CS6923 Machine Learning (Spring 2015)
Homework 2

Due March 9 in class, but you can hand this homework in to me in my office or mailbox until Wed. March 11 at 2:30, without penalty.

Answers to questions must be submitted in hardcopy. For problem 6, hand in hard-copy of your code and upload your code to NYU Classes also. Do not upload the other parts of the homework.

HOMEWORK COLLABORATION POLICY:
You may discuss the general concepts in this homework with other students, but you must write up the solutions and write the programs on your own. NO SHARING OF CODE IS ALLOWED. If you are not sure whether you are crossing the line between general discussion and inappropriate collaboration, speak to me first. Violation of this policy can result in disciplinary and grade penalties, including receiving a grade of F in this course.

1. Below is a table giving a training set for a binary (two-class) classification problem with 3 attributes, in which one of the attributes has three possible values. The rest have only two possible values.

X1	X2	X3	label
Yes	Normal	A	+
No	Normal	B	+
Yes	Abnormal	C	-
No	Abnormal	B	-
No	Abnormal	A	+

Use Naïve Bayes on this training set to predict the label for the example (Yes, Abnormal, B). Show your work. (Note that you do not have to calculate e.g., P(X1=No|+) to answer this particular question, because X1=Yes in the example you are predicting.

2. Suppose we are given a random sample $\mathcal{X}$ of size N drawn iid from a Bernoulli distribution defined by parameter θ (so $P[x=1] = \theta$). Suppose we have a continuous prior distribution on θ, with pdf $p(\theta) = 3\theta^2$. Answer the following questions, given that the random sample $\mathcal{X} = \{1,0,1\}$. (You may want to look at p. 63 of the textbook.)
    a. What is the log likelihood $\mathcal{L}(\theta=1/2|\mathcal{X})$?
    b. What is the ML estimate for θ?
    c. What is the MAP estimate for θ? Show your work.

d. What is P[{1,0,1}]? Remember to take the prior into account in this calculation. Show your work.
e. What is the Bayes estimate for θ? Show your work.

3. In class, we mentioned that in implementing Naïve Bayes, it is common to do ``smoothing'' to avoid having probability estimates that are equal to 0. One smoothing technique is called "add-k" smoothing (sometimes called additive smoothing), where we choose the value of k. Suppose that D is an unknown discrete distribution defined on the finite domain {1,..,t}, where p_j equals P[X = j], for X drawn from D. Given an iid sample $\mathcal{X}$ of size N drawn from this distribution, if the value j in {1,..,t} occurs N_j times in that sample, the estimate of p_j using add-k smoothing is

$$\frac{N_j + k}{N + kt}$$

t number of unique case - here 3 balls

When k=1, this technique is sometimes called Laplace smoothing. In practice, add-1 smoothing often does not work well, and it is usually better to do additive smoothing with a smaller value of k. An easy way to remember the formula for add-k smoothing is to imagine adding k additional occurrences of each value j $\in$ {1,…,t} to the sample $\mathcal{X}$, and then calculating the frequency estimates using this new, augmented sample of size N+kt.

[Note: Additive smoothing has a Bayesian interpretation. It is equal to the posterior probability that the next element sampled from D will be equal to j, given the sample $\mathcal{X}$, assuming a certain prior distribution on D. Equivalently, it is equal to the Bayes estimate of p_j assuming this prior. For k=1, this prior is just the uniform distribution over all possible distributions on {1,..,t}. More generally, for k >= 1, the prior is a distribution called the symmetric Dirichlet distribution with parameter k.]

When K = 0 then Nj/N =4/10
K = 0.1 then 4.1/10.3
K = 1 then 5/13

a. Suppose we pick 10 balls randomly, with replacement from an urn containing red, green, and yellow balls. Suppose 4 of the balls we picked were red, 5 were green, and 1 was yellow. Give 3 estimates of the fraction of red balls in the urn, using additive smoothing with k=0, k=0.1, and k=1.

b. As described at the start of this problem, add-k smoothing is a method for estimating the probability P[X=j] from the iid sample $\mathcal{X}$. Consider the above formula for add-k smoothing. Depending on how you set k, you get a different estimate for P[X=j]. Which value of k yields the maximum likelihood estimate for P[X=j]?

