CSE343/ ECE363/ ECE563: Machine Learning W2021
Assignment: Linear / Logistic Regression and Näive Bayes
Max Marks: Programming:85, Theory:17/31 (for UG/PG)
Due Date: 31/3/2021, 11:59 PM

Instructions

- This is a group-individual hybrid assignment. Each member should attempt one programming question, and *all* theory questions (where applicable). If your group has only two members, your team may skip Programming Question 3. The final grading for the assignment will be based on a combination of group score and individual score.
- Try to attempt all questions.
- The theory questions should be your individual effort. You may consult within your group for programming questions, however, your code should be an individual effort. Copying/Plagiarism will be dealt with strictly.
- Start early, solve the problems yourself.
- Late submission penalty: As per course policy.
- Your submission would be a single .zip file (rollno_HW1.zip) file, that would contain two items (codes + .pdf file). You have to include all your plots, results, analysis, conclusion, and solutions for the theory questions in the pdf file.
- Anything not written in the report will fetch 0 marks.
- It is preferred that you write LaTeX reports.
- Remember to **turn in** after uploading on Google Classroom. No excuses or issues would be taken regarding this after the deadline.
- Start the assignment early. Resolve all your doubts from TAs in their office hours **two days before the deadline.**

PROGRAMMING QUESTIONS

1. (35 points) **Linear Regression**

    1. (15 points) : Implement Linear Regression for the Abalone Dataset. The dataset contains 9 variables out of which the last column is the output variable and the other 8 are input attributes. You need to implement gradient descent from scratch i.e. you cannot use any libraries for training the model (You may use numpy, but libraries like sklearn are not allowed). Choose an appropriate learning rate. You may need feature normalisation.

(a) Include plots for the root mean squared error (RMSE) vs. gradient descent iterations for both training as well as validation set (one for each set computed over the 5 folds (2 plots in total). The curves should show the mean RMSE.

(b) Also implement the normal equation (closed form) for linear regression and get the optimal parameters directly. Compute the RMSE (after getting the optimal parameters) for both training and validation set and report them for all the 5 folds.

(c) Compare the final RMSE obtained from (a) after convergence and the RMSE from (b) and make a note of any observations you might have.

2. (15 points) **Regularization**
From the previous part, identify the validation set that has the lowest RMSE, and hold it out as the test set. Use the remaining 80% of the data as the new training + validation set. Note that your model should never see the test set (neither for training, nor for validation).
Use 5-fold cross validation with grid search on the train + val set (without using the test set) to find the appropriate regularization parameter (hyperparameter). You may use Ridge, Lasso and GridSearchCV routines from the sklearn library to perform the following:

   (a) Find the optimal regularization parameter for L2

   (b) Find the optimal regularization parameter for L1

   Write the above two regularization parameters in the report. Once the optimal regularization parameters for both L2 and L1 have been found out, modify the gradient descent algorithm implemented in (1) to accommodate for the L2 and the L1 regularization term. Use the values of the 'optimal' regularization parameter found out in (a) and (b) and use it with the modified gradient descent algorithm and plot the RMSE error vs iterations curve. Also report the RMSE on the test set. [Two plots - one for L1, one for L2]

3. (5 points) **Best Fit Line**
Use the following data that contains only 1 input variable and 1 output variable i.e the brain-weight to the body-weight proportion for varying species. Consider the dataset as a whole i.e. do not split it into train, val or test. Perform the following tasks:

   (a) Plot the data points using a scatter plot along with the best fit line found out using linear regression (without regularisation).

   (b) Use L2 regularisation and plot the data points and the new best fit line.

   (c) Use L1 regularisation and plot the data points and the new best fit line.

   Use the gradient descent algorithm implemented in (1) to perform regression on the whole dataset in all these 3 parts.
   Compare how the best fit line changes visually with adding different kinds of regularisation, was it a better fit, worse fit than the regression performed without regularisation?

2. (35 points) **Logistic Regression**

1. (15 points) Download the dataset from here. It's a standard dataset, where you have to predict whether the income of a person is above or below 50k$. Implement a classifier using logistic regression **from scratch** (you are allowed to use only NumPy and Pandas). Also implement L1 and L2 regularization. Report the accuracy on the train, val and test set while using L1 and L2 regularization separately. You may use scikit-learn only for encoding categorical data. You can play around with the data - it may or may not be necessary to use all the features. Can you come up with a reason as to why L1 regularization works better than L2 regularization, or vice versa? Also plot a loss vs iteration and accuracy vs iteration graphs for both models (L1 and L2).
2. (15 points) Download the MNIST dataset from here. Implement an L1 and L2 regularized logistic regression model using the scikit-learn library. Compute and report the accuracy obtained using one-vs-rest approach for each of the 10 classes, for both the training and test sets. See this tutorial to understand how the one-vs-rest approach works while using a binary classifier for a multi-class classification problem. Report the train and test accuracy for both L1 and L2 regularized logistic regression. Comment on whether or not it is a good fit, i.e, underfitting or overfitting.
3. (5 points) For the MNIST dataset, plot the Receiver Operating Curve (ROC) curve for each class (Do this just for L2 regularized Regression). Plot all ROC curves on the same graph. A tutorial on ROC curves is given here and here.

