CPSC 340 Assignment 1 (due Friday September 12 at 11:55pm)

Commentary on Assignment 1: CPSC 340 is tough because it combines knowledge and skills across several disciplines. To succeed in the course, you will need to know or very quickly get up to speed on:

- Math to the level of the course prerequisites: linear algebra, multivariable calculus, some probability.
- Basic Julia programming, and the ability to translate from math to programming and back.
- Statistics, algorithms, and data structures to the level of the course prerequisites.
- Some basic LaTeX skills so that you can typeset equations and submit your assignments.

The purpose of this assignment is to make sure you are prepared for this course. We anticipate that each of you will have different strengths and weaknesses, so don't be worried if you struggle with *some* aspects of the assignment. But if you find this assignment to be very difficult overall, that is an early warning sign that you may not be prepared to take CPSC 340 at this time. Future assignments will be more difficult than this one (and probably around the same length).

Questions 1-4 are on review material, that we expect you to know coming into the course. The rest is related to the first few lectures.

IMPORTANT!!!!! Before proceeding, please carefully read the homework instructions:
`www.cs.ubc.ca/~schmidtm/Courses/340-F19/assignments.pdf`

You may receive a 50% deduction on the assignment if you don't follow these instructions.

We use blue to highlight the deliverables that you must answer/do/submit with the assignment.

You may also want to read the answers to this Quora question as motivation:
`https://www.quora.com/Why-should-one-learn-machine-learning-from-scratch-rather-than-just-learning-to-use-the-available-libraries`

Basic Information

1. Name:

   Answer: Zijia Zhang

2. Student ID:

   Answer: 42252965

1 Linear Algebra Review

For these questions you may find it helpful to review these notes on linear algebra:
`http://www.cs.ubc.ca/~schmidtm/Documents/2009_Notes_LinearAlgebra.pdf`

1.1 Basic Operations

Use the definitions below,

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

and use x_i to denote element i of vector x. Evaluate the following expresions (you do not need to show your work).

1. $\sum_{i=1}^{n} x_i y_i$ (inner product).

   Answer: 14

2. $\sum_{i=1}^{n} x_i z_i$ (inner product between orthogonal vectors).

   Answer: 0

3. $\alpha(x + y)$ (vector addition and scalar multiplication).

   Answer: $\begin{bmatrix} 6 \\ 10 \\ 14 \end{bmatrix}$

4. $\|x\|$ (Euclidean norm of x).

   Answer: $\sqrt{5}$

5. x^T (vector tranpose).

   Answer: $\begin{bmatrix} 0 & 1 & 2 \end{bmatrix}$

6. Ax (matrix-vector multiplication).

   Answer: $\begin{bmatrix} 6 \\ 5 \\ 7 \end{bmatrix}$

7. $x^T Ax$ (quadratic form).

   Answer: 19

1.2 Matrix Algebra Rules

Assume that $\{x, y, z\}$ are $n \times 1$ column vectors and $\{A, B, C\}$ are $n \times n$ real-valued matrices, 0 is the zero matrix of appropriate size, and I is the identity matrix of appropriate size. State whether each of the below is true in general (you do not need to show your work).

1. $x^T y = \sum_{i=1}^{n} x_i y_i$.

   Answer: true

2. $x^T x = \|x\|^2$.

   Answer: true

3. $x^T x = xx^T$.

   Answer: false

4. $(x - y)^T (y - x) = \|x\|^2 - 2x^\top y + \|y\|^2$.

   Answer: false

5. $AB = BA$.

   Answer: false

6. $A(B + C) = AB + AC$.

   Answer: true

7. $(AB)^T = A^T B^T$.

   Answer: true

8. $x^T Ay = y^T A^T x$.

   Answer: false

9. $A^\top A = I$ is the columns of A are orthonormal.

   Answer: true

2 Probability Review

For these questions you may find it helpful to review these notes on probability:
`http://www.cs.ubc.ca/~schmidtm/Courses/Notes/probability.pdf`
And here are some slides giving visual representations of the ideas as well as some simple examples:
`http://www.cs.ubc.ca/~schmidtm/Courses/Notes/probabilitySlides.pdf`

2.1 Rules of probability

Answer the following questions. You do not need to show your work.

1. You are offered the opportunity to play the following game: your opponent rolls 2 regular 6-sided dice. If the difference between the two rolls is at least 3, you win \$15. Otherwise, you get nothing. What is a fair price for a ticket to play this game once? In other words, what is the expected value of playing the game?

