Homework 4
MSCS 550 | CMPT 404

Pablo Rivas

Assigned: Oct/26/16;   Due: Nov/1/16;   Points: 90

1 Instructions

This assignment could be written in LaTeX, just as the last homework assignment. Write in understandable, easy to follow English. Make sure you provide good illustrations and figures; **however, do not just throw in figures withouth a proper explanation and discussion. Figures are meant to prove a point you are making verbally; figures are a resource and not the main point.** Remember to include your Python programs in your assignment (**in GitHub only please**).

Your assignment should be submitted in two ways: through GitHub, and in hardcopy (in class). Use the **same** repository you have been using and submit your work in a folder named "`lastname-xx`", where lastname is your last name xx is the number of the assignment.

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. Use `sklearn`'s implementation of k-Nearest Neighbors for regression purposes, which is found in `sklearn.neighbors.KNeighborsRegressor`. You will find the best value of k using 10-fold Cross-Validation (CV), which is found in `sklearn.model_selection.KFold`.

   (a) You will modify the python code below to generate 1000 data points, or alternatively you could use part of your semester project dataset if it is related to regression.

   ```
   import numpy as np
   from matplotlib import pyplot as plt

   def genDataSet(N):
       x = np.random.normal(0, 1, N)
       ytrue = (np.cos(x) + 2) / (np.cos(x * 1.4) + 2)
       noise = np.random.normal(0, 0.2, N)
       y = ytrue + noise
       return x, y, ytrue

   x, y, ytrue = genDataSet(100)
   plt.plot(x,y,'.')
   plt.plot(x,ytrue,'rx')
   plt.show()
   ```

   (b) Using 10-fold CV, you will report the three best values of k-neighbors that yield the best CV E_{out}. You will vary the values of k in the following range: $k = 1, 3, 5, \ldots, 2\lfloor \frac{N+1}{2} \rfloor - 1$.

(c) You will report the best CV E_{out}.

2. (graduate students) Using the same dataset you just tried in the previous problem, repeat the experiment 100 times storing the best three k number of neighbors in every single trial, and at the end of all trials plot a histogram of all the values of k that you saved.