Problem Set

The assignment is worth 10% of your final grade.

Why?

Now it's time to check your understanding in machine learning, theory and Bayesian statistics.

The Problems Given to You

Question 1:

(1) Implement a Naïve Bayes Classifier [30 pt].

In this part, you will write a Naive Bayes classifier and verify its performance on a news-group data-set. As you learned in class, Naive Bayes is a simple classification algorithm that assumes about conditional independence of features, but it works quite well in practice. You will implement the Naive Bayes algorithm (Multinomial Model) to classify a news corpus into 20 different categories.

Data Download [Link].

You have been provided with the following data files:

- train.data - Contains bag-of-words data for each training document. Each row of the file represents the number of occurrences of a particular term in some document. The format of each row is (docId, termId, Count).
- train.label - Contains a label for each document in the training data.
- test.data - Contains bag-of-words data for each testing document. The format of this file is the sameas that of the train.data file.
- test.label - Contains a label for each document in the testing data.

For this assignment, you need to write code to complete the following functions:

- logPrior(trainLabels) - Computes the log prior of the training data-set. (7 pts)
- logLikelihood(trainData, trainLabels) - Computes the log likelihood of the training data-set. (8 pts)
- naiveBayesClassify(trainData, trainLabels, testData) - Classifies the data using the Naive Bayes algorithm. (15 pts)

Implementation Notes

1. You are still allowed to “steal” other’s functions. You can still build your classifier on top of others’ code. However, you can’t wrap other’s classifier in your function.
2. If you indeed borrowed others’ code, comment your code intentionally to demonstrate your understanding of that specific code snippet.
3. We compute the log probabilities to prevent numerical underflow when calculating multiplicative prob-abilities. You may refer to this article on how to perform addition and multiplication in log space.
4. You may encounter words during classification that you haven’t during training. This may be for a particular class or overall. Your code should deal with that. Hint: Laplace Smoothing
5. Be memory efficient and please do not create a document-term-matrix in your code. That would require upwards of 600MB of memory.

(2) Challenge component [10pt]

In the above question, we are using all the terms from the vocabulary to make a prediction. This would lead to a lot of noisy features. Although it seems counter-intuitive, classifiers built from a smaller vocabulary perform better because they generalize better over unseen data. Noisy features that are not well-represented often skew the perceived distribution of words, leading to classification errors. Therefore, the classification can be improved by selecting a subset of extremely effective words. Write a program to select a subset of the words from the vocabulary provided to you and then use this subset to run your naive bayes classification again. Verify changes in accuracy. TF-IDF and Information Theory are good places to start looking.

Question 2 [10pt].

Imagine you had a learning problem with an instance space of points on the plane and a target function that you knew took the form of a line on the plane where all points on one side of the line are positive and all those on the other are negative. If you were constrained to only use decision tree or nearest-neighbor learning, which would you use? Why?

Question 3 [10pt].

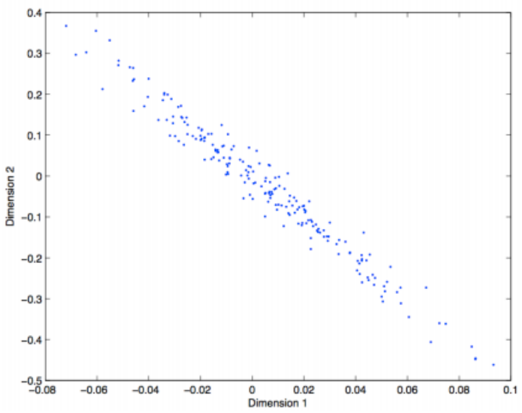
Give the VC dimension of the following hypothesis spaces. Briefly explain your answers.

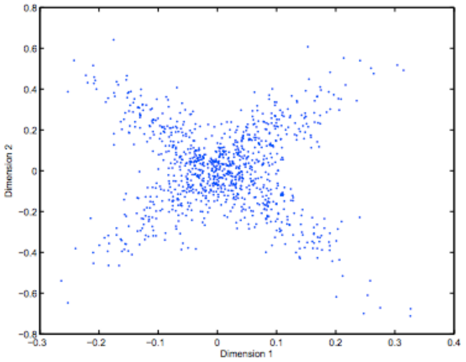
1. An origin-centered circle (2D)

2. An origin-centered sphere (3D)

Question 4 [10pt].

Plot the direction of the first and second PCA components in the figures given



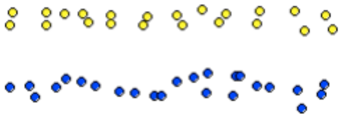


Question 5 [10pt].

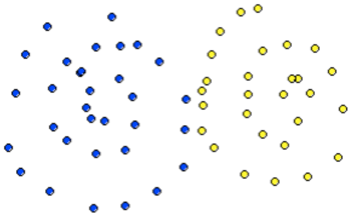
Which clustering method(s) is most likely to produce the following results at k = 2? Choose the most likely method(s) and briefly explain why it/they will work better where others will not in at most 3 sentences. Here are the five clustering methods you can choose from:

1. Hierarchical clustering with single link
2. Hierarchical clustering with complete link
3. Hierarchical clustering with average link
4. Kmeans
5. EM

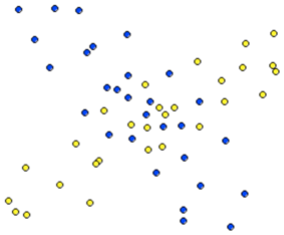
a.



b .



c.



Question 6 [10pt].

Explain how you can use Decision Trees to perform regression? Show that when the error function is squared error, then the expected value at any leaf is the mean. Take the Boston Housing dataset (https://archive.ics.uci.edu/ml/datasets/Housing) and use Decision Trees to perform regression.

Question 7 [10pt].

Suggest a lazy version of the eager decision tree learning algorithm ID3. What are the advantages and disadvantages of your lazy algorithm compared to the original eager algorithm?

What to Turn In

You must submit a tar or zip file named *firstname_lastname_NUID*.{zip,tar,tar.gz} that contains a single folder or directory named *firstname_lastname_NUID* that in turn contains: -->

1. A file named *README.txt* that contains instructions for running your code
2. Your code
3. A file named firstname_lastname_NUID-*classifier.pdf* that contains your writeup.
4. Another file named firstname_lastname_NUID-*solutions.pdf* that answers questions 2 to 7.
5. Any supporting files you need (for example, your datasets).