CISC 5800 – Machine Learning
Homework 0
Due September 5 and 8
Submit Parts A and B on paper at the start of class September 5;
Submit Part C on your erdos account by 11:59pm September 8 (see Part C instructions below).

A. Probability:

1. Consider the following joint probability table:

<table>
<thead>
<tr>
<th>A</th>
<th>B</th>
<th>C</th>
<th>P(A,B,C)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0.07</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.12</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0.16</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0.06</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0.13</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0.24</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0.10</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0.12</td>
</tr>
</tbody>
</table>

a) What is P(A=0, B=0, C=1)?
b) What is P(C=1)?
c) What is P(A=0 or C=1)?
d) What is P(A=1, B=1)?
e) What is P(A=1|B=0)?
f) If B=1, is A independent of C? In other words, does P(A,C|B=1) = P(A|B=1) P(C|B=1)?
g) If A=0, is B independent of C?
2. Consider four multi-valued random variables $A$ (age), $H$ (height), $W$ (weight), and $S$ (salary). **We know that $S$ is independent of $A$, $H$, and $W$; $A$, $H$, and $W$ are NOT independent of one another.** We are provided the probability tables/functions for the following six joint, marginal, and conditional probabilities.

**The six probabilities:**

<table>
<thead>
<tr>
<th>Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>$P(S)$</td>
</tr>
<tr>
<td>$P(W)$</td>
</tr>
<tr>
<td>$P(H, S)$</td>
</tr>
<tr>
<td>$P(S</td>
</tr>
<tr>
<td>$P(A</td>
</tr>
<tr>
<td>$P(S</td>
</tr>
</tbody>
</table>

For example, we are told:

$P(W=$light$)=0.2$, $P(W=$medium$)=0.5$, $P(W=$heavy$)=0.3$

$P(H=$short$,$S=10K$)$=0.03$, $P(H=$medium$,$S=10K$)$=0.12$, ... $P(H=$tall$,$S=10K$)$=0.08$

We are **not** provided any other probability tables; for example, we are not given values for:

$P(H=$short$)$ or $P(W=$psych$,$S=200K$)$

**Explain how to combine the six probabilities from above (and the knowledge that $S$ is independent) to compute each probability below, or write “not possible” if it is not possible.**

For example: $P(S) = \sum_h P(H = h, S)$ (we sum over values of $P(H,S)$ which are known to get $P(S)$)

a) $P(W=$medium$)$
b) $P(H=$medium$,$W=$heavy$ | A=$old$)$
c) $P(A=$young$ | S=10K)$
d) $P(H=$tall$,$W=$light$)$
e) $P(A=$young$,$W=$medium$,$S=50K)$
f) $P(W=$heavy$,$A=$young$)$

**B. Algebra/Calculus**

Express $x$ as a function of $y$.

Example question: $3y=6x+7$

Example answer: $x = \frac{3y-7}{6}$

1. $4+x=7(y^2-x)$

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

3. $z = \sum_i 5(y - x)$
Consider the function \( f(x) = \log(2x^3 - 5x^2 - 3k^2) \)

4. What is the derivative of \( f(x) \) (the derivative with respect to \( x \), \( \frac{df}{dx} \))?

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

6. At the value you found above, will \( f(x) \) have its largest 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 \( x=1, y=2 \)?

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

C. Programming:
Use Python to solve the following tasks. For this homework, the function inputs must be either 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 website). 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 September 8 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 your programs by **September 8, 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 based on animal height and speed, and assign values for missing feature values.

You must complete questions 1-3. Question 4 is optional.

1. Write a function called `treeClassify` that takes in 2D numpy array of numbers (the heights and speeds of our animals), and three threshold values \( x_{\text{threshHeight}} \), \( x_{\text{threshSpeed1}} \), \( x_{\text{threshSpeed2}} \). 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**:
   ```python
   threshClassify(featureArray, thresholds)
   ```
   `featureArray` will be a numpy array of shape \((2,n)\) (where there are arbitrary number \(n\) animals to classify) and `thresholds` is a numpy array of shape \((3,)\) containing \([x_{\text{threshHeight}}, x_{\text{threshSpeed1}}, x_{\text{threshSpeed2}}]\) values in that order. If
   ```python
   featureArray=np.array([[2,1], [4,2], [8,12], [6,5], [3,5]])
   thresholds=np.array([5,3,9])
   ```
   the function will return the list \([0, 0, 1, 0, 1]\). *Classification is accomplished according to the decision tree:*

   ```
   If animalHeight < \( x_{\text{threshHeight}} \) then
     If animalSpeed < \( x_{\text{threshSpeed1}} \) then Not Giraffe; otherwise Giraffe
     If animalHeight >= \( x_{\text{threshHeight}} \) then
       If animalSpeed < \( x_{\text{threshSpeed2}} \) then Not Giraffe; otherwise Giraffe
   ```

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**:
   ```python
   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 fillBlanks that takes in a featureArray (2,n) numpy array and returns a (2,n) numpy array where all missing values are filled in with the number 5. “Missing values” are any entry that is 0.

Specifically, the function call must look like this:

```python
featArrayCorrected=fillBlanks( featArray)
```

If `featArray=np.array([[0,5],[0,3],[1,8],[10,0]])` the function will return `np.array([[5,5],[5,3],[1,8],[10,5]])` into `featArrayCorrected`

Optional Question 4: write fillBlanksAve where 0 is replaced by the AVERAGE value of all the non-zero numbers in the same row. If `featArray=np.array([[0,5],[0,3],[1,8],[10,0]])` the function fillBlanksAve will return `np.array([[5.5,5],[5.5,3],[1,8],[10,5.33]])` into the output `featArrayCorrected`

Part C will be graded out of 50 points. If you skip question 4 or do poorly on question 4, Questions 1-3 will be worth 50 points and I will ignore question 4. If you do well on question 4 and do poorly on an earlier question in Part C, the points you earn for question 4 will raise your grade for Part C.