CS 534: Machine Learning
Homework 1

(Due Sep 28th at 11:59 PM on Gradescope)

Submission Instructions:

The homework must be submitted electronically on Gradescope as a single submission. Each question must be tagged appropriately (you need to set which page the problem is on), otherwise you may lose credit. The answer must also be explicit, if you are exporting an iPython notebook, you will need to add remarks to denote your final answer for the question.

The code can be done in Python, R, or MATLAB. The use of any other programming language must be confirmed with Huan (TA). Source code must be submitted separately on Canvas (the deadline is 15 minutes after the Gradescope submission) and zipped into a single file. There is no need to attach the data, and you can assume the data will be in the same directory as the code. The source code should be executable and there should be no explicit path names for the data files (`joyce/CS534/HW1/train.csv` is an example of hard-coding).

1. **(10 pts) (Faking) Ridge Regression**

   Show that the ridge regression estimates can be obtained by ordinary least squares regression on an augmented data set. We augment the centered matrix $\mathbf{X}$ with k additional rows $\sqrt{\lambda}\mathbf{I}$ and augment $\mathbf{y}$ with k zeros. The idea is that by introducing artificial data having response value zero, the fitting procedure is forced to shrink the coefficients towards zero.

2. **(3 + 3 + 10 + 5 + 10 + 4 = 35 pts) Predicting Appliance Energy Usage using Linear Regression**

   Consider the Appliances energy prediction Data set (`energydata.zip`), which contains measurements of temperature and humidity sensors from a wireless network, weather from a nearby airport station, and the recorded energy use of lighting fixtures to predict the energy consumption of appliances in a low energy house. The dataset is split into three subsets: training data from measurements up to 3/20/16 5:30, validation data from measurements between 3/20/16 5:30 and 5/7/16 4:30, and test data from measurements after 5/7/16 4:30. There are 27 attributes[1] for each 10 minute interval, which are described in detail on the UCL ML repository, Applicances energy prediction dataset.

   For this problem, you can use an existing toolbox / implementation for linear regression, ridge regression, and lasso regression.

   (a) Train a standard linear regression model *only on* the training data. What are the RMSE and R^2 on the training set, validation set, and test set?

   (b) Train a standard linear regression model using training and validation together. What are the RMSE and R^2 on the training set, validation set, and test set? How does this compare to the previous part? And what do the numbers suggest?

   (c) Train ridge regression and lasso regression *only on* the training data. You will want to consider a range of parameter values (λ) to find the optimal regularization parameter that gives you the lowest RMSE or R^2 on the *validation dataset.* Report (using a table) the RMSE and R^2 for training, validation, and test for all the different (λ) values you tried.

[1] We have removed the last two random variables as they aren't relevant for this class.

(d) Similar to part (b), train ridge and lasso using both the training and validation set (with your optimal regularization parameter from (c)). What are the RMSE and R^2 on the training set, validation set, and test set? How does this compare to the previous part? What do the numbers suggest?

(e) Generate the coefficient path plots (regularization value vs. coefficient value) for both ridge and lasso. Also, note (line or point or star) where the optimal regularization parameters are on their respective plots. Make sure that your plots encompass all the expected behavior (coefficients should shrink towards 0).

(f) What are 3 observations you can draw from looking at the coefficient path plots, and the metrics? This should be different from your observations from (b) and (d).

3. **(30 + 10 + 5 + 5 + 5 = 55 pts) Predicting Appliance Energy Usage using SGD**

Consider the Appliances energy prediction Data set from the previous problem. For this problem, you *ARE NOT* allowed to use any existing toolbox / implementation.

(a) Implement elastic net regression using stochastic gradient descent. As a reminder, the optimization problem is:

$$\min f_o(\mathbf{x}) = \frac{1}{2}||\mathbf{y} - \mathbf{X}\boldsymbol{\beta}||_2^2 + \lambda\left(\alpha||\boldsymbol{\beta}||_2 + (1-\alpha)||\boldsymbol{\beta}||_1\right), 0 \leq \alpha \leq 1$$

You will want to derive the update for a single training sample. As a hint, you will want to consider proximal gradient descent for the $||\boldsymbol{\beta}||_1$ portion of the objective function. Your implementation should allow the user to specify the regularization parameters (λ, α), the learning rate (η) and the mini-batch size ($n \in [1, N]$).

(b) For the optimal regularization parameters from ridge (λ_{ridge}) and lasso (λ_{lasso}), and $\alpha = \frac{1}{2}$, what are good learning rates for the dataset? Justify the selection by trying various learning rates and illustrating the objective value ($f_o(\mathbf{x})$) on a graph for a range of epochs (one epoch = one pass through the training data)[2]. For the chosen learning rates (ridge and lasso may have different values), what are the RMSE and R^2 for the elastic net model trained on the entire training set on the training, validation, and test sets?

(c) Using the learning rate from the previous part, train elastic net (using only training data) for different values of α (it should encompass the entire range and include $\alpha = 0, 1$). Report the RMSE and R^2 for the models on training, validation, and test set.

(d) Based on the results from (c) and 2(a) and 2(c), what conclusions can you draw in terms of RMSE and R^2? Which model is the best? Also discuss the differences between the SGD-variants of Ridge and LASSO and the standard implementations (Problem 2).

(e) What are the final coefficients that yield the best elastic net model on the test data? Compare these with the final coefficients for the best performing model on the validation dataset. Are there noticeable differences? If so, discuss the differences with respect to the impact on the performance.

[2] You do not need to use the entire training set. Stochastic gradient descent, in theory, is not sensitive to the dataset. Thus, you can subsample a reasonable percentage of data to tune the learning rate.