Homework 1 - Coding Problems

2. (30 pts) Predicting Appliance Energy Usage using Linear Regression

Consider the Appliances energy prediction dataset (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 (Appliances attribute) in a low energy house. The data has been 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 26 attributes[1] for each 10-minute interval, which is described in detail on the UCL ML repository, Appliances energy prediction dataset. Your goal is to predict the Appliances.

For this problem, you will use R or python packages for linear regression, ridge regression, and lasso regression. All the specified functions should be in the file 'q2.R (or py, ipynb,)'. The functions in 'q2.py' will be tested against a different training, validation, and test set, so it should work for a variety of datasets and assume that the data has been appropriately pre-processed (i.e., do not do any standardization or scaling or anything to the data prior to training the model). Any additional work such as loading the file, plotting, and required analysis with the data (e.g., parts 2e, 2h, 2j, etc.) should be done in a separate file and submitted with the Code.

(a) How would you preprocess the data? Explain your reasoning for using this pre-processing.

(b) Write code to preprocess data (trainx, valx, testx) that does what you specified in 2a above. Your should return the preprocessed trainx, valx, and testx.

(c) Apply the standard linear regression. Your must return following metrics and the associated values are the numeric values (a dictonary for example: {'train-rmse': 10.2, 'train-r2': 0.3, 'val-rmse': 7.2, 'val-r2': 0.2, 'test-rmse': 12.1, 'test-r2': 0.4}).

You can write a function eval_linear1(trainx, trainy, valx, valy, testx, testy) that takes in a training set, validation set, and test set, respectively, and trains a standard linear regression model only on the training data and reports metrics on the training set, validation set, and test set.

(d) Apply ridge. Write a function eval_ridge(trainx, trainy, valx, valy, testx, testy, alpha) that takes the regularization parameter, alpha, and trains a ridge regression model only on the training data.

(e) Apply lasso Write a function eval_lasso(trainx, trainy, valx, valy, testx, testy, alpha) that takes the regularization parameter, alpha, and trains a lasso regression model only on the training data.

(f) Report (using a table) the RMSE and R^2 for training, validation, and test for all the different λ values you tried. What would be the optimal parameter you would select based on the validation data performance?

(g) Generate the coefficient path plots (regularization value vs. coefficient value) for both ridge and lasso. Make sure that your plots encompass all the expected behavior (coefficients should shrink towards 0).

(h) What are 3 observations you can draw from looking at the coefficient path plots, and the metrics?

3. (40 pts) Predicting Appliance Energy Usage using SGD (20+10+10)

Consider the Appliances energy prediction Data set from the previous problem. You are tasked with implementing a stochastic gradient descent (SGD) algorithm for predicting appliance energy usage using the Appliances Energy Prediction dataset. The goal of this task is to help you understand the effect of different hyperparameters, including batch size, learning rate (η), and the maximum number of epochs, on the performance of the model.

(a) Implement the SGD algorithm to solve ridge regression. As a reminder, the optimization problem:

$$\min f\_o(x) = (1/2)||y - X\beta||^2_2 + \lambda||\beta||^2_2 \quad (1)$$

You will need to write code to evaluate the loss function and compute the gradient. You will also need write a function train(self, x, y) that trains a ridge regression model using stochastic gradient descent. Your function should return a dictionary (paired value) where the key denotes the epoch number and the value of the loss associated with that epoch.

(b) Tuning the learning rate: For the optimal regularization parameters you observed from ridge from Q2, what are good learning rates for the dataset? Justify the selection by trying various learning rates and illustrating the objective value on a graph for a range of epochs (one epoch = one pass through the training data)

(c) Tuning the batch size: For the learning rate you selected for Q3b, how does the batch size affect training the model? Justify the selection by trying various batch size and illustrating the objective value on a graph for a range of epochs (one epoch = one pass through the training data)