CPSC 340 Assignment 0 (due Friday September 15 ATE)

Rui Yang, Student ID: 75497115

September 14, 2017

IMPORTANT!!!!! Before proceeding, please carefully read the homework instructions: `www.cs.ubc.ca/~schmidtm/courses/340-F17/assignments.pdf`.

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`

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

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. A^T (matrix transpose). Answer:

$$\begin{bmatrix} 3 & 1 & 1 \\ 2 & 3 & 1 \\ 2 & 1 & 3 \end{bmatrix}$$

7. Ax (matrix-vector multiplication).

Answer:

$$\begin{bmatrix} 6 \\ 5 \\ 7 \end{bmatrix},$$

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. State whether each of the below is true in general (you do not need to show your work).

1. $x^T y = \sum_{i=1}^{n} x_i y_i$. True
2. $x^T x = \|x\|^2$. True
3. $x^T(y + z) = z^T x + x^T y$. True
4. $x^T(y^T z) = (x^T y)^T z$. False
5. $AB = BA$. False
6. $A(B + C) = AB + AC$. True
7. $(AB)^T = A^T B^T$. False
8. $x^T Ay = y^T A^T x$. True
9. $\det A \neq 0 \iff A$ is invertible. True

1.3 Special Matrices

In one sentence, write down the defining properties of the following special types of matrices:

1. Symmetric matrix. Answer: Matrix that does not change under transpose transformation.
2. Identity matrix. Answer: A matrix with 1 on the diagonal line and 0 everywhere else. Such as:

$$I = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix},$$

3. Orthogonal matrix. Answer: If matrix A multiplies matrix B equals 0, A and B are orthogonal to each other.

2 Probability Review

2.1 Rules of probability

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

1. You flip 4 fair coins. What is the probability of observing 3 heads?

   Answer: $C_4^3(1/2)^3 \cdot (1/2) = 1/4$

2. 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 $12. 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: $1vs4, 5, 6; 2vs5, 6; 3vs6$, in total we have 12 possibilities. The total combinations are 36. Thus the odds is $12/36 = 1/3$, thus a fair price should be $1/3 \cdot 12 = 4$, 4 dollars

3. Consider two events A and B such that $\Pr(A, B) = 0$. 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: $\Pr(B) = \Pr(A \cup B) - \Pr(A) = 0.55$

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: $P(T = 1) = P(D = 0) \cdot P(T = 1|D = 0) + P(D = 1) \cdot P(T = 1|D = 1) = (1 - P(D = 1)) \cdot (1 - P(T = 0|D = 0)) + P(D = 1) \cdot P(T = 1|D = 1) = (1 - 0.0001) \cdot (1 - 0.99) + 0.0001 \cdot 0.95 = 0.009999 + 0.000095 = 0.010094$

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

   Answer: Most of the positive tests come from the false positive.

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

   Answer:

   $P(D = 1|T = 1) = \frac{P(D=1) \cdot P(T=1|D=1)}{P(T=1)} = \frac{0.0001 \cdot 0.95}{0.010094} = 0.0094115316$

4. Are your answers from part 2 and part 3 consistent with each other?

   Yes, consistent with each other. Part 2 means that most of the positive tests are false positive, thus do not panic when a person is told positive test result. Part 3 means that if a random person were tested as positive, don't panic, the chance of being real positive is very low.

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

Answer: increase ratio between true positive and false positive, thus making false positive odds much lower than true positive odds. Which can be realized by greatly improving $P(T = 0|D = 0)$.

2.3 Bayes Rule and Independence

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, that the host always opens a door that doesn't contain a prize, and that if the prize is behind the selected door that the host is equally likely to choose door 2 or 3.

Answer: the contestant should switch to door 2. Because at the first guess, there was 2/3 odds that the contestant chose the wrong door, thus the first guess ruled out a no-prize door with 2/3 probability. Then, the host ruled out another no-prize door with probability 1 by intentionally opening a no-prize door, thus, the contestant and the host ruled out 2 no-prize doors with $\frac{2}{3} \cdot 1 = \frac{2}{3}$ odds, that means the probability that the only remaining door has prize behind it is $\frac{2}{3}$, this probability is much higher than the probability of the initial guess (1/3). Similar probability values can also be calculated with Bayesian law.

