Homework 0
DATA 440

Pablo Rivas

Assigned: 1/23/19; Due: The following week; Points: 22

1 Instructions

This assignment should be written in LaTeX. Please check the template that has been provided to you in `hw.template.tar.gz` that contains the basic elements for you to get started. You must show your work and how you derivated or arrived at your solution. Write in understandable, easy to follow English.

Write all your programs in Python and make sure that your programs are included in your assignment. Also include any and all figures and graphs produced using Python in your writeup.

Your assignment should be submitted in two ways: through GitHub, and in hardcopy (in class). You must create a **single** repository, that will be organized in folders named "`lastname-xx`", where lastname is your last name (or the first word of your last name if you have multiple words in it), and xx is the number of the assignment. For example my GitHub folder for this assignment would be `rivas-00` and I would put all my sources there. Writeup: latex file (.tex), main pdf file (.pdf), and any graphics (.pdf or .eps). Code: Python files (.py).

2 Course Setup

2.1 Python Requirements

Python is the language we will use in this course. Python is easy to use and has most of the tools we will use in class already implemented. If you are new to python please make sure you have it installed (most computers do). Here is a good resource about its installation on different platforms:

`https://wiki.python.org/moin/BeginnersGuide/Download`

Once you are sure that Python is installed, make sure you have the latest versions of the following packages installed in your system:

- NumPy
- SciPy
- scikit-learn
- matplotlib
- Pandas

To test your installation, and to begin your understanding of Python for machine learning, try running some of the examples provided here:

`http://scikit-learn.org/stable/auto_examples/index.html`

To Deliver: Submit proof that you have python, and all the above packages installed. Simple `import` statements should suffice. You may use this code if you want:

```
import sys
import numpy
import scipy
import sklearn
import matplotlib
import pandas

print sys.version                #prints python version
print numpy.__version__          #prints numpy's version
print scipy.__version__          #prints scipy's version
print sklearn.__version__        #prints sklearn's version
print matplotlib.__version__ #prints matplotlib's version
print pandas.__version__         #prints pandas' version
```

Submit the output of this code from your computer as proof.

2.2 GitHub Class Repository

As stated earlier, this class requires you to have a **single** GitHub repository. Just one. Your repository will be organized in folders corresponding to the assignments. See the instructions above for the naming conventions of such folders. You will place your sources and files for your assignments there and you will receive your grade and feedback there.

If you do not have a GitHub account, it is important that you get your account right away. Usually, GitHub charges seven dollars, but they have a student discount. Using your Marist email (.edu) you should get a student account here:

`https://education.github.com/pack`

Once you have your account, create a repository for the course. Create a test folder with something in it (GitHub doesn't sync empty folders) and sync it between your computer and the GitHub "cloud".

Now, this is very important. You must add `pablorp80` as a contributor, otherwise, you will not receive a grade in the assignment. Here is how to do it:

`http://stackoverflow.com/a/7920363/4992019`

Now, lets make something clear. GitHub is a "cloud" service for source control. The actual source control program is simply called `git`. Therefore, in order to be GitHub savvy, you must learn the basics of git. This is a great tutorial for you to learn the basics of git:

`https://try.github.io/levels/1/challenges/1`

To Deliver: Submit your GitHub username. Submit the link to your class repository. Submit proof that you have added `pablorp80` as a collaborator.

2.3 Kaggle Account

In this class we will also compete with our machine learning models against others. Sometimes for a prize, and sometimes for knowledge. Please create an account in Kaggle using your Marist (.edu) email address because we will have both in-class and outside-class competitions. Register here:

`https://www.kaggle.com/account/register`

To Deliver: Submit your Kaggle username. Submit proof that you have created your account.

3 Mathematics Requirements

Artificial intelligence and machine learning rely heavily on mathematics. It involves several topcis such as linear algebra, probability, statistics, optimization, and multivariate calculus.