   Answer: \$5

2. Consider two events A and B such that $\Pr(A, B) = 0$ (they are mutually exclusive). If $\Pr(A) = 0.4$ and $\Pr(A \cup B) = 0.95$, what is $\Pr(B)$? Note: $p(A, B)$ means "probability of A and B" while $p(A \cup B)$ means "probability of A or B". It may be helpful to draw a Venn diagram.

   Answer: 0.55

3. Instead of assuming that A and B are mutually exclusive ($\Pr(A, B) = 0$), what is the answer to the previous question if we assume that A and B are independent?

   Answer: 0.917

2.2 Bayes Rule and Conditional Probability

Answer the following questions. You do not need to show your work.

Suppose a drug test produces a positive result with probability 0.95 for drug users, $P(T = 1|D = 1) = 0.95$. 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 0.0001, so $P(D = 1) = 0.0001$.

1. What is the probability that a random person would test positive, $P(T = 1)$?

Answer:

$$\begin{aligned} P(T=1) &= P(T=1|D=1)P(D=1) + P(T=1|D=0)P(D=0) \\ &= 0.95 * 0.0001 + 0.01 * 0.9999 \\ &= 0.000095 + 0.009999 \\ &= 0.010094 \end{aligned}$$

2. In the above, do most of these positive tests come from true positives or from false positives?

   Answer: from false positives

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

   Answer: $P(D=1|T=1) = \frac{0.95 \times P(D=1)}{P(T=1)} = 0.0094115316$

4. Suppose you have given this test to a random person and it came back positive, are they likely to be a drug user?

   Answer: no

5. What is one factor you could change to make this a more useful test?

   Answer: Increase $P(T=0|D=0)$.

3 Calculus Review

For these questions you may find it helpful to review these notes on calculus:
`http://www.cs.ubc.ca/~schmidtm/Courses/Notes/calculus.pdf`

3.1 One-variable derivatives

Answer the following questions. You do not need to show your work.

1. Find the derivative of the function $f(x) = 3x^2 - 2x + 5$.

   Answer: $f'(x) = 6x - 2$

2. Find the derivative of the function $f(x) = x(1-x)$.

   Answer: $f'(x) = -2x + 1$

3. Let $p(x) = \frac{1}{1+\exp(-x)}$ for $x \in \mathbb{R}$. Compute the derivative of the function $f(x) = x - \log(p(x))$ and simplify it by using the function $p(x)$.

   Answer: $f'(x) = 1 - \frac{1}{p(x)}p'(x) = 1 - \frac{e^{-x}}{1+\exp(-x)} = p(x)$

Note that in this course we will use $\log(x)$ to mean the "natural" logarithm of x, so that $\log(\exp(1)) = 1$. Also, obseve that $p(x) = 1 - p(-x)$ for the final part.

3.2 Multi-variable derivatives

Compute the gradient $\nabla f(x)$ of each of the following functions. You do not need to show your work.

1. $f(x) = x_1^2 + \exp(x_2)$ where $x \in \mathbb{R}^2$.

   Answer: $\nabla f(x) =< 2x_1, \exp(x_2) >$

2. $f(x) = \exp(x_1 + x_2 x_3)$ where $x \in \mathbb{R}^3$.

   Answer: $\nabla f(x) =< \exp(x_1 + x_2 x_3), x_3 \exp(x_1 + x_2 x_3), x_2 \exp(x_1 + x_2 x_3) >$

3. $f(x) = a^T x$ where $x \in \mathbb{R}^2$ and $a \in \mathbb{R}^2$.

   Answer: $f(x) = a_1 x_1 + a_2 x_2$
   $\nabla f(x) =< a_1, a_2 >$

4. $f(x) = x^\top A x$ where $A = \begin{bmatrix} 2 & -1 \\ -1 & 1 \end{bmatrix}$ and $x \in \mathbb{R}^2$.

   Answer: $f(x) = [2x_1 - x_2, -x_1 + x_2] \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = 2x_1^2 - x_2 x_1 + -x_1 x_2 + x_2^2 = 2x_1^2 - 2x_1 x_2 + x_2^2$
   $\nabla f(x) =< 4x_1 - 2x_2, 2x_2 - 2x_1 >$

5. $f(x) = \frac{1}{2}\|x\|^2$ where $x \in \mathbb{R}^d$.

   Answer: $f(x) = \frac{1}{2}(x_1^2 + x_2^2 + ...)$
   $\nabla f(x) =< x_1, x_2, x_3... >= x$

Hint: it is helpful to write out the linear algebra expressions in terms of summations.