3 Calculus Review

3.1 One-variable derivatives

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

1. Find the minimum value of the function $f(x) = 3x^2 - 2x + 5$ for $x \in \mathbb{R}$.

   Answer: $f(x) = 3(x - 1/3)^2 + 14/3$, thus the minimum is 14/3

2. Find the maximum value of the function $f(x) = x(1 - x)$ for x in the interval $[0, 1]$.

   Answer: $f(x) = x - x^2 = -(x^2 - x) = -((x - 1/2)^2 - 1/4) = -(x - 1/2)^2 + 1/4$, thus the maximum is 1/4

3. Find the minimum value of the function $f(x) = x(1 - x)$ for x in the interval $[0, 1]$.

   Answer: $f(x) = x - x^2 = -(x^2 - x) = -((x - 1/2)^2 - 1/4) = -(x - 1/2)^2 + 1/4$, plotting it, we get the minimum is 0

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

Remember that in this course we will $\log(x)$ 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.

Answer: $f^{'}(x) = -p^{'}(x)/p(x)$, $p^{'}(x) = -\frac{-exp(-x)}{(1+exp(-x))^2} = (1/p(x) - 1) \cdot p(x)^2$, thus $f^{'}(x) = -1 + p(x) = -p(-x)$

3.2 Multi-variable derivatives

Compute the gradient $\nabla f(x)$ of each of the following functions.

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: $\nabla f(x) = [\exp(x_1 + x_2x_3), x_3 \exp(x_1 + x_2x_3), x_2 \exp(x_1 + x_2x_3)]$

3. $f(x) = a^T x$ where $x \in \mathbb{R}^2$ and $a \in \mathbb{R}^2$. Answer: $f(x) = a_1x_1 + a_2x_2$ $\nabla f(x) = [a_1a_2] = a^\top$

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

   Answer: $f(x) = x_1^2 + (x_1 - x_2)^2$, thus $\nabla f(x) = [2x_1 + 2(x_1 - x_2), -2(x_1 - x_2)] = [4x_1 - 2x_2, -2(x_1 - x_2)]$

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

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

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

3.3 Derivatives of code

The zip file `a0.zip` contains a Julia file named `grads.jl` which defines several functions. Complete the functions `grad1`, `grad2`, and `grad3` (which compute the gradients of `func1`, `func2`, and `func3`). Include the code in 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. 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.

```
### An example to get you started:

# Function
func0(x) = sum(x.^2)

# Gradient
grad0(x) = 2x

### 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]=3x[i]^2;     #g[i]=3x[i]^2; # Put gradient code here

        end
        return g
end

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

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

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

function grad3(x)
        # Put gradient code here
return exp(-x)/( 1+exp(-x)  )   #--
end

### A function to compute the derivative numerically
function numGrad(func,x)
        n = length(x);
        delta = 1e-6;
        g = zeros(n);
        fx = func(x);
        for i = 1:n
                e_i = zeros(n);
                e_i[i] = 1;
                g[i] = (func(x[:] + delta*e_i) - fx)/delta;
        end
        return g
end
```

4 Algorithms and Data Structures Review

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: $1+2+4+8+...+2^l = 2^{l+1}-1$

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: The time cost is $O(n \log n)$, the space cost is $O(n)$
2. What is the cost of finding the third-largest element of an unsorted list of n numbers? Answer: The time cost is $O(n^3)$
3. What is the cost of finding the smallest element greater than 0 in a *sorted* list with n numbers.

   Answer: $O(1)$
4. What is the cost of computing the matrix-vector product Ax when A is $n \times d$ and x is $d \times 1$. Answer:$O(nd)$
5. How does the answer to the previous question change if A has only z non-zeroes.

   Answer: Does not change

4.3 Running times of code

Included in `a0.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.

func1(n): $O(n)$

func2(n): $O(1)$

func3(n): $O(n)$

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