Machine Learning (CS60050): 2023-2024 Autumn

Assignment 1

Submission deadline: 16-Aug-2023, 11:55 PM

Ashwamegh Rathore * and Tarun Mohanty *

*Indian Institute of Technology, Kharagpur

Instructions

1. The submission deadline is hard. There may be unforeseen glitches during submission. So, for safety, submit your files well ahead.
2. All submissions should be on microsoft teams only. No email submission will be accepted. Special consideration may be made for medical emergencies.
3. Please download the dataset and directory structure to be maintained from here.

   ```
   dataset
   ├── linear-regression.csv
   └── logistic-regression
       ├── Pumpkin_Seeds_Dataset.xlsx
       └── Pumpkin_Seeds_Dataset_Citation_Request.txt
   ```

4. For each part of the assignment implement the model in a **code.ipynb** file. Report the model performance by attaching a **report.pdf** which contains the output of all cells of the jupyter notebook. Refer to this article for converting the ipynb file to pdfs. Finally attach the list of package dependencies for each model in the **requirements.txt** file.
5. After implementing the required machine learning models, you must submit only the `<rollno>_<name>` directory. No need to submit the dataset directory. Directory structure to be submitted-

   ```
   rollno_name
   ├── linear-regression
   │   ├── code.ipynb
   │   └── report.pdf
   ├── logistic-regression
   │   ├── code.ipynb
   │   └── report.pdf
   └── requirements.txt
   ```

   The name of the zip file should be your roll number, followed by an underscore, followed by your name. For example, if your roll number is `22CS0100` and name is `John Doe`, then the zip file should be named as `22CS0100_JohnDoe.zip`.

   Failing this, your assignment will not be evaluated.

6. **Unless explicitly asked,you cannot use any library/module meant for Machine Learning or Deep Learning.** You can use libraries for other purposes, such as formatting and pre-processing of data, but NOT for the ML part. Also you should not use any code available on the Web. Submissions found to be plagiarised or having used ML libraries will be awarded zero marks for all the students concerned.

1 Linear Regression

25 marks

Linear regression is a statistical technique that seeks to establish a linear relationship between a dependent variable (Y) and one or multiple independent variables (X), by minimizing the sum of squared differences between actual Y values and the values predicted by a linear equation. It aims to estimate the coefficients (slope and intercept) of the linear equation, enabling prediction and interpretation of the dependent variable's variations based on changes in the independent variables.

For this assignment,

1. Split the dataset into 50% for training, 30% for validation and 20% for testing. Normalize/Regularize data if necessary. Encode categorical variables using appropriate encoding method if necessary.

2. Formulate a **linear regression model** between the observations $x_j^{(i)}$ and target parameter $y^{(i)}$.
$$y^{(i)} = \theta_0 + \sum_j \theta_j * x_j^{(i)} \tag{1}$$

3. Use the **mean squared error** loss function to fit the 2 following models.

4. For the first model use the **analytic solution** as discussed in class. Report the R-squared and RMSE score for the test set after training the model.

5. While for the second model use the an iterative solution via **gradient ascent.** Use 3 different learning rates- 0.01, 0.001 and 0.0001. Plot the Loss function for the training set and validation set at each iteration. Finally report the R-squared and RMSE score for test set.

2 Logistic Regression

25 marks

Logistic regression is a statistical method used for binary classification. It models the probability of an event occurring by fitting a logistic curve to data, assigning values between 0 and 1. It's a vital tool for predicting outcomes like yes/no or true/false in various fields.

For this assignment,

1. Split the dataset into 50% for training, 30% for validation and 20% for testing. Normalize/Regularize data if necessary. Encode categorical variables using appropriate encoding method if necessary.

2. Implement a **standard Logistic Regression Classifier** as discussed in class. **Do NOT use scikit-learn for this part.**

3. Report the **mean accuracy, precision and recall** for the class 1(good)for the classifiers. You may or may not use the scikit-learn implementations for computing these metrics.

References

1. Blog on metrics used for bench marking linear regression performance.
2. Blog on metrics used for bench marking classification model performance.
3. Article on exporting package dependencies to requirements.txt
4. Article on setting up a virtual environment using virtualenv
5. Article on generating pdf snapshots from jupyter notebooks.