Assignment 9
Natural Language Processing, Naive Bayes, Support Vector Machines and Neural Networks

Applied Data Science

Instructions

This assignment has three (3) parts.

The first part will test your understanding of the algorithms learned in the tutorial. You will be implementing Naive Bayes on a small dataset; answering questions about assumptions of Naive Bayes, SVMs and Neural nets and designing a neural network to solve a logic gate problem.

The second part requires you to perform all steps involved in developing a classification model, starting from collection of data, feature engineering, data pre-processing, model fitting and analyzing model performance. For the Twitter data that you will be extracting, build classification models using Naive Bayes, SVMs and Neural nets. You will not be graded based on accuracy of the model. Points will be awarded based on clarity of reasoning and proper interpretation of results.

Lastly we expect you to determine which algorithm can be used to best model the given scenario. Provide detailed reasoning for choosing that algorithm.

Submit your assignment as a .pdf file with your name and email in the header, single-spaced and font no-larger than 12pt. For the questions required to be solved with pen and paper, take pictures of your answers and insert them to the PDF file. Attach the code used for part II in the appendix of the PDF file. The assignment will be graded on the quality of the descriptions and overall quality.

Part I

Naive Bayes

1. Build a Naive Bayes classifier for the dataset given below and find the probability of it raining given that the temperate is High, it is windy and the Humidity is high.

Temperature	Windy	Humidity	Rain
High	Yes	Low	No
High	Yes	Moderate	Yes
Low	Yes	High	Yes
Low	Yes	Low	No
High	No	Low	No
High	No	Moderate	Yes
Low	No	Low	No
Low	No	Moderate	No

Your are required to finish this question using pen and paper. Please include all the details of how you calculate the probabilities.

2. What is the basic assumption of the Naive Bayes classifier and why is it a drawback when it comes to solving real life problems?

Support Vector Machines

1. For the classification space given below, identify the support vectors that help define the margin of separation:

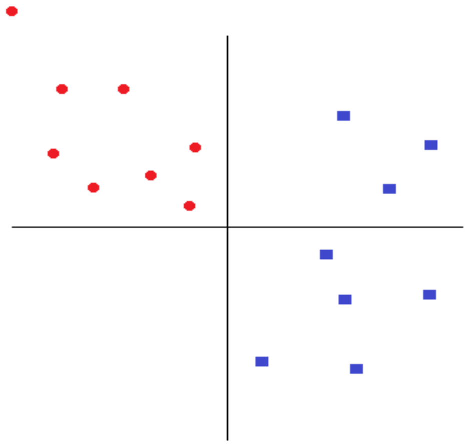

Figure 1: 2-D Classification space

Print this plot and do the best you can to draw the vector by hand since the actual coordinates are not provided.

2. What is the underlying assumption of SVMs. Explain briefly how SVMs can be applied in non-linearly separable spaces.

Neural Networks

1. Given the following 3-class classification problem:

C1: (4,1), (2,3), (3,5), (5,4), (1,6)
C2: (0,2), (-2,2), (-3,2), (-2,4)
C3: (1,-2), (3,-2)

(a) Will a single layer neural network be sufficient to predict the class of the input data? Justify your answer.

(b) Add the sample (-1,6) to C1. Repeat part (a).

2. Design a neural network to determine ouput 'd' based on the inputs a, b and c. Choose the appropriate activation function and number of layers to solve the neural net problem.

a	b	c	d
0	0	0	0
0	0	1	1
0	1	0	1
0	1	1	1
1	0	0	0
1	0	1	0
1	1	0	0
1	1	1	0

Part II

Extract tweets using the Twitter API as demonstrated in the tutorial for two specific hashtags: '#dog' and '#cat'. Collect equal number of tweets for each of the hashtags and it is not necessary to collect more than 5000 tweets in total. A given tweet belongs to class 'dog' if it contains '#dog'(even if the tweet contains both '#dog' and '#cat' it will belong to class 'dog') and belongs to class 'cat' otherwise.

In this part of the assignment, you will be using the dataset created above to develop classification models. Create appropriate features and develop models using Naive Bayes, SVMs and Neural nets and compare their accuracies. Make sure to split the data into 80-20 for training and validation.

You DO NOT have to submit the dataset you used for the analysis. Provide detailed explanation on how you created the model. Reflect on the performance of each of the models in correctly predicting the class of the tweet.

Part III

For each question, determine which algorithm is best. Discuss why you chose the algorithm you did and why you did not select the others.

1. You have 100 observations of users top 5 favorite songs on a popular music streaming site, what algorithm is best to predict the favorite song of 20 users based on their other four favorite songs?
2. You have 5 billion observations with 300 million variables from your most recent self-driving car testing, but you're missing 5% of your data because of an intermittent sensor failure. Which algorithm can help you complete your dataset?
3. You have 5 billion observations with 300 million variables from your recent self-driving car testing. Each observation has been labeled as either "hazard" or "safe", depending on the safety of the situation. Which algorithm can help you predict whether new observations are occurring in hazardous or safe situations?