Preparation

1. Upload HW1_dataset folder to your working directory

- HW1_Q2.py and HW1_Q3.py files are python code to answer questions 2 and 3
- 6 datasets (3 sets of training and 3 sets of testing sets)
- Please update the path highlighted in orange in '#Read Data' section to your working directory. For example:
- dataset1_train = pd.read_csv('u/erdos/csga/yyu149/dataset/train-100-10.csv')
- Three additional training sets from the train-1000-100.csv have been created by taking the first 50, 100, and 150 instances respectively. Call them: train-50(1000)-100.csv, train100(1000)-100.csv, and train-150(1000)-100.csv. The corresponding test file for these 3 datasets is test-1000-100.csv. Please see #Data Experiment section for details
- Readme file
- Written answer for question 1
-

Question & Answers:

Question 2

Answers:

- 2(a) which λ value gives the least test set MSE?
- Run HW1_Q2.py file, the least test set MSE and the corresponding lambda value for each dataset will be printed out
- 2 (b) creating fit, validation, and test sets' MSEs graphs with lambda ranging from 1 to 150 for all 6 datasets
- graphs will be saved in the current working directory folder
- (c) why λ = 0 (i.e., no regularization) gives abnormally large MSEs for those three datasets in (b)
- Answer will be printed out by running

Please see the below for details of model selection and validation and python code as needed:

Steps:

1. Split training data into fitting and validation datasets with 80% vs 20% ratio
2. Get the fitted model that has the w vector as a function of lambda using fitting dataset (80% of train set) and L2 regularization closed form
   - calculate fit set's MSE using each lambda ranges in 1 to 150 integer
3. Apply the fitted models that have different lambda and w parameters on validation and test sets to get the MSEs separately
4. Create a table to compare fitting, validation, and test sets' MSEs change with lambda values side by side
5. Select the lambda value that gives the smallest MSE
   - In this case, I would recommend selecting the least valuation set MSE so that I can use the test set to check whether the lambda is optimal. However, if there is another out sample dataset existing, we can pick the lambda value which corresponding to the smallest test set MSE
   - Take out the comment sign in #Test MSE with selected lambda section, the corresponding test MSE from the selected lambda using valuation set will be printed out. The valuation and test set MSEs are close, which indicate the selected

lambda is good.

6. Plot fit, validation, and test sets' MSEs with changing lambda values
7. Do above steps on all 6 datasets to get answers question 2

-

Question 3

Answers:

3(a):

- Run HW1_Q3.py file, the best choice of λ value and the corresponding test set MSE for each of the six datasets will be printed out

3 (b,c,d):

- Answer will be printed out by running HW1_Q3.py

Steps:

1. Split train dataset into 10 folds, each fold has both fitting and validation sets with 90% and 10% respectively
2. Do below steps for all lambda values
   - Do the below steps for all 10 folds, and get the average MSE
     - Pick 1 fold out of 10 randomly to:
     - Get the fitted model that has the w as a function of a lambda value using fitting set and L2 regularization closed form
     - Apply the fitted model on validation set to get the validation set MSE
     - average the validation set MSEs from 10 disjoin folds for a lambda value as the train set MSE
3. Select the lambda that givens the least training MSE value
4. use the selected lambda value to retrain the entire training set to get the new w vector
5. Calculate test MSE using the selected lambda and the new w vector
6. Create a table to compare training and test set MSEs side by side
7. Check the MSE from train set and test set is closer using the selected lambda to ensure the parameter selection is good
8. Do above steps again on all 6 datasets

-