Zach Jagoda

Dr. Erik Linstead

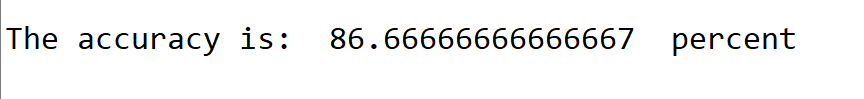
Introduction to Data Science

**Assignment 3: KNN (K Nearest Neighbor)**

1. What missing values and outliers did you identify? How did you fix them? If you discover outliers, you must impute values to replace them, not delete them.

In regard to the missing values, the sample\_train.csv that we were provided with was missing three values and contained two outliers. The missing values resided in (Column, Row) D 22, B 65, and C 97. The outlying data resided in D 44 and A 148. I fixed the values by calculating the Average of all the numbers per type of flower. I also calculated the Standard Deviation and worked to find the High and Low of the numbers I calculated. For missing values, I simply added in the Average number that was calculated. For the outlier data, defined as being outside of the range of the High and Low numbers calculated, they were also replaced with the Average number.

2. What is the accuracy of nearest neighbor on the provided data? (Copy and paste the output of your program to answer this question)



3. What lines of code are the most computationally intense? (You can copy your py file into word and highlight the lines)

def knn(train\_data,train\_labels,test\_data):

predictions = []

greatestDist = -2.0

distance = -1.0

c = 0

for test in test\_data:

for index, train in enumerate(train\_data):

distance = cosDistance(test, train)

if(distance > greatestDist):

greatestDist = distance

c = index

predictions.append(train\_labels[c])

greatestDist = 0.0

return predictions

4. How would your distance metric have to change if the data contained categorical variables in addition to the continuous variables?

If you were to measure categorical variables in addition to the continuous variables, you would have to hard code a similarity function in order to weight the differences between the two attributes.