3.1 Linear Algebra

Since data is often expressed as vectors or matrices, we will need to be good in linear algebra. Look here for a good refresher:

`http://www.ee.ic.ac.uk/hp/staff/dmb/matrix/intro.html`

The most important sections of the website, for our course, would be the following two:

`http://www.ee.ic.ac.uk/hp/staff/dmb/matrix/property.html`

`http://www.ee.ic.ac.uk/hp/staff/dmb/matrix/eigen.html`

3.2 Probability and Statistics

We also need to understand probability and statistics because matrices or vectors with data have specific statistical properties. Often times we want to treat them as random variables. Make sure you understand what probability density functions are, their different families, and the operations we can do with them. Here is good resource:

`http://web.stanford.edu/class/msande223/handouts/psrev.pdf`

3.3 Calculus and Optimization

Calculus is required a lot, especially when we try to optimize values of functions. We would want to take derivatives to find maximum or minimum values of functions. But we rarely will do integrals, unless we are trying to prove things in probability density functions. Make sure you understand the basic of derivative calculus.

Here is a good resource:

`http://www.stat.wisc.edu/~ifischer/calculus.pdf`

4 Problems

Before you commit to this class, make sure you can easily solve the following problems. You will submit the solutions to this problems for credit, but you will not receive a lot of feedback on this assignment since it is just a refresher of what you already should know.

Write your derivations and reasoning in LaTeX. User Python with Numpy or Scipy to verify your solutions to problems marked with a $\star$.

1. For the function $g(x) = -3x^2 + 24x - 30$, find the value for x that maximizes $g(x)$.

2. Consider the following function:

$$f(x) = 3x_0^3 - 2x_0x_1^2 + 4x_1 - 8 \tag{1}$$

what are the partial derivatives of $f(x)$ with respect to x_0 and x_1.

3. $\star$ Consider the matrix $A = \begin{bmatrix} 1 & 4 & -3 \\ 2 & -1 & 3 \end{bmatrix}$ and $B = \begin{bmatrix} -2 & 0 & 5 \\ 0 & -1 & 4 \end{bmatrix}$, then answer the following and verify your answers in Python:

(a) can you multiply the two matrices? ellaborate on your answer

(b) multiply A^T and B and give its *rank*.

(c) (extra credit) let $C = \begin{bmatrix} 1 & 0 \\ 0 & 2 \end{bmatrix}$ be a new matrix; what is the result of $AB^T + C^{-1}$?

4. Give the mathematical definitions of the simple Gaussian, multivariate Gaussian, Bernoulli, binomial, and exponential distributions.

5. (extra credit) What is the relationship between the Bernoulli and binomial distributions?

6. Suppose that random variable $X \sim N(2, 3)$. What is its expected value?

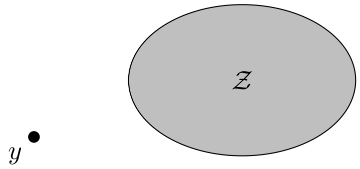
7. An *euclidean projection* of a d-dimensional point $y \in \mathbb{R}^d$ to a set $\mathcal{Z}$ is given by the following optimization problem:

$$x^* = \arg\min_x ||x - y||_2^2, \quad \text{subject to: } x \in \mathcal{Z} \tag{2}$$

where $\mathcal{Z}$ is the feasible set, $|| \cdot ||_2$ is the ℓ_2-norm (euclidean) of a vector, and $x^* \in \mathbb{R}^d$ is the projected vector.

(a) What is x^* if $y = 1.1$ and $\mathcal{Z} = \mathbb{N}$, where $\mathbb{N}$ is the set of natural numbers?

(b) Locate x^* in the following picture:



8. Suppose that random variable Y has distribution:

$$p(Y = y) = \begin{cases} e^{-y} & \text{if } y \geq 0 \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

(a) Verify that $\int_{y=-\infty}^{\infty} p(Y = y) = 1$

(b) What is $\mu_Y = E[Y] = \int_{y=-\infty}^{\infty} p(Y = y) y \, dy$; that is, the expected value of Y?

(c) What is $\sigma^2 = \text{Var}[Y] = \int_{y=-\infty}^{\infty} p(Y = y)(y - \mu_Y)^2 \, dy$; that is, the variance of Y?

(d) What is $E[Y | Y \geq 10]$; that is, the expected value of Y, given that (or conditioned on) $Y \geq 10$?