. There will be four (4) such scatter plots.

Produce the quantitle-quantile plots as well.

Q6 (15 marks)

Perform an intuitive partitioning of the output of Q4 up to the second level.