CPSC 540 Assignment 1 (due September 10)

Review of prerequisites and first taste of supervised learning

This assignment has two purposes:

1. To refresh your knowledge on some of the prerequisite topics that we will use. It's ok if you are not familiar with all of these; the goal is to help you identify areas that you may need to read about (or practice) to keep up with the course.

2. To give you a first taste of the training/testing supervised learning framework applied to a real data set.

It is highly-encouraged to work in groups, since there are people with many different backgrounds in the class. However, please hand in your own solutions. Also, please write legibly and divide your submission into the numbered sections used in this document, think of the marker!
Updates/clarifications to the assignment after its initial release (based on discussions on Piazza) are marked in red.

1 Logistics Survey

Write your name, student number, faculty, home department, year of study, and whether you are an undergraduate or a graduate student.

2 Calculus Refresher

Hint: You can find the answers to all these questions in a calculus textbook or using Wolfram-Alpha.
Compute the derivative with respect to x of the following functions:

1. $f(x) = ax + b$ (linear)
2. $f(x) = ax^2 + bx + c$ (quadratic)
3. $f(x) = \log(x)$ for $x > 0$ (logarithm)
4. $f(x) = ae^{bx}$ (linear times exponential)
5. $f(x) = \log(\sum_{i=1}^{n} \exp(a_i x))$ (log-sum-exp)

In this course we will use $\log(x)$ to mean the 'natural' logarithm of x, so that $\log(e) = 1$.
Write down the values of following sequences/series (or say that they diverge):

6. $\sum_{k=1}^{N} k$
7. $\sum_{k=1}^{\infty} r^k$ for $|r| < 1$ (geometric series)
8. $\sum_{k=1}^{\infty} \frac{1}{k}$ (harmonic series)
9. $\sum_{k=1}^{\infty} \frac{1}{k^2}$
10. $\sum_{k=1}^{N} (a_k - a_{k-1})$ (telescoping series)
11. $\lim_{n\to\infty} \left(1 + \frac{1}{n}\right)^n$ (compound interest)

3 Linear Algebra Refresher

Hint: you can find the answers to all the questions in this section from a friend who has taken a few linear algebra courses, or here: `http://www.cs.ubc.ca/~schmidtm/Documents/2009_Notes_LinearAlgebra.pdf`
Using the definitions below,

$$\alpha = 2, \quad x = \begin{bmatrix} 0 \\ 1 \\ 2 \end{bmatrix}, \quad y = \begin{bmatrix} 3 \\ 4 \\ 5 \end{bmatrix}, \quad z = \begin{bmatrix} 1 \\ 2 \\ -1 \end{bmatrix}, \quad A = \begin{bmatrix} 3 & 2 & 1 \\ 2 & 3 & 1 \\ 1 & 1 & 3 \end{bmatrix},$$

evaluate the following expresions:

1. x^T (transpose)
2. $\alpha(x+y)$ (vector addition and scalar multiplication)
3. $x^T y + x^T z$ (inner product and inner product between orthogonal vectors)
4. A^T (transpose of symmetric matrix)
5. Ax (matrix-vector multiplication)

If $\{x, y, z\}$ are real-valued column vectors (of the same length) and $\{A, B, C\}$ are real-valued matrices such that the additions/multiplications below have the right dimensions, which two of the following are not true in general?:

$$x^T y = y^T x$$
$$x^T A y = y^T A^T x$$
$$x^T (y + z) = x^T y + x^T z$$
$$x^T (y^T z) = (x^T y)^T z$$
$$A + (B + C) = C + (A + B)$$
$$A(BC) = (AB)C$$
$$A(B + C) = AB + AC$$
$$AB = BA$$
$$(AB)^T = B^T A^T$$

Write down the defining property of the following special types of matrices:

1. Identity matrix.
2. Tridiagonal matrix.
3. Rank-1 matrix.
4. Orthogonal matrix.
5. Positive semi-definite matrix.

4 Algorithm Runtimes Refresher

Hint: You can find the answers to these questions on Wikiepdia or in *Introduction to Algorithms* by Cormen, Leiserson, Rivest, and Stein.
In big O notation, what are the costs of the following operations (e.g., finding the maximum in a list of n numbers costs $O(n)$ since we must check each element):

1. Sorting a list of n numbers.
2. Finding the median in a list of n numbers.

3. Finding the smallest element greater than 0 in a sorted list with n numbers (binary search).
4. Computing a matrix-vector product with an m-by-n matrix, Ax.
5. Computing the longest common subsequence between a string of length m and a string of length n (dynamic programming).
6. Performing a breadth-first search through a graph with V vertices and E edges.

5 Probability Exercises

We review basic concepts in probability in class. Use probability arguments to address the following famous problems. Show your calculations in addition to giving the final result. The latter three problems come from MLAPP Chapter 2.

