Nicholas Fang

Assignment 3

Question 1

1. Create a logistics regression model to predict the class label from the first eight attributes of the question set.
2. Try doing the same using two different subsets (your choice) of those eight attributes.
3. Report the accuracies of each of these three models.
    I. Accuracy of the first model when using all 8 attributes is reported to be 0.555, meaning an accuracy score of ~55.55%.
    II. Accuracy of the second model when using attributes (num_words, num_misspelled, bin_end_qmark, and 'num_interrogative) is 66.66%
    III. Accuracy of the third model when using attributes ('num_misspelled, bin_start_small, and num_punctuations) is 44.44%.
4. For the two subsets that you use, provide some justification (why you chose those features in a given subset).
    I. The automated answer rating marks posts as good or bad based on the quality of the post. Following this logic, a good post would likely be descriptive and few mistakes. Furthermore, if the post was asking a question, then adding whether or not the post ends with a question mark and whether or not the post properly used interrogative words would be beneficial.
    II. The second subset follows the same logic. Began by choosing attributes which I thought that a quality post would incorporate but without being in a question format.
    III. Mixed results from logistic regression. Results ranging from 44-66% after running the 2 models with chosen attributes multiple times

Question 2

1. Download the wine dataset. Download wine dataset.It contains information about several wines—their characteristics (features) and if it's considered high quality or not (1 or 0).
2. First, do some experiments (trial-and-error) to figure out a good subset of features to use for learning wine quality (last column). Report these features.

I. Features chosen for this experiment are alcohol, density, chlorides, residual sugar, and volatile acidity.

3. Then, use 70% data for training to build a kNN classifier with different values of k ranging from 2–10.

K(2) = 83.08, K(3) = 82, K(4) = 82.05, K(5) = 81.38, K(6) = 81.74, K(7) = 81.28

K(8) = 80.87, = , K(9) = 81.28 , K(10) = 81.28

4. Plot your accuracies with each of these. In other words, your final result will be a line chart with k on the x-axis and accuracy on the y-axis.