CS6923 Machine Learning, Fall 2019
Prof. Linda Sellie, NYU School of Engineering

Homework 4

Submit on NYU Classes by Oct. 4th at 8:00 p.m. You may work together with one other person on this homework. If you do that, hand in JUST ONE homework for the two of you, with both of your names on it. You may *discuss* this homework with other students but YOU MAY NOT SHARE WRITTEN ANSWERS OR CODE WITH ANYONE BUT YOUR PARTNER.

IMPORTANT SUBMISSION INSTRUCTIONS: Please submit your solutions in 3 separate files: one file for your written answers to Part I, one file for your written answers/output for the questions in Part II, and one file with your code (a zip file if your code requires more than one file).

Part I: Written Exercises

1. The cost function for ridge regression (if we zero centered the data and we ignore the intercept term) is: $\sum_{i=1}^{N}(y^{(i)} - \mathbf{w}^T\mathbf{x}^{(i)})^2 + \lambda\|\mathbf{w}\|_2^2$

   - What is the gradient of the ridge regression cost function?
   - Derive the closed form solution of ridge regression.

2. (Do not turn in this question)

   Consider a binary classification problem ($y \in \{0, 1\}$), where the iid examples $D = \{(\mathbf{x}^{(1)}, y^{(1)}), (\mathbf{x}^{(2)}, y^{(2)}), \ldots, (\mathbf{x}^{(N)}, y^{(N)})\}$ are divided into two disjoint sets D_{train} and D_{val}.

   - Suppose you fit a model h using the training set, D_{train}, and then estimate its error using the validation set, D_{val}.

     If the size of D_{val} was 100 (i.e. $|D_{\text{val}}| = 100$), how confident are you the true error of h is within 0.1 of its average error on D_{val}?
   - Repeat the previous question where now $|D_{\text{val}}| = 200$ (i.e you have 200 examples in your validation set).
   - Now, suppose you fit two models h_1 and h_2 (both fit using D_{train}) and then you selected the model that had the smallest error on your validation set, D_{val}.

     If $|D_{\text{val}}| = 100$, how confident are you that the model you selected is within 0.1 of its average error on D_{val}?

   In solving this problem, use the Hoeffding bound we discussed in class. An additional resource is `https://www.cs.cmu.edu/~avrim/ML14/inequalities.pdf`

3. (Do not turn in this question)

   Consider a linear regression model $y^{(i)} = \mathbf{w}^T\mathbf{x}^{(i)} + \epsilon^{(i)}$, where instead of assuming the noise $\epsilon^{(i)} \sim N(0, \sigma^2)$ we assume $\epsilon^{(i)} \sim \text{Laplace}(0, b) = \frac{1}{2b}\exp\left(-\frac{|\epsilon^{(i)}|}{b}\right)$

   - Show that the maximum likelihood estimate $\mathbf{w}$ is the one that minimizes $\sum_{i=1}^{N}|y^{(i)} - \mathbf{w}^T\mathbf{x}^{(i)}|$.
   - Informally discuss why this model will be more robust to noise as compared to the model where we assume the noise $\epsilon^{(i)} \sim N(0, \sigma^2)$.

Part II: Programming Exercise

In the first exercise, you will write the code to predict housing prices in Boston.

1. In this problem you will experiment with a linear regression problem based on real world data. The data is from the Boston Housing dataset in scikit-learn. Your task is to estimate the price of a house in Boston using 13 attributes. Your program should do the follow:

   (a) fit a linear regression model using the closed form solution presented in class. Use k-fold cross validation to estimate the performance of this model. Print the average of your recorded scores for both the test set and training set.

   (b) fit a ridge regression model using the closed solution from written question 1. Use k-fold cross validation to find the best $\lambda \in [10^1, 10^1.5, 10^2, 10^2.5, ..., 10^7]$. Use the Numpy function: `np.logspace(1, 7, num=13)` to get the different values for λ.

   For the best λ you found, use k-fold cross validation to estimate the performance of this model with this λ. Print the average of your recorded scores for both the test set and training set.

   In the Jupyter Notebook we will use `alpha` instead of `lambda` for λ.

2. Repeat the previous exercise, but this time, by creating a ~~polynomial~~ transformation of degree 2 on the features of the dataset.

3. If you are given a choice of predicting future housing prices using one of the models you have learned above, which one would you choose and why?

   State the parameters of that model.

   Using this model predict the price of a house with features: $[5, 0.5, 2, 0, 4, 8, 4, 6, 2, 2, 2, 4, 5.5]$. Make sure to scale the features before predicting the values.