5.1 Two sons problem

I independently toss two fair coins (each having a 0.5 probability of landing 'heads' and 0.5 probability of landing 'tails'). If I tell you that the first coin landed 'heads', what is the probability that the second coin landed 'heads'. If I instead tell you that at least one coin landed 'heads', what is the probability that both coins land heads?

5.2 Bayes rule for medical diagnosis

After your yearly checkup, the doctor has bad news and good news. The bad news is that you tested positive for a serious disease, and that the test is 99% accurate (i.e., the probability of testing positive given that you have the disease is 0.99, as is the probability of testing negative given that you don't have the disease). The good news is that this is a rare disease, striking only one in 10,000 people. What are the chances that you actually have the disease?

5.3 Prosecutor's fallacy

A crime has been committed in a large city and footprints are found at the scene of the crime. The guilty person matches the footprints. Out of the innocent people, 1% match the footprints by chance. A person is interviewed at random and his/her footprints are found to match those at the crime scene. Determine the probability that the person is guilty, or explain why this is not possible.

5.4 The Monty Hall problem

On a game show, a contestant is told the rules as follows:

> There are three doors, labelled 1, 2, 3. A single prize has been hidden behind one of them. You get to select one door. Initially your chosen door will *not* be opened. Instead, the gameshow host will open one of the other two doors, and *he will do so in such a way as not to reveal the prize.* For example, if you first choose door 1, he will then open one of doors 2 and 3, and it is guaranteed that he will choose which one to open so that the prize will not be revealed.
>
> At this point, you will be given a fresh choice of door: you can either stick with your first choice, or you can switch to the other closed door. All the doors will then be opened and you will receive whatever is behind your final choice of door.

Imagine that the contestant chooses door 1 first; then the gameshow host opens door 3, revealing nothing behind the door, as promised. Should the contestant (a) stick with door 1, or (b) switch to door 2, or (c) does it make no difference? Assume that initially the prize is equally likely to be behind any of the 3 doors, and that if the prize is behind door one that the host is equally likely to choose door 2 or 3.

6 Naive Bayes and K-Nearest Neighbours

Expand the file *ass1.zip*, switch to the created directory, and (in Matlab) run the script *demo*. This script

1. Loads the data from Figure 1.2 of MLAPP (which I've split into training data (X,y) and testing data (Xtest,ytest)).
2. Trains an extremely naive model (that always predicts the most common class).
3. Tests this model on the test examples.

Modify this demo so that it implements a naive Bayes classifier (note that there are binary features but 4 possible class labels). Report the test accuracy with naive Bayes and include the modified file.
What are the 10 most predictive words for each of the 4 classes?
(Use the 'wordlist' and 'groupnames' vectors.) You can define the 'most predictive' words for class c as those where $p(x_i = 1|y_c = c, \theta_c^j)$ is the largest, or you can use a different estimate and give an argument why it should be used.

6.1 Naive Bayes with Dirichlet Prior

Naive Bayes can be problematic if a feature is always zero (or one) in the training data (if it is different in the test data, then we predict that all classes have zero probability). We can avoid this problem (and others) by adding one to each of the probabilities estimated by the model before normalizing them. Modify your naive Bayes implementation to include this modification.
Report the test accuracy with the prior added. (Hint: Section 3.3.4 of MLAPP)
Can we give an interpretation to this modification? (Hint: Section 3.5.1.2 of MLAPP)
Now what are the 10 most predictive words for each of the 4 classes?

6.2 Extensions of Naive Bayes

What is one way to modify naive Bayes to allow continuous features?
What is one way to make naive Bayes less 'naive'?

6.3 K-Nearest Neighbours

Make a new file that implements the k-nearest neighbours algorithm as discussed in class. Use Euclidean distance as the distance metric. Report the test error for $k = 1$ up to $k = 10$, and include the new function.

Matlab is not very fast at executing 'for' loops but is very fast at matrix multiplication operators. You may find the following command useful for more quickly computing the squared Euclidean distances between all rows of two matrices X and Xtest (of sizes N-by-P and T-by-P):

```
D = X.^2*ones(P,T) + ones(N,P)*(Xtest').^2 - 2*X*Xtest';
```

Assume ties are broken in order of the training examples: E.g., if $k = 1$ and test example i is equally close to training examples 65 and 102, choose example 65 as the neighbour. Similarly, if two or more classes appear equally often among the k closest training examples then choose the smaller class number.

6.4 Extensions of K-Nearest Neigbours

How might we modify KNN if we had a continuous label?
What might a smart distance metric take into account?

6.5 Bias vs. Variance

The K-nearest neighbours classifier has a high 'variance' but a low 'bias', while naive Bayes has a low 'variance' but a high 'bias'. What do you think these terms means?