CS6923 Machine Learning, Spring 2020
Prof. Linda Sellie, NYU School of Engineering

Homework 4

Submit on NYU Classes by February 29th at 8:00 p.m.

Submission is required only on GradeScope. There will be 2 submission links on Gradescope, namely HW4_Written_Part and HW4_Coding_Part. As the names suggest, one will be the submission link for theory part and the other for the programming part.

This is what the submission would look like:

1. You will submit a pdf called part1.pdf which will have the theory answers. This will be submitted only on HW4_Written_Part link on Gradescope.
2. You will submit a zip file called part2.zip. This will be submitted only on HW4_Coding_Part link on Gradescope. Within zip file, you will be having 2 files:
   - proganswers.pdf
   - ipynb notebook

You do not have to typeset your written answers: as long as your handwriting is readable, you can write out the answers by hand and scan your answers. If you do that, use a utility like camscanner to make sure that the scan is clear.

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.

Part I: Written Exercises

1. If we modified the cost function for ridge regression so we didn't include a penalty for the intercept term, the cost function would be $\sum_{i=1}^{N}(y^{(i)} - \mathbf{w}^T\mathbf{x}^{(i)})^2 + \lambda\sum_{j=1}^{d} w_j^2$
   - What is the gradient of the ridge regression cost function?
   - Derive the closed form solution of ridge regression.[1]
2. 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}?

[1] How could you modify the identity matrix so you don't restrict the size of the intercept term. If you are having difficulty discovering how to do this, please talk to a GA for help.

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) This question explores the Bayesian connection to ridge regression. We will use the probabilistic view we discussed in class (slide 22 and 23 in lecture 3) where we assumed $y^{(i)} = \mathbf{w}^T\mathbf{x}^{(i)} + \epsilon^{(i)}$ with noise iid and $\epsilon^{(i)} \sim \mathcal{N}(0, \sigma^2)$.

   But we will now specify a *prior distribution* $p(\mathbf{w})$ on the parameters $\mathbf{w}$. Adding a prior allows us to specify that some $\mathbf{w}$ are more likely than others and this can work as a regularizer for the parameters.

   As we discussed in the second lecture, we can use Bayes rule to combine the prior $p(\mathbf{w})$ with the likelihood $\prod_{i=1}^{N} p(y^{(i)} \mid \mathbf{x}^{(i)}; \mathbf{w})$. Thus $\mathbf{w}_{\text{MAP}} = \arg\max_{\mathbf{w}} \prod_{i=1}^{N} p(y^{(i)} \mid \mathbf{x}^{(i)}; \mathbf{w}) p(\mathbf{w})$

   In this question assume that w_i for $i = 1..d$ are independent, identically and distributed according to a gaussian distribution, $w_i \sim \mathcal{N}(0, \rho^2)$ for $i = 1..d$. Here I am using ρ instead of σ since we are already using σ for the noise.

   What is $\mathbf{w}_{\text{MAP}}$? (You do not need to re-derived the results on slides 22 and 23 in lecture 3.)

   Argue that the ridge regression estimate is the MAP (maximum a posteriori probability) estimate, when $\mathbf{w}$ has the prior distribution described above.

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 10-fold cross validation to estimate the performance of this model using MSE (mean squared error). Print the average of your recorded MSE for both the test set and training set.

   (b) fit a ridge regression model using the closed solution from written question 1. Use 10-fold cross validation to find the best $\lambda \in \{1.023, 1.237, 1.496, \ldots, 6.83916, 8.270, 10.\}$ Use the Numpy function: `np.logspace(.01, 1, num=13)`to get the different values for λ.[2]

   In the Jupyter Notebook we will use `alpha` instead of `lambda` for λ. In the notebook, some of the code has already been written for you.

   Print the average of your recorded MSE for both the test set and training set for each value of λ.

   (c) experiment. Choose different λ's than the ones given above, and different transformations of the data

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

3. Answer the following questions in a pdf file called proganswers.pdf:

   (a) What were the estimated MSE printed by your code for the different values of λ? (Include the case where $\lambda = 0$. You computed this value in programming question 1a)

[2] If you choose to use the penalty $\lambda/N\mathbf{w}^T\mathbf{w}$ then multiply your λ values by N.

(b) 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?

For this model, refit the parameters using all the data. 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.

(c) What did you discover from question 1c above?