CPSC 340 Assignment 1 (due September 18)

Summary Statistics and Data Visualization, Decision Tress and Cross-Validation, Probability

- You can work in groups on the assignments. However, please hand in your own assignments and state the group members that you worked (as well as other sources of help like online material).
- It preferred that you type up the solution to your assignment. If you are going to submit hand-written parts, please write legibly. Marks will be taken off for unreadable or unclear solutions.
- Please organize your submission according to the sections used in this document.
- Place your name and student on the first page, and (if submitting a paper company) then please staple your document together.
- All Sections (1-4) are equally weighted.
- There may be updates/clarifications to the assignment after the first version is put online. Any modifications will be marked in red.
- We may change from paper submission to an electronic submission as a PDF file. If that change takes place, instructions will be placed here and the assignment will still be due at the start of Friday's class.

1 Logistic Survey

Please fill out the survey located here:
`https://survey.ubc.ca/surveys/37-7d0090012c11ea5c07f0bca610f/cpsc-340-logistic-survey`

2 Summary Statistics and Data Visualization

Download and expand the file *a1.zip*, which contains data on the athletes from the last summer olympics.[1] You can load this data into Matlab from the directory containing the file using:

```
load london2012.csv
```

This creates a matrix 'london2012', where each row corresponds to an athlete and the columns correspond to:

1. Age.
2. Height.
3. Weight.
4. Gender (1 - female, 0 - male).

[1] Data obtained at:`http://www.theguardian.com/sport/datablog/2012/aug/07/olympics-2012-athletes-age-weight-height#data`, and I removed athletes with missing values.

5. Number of bronze medals.
6. Number of silver medals.
7. Number of gold medal.

2.1 Summary Statistics

Report the following statistics:

1. Range of age values (i.e., minimum and maximum).
2. Median of age value for each gender.
3. The 10%, 25%, 50%, 75%, and 90% age quantiles.

2.2 Data Visualization

Show the following figures:

1. Histogram of age values.
2. Scatterplot of height and weight values, coloured by gender.
3. Boxplot of weight values for each age value.

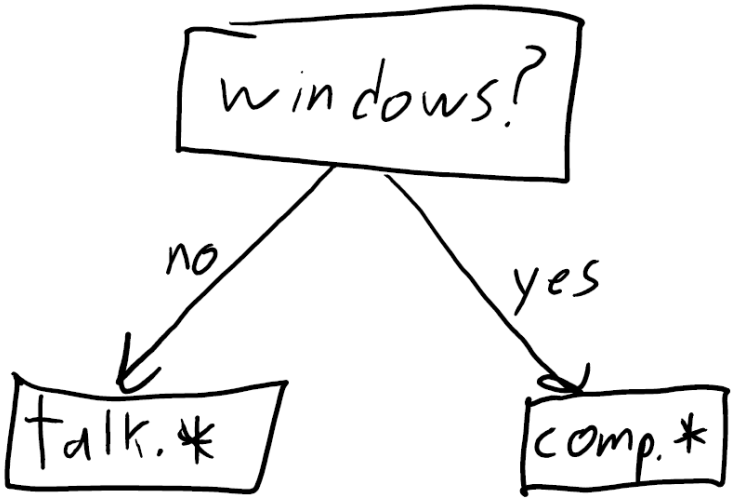
3 Decision Trees and Cross-Validation

The file *newsgroups.mat* is a Matlab file containing the following objects:

1. *groupnames*: The names of four newsgroups.
2. *wordlist*: A list of words that occur in posts to these newsgroups.
3. X: A sparse binary matrix. Each row corresponds to a post, and each column corresponds to a word from the word list. A value of 1 means that the word occured in the post.
4. y: A vector with values 1 through 4, with the value corresponding to the newsgroup that the post came from.
5. $Xtest$ and $ytest$: the word lists and newsgroup labels for additional newsgroup posts.

3.1 Decision Stumps and Decision Trees

The function *example_decisionStump* shows how to load the data, fit a decision stump to the training data, and then evaluate the error of the model on the training data. Running the demo shows that decision stumps have a classification error of 0.60, which is a bit better than just predicting the most common label (which obtains an error of 0.66). The image below gives an interpretation of the decision stump that is learned:



This is an interpretable but not very accurate model. Modify this demo so that it uses the provided *decisionTree* function rather than the *decisionStump* function. By looking through the decision tree code, draw a picture (similar to the one above) showing the learned decision tree when the maximum depth is 2, and report the accuracy when using a decision tree of depth 10.

3.2 Cost of Fitting Decision Stumps and Decision Trees

The bottleneck in fitting a decision stump with binary features is the loop over the features. If we have D features, we pass through this loop D times. Inside this loop, the bottlenecks are computing operations that involve going through all N examples, and applying a simple $O(1)$ operation to each example. Thus, the total cost of fitting a decision stump is $O(ND)$, which is the size of the dataset. This indicates that fitting decision stumps is very fast.[2] What is the cost of fitting a decision tree of depth M in terms of N, D, and M? (Hint: even thought there could be 2^{M-1} decision stumps, keep in mind not every stump will need to go through every example. Note also that we stop growing the decision tree if a node has no examples, so we may not even need to do anything for many of the 2^{M-1} decision stumps.)

3.3 Training Error vs. Testing Error

The function *example_trainTest* shows how to evaluate the training and testing error of a decision tree fit with the information-gain criterion on a dataset with 10 features. Modify this demo to make a plot with the depth of the decision tree on the x-axis (varying it from 1 through 15) and the error on the training data $\{X, y\}$ on the y-axis. Make the same plot, but instead of the training error (on $\{X, y\}$ plot the testing error on $\{Xtest, ytest\}$).

3.4 Cross-Validation

On the 10-feature dataset from Question 3.3 (DTdata.mat), compute the 2-fold cross-validation scores on the training data alone (using decisionTree_infoGain). To split the data, use examples 1 to 2500 as the first fold and examples 2501 through 5000 as the second fold. Report the cross-validation error for all depths

[2] If you aren't familiar with big-O notation for analyzing algorithms, some useful links are: `https://rob-bell.net/2009/06/a-beginners-guide-to-big-o-notation` and `https://www.interviewcake.com/article/big-o-notation-time-and-space-complexity`

1 through 15 (averaging over the error for the two folds), and report the depth that would be chosen by cross-validation.

4 Probability Exercises

Please read the *Notes on Probability* on the course webpage, if you need a refresher on probabilities. Use probabilistic arguments to address the following problems. Show your calculations in addition to giving the final result. The last problem comes from Chapter 2 of Murphy's Machine Learning book.

4.1 Bayes rule for drug testing

Suppose a drug test produces a positive result with probability 0.99 for drug users, $P(T = 1|D = 1) = 0.99$. It also produces a negative result with probability 0.99 for non-drug users, $P(T = 0|D = 0) = 0.99$. The probability that a random person uses the drug is 0.001, so $P(D = 1) = 0.001$.

What is the probability that a random person who tests positive is a user, $P(D = 1|T = 1)$?

4.2 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?

4.3 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.