TA Notes

# Section Overview

1. Student should understand what is Machine Learning, when we need it and how to formulate the problem
2. Feature Selection-Student should understand which features are important for different datasets
3. Feature Engineering
   1. Student should understand how to extract meaningful information from the data.
   2. They should able to handle both the textual and non-textual data sets
4. Algorithms
   1. Students should understand a brief theory of the different algorithms that are provided in the lecture notes
   2. They should be able to import sklearn libraries and use them
5. Evaluation
   1. They should know how to evaluate the algorithms and determine which algorithms work best on different data sets

# Exercise Solution:

Textual Dataset

Q. What words are likely to appear in a positive review, and in a negative review?

A: 'good', 'well', 'excellent', 'amazings' and etc are supposed to appear in positive reviews, and 'bad', 'disappointing', 'worst' and etc are supposed to appear in negative reviews. Studetns should be able to intuitively understand the bag-of-words model from this question.

Q. What is the problem with unigram model?

A: 1. There are phrases that can represent the opposite meaning than unigrams, such as 'not good' is in negative reviews however unigram will classify this review as positive. 2. Many words appears frequently in all reviews such as 'the', 'a', 'and' and etc, those should not be considered as useful features. Students should be able to intuitively understand the motivation of n-gram models and tf-idf values.

Numerical Data

Q. Which features are important for this dataset?

A. Base Pay, Over time Pay, Benefits

Q. How have salaries changed over time between different groups of people?

A. Data exploration and data visualization is important. Will differ from different groups.

Q. How are base pay, overtime pay, and benefits allocated between different groups?

A. Study the age attribute and plot appropriate graphs

Q. Predict whether there is pay discrimination based on gender

A. Depends on the dataset and implementing algorithms

Linear Regression

Q: In the above linear regression implementation, check what is the correlation between features by including latitude and longitude features. Does including location variables change the accuracy of linear fit?

A: Expect the correlation plot between location variables and Y variable (stars in this case)

The accuracy of the model depends on the correlation plot generated above. The accuracy of the fit is expected to increase by including more granular details like location of the business.

kMeans Clustering

Q: When would the Means algorithm stop?

A:The algorithm halts when no points are reassigned to different clusters.

Q: Given the following data points cluster them in 3 clusters.

3,10,15,25,22,9,34,18,19

A: Stepwise implementation of algorithm for allocating the points to 3 clusters.

Cluster 1 - 3,9,10

Cluster 2 - 15,18,19

Cluster 3 - 22,25,34

Q: For the given yelp reviews data set determine K using elbow method.

Features – latitude, longitude and business\_avg\_stars

A: Use elbow method.

Q: Using the k determined above, cluster the businesses in K clusters using k-means clustering.

A: Python implementation using scikit-learn