3. (15 points) **Naïve Bayes**
Use sklearn wine dataset, divide the data into 70:30 ratio for train and test set using sklearnś train_test_split and perform the following tasks:

1. (10 points) Implement Gaussian Naïve Bayes from scratch for the given dataset.
2. (2 points) Report F1-Score, Accuracy and confusion metrics for your test set with any observation that you may have.
3. (3 points) Implement Gaussian Naïve Bayes from sklearn and compare the evaluation metrics with your implementation. Report any discrepancy if any and provide a suitable reason.

THEORY QUESTIONS

4. (10 points) The linear model with several explanatory variables is given by equation:

$$y_i = \beta_1 + \beta_2 x_{2i} + \beta_3 x_{3i} + \ldots + \beta_k x_{ki} + \epsilon_i \tag{1}$$

For the purpose of analysis, it is convenient to express the above linear model in matrix form as shown in equation 2:

$$Y = X\beta + \epsilon \tag{2}$$

Where,

$$Y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}, X = \begin{pmatrix} 1 & x_{21} & ... & x_{k1} \\ \vdots & \vdots & & \vdots \\ 1 & x_{2n} & ... & x_{kn} \end{pmatrix}, \beta = \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_k \end{pmatrix}, \epsilon = \begin{pmatrix} \epsilon_1 \\ \vdots \\ \epsilon n \end{pmatrix} \tag{3}$$

Write the matrix expression for the sum of squared errors loss function. Derive an expression to find the β^* that minimizes this loss for the above linear regression problem. In order to derive the least squares solution, you would need to differentiate the matrix form directly. Once you have the solution expression, write the conditions under which the solution (in the matrix form) will exist.

5. (7 points) Suppose we have a data set with five predictors, X1 = GPA, X2 = IQ, X3 = Gender (1 for Female and 0 for Male), X4 = Interaction between GPA and IQ, and X5 = Interaction between GPA and Gender. The response is starting salary after graduation (in thousands of dollars). Suppose we use least squares to fit the model, and get estimates of $\beta_0, \beta_1, \beta_2, \beta_3, \beta_4, \beta_5, as$ =50, 20, 0.07, 35, 0.01, -10, *respectively*.

    1. (3 points) Which answer is correct, and why?
        (a) For a fixed value of IQ and GPA, males earn more on average than females.
        (b) For a fixed value of IQ and GPA, females earn more on average than males.
        (c) For a fixed value of IQ and GPA, males earn more on average than females provided that the GPA is high enough.
        (d) For a fixed value of IQ and GPA, females earn more on average than males provided that the GPA is high enough.
    2. (2 points) Predict the salary of a female with IQ of 115 and a GPA of 3.5.
    3. (2 points) True or false: Since the coefficient for the GPA/IQ interaction term is very small, there is very little evidence of an interaction effect. Justify your answer.

6. (14 points) (**For PG Students only**) Bias-variance decomposition of training loss helps in understanding the machine learning algorithms.
Given a training set $\{(x_1, t_1), ..., (x_n, t_n)\}$ a machine learning algorithm produces a model M. Here, x is the input data sample and t is its target value. Given a test sample x, the model M provides a prediction $y = M(x)$. A loss function $L(t, y)$ measures the cost of predicting y when the true value is t. Depending on the type of learning problem the loss functions may differ, a few of the loss functions are, squared loss $L(t, y) = (t - y)$, zero-one loss ($L(t, y) = 0$ if $y = t$, $L(t, y) = 1$ otherwise), absolute loss ($L(t, y) = |t - y|$). The goal of a machine learning algorithm is to produce a model that minimizes the average loss $L(t, y)$ over all possible samples.
Let S be a set of training sets. Consider an example x and suppose a model M predicts y given a training set in S. For a certain loss functions L one can decompose the expected loss $E_{D,t}[L(t, y)]$ into three components Bias $B(x)$, Variance $V(x)$ and Noise $N(x)$ as shown in equation 4 :

$$E_{D,t}[L(t, y)] = B(x) + c_1 V(x) + c_2 N(x) \tag{4}$$

1. (7 marks) Verify that for squared loss the value of $c1 = c2 = 1$.
2. (7 marks) Find the values of all the three components , $c1$ and $c2$ for two class classification problem using zero-one loss.

Hint: Refer A Unified Bias-Variance Decomposition for Zero-One and Squared Loss.