15.095, Homework 2

Due date: October 10, 2018, 4pm

1. **Algorithmic Framework for Linear Regression**

   Using JuMP, implement an algorithm to build a linear regression model that

   (a) ensures robustness (using a Lasso penalty);

   (b) restricts sparsity;

   (c) avoids pairwise collinearity greater than 0.9;

   (d) allows the nonlinear transformations x^2, $\log x$, and $\sqrt{x}$; and

   Run your algorithm on the data sets `data1.csv` and `data2.csv` to find the best regression model you can. You will need to tune all of the hyperparameters in the algorithm to achieve the best results. In each file, $\mathbf{y}$ is the last column and $\mathbf{X}$ is everything except for the last column. For your final model, please report:

   - variables with non-zero coefficients,
   - hyperparameter values, and
   - training, validation, testing R^2 values..

2. **Convex Regression**

   Part A:

   Boyd and Vandenberghe showed that Convex Regression may be formulated as the following quadratic optimization problem:

$$\begin{aligned} \min_{\boldsymbol{\theta}, \{\boldsymbol{\xi}_i\}_{i=1}^n} \quad & \sum_{i=1}^n (y_i - \theta_i)^2 \\ \text{subject to} \quad & \theta_i + \boldsymbol{\xi}_i^T(\mathbf{x}_j - \mathbf{x}_i) \leq \theta_j \quad \forall i, j, \\ & \boldsymbol{\theta} \in \mathbb{R}^n, \\ & \boldsymbol{\xi}_i \in \mathbb{R}^p \qquad\qquad\qquad \forall i, \end{aligned}$$

   where $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$ is the data for the regression problem with n observations and p features.

(a) Using JuMP, implement a function to solve this problem using the delayed constraint generation approach outlined in lecture. For computational purposes, assume that each constraint may be violated by at most $tol = 0.1$. The function should take in as input $(\mathbf{X}, \mathbf{y}, \lambda, tol)$, and output an optimal value for $(\boldsymbol{\theta}, \boldsymbol{\xi})$. This function should also print out the number of iterations and constraints added.

(b) Generate some random data $\mathbf{X} \in \mathbb{R}^{25\times 5}$, $\mathbf{y} \in \mathbb{R}^{25}$ and solve the convex regression problem for $\lambda = 10$ and $tol = 0.1$. How does your solution scale as we increase the number of observations? How about if we increase the number of features? Compare the number of lazy constraints added to the number of total constraints in this optimization problem. What do you observe?

(c) Write a second function in Julia to obtain the model predictions for a new data matrix $\mathbf{X}_2$ given as input the optimal values for $(\boldsymbol{\theta}, \boldsymbol{\xi})$ and the original data matrix $\mathbf{X}$. Is there a difference between these model predictions on $\mathbf{X}$ and the optimal value of $\boldsymbol{\theta}$? How does this difference change as we vary tol?

Hint: Approximate the convex function value as the maximum of the subdifferentials.

(d) We say that a function f is **concave** if $-f$ is a convex function. We may also consider the **concave regression problem**, which is to find the best concave function that predicts y_i as a function of $\mathbf{x}_i$. How can we use the functions that we have already written to solve the concave regression problem and compute the model predictions?

Part B:

In this problem, we will apply our convex regression algorithm to a real-world data set. We have provided you with the data set `15095_kc_house_data.csv`, which comes from Kaggle (`https://www.kaggle.com/harlfoxem/housesalesprediction`). In this file, the last column is home price, which we would like to predict, and the rest of the columns are numeric features of the houses. For this problem, let the first 50 rows be the training set and the remaining rows be the testing set.

(a) Fit a concave regression model to predict price, using all of the features in the data set.

Report the training / testing R^2. Is this model overfitting? Describe a modification to improve the out-of-sample accuracy of this method.

Hint: Due to the large scale of the y variable, you may need to normalize the data or increase *tol* for this problem to solve quickly.

(b) Fit a new concave regression model to predict price, using only the variable "sqft_living" as a covariate. Plot the estimated house price as "sqft_living" varies from 0 to 5,000 square feet. What is your interpretation of these results?

(c) Fit concave and linear models regressing price on latitude ("lat") only, and plot the estimated house prices as latitude varies from its min to max value (47.16 to 47.78). Is there a purely monotonic relationship between latitude and predicted price? What is your interpretation of these results?

(d) Our formulation for Convex Regression is a (convex) Quadratic Optimization problem, so it is solvable in polynomial time. There is a related problem that we cover in this class, Sparse Convex Regression, which is a Mixed-Integer Quadratic Optimization problem. Typically, MIQOs are solvable in exponential time, however in this case our algorithms for Convex Regression and Sparse Convex Regression scale to similar sizes of n in the 100s. Discuss what is going on here, and why the MIQO is competitive for this problem.

Hint: Think about the number of constraints.

3. **Primal and Dual Perspective on Sparse Regression**

In this problem, we will contrast the primal and dual approaches to sparse linear regression. (The dual approach is the one covered in the end of Lecture 2's slides.) Let us consider the problem

$$\begin{aligned} \min_{\boldsymbol{\beta}} \quad & \|\mathbf{y} - \boldsymbol{\beta}\|_2^2 + \|\boldsymbol{\beta}\|_2^2 \\ \text{subject to} \quad & \|\boldsymbol{\beta}\|_0 \le k. \end{aligned}$$

In other words, this is the problem when the data matrix $\mathbf{X}$ is the identity matrix.

(a) This problem can be solved by hand. Solve it in two ways:

i. the "primal" way where you solve directly using the variable $\boldsymbol{\beta}$;

ii. and the “dual” way using the dual reformulation that eliminates the variable $\boldsymbol{\beta}$, leaving only binary variables and a convex objective.

Comment on how the two solution methods compare.

(b) Now solve this problem by applying the cutting plane method discussed in class for the dual reformulation. In particular, start with a single cutting plane (corresponding to all binary variables equal to zero) and perform at least one iteration of the cutting plane algorithm by hand. Comment on what is happening. In this case, how many cuts would you need to add in order to find the optimal solution?

Note: because the solution value obtained using the three approaches will be the same, it is important that you show your work and/or rationale in each of the approaches.