CISC 5800 – Machine Learning

Homework 0

Due January 23 and 26

Submit Parts A and B on paper at the start of class January 23;
Submit Part C on your erdos account by 11:59pm January 26 (see Part C instructions below).

A. Probability:

1. Consider the following joint probability table:

A	B	C	P(A,B,C)
1	0	0	0.03
0	0	0	0.15
1	1	0	0.02
0	1	0	0.11
1	0	1	0.14
0	0	1	0.28
1	1	1	0.09
0	1	1	0.18

a) What is P(A=1, B=0, C=1)?
b) What is P(B=1)?
c) What is P(A=0, C=1)?
d) What is P(A=1 or B=0)?
e) What is P(A=1|B=0)?
f) If C=1 (consider only the last four columns of the table), is A independent of B? In other words, does P(A,B|C=1) = P(A|C=1) P(B|C=1) ?
g) If B=0, is A independent of C?

7. Consider four multi-valued random variables C (campus), G (grade), M (major), and Y (year). We know that **none of these variables are independent**. We are provided the probability tables for the following joint, marginal, and conditional probabilities.

P(Y)	P(M)	P(G,Y)
P(C|Y)	P(C,M)	P(Y|M)

For example, we are told:
P(M=compSci) = 0.3, P(M=psych) = 0.2, P(M=bio)=0.2, P(M=business)=0.1
P(G=A,Y=freshman)=0.03, P(G=B,Y=freshman)=0.12 , … P(G=F , Y=senior)=0.08
[*corrected Jan 18, 11am*]

We are **not** provided any other probability tables; for example, we are not given values for: P(G=B) or P(M=psych, Y=junior)

Explain how to combine probabilities from above to compute each probability below, or write "not possible" if it is not possible.
For example: P(Y) = $\sum_g P(Y, G = g)$
a) P(M=compSci, Y=sophomore)
b) P(G=B | C=LC, M=business)
c) P(C=RH | Y=freshman) I **meant** to write: P(Y=freshman | C=RH)
d) P(G=B, M=bio)

B. Algebra/Calculus

Express x as a function of *y.*
Example question: 3y=6x+7
Example answer: $x = \frac{3y-7}{6}$

1. $4x+3x^4= 2(8y^2+2x)$

2. $6y+3x= y^2+6$

3. $3x-4=5(y^2-x)$

Consider the function f(x)=~~$\log(4x^2-6)$~~ -> $\log(4x^2+6)$
4. What is the derivative of f(x)?

5. For what value of x is f'(x)=0?

6. At the value you found above, will f(x) have its largest possible value or its smallest possible value?

Consider the function g(x,y)= $\sum_{i=0}^{3} x^i y^{2i}$ (Note, for example, $y^4 = y \times y \times y \times y$
7. What is the value of g(x,y) when x=3, y=2?

8. What is the derivative of g(x,y) with respect to y : $\frac{d}{dy} g(x,y)$?

C. Programming:

Use Python to solve the following tasks. For this homework, the function inputs must be lists, not panda dataframes and not numpy arrays. You may use numpy or pandas **inside** the function definition if you wish. (If you are rusty on Python, you may use Matlab or C++ for this first assignment, but **not** on future assignments.)

Submission instructions for Part C: Log into your erdos account (erdos.dsm.fordham.edu) – you can use Terminal on Mac or Putty on Windows (see Resources section on our course web site). Inside your folder called "private"

Linux command: `cd private`

create a folder called "CIS5800".

Linux command: `mkdir CIS5800`

Save the three programs, inside private/CIS5800/ in the file hw0.py . As course instructor, I will be able to access your files inside private/CIS5800/. You must have the necessary files in the proper directory by January 26 at 11:59pm.

You are welcome to write your programs on your local computer (or on erdos). To transfer files from your local computer to erdos, you may use a program such as FileZilla `https://filezilla-project.org/` . **Make sure you transfer your files into your private/CIS5800/ directory!** Connect to erdos using port 22.

If you have trouble accessing erdos for this assignment, you may e-mail me and Amy your programs by January 26, 11:59pm – however, we will use erdos for code submission throughout the rest of the semester, so you must resolve your erdos troubles by the time the next homework is due!

We will consider the world-famous problem of giraffe classification, discussed in the first lecture. We will make a very simple classifier and partition the data into three sets.

1. Write a function called **`threshClassify`** that takes in a list of numbers (the heights of our animals), and a threshold value x_{thresh}. The function will return a list of 0s and 1s – a 0 for each non-giraffe input and a 1 for each giraffe input. Specifically, the function call **must look like this**:

```
threshClassify(heightList, xThresh)
```

If `heightList=[2, 5, 8, 1, 7]` and `xThresh=5`, the function will return the list `[0, 0, 1, 0, 1]` (any entry GREATER than the threshold is assigned a value of 1).

2. Write a function called **`findAccuracy`** that takes in a list of approximated class labels output by the classifier (threshClassify) and a list of true labels provided in the training set, and calculates the accuracy of the classifier as a number between 0 and 1. Specifically, the function call **must look like this:**

```
findAccuracy(classifierOutput, trueLabels)
```

If `classifierOutput=[1,1,1,0,1,0,1,1]` and `trueLabels=[1,1,0,0,0,0,1,1]`, the function will return the number `0.75` (2 out of 8 values were incorrect).

3. Write a function called **`getTraining`** that takes in a two-dimensional list with 2 rows and C columns, and returns a new two-dimensional list with the first C/3 columns . In other words, we extract the training data as the first third of the entries in the data list. Specifically, the function call **must look like this:**

```
getTraining(fullData)
```

If `fullData=[[4,5,1,2,8,3], [1,1,0,0,1,0]]` the function will return `[[4,5],[1,1]]`