HW #1 – due Dec. 16th

This HW is done SOLO ☺, (pairs next time)

Upload:

- **A matlab file (one or more) with the code.**
- **A documentation and answer file in word or pdf.**

Use the .csv file ENB2012_data.csv.

This data was download from UCI, it is named "Energy efficiency Data Set". Originally there are two variables to be predicted – y_1 and y_2, here we only use y_1. Originally there were 768 data rows, here we only use 760 rows.

Split the data into X and Y the following way:

X = ENB2012_data(:,1:8);

Y = ENB2012_data(:,9);

1. Use a 10-fold cross validation to split the data into train and test.
2. Add a column of ones to the data and run the matlab function *regress*
3. For each iteration save the value of R^2, the Mean Square Error (MSE) for the train point and the Mean Square Error for the test points.
4. At the end of the 10-fold loop, plot the results of the MSE's and the R^2.
5. For the last iteration (you can do this at the end of the 10-fold loop) plot Y_test in blue and Y_test_hat in pink in the same figure.

6. LASSO - In a different file, run LASSO on this data, no need to split into train and test, we just want to see the effect of the Lasso.

Use the following lines:

```
%linear regression

XX = [ones(760,1), X];

[B,BINT,R,RINT,STATS] = regress(Y,XX);

%lasso regression

k = 0:1e-3:1;

[B_Lasso,STATS_Lasso] = lasso(X,Y,'lambda', k);

figure; plot(k,B_Lasso,'LineWidth',2)
```

5a. Look at the values of B_Lasso(:,1) and the values of B. Are they similar or different? What are the values in B_Lasso(:,1)?

5b. The variable STATS_Lasso.MSE holds the MSE for each run of the lasso (Lasso with different regularization values k).

Plot the graph: **figure; plot(k, STATS_Lasso.MSE, 'b').**

How does MSE behave as k grows? Explain shortly.

5c. One would like to select a subset of the original variables while keeping the MSE relatively close to the original MSE value. According to the above graph and the values in B_Lasso, recommend a subset of variables (columns of X) to be used in a reduced model with a reasonable error (Look at the values of B_Lasso for different k's). Explain your answer.

Good luck !