Assignment 1

Advanced Topics in High-Performance Computing (MCSC 6230G/7230G)

Faisal Qureshi

http://faculty.uoit.ca/qureshi

Computer Science, Faculty of Science

University of Ontario Institute of Technology

Due Back Oct. 11, 11:59 pm

Part 1 - Linear Regression

Your goal is to set up a linear regression model for predicting house prices using housing data.

Data Whitening

To achieve this you'll first need to *whiten* the data. This is also sometimes referred to as *standardizing* the data. This is needed to make different input attributes comparable. Say one attribute is the number of rooms in the house, and a second attribute is the its area in square feet. In this case the value for the first attribute will vary between, say, 1 and 10; where as, the value for the second attribute may vary between 1500 and 5000 square feet.

Ridge Regression

You are asked to construct the model using ridge regression to predict the house prices. Specifically, you should complete a function as shown below that estimates model parameters θ for different values of δs.

```
def estimate_theta(X, y, d2):
    ??
    return theta
```

Hints

- You should treat bias differently. The regularization δ^2 shouldn't affect bias.
- A linear regression model can do more than fit straight lines (planes).

Tasks

- Plot the values of θ in the y-axis against values for δ^2 in the x axis. You will have to compute θ for different values of δ^2 to get this plot. This set of plotted value are referred to as *regularization paths*. **(Plot A)**
- Compute test and train error for different values of θ. Plot the test and train error for different values of δ^2. **(Plot B)**
- Choose the *best* value of δ^2 using cross-validation.

Dataset

For this exercise, you will use the California Housing dataset. You can load this dataset as follows:

```
import numpy as np
import sklearn.datasets as ds
dataset = ds.fetch_california_housing()
```

- `dataset.data`: 8 feature values ("MedInc", "HouseAge", "AveRooms", "AveBedrms", "Population", "AveOccup", "Latitude", "Longitude")
- `dataset.target`: Average house value in units of 100,000.
- `dataset.feature_names`: Array of ordered feature names used in the dataset.

This dataset consists of 20640 samples.

Submission

Submit a 2 to 4 page report outlining your findings. The report should include the following:

- Plot A
- Plot B
- Python code for `estimate_theta(X, y, d2)`
- Your strategy for cross-validation for selecting the best model parameters
- Rationale
  - Which model did you choose and why? Is this the best model? Did you try any other models.
- Conclusions
  - What features did you find have the most (or least) impact on the housing prices

Grades will reflect the clarity of your report, the thoughtfulness that you have put into your experiments, and of course the correctness of your experiments.

Submit via Blackboard.