Test-1

COMP 7150/8150

FALL 2016

Instructor: Deepak Venugopal

Due Date: 11:59 P.M. November 11, 2016 (submit code, answers to ecourseware)

NOTE: No collaboration of any kind with fellow students is allowed. You may not refer to or take code from the internet. Plagiarism of any kind will be treated very seriously.

For visualizations, you will not need anything more complex than scatter-plots, histograms or line plots. You will provide a single python notebook that contains the code for all the answers. Use a separate tab for each question. For each question, also write your appropriate answers in a .txt, .doc or .pdf and submit this along with your code.

1. (20 points) I have provided you with a dataset called dataset-q-1.csv. Analyze this dataset and use a suitable Machine Learning algorithm to predict the "Class" given the feature values. You need to answer the following:
    a. Why is the algorithm you chose appropriate for this dataset based on your analysis? (Hint: Present some appropriate visualization of the dataset)
    b. Evaluate your method appropriately and present the results.

2. (20 points) I have provided you with a dataset called movie.zip. It contains multiple files that describe movie ratings. Your task is to perform the following:
    a. For each genre, compare the average rating by female users between ages 18-25 (including 18, 25) to the average rating by male users between ages 18-25 (including 18,25). (Hint: You would need to extract the required data from the files and visualize the results)
    b. We wish obtain the 5 highest rated movies (based on average rating) released in 1995?

3. (20 points) I have provided you with a dataset called dataset-q-3.zip. It contains a train and test dataset. Use a suitable method to predict the "Value" given the features (there are 100 features) (Hint: there are a number of redundancies in the features). Evaluate and present your results using an appropriate error measure.

4. (20 points) I have provided you with two datasets in dataset-q-4.zip. For each dataset:
    a. Analyze the data using an appropriate visualization
    b. Use an appropriate method to cluster similar data-points together. Justify why you picked the specific method for each dataset.
    c. Output the clustered points using an appropriate visualization.

5. (20 points) Use the 'sci.space' category from the training set of fetch_20newsgroups as follows. newsgroups_train = fetch_20newsgroups(subset='train', categories=['sci.space'])

Compare how similar is the first document in newsgroups_train to the rest of the documents in newsgroups_train. Use an appropriate metric to measure similarity and also visualize the output similarities.