3.3 Derivatives of code

For these questions you may find it helpful to review this list of useful Julia commands:
`http://www.cs.ubc.ca/~schmidtm/Courses/Notes/juliaCommands.txt`

The zip file `a1.zip` contains a Julia file named `grads.jl` which defines several functions that take in a vector as input. Complete the functions `grad1`, `grad2`, and `grad3` (which compute the gradients of `func1`, `func2`, and `func3`). Include the code in the PDF file for this section, and also in your zip file.

Hint: for many people it's easiest to first understand on paper what the code is doing, then compute the gradient, and then translate this gradient back into code. We've given you `func0` and `grad0` as an example. Also, we've provided the function `numGrad` which approximates the gradient numerically to help you debug. Below is an example of using these functions:

Note: do not worry about the distinction between row vectors and column vectors here. For example, if the correct answer is a vector of length 5, we'll accept vectors of size 5×1 or 1×5. In future assignments we will be more careful to always use column vectors.

```
### Function 1

function func1(x)
    f = 0;
    for x_i in x
        f += x_i^3;
    end
    return f
end

function grad1(x)
    n = length(x);
    g = zeros(n);
    for i in 1:n
        g[i] = 3*x[i]^2;
    end
    return g
end

### Function 2
func2(x) = prod(x)

function grad2(x)
    n = length(x);
    g = zeros(n);
    for i in 1:n
        g[i] = prod(x)/x[i];
    end
    return g
end

### Function 3
func3(x) = -sum(log.(1 .+ exp.(-x)))

function grad3(x)
    n = length(x);
    g = zeros(n);
    for i in 1:n
        g[i] = -(-exp(-x[i])/(1+exp(-x[i])));
    end
    return g
end
```

4 Algorithms and Data Structures Review

For these questions you may find it helpful to review these notes on big-O notation:
`http://www.cs.ubc.ca/~schmidtm/Courses/Notes/bigO.pdf`

4.1 Trees

Answer the following questions You do not need to show your work.

1. What is the maximum number of *leaves* you could have in a binary tree of depth l?

   Answer: 2^l

2. What is the maximum number of *internal nodes* (excluding leaves) you could have in a binary tree of depth l?

   Answer: $2^l - 1$

Note: we'll use the standard convention that the leaves are not included in the depth, so a tree with depth 1 has 3 nodes with 2 leaves.

4.2 Common Runtimes

Answer the following questions using big-O notation You do not need to show your work.

1. What is the cost of running the mergesort algorithm to sort a list of n numbers?

Answer: $O(n \log n)$

2. What is the cost of finding the third-largest element of an unsorted list of n numbers?

   Answer: $O(n)$

3. What is the cost of finding the smallest element greater than 0 in a *sorted* list with n numbers?

   Answer: $O(\log n)$

4. What is the cost of finding the value associated with a key in a hash table with n numbers? (Assume the values and keys are both scalars.)

   Answer: $O(1)$

5. What is the cost of computing the matrix-vector product Ax when A is $n \times d$ and x is $d \times 1$?

   Answer: $O(nd)$

6. What is the cost of computing the quadratic form $x^T Ax$ when A is $d \times d$ and x is $d \times 1$?

   Answer: $O(d^2)$

7. How does the answer to the previous question change if A has only z non-zeroes? (You can assume $z \geq d$)

   Answer: $O(z)$

4.3 Running times of code

Included in `a1.zip` is file named `bigO.jl`, which defines several functions that take an integer argument n. For each function, state the running time as a function of n, using big-O notation.

Answer: func1: $O(n)$

Answer: func2: $O(1)$

Answer: func3: $O(n)$

Answer: func4: $O(n^2)$

5 Summary Statistics and Data Visualization

The file *a1.zip* contains estimates of the influenza-like illness percentage over 52 weeks on 2005-06 by Google Flu Trends in a comma-separated values (CSV) file. You can open this with Excel or other spreadsheet programs; the first row gives the abbreviation of the region names for each column, and each row gives the estimate for a week. After you change to the a0 directory, you can load this data in Julia using:

```
using DelimitedFiles
dataTable = readdlm("fluTrends.csv",',')
```

This creates an two-dimensional array of type "Any" populated with all the information in the CSV file.

5.1 Summary Statistics

Report the following statistics: the minimum, maximum, mean, median, and mode of all values across the dataset. In light of thea above, is the mode a reliable estimate of the most "common" value? Describe another way we could give a meaningful "mode" measurement for this (continuous) data.

Answer: Min value = 0.352
Max value = 4.862

Mean = 1.3246249999999984
Median = 1.1589999999999998
Mode = 0.77
A more meaningful Mode can be obtained by setting intervals and count the most common intervals that values lands.