In given problem, we have to know the value of K which yields the maximum likelihood estimate. Now MLE according to given equation , we have to set the value of K such away that N + Kt ~= Nj + K
in other words we should select value of K(minimum) in such away that it will not impact much on estimation.

4. Let D be a Bernoulli distribution, defined by the parameter θ. So, if X is a random variable from this distribution, θ = P[X=1]. Let S be an iid sample of size N drawn from D. Let N_1 denote the number of occurences of 1 in the sample, so $N_1 = \sum_{x \in S} x$. Let d=N_1/N. Suppose we use d as our estimator of θ.

   a. What is the bias of the estimator d? You may use θ in your answer if necessary. Justify your answer. N1/N - 1/2
   b. What is the variance of the estimator d? You may use θ in your answer if necessary. Justify your answer. (Hint: Given two independent random variables X and Y, with variances σ_X and σ_Y respectively, what is the variance of X+Y? Consult a statistics textbook if you don't want to figure this out for yourself.)
   c. Does the variance of the estimator d increase or decrease as N increases?

N Increase then variance of d decrease.

5. Consider the following linear regression dataset:

x	R
3	4
5	12
8	25
11	76

Recall that the formula for the line minimizing the squared error is $w = A^{-1}y$.

   a. For the above dataset, list the values of A, A^{-1}, and y.
   b. Give the equation for the line minimizing the squared error on the above dataset. It should be in the form

$$g(x) = w_1 x + w_0$$

6. Download the zipped folder **enronfiles.zip** (posted on NYU Classes with this homework). This folder contains a preprocessed dataset of emails, labeled according to whether they are spam or ham (non-spam). The emails were written by employees at the Enron corporation, and were collected during the legal investigation into the Enron scandal. The emails became public because they were part of the court record. There are very few datasets containing real emails that are publicly available (because most people don't want to release their private emails), so researchers in text processing began using the Enron dataset for research. The folder also contains some Matlab template files to help you answer the questions below.

   The data files in enronfiles.zip do not include the original emails. The emails were processed and a vocabulary of terms appearing in the emails was compiled. Let W be the number of terms in the vocabulary. One file contains a Wx1 character array containing the terms in the vocabulary, stored as strings.

The emails are divided into three sets, train, validation, and test. (We won't use the validation set in this exercise.) For each of these sets, there are two files, one with information about the words appearing in the emails in that set, and one with information about the labels of those emails. The file with the term (feature) information contains a DxW matrix, in sparse format, where D is the number of emails in the set. Each row of the matrix corresponds to an email in the set, each column corresponds to a vocabulary term, and entry [i,j] of the matrix contains the number of occurrences of vocabulary term j in email number i in the set. The file with the label information contains a Dx1 matrix where the ith entry is 1 if email number i is spam, and 0 if it is ham.

We will use a simple binary Naïve Bayes approach to classify these emails.

a. Open the files in Matlab.
   i. How many words are in the vocabulary? (The vocabulary is stored in the matrix `vocab`, so you can run `length(vocab)` to answer this question.) 158963
   ii. How many examples (emails) are in the training set? 6744
   iii. How many examples are in the test set? spam - 3435, size of testLable matrix = 6744 X 1
   iv. Look at the first 50 terms in the vocabulary. Are they all words?
   v. What percentage of training documents are spam? 10302/20229 = 50.9268%
   vi. What percentage of test documents are spam? 3435/6744 = 50.934164 %

