MAS Data Science and Engineering – DSE 220

Assignment #1

1. **Problem #1**
   Download the Weather data set, a simple data set describing whether or not to play tennis based on the weather conditions.

Day	outlook	temperature	humidity	windy	play
1	sunny	85	85	FALSE	no
2	sunny	80	90	TRUE	no
3	overcast	83	86	FALSE	yes
4	rainy	70	96	FALSE	yes
5	rainy	68	80	FALSE	yes
6	rainy	65	70	TRUE	no
7	overcast	64	65	TRUE	yes
8	sunny	72	95	FALSE	no
9	sunny	69	70	FALSE	yes
10	rainy	75	80	FALSE	yes
11	sunny	75	70	TRUE	yes
12	overcast	72	90	TRUE	yes
13	overcast	81	75	FALSE	yes
14	rainy	71	91	TRUE	no

- Represent the following table using a data structure of your choice
- Given that data structure
  - Calculate the mean temperature and mean humidity
  - Print outlook and play for those days where the temperature is greater than the average temperature
  - Print outlook and play for those days where the humidity is greater than the average humidity
  - Convert the temperature to Celsius and add a new column therefore in the table. Use the following conversion equation

$$C=(F-32)*5/9$$

1. How often do you play tennis independent of the other attributes?
2. How often do you play tennis when it is "sunny"?
3. Compare the average, minimum and maximum temperature when you play tennis?
4. Compare the average, minimum and maximum humidity when you play tennis?
5. Plot the an scatter plot (x,y diagramm) of humidity (x) and temperature (y) when you play tennis compared to when you do not play tennis.

2. Problem #2

Included with the assignment are several files (stxxxxts). These files track the historical population of US states by year for 1900-1990. Write a script to process these data and load them into a data structure you can work with. What problems did you have to deal with when working with these files?

Plot the populations of Alaska and California over time. Plot the population of New England and the Southwest over time. Don't forget to label your axes.

What state showed the greatest change in population? Note that there is more than one way to quantify this - provide at least two (meaningful) ways in your iPython Notebook.

3. Problem #3

The IRI data is set available on the AWS instance. Several additional documents describing the data set and pointing to additional Bibliography are made available (including the white paper: Bronnenberg, Bart J., Kruger, Michael W, and Mela, Carl F. The IRI Marketing Dataset.). Getting to know this data set is time well spent as it will be very applicable towards the Final project. The data set is of medium size and complexity – however understanding the data and how file/folders/datasets related to each other will be of crucial importance for creating an appropriate training data set.

The data set has been purchase by UCSD, it is only for internal use and comes with restrictions. Please read and follow the NDA specified at the very end of the "IRI_database_technical_appendix" document provided with the data set.

I order to answer the questions below – document and describe what kind of data preparation and cleaning techniques did you have to apply? What kinds of anomalies and missing data did you encounter? Document how what assumption you made and how you approach the issues.

` 1. What are the most popular item(s) for a chosen product/category?
2. How does the popularity of an item change with the cost associated with it for a chosen year?
4. How is the popularity affected over the years?

Extra credit:

Apply K-nearest neighbor algorithms to a chose subset of data for a specific set of products/categories.

4. Problem #4 - Decision Trees

Classification trees, either binary or multi-class, are implemented in scikit-learn in the DecisionTreeClassifier class. Build, plot and evaluate a decision tree on the wine dataset. Split the dat set into 75% for training and 25% for testing. Evaluate based on

confusion matrix how well the model performed on training vs. testing. Document the steps taken.

5. **Problem #5 – optional**
   Download Dataset of London 2012 Olympians called AHW_1.CSV. Perform Data preparation and cleaning of the data set.

- What are the statistical distributions of variables using no class? How much missing data is there? How do distributions differ by each gender? Describe summary statistics for each attribute. Plot each one of the attributes distributions. Are any of the variables different for male vs. female athletes? Visualize potential difference via the scatter plots. Are there any 'high' correlations between variables? Create a new variable for the weight in lbs. Check out the correlations again. Do you notice any changes? Remove one of the weight variables. Add new variable *weight + height*. Visualize scatter plot. Is this a useful variable? Repeat the same exercise for Body Mass Index defined as Mass (kg)/Height(m) ^{2} (Note: Weight already in Kg. and Height is in cm). Is this a useful variable? Plot the BMI of the athletes. Are there any obese athletes? Male of Female? Visualize scatterplots of Total Class with Height, Weight, Sex and BMI. Split data by sport. What can you conclude based on the split?