Hint: Since the first row of the CSV file is just the names of the columns, we can create a matrix X containing the data stored as real numbers using:

```
X = real(dataTable[2:end,:])
```

You can make Julia dispaly the matrix X using

```
@show X
```

The *show* macro can be used to display the result of any expression, like showing the tenth row of X:

```
@show X[10,:]
```

Note that this can be run inside functions, so it's helpful for debugging.

Julia has a mean and median function available, if you include the Statistics package. This package does not have a mode command, so I've included one in 'misc.jl'.

5.2 Data Visualization

Consider the figure on the next page. The figure contains the following plots, in a shuffled order:

1. A histogram showing the distribution of all values in the matrix X.
2. A boxplot grouping data by weeks, showing the distribution across regions for each week.
3. A scatterplot between the two regions with highest correlation.
4. A single histogram showing the distribution of *each* column in X.
5. A scatterplot between the two regions with lowest correlation.
6. A plot containing the weeks on the x-axis and the percentages for each region on the y-axis.

Match the plots (labeled A-F) with the descriptions above (labeled 1-6), with an extremely brief (a few words is fine) explanation for each decision.

Answer: B-2 : Only boxplot.
3-F : The correlation of the two values is higher than E.
5-E : The correlation of the two values is lower than F.
6-A : The name of the Axis matches the Disctiption.
4-D : In the margin, it shows the names of each column.
1-C : Only one class, which repersents all values in X.
Hint: you can generate similar plots by adding the PyPlot package. To add this package use:

```
using Pkg # Loads the package manager
Pkg.add("PyPlot") # Only needs to be done once  (installs a Julia-callable Python build)
using PyPlot # Do this once per session
plot(1:52,X[:,1]) # Plot the first row
```

To generate similar-looking plots you can use the functions 'plot', 'boxplot', 'plt.hist', and 'scatter'.

5.3 Decision Surfaces

Consider the figure below, which plots a set of two-dimensional training examples and the decision surface produced by a "neural network" classifier (a model we'll see later in the course). How many training examples has the neural network mis-classified? (This figure is best viewed in colour.)

Answer: 17

6 Decision Trees

If you run the file *example_decisionStump.jl*, it will load a dataset containing longtitude and latitude data for 400 cities in the US, along with a class label indicating whether they were a "red" state or a "blue" state in the 2012 election.[1] Specifically, the first column of the variable X contains the longitude and the second variable contains the latitude, while the variable y is set to 1 for blue states and 2 for red states. After it loads the data, it plots the data and then fits two simple classifiers: a classifier that always predicts the most common label (1 in this case) and a decision stump that discretizes the features (by rounding to the nearest integer) and then finds the best equality-based rule (i.e., check if a feature is *equal* to some value). It reports the training error with these two classifiers, then plots the decision areas made by the decision stump. The plot should look like this:

Note that these functions use the "JLD" package for loading the data and the "PyPlot" and "PyCall" package to do the plotting. You can install these packages using:

```
using Pkg
Pkg.add("JLD")
Pkg.add("PyPlot")
Pkg.add("PyCall")
```

6.1 Equality vs. Inequality Splitting Rules

In class we discussed splitting rules based on inequalities rather than equalities. Is there a type of feature where it makes sense to use an equality-based splitting rule?

Answer: Yes, if the feature that is discrete (not continuous) like city or country, we might use Equality rules.

6.2 Decision Stump Implementation

The file *decisionstump.jl* contains a function that finds the best decision stump using the equality rule ("decisionStumpEquality"), and then returns a function that can apply this decision stump to new data. Instead of discretizing the data and using a rule based on testing an equality for a single feature, we want to check whether a feature is above a threshold and split the data accordingly (this is the more sane approach, which we discussed in class). Add a new function "decisionStump" to *decision_stump.jl* that finds the best inequality-based rule, and report the updated error you obtain by using inequalities instead of discretizing and testing equality.

Hint: you may want to start by copy/pasting the contents of of the "decisionStumpEquality" function and then make modifications from there. Note that you should remove the calls to the "round" function for the inequality case. Make sure that you maintain the same input/output format in your function, since otherwise subsequent questions will not work (it should produce a plot that divides the US into a northern blue and a southern red area). If you are new to Julia, you may also want to look at *majorityPredictor.jl* to get an idea of the syntax in a simpler case.

[1] The cities data was sampled from `http://simplemaps.com/static/demos/resources/us-cities/cities.csv`. The election information was collected from Wikipedia.

