DSBA 6156 – Fall 2019

Quiz- 2

Due: 11/13/19 (7:59 am)

Instructions:

1) Please submit a PDF file with all the answers for your quiz.
2) Please submit all supporting Jupyter notebooks associated with the quiz
3) The first three questions are programming based and the last one requires you to answer with critical reasoning.
4) All submissions must be on Canvas.
5) The teaching staff will most likely request a 5-7-minute time slot to discuss your solutions during evaluation and ensure academic integrity.

The data set for this quiz: the 'pizza dataset' (available on Canvas) consists of the following features:

brand -- Pizza brand (class label)
id -- Sample analysed
mois -- Amount of water per 100 grams in the sample
prot -- Amount of protein per 100 grams in the sample
fat -- Amount of fat per 100 grams in the sample
ash -- Amount of ash per 100 grams in the sample
sodium -- Amount of sodium per 100 grams in the sample
carb -- Amount of carbohydrates per 100 grams in the sample
cal -- Amount of calories per 100 grams in the sample

Question 1 (25 points):

The exercise is to perform clustering, a machine learning technique that involves grouping of data points based on similarity or distance between them. Clustering is a method of unsupervised learning and is a common technique for statistical data analysis.

For this problem, use all the columns.

- Conduct k-means clustering on the dataset above.
- Determine the optimal number of clusters for k-means and discuss the reasoning behind your choice.

Question 2 (25 points):

The second exercise is to perform PCA, a dimensionality reduction technique, aimed at detecting correlation among features and merging/shrinking them to fewer feature sets, while maximizing the information contained in them. (Refer to the PCA tutorial on Canvas if you have questions)

- What's the optimal number of components for the pizza dataset?
- What is the explained variance for every new PCA feature?
- What does the feature distribution look like?
  (Hint: Compare original features to New PCA components)

Question 3 (20 points):
The third exercise is to perform k-means clustering on the features obtained from PCA. (For this experiment, exclude all original features of the dataset and use only the PCA components)

- Do you see a change in the ideal number of clusters that are to be used? Explain your intuition based on the results observed.

Question 4 (30 points):
Suppose you're a data scientist consultant, and one of your clients is Goodreads.com. Goodreads is a social cataloging website that allows individuals to freely search its database of books, annotations, and reviews. Users can sign up and register books to generate library catalogs and reading lists. With a model in place, the company recommends books to new and existing uses. They want you to work on improving their model. How would you approach this problem? Based on your understanding of how recommender systems are built using a combination of how content and collaborative filtering is used, what data would you need specifically from Goodreads to build a recommendation engine? Justify why you would need every data point/feature using the concepts from the theory of recommender systems.

For example, one question you may ask is if they have a rating system? If they do, what's the scale?

For this question, we are expecting a "specification" and "model description" document. Since this is an examination, one page would suffice. But please be as complete as you can. Once a developer reads this document, they should have a clear understanding of the kind of data that is available, why a certain data point was requested from GoodReads, what models are going to be used, and how that would feature in your recommendation engine?

Secondly, given that you can construct a social network from the user database of Goodreads and that you have a list of genre preferences for each user, explain how would you construct a topic-specific PageRank algorithm to design a recommender engine.