Final Exam
DATA 440

Pablo Rivas

Assigned: May/8/19; Due: May/15/19 by 5pm; Points: 100

1 Instructions

This test could be written in LaTeX, just as all homework assignments. Write in understandable, easy to follow English. Make sure you provide good illustrations and figures. Remember to include all your Python programs in your GitHub repository.

Your test should be submitted in two ways: through GitHub, and in hard copy (slide it under my office's door before the deadline). Use the **same** repository you have been using and submit your work in a folder named "`lastname-final`", where lastname is your last name.

2 Problem Set

The following is a list of problems you will work on. When providing your solutions (hopefully using LaTeX), do not simply give the final answer, show how you arrived to the solution, justify your assumptions, and explain your results clearly.

1. **Support Vector Machines**

   (a) Download the python program `final.SVM.sinc.py` which implements a 10-fold cross-validation approach to find the best set of hyper-parameters C, ϵ (epsilon), and γ (gamma), in an Support Vector Machine for Regression (SVR).

   (b) Download the python program `finalGenData.py` which generates data points of the "sinc" function contaminated with random noise.

   (c) Modify the program in 1.(a) to run for 1,000 samples, and then report the best set of hyper-parameters found. Go here `https://goo.gl/forms/X1maf8zSIjPhoyNA2` and report your results. You can do it as many times as you want, but at least one is required.

   (d) **Explain** your results. How do you interpret the hyper-parameters? Is there a great penalty? Is there room for errors without penalty? Is there a relationship between C and ϵ? Any other thoughts?

   (e) (**Extra credit** +10) Repeat 1.(c)-(d) but for 10,000 samples.

2. **More Support Vector Machines**. For this part you will use the digits dataset. For reading the dataset, you should use the following Python code in a file named `finalGetDigits.py`:

```
import numpy as np
from numpy import genfromtxt

def getDataSet():
  # read digits data & split it into X and y for training and testing
  dataset = genfromtxt('features.csv', delimiter=' ')
  y = dataset[:, 0]
  X = dataset[:, 1:]

  dataset = genfromtxt('features-t.csv', delimiter=' ')
  y_te = dataset[:, 0]
  X_te = dataset[:, 1:]
  return X, y, X_te, y_te
```

Once saved do the following:

(a) Download the python program `final.SVM.dig.py` which implements a 10-fold cross-validation approach to find the best set of hyper-parameters C, ϵ (epsilon), and γ (gamma), in an Support Vector Machine for Regression (SVR) in the digits dataset.

(b) Run the program in 2.(a) and modify it if you need to, to find and report the best set of hyper-parameters and the final validation score. Save the plot and interpret it, give your comments about it.

(c) **Explain** your results. How do you interpret the hyper-parameters? Is there a great penalty? Is there room for errors without penalty? Is there a relationship between C and ϵ? Any other thoughts?

(d) (**Extra credit** +30) Prepare and share a single Colaboratory for both 1.(c)-(d) and 2.(b)-(c). You should use proper headings and comments and explanations. *Pro Tip*: you may have to combine all files into a single program or upload files on demand.