Answer:

```
function decisionStump(X,y)
    (n,d) = size(X);
    y_mode = mode(y);
    minError = sum(y .!=y_mode);
    splitVariable = [];
    splitValue = [];
    splitYes = y_mode;
    splitNo = [];
    # Search for the best rule
    # (Uses O(n^2d) approach to keep code simple)
    yhat = zeros(n)
    for j in 1:d
        # Try unique values of column as split values
        for val in unique(X[:,j])

            # Test whether each object satisfies equality
            yes = X[:,j] .> val

            # Find correct label on both sides of split
            y_yes = mode(y[yes])
            y_no = mode(y[.!yes])

            # Make predictions
            yhat[yes] .= y_yes
            yhat[.!yes] .= y_no

            # Compute error
            trainError = sum(yhat .!= y)

            # Update best rule
            if trainError < minError
                minError = trainError
                splitVariable = j
                splitValue = val
                splitYes = y_yes
                splitNo = y_no
            end
        end
    end
    # Now that we have the best rule,
    # let's build our splitting function
    function split(Xhat)
        (t,d) = size(Xhat)
        Xhat = round.(Xhat)
        if isempty(splitVariable)
            return fill(true,t)
        else
            return (Xhat[:,splitVariable] .> splitValue)
        end
    end

    # Now that we have the best rule,
    # let's build our predict function
    function predict(Xhat)
        (t,d) = size(Xhat)
        yes = split(Xhat)
        yhat = fill(splitYes,t)
        if any(.!yes)
            yhat[.!yes] .= splitNo
        end
        return yhat
    end

    return StumpModel(predict,split,isempty(splitNo))
end
```

6.3 Constructing Decision Trees

Once your *decisionStump* function is finished, the script *example_decisionTree* will be able to fit a decision tree of depth 2 to the same dataset (which results in a lower training error). Look at how the decision tree is stored and how the (recursive) *predict* function works. Using the same splits as the fitted depth-2 decision tree, write out what an alternate version of the predict function would be for classifying one training example as a simple program using if/else statements (as in the first slide of L3 that has the title "Decision Trees").

Hint: you can use use the "@show" macro to print the values of various expressions during the execution of a .jl file.

Answer:

```
function predict2(Xhat)
    if Xhat[2]>37.669007
        if Xhat[1] > -96.090109
            return 1
        else
            return 2
        end
    else
        if Xhat[1] > -115.577574
            return 2
        else
            return 1
        end
    end
end
```

6.4 Cost of Fitting Decision Trees

In class, we discussed how in general the decision stump minimizing the classification error can be found in $O(nd \log n)$ time. Using the greedy recursive splitting procedure, what is the total cost of fitting a decision tree of depth m in terms of n, d, and m?

Answer:

1. Assume $T(n) = T(n-k) + T(k) + n \log n$ for some constant k.

$$\begin{aligned} T(n) &= T(n-k) + T(k) + dn \log n \\ &= T(n-k) + c + dn \log n \\ &= T(n-2k) + 2c + d(n-k)\log(n-k) + dn \log n \\ &= ... \\ &= dmn \log n + m * c \in O(dmn \log n) \end{aligned}$$

2. $T(n) = T(\frac{n}{k}) + T(\frac{(k-1)n}{k}) + n \log n$ for some constant k.

$$
\begin{aligned}
T(n) &= T(\frac{n}{k}) + T(\frac{(k-1)n}{k}) + dn \log n \\
&= T(\frac{n}{k^2}) + 2T(\frac{(k-1)n}{k^2}) + T(\frac{(k-1)^2 n}{k^2}) + \frac{dn}{k} \log(\frac{n}{k}) + \frac{d(k-1)n}{k} \log(\frac{(k-1)n}{k}) + dnlogn \\
&\leq T(\frac{n}{k^2}) + 2T(\frac{(k-1)n}{k^2}) + T(\frac{(k-1)^2 n}{k^2}) + dn \log(\frac{(k-1)n}{k}) + dnlogn \\
&\leq \dots \\
&= dn(\log n + \log(\frac{(k-1)n}{k}) + ... + \log(\frac{(k-1)^m n}{k^m})) \\
&= dn(m \times \log n + \log(\frac{k-1}{k} \times \frac{(k-1)^2}{k^2} \times \dots \times \frac{(k-1)^m}{k^m})) \\
&= dnm \log n + dn(\log(\frac{k-1}{k} \times \frac{(k-1)^2}{k^2} \times \dots \times \frac{(k-1)^m}{k^m})) \\
&\in O(dnm \log n)
\end{aligned}
$$

Therefore, the runtime will be in $O(dmn \log n)$

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.

HAVE YOU DOUBLE CHECKED THAT YOU'RE FOLLOWING ALL THE ASSIGNMENT SUBMISSION INSTRUCTIONS???

`www.cs.ubc.ca/~schmidtm/Courses/340-F19/assignments.pdf`