No some of them are comma,underscore, dot and inverted comma

b. Write a Matlab program that implements Naïve Bayes on a "binary" version of the dataset. In the binary version of the dataset, we replace each positive entry in the train and test matrix with the number 1. We view each email as an example, and each term as a feature. Each email thus corresponds to a binary feature vector, whose entries correspond to the terms in the vocabulary, where the term corresponding to an entry is set to 1 if the term appears in the document, and is set to 0 otherwise. (This corresponds to the "Bernoulli" document model for text, and Naïve Bayes applied in this way to text classification is sometimes called Bernoulli Naïve Bayes. For more information about the difference between Multinomial and Bernoulli Naïve Bayes for text classification, see the textbook Christopher D. Manning, Prabhakar Raghavan and Hinrich Schütze, *Introduction to Information Retrieval*, Cambridge University Press. 2008. The relevant chapter is available online at http://nlp.stanford.edu/IR-book/pdf/13bayes.pdf)

   Use add-k smoothing, for k=0.1, when calculating the conditional probability estimates P(D|C). When applying the smoothing formula, remember that each feature can only have two possible values, 0 or 1.

   To avoid underflow problems, compute log likelihoods, log P(D|C), instead of computing P(D|C) directly. To get you started, we provide a template with partial code in the file "bernoulliNB.m".

The training and test data for this problem are stored in sparse Matlab matrices. These matrices are very large, but contain many 0's. Matlab sparse matrices store only the non-0 entries, resulting in a significant savings in space, and reducing the time required to perform common operations on these matrices.

If your program is running so slowly that you cannot get results on the full dataset, you should check your code. Avoid loops and use matrix operations instead where possible. See if you can use operations that take advantage of the sparsity (small number of non-zero entries) of the matrices.

Print out hardcopy of your code and include it when you hand in your homework in class. You also need to upload your code on NYU Classes.

c. Using your program, train on the examples in the training set and test on the examples in the TEST set. Also test it on the documents in the TRAINING set.
   i. What percent accuracy did you obtain on the TRAINING set?
   ii. What percent accuracy did you obtain on the TEST set?
   iii. In general, we expect to get higher accuracy on the TRAINING set than on the TEST set. Is this what you observed?

d. The accuracy that can be achieved on this dataset is extremely high. Most classification tasks are not this easy. Why do you think such a high accuracy is achievable on this particular dataset? If you have time, you may want to look more carefully at the data, and run further experiments, to try to figure out what is happening. Alternatively, try to come up with a plausible reason.

7. In this exercise, you will briefly explore linear regression, using Matlab, by generating some artificial linear data with additive Gaussian noise.

a. First, generate and plot some linear data, with additive Gaussian noise, using the following sequence of Matlab commands:

```
% Let y = .75x + 2 be our line. Assume additive noise generated from the standard Gaussian distribution N(0,1).
% Generate 15 points (x,y+e) where x is a random number between 0 and 5, and e is the random noise.
w = .75;
b=2;
x = 5*rand(15,1);
t = w*x+b;
t_noisy = t + (.5*randn(length(t),1));
```

plot(x,t_noisy,'ms');

% Now, at the top of the plot window, go to the Tools menu and select Basic Fitting.
% This opens a menu. On the menu, check the 'linear' checkbox. You will see the line that minimizes squared error.
% Now also check the boxes for cubic. Then check the box for degree 7. Finally, check the box for degree 10.
% Click on the right arrow at the bottom of the Basic Fitting menu, which opens the numerical results.

For each of the 4 lines/curves generated, list the "Norm of residuals". (If an error message was received, indicate that instead.) The Norm of residuals number is equal to the square root of the sum of the squared errors. More precisely, if X is the dataset containing N points, "residuals" is the vector of length N with an entry for each point x in X. The entry for x in X has the value (r – y), where r is the given output for x in the dataset, and y is the output for x that is predicted by the curve. Letting R be the residuals vector, "Norm of residuals" reports the value norm(R), which is the L_2 norm of the vector R (=$\sqrt{(R(1))^2 + (R(2))^2 + \cdots (R(n))^2}$).

b. Repeat the above, but this time generate 500 points. Show the results.
c. If you received an error message above, why do you think that happened?