Assignment 2

September 2021

Instructions

- This assignment should be completed individually.
- Do not look at solutions to this assignment or related ones on the Internet.
- The files related to the assignment are present in `lab2-rollno.zip` folder. Extract it and upload it on moodle in the same .zip format after completion and after replacing the string "rollno" with your actual roll number. For example, if you roll number is 00405036, then single zip folder that you will upload will be named "lab2-00405036.zip". Also collate all the CS337 based theory solutions into ONE pdf file named `answers.pdf`. Include `answers.pdf` inside the zip folder mentioned above and submit the zip folder.
- Answers to all subjective questions need to be placed in single pdf `answers.pdf` including all plots and figures and uploaded.
- Only add/modify code between `TODO` and `END TODO` unless specified otherwise
- The code for all the questions should be written in provided ipython notebook only. Don't modify the name/directory structure of the provided `.zip` file.
- Please make sure that your code runs in **python 3.x**. You shuld not import any new python libraries.
- Code should be written in the provided `assignment_2.ipynb` file only.
- This Assignment carries a total of 11 marks for CS337 Theory and 12 marks for CS335 Lab.

1 Perceptron

A Perceptron keeps a weight vector w_y corresponding to each class y. Given a feature vector f, we score each class y with

$$score(f, y) = \sum_i f_i w_{y_i}$$

where i iterates over the dimensions in the data. Then we choose the class with the highest score as the predicted label for data instance f.

f_1	f_2	y
0	0	0
0	1	0
1	0	0
1	1	1

Table 1: AND Dataset

Learning the weights of the Perceptron

In the basic multi-class Perceptron, we scan the entire training data, one instance at a time. When we come to an instance (f, y), we find the label with the highest score: $y' = \arg\max_{y''} score(f, y'')$, while breaking ties arbitrarily. We compare y' to the true label y. If $y' = y$, we have correctly classified the instance, and we do nothing. Otherwise, we have wrongly predicted y' instead of the true label y. This means that w_y has scored f lower, and/or $w_{y'}$ has scored f higher, than what would have been ideally desirable. To avert this error in the future, we update these two weight vectors accordingly: $w_y = w_y + f$ and $w_{y'} = w_{y'} - f$

1.1 CS 337: Theory

1. An alternative to the 1-vs-rest perceptron defined above is to have a 1-vs-1 perceptron, one for each pair of classes. Compare this proposed 1-vs-1 approach to the previously described 1-vs-rest approach for multi-class classification. Outline and briefly justify the advantages and disadvantages of each. (1 mark)

2. Assume a data point (f, y) which got misclassified as y'. Show that the above update rule increases $score(f, y)$ and decreases $score(f, y\prime)$. (1 mark)

3. We have given the update rule for perceptron above and also learned in class that the algorithm converges in finite number of steps for linearly separable datasets. Assume a $2-$class dataset with $y \in \{-1, +1\}$. Lets define a loss function $\mathcal{L}(f, y) = max(0, -yf^T w)$ and learn w using gradient descent algorithm. Mention the gradient descent update rule and show that this method also converges for linearly separable datasets. You may use the fact that the perceptron update rule converges for linearly separable dataset. (1.5 marks)

4. For a dataset, assume that the upper bound on number of iterations required for convergence of perceptron is M. If we change the perceptron update rules as follows:

$$w_y = w_y + 0.5f$$
$$w_{y\prime} = w_{y\prime} - 0.5f$$

   Comment on the upper bound on number of iterations under the modified algorithm. (0.5 mark)

5. For `AND` dataset given by the AND function as mentioned in the Table 1, Compute the upper bound on the number of iterations required by the Perceptron algorithm. (1 mark)

1.2 CS 335: Lab

Implement 1-vs-rest perceptron algorithm described above. Complete the function `perceptron()` in the notebook. (2.5 mark)

Data Description

For this assignment, we assume a ground truth function $y = f(x) = 1 - 3x - 2x^2 + 2.5x^3$. However, to exploit the feature selection property of LASSO we have added few redundant features to the dataset. You have to learn a lasso model which should be able to detect the redundant features from the dataset.

2 LASSO and ISTA

Lasso regression uses the ℓ_1 penalty term and stands for Least Absolute Shrinkage and Selection Operator. It is a widely used tool for achieving sparse solutions to optimization problems. The penalty applied for ℓ_1 is equal to the absolute value of the magnitude of the coefficients.

$$L(w) = \sum_{i=1}^{n} \|y_i - x_i.w\|_2^2 + \lambda \|w\|_1^1$$

In matric form (using matrix X here instead of Φ as used in the class, LASSO optimizes

$$\min_w \|y - Xw\|_2^2 + \lambda \|w\|_1$$

In this problem, we will implement the Iterative Soft-Thresholding Algorithm (ISTA) for LASSO.

2.1 CS 337: Theory

1. Prove that the solution to LASSO is the MAP estimate of Linear regression subject to the Laplacian prior on weights. (1.5 mark)
2. Similar to ridge regression, a lambda value of zero results in the basic OLS equation, however given a suitable lambda value, lasso regression can give more sparse solution as compared to ridge regression. Discuss the reason behind this property of lasso. (1 mark)
3. Does the closed form solution for Lasso exist? If yes, then mention the conditions for which closed form solution will be possible. Give your answer with explanations. (1 mark)

2.2 CS 337: Lab

1. Complete the `ista()` function in file `assignment_2.ipynb`. You can reuse the relevant functions you had implemented in Assignment 1. You can modify `lr`, `epochs` and λ, if needed. Add the generated figures in the report. (2.5 mark)
2. Complete the `mse_multi_var()`, `mse_regularized()`, `split_data()`, `multivar_reg_closedform()` functions. (1 mark)

3. Analyze how LASSO does feature selection using the ista() function. Compare the weights obtained from Lasso and Ridge Regression by creating a single plot. For Ridge, you can use the code from previous assignment. (2 mark)

3 Bias Variance Trade-off

3.1 CS 337: Theory

1. Indicate the effect of each of the following on the bias and/or variance of the learned model along with an informal reasoning. (1.5 marks)
   (a) Increasing the value of λ in lasso regression.
   (b) Increasing model complexity by adding more features of high degree.
   (c) Reducing dimension by choosing only those subset of features which are of more importance.
2. Given irreducible noise, $\epsilon \sim \mathcal{N}(\mu = 0, \sigma^2)$. Let $\hat{f(x)}, x \in$ Test Set, be the function which approximates underlining true function, $f(x)$. Show that the Mean Squared Error can be written as sum of $Variance(\hat{f(x)}$ and $(Bias(\hat{f(x)}))^2$ (1 marks)

3.2 CS 337: Lab

1. Complete `gen_bias_variance()` by computing bias and variance. (2 marks)
2. Complete the function `driver()` by computing the true mean of our linear model which is given by y = mx + c + error. Add the generated figures to the report and discuss the bias variance trade-off for OLS, Ridge and Lasso. (1 marks)
3. Complete the function `ridge(),lasso(),ols()` for computing w, b and return the predictions. (1 marks)