Assignment 4
Due Date: 1400/Dey/10

Supervised Learning

Foundations of Data Science
Supervisor: Dr. SaeedReza Kheradpisheh
Teaching Assistants: Hesam Damghanian, Ali Rahimi

Shahid **B**eheshti **U**niversity

* This assignment is divided into two parts, Regression and Classification *

Part 1

Supervised Learning – Continuous Variable Prediction

In this exercise you will implement multivariate linear regression. This may seem like a long assignment, but in practice you only implement about 10 lines of code and become familiar with implementing machine learning algorithms in Python.

Files and Functions

In this assignment you are given dataset and python codes:

- linreg.py
  - Linear regression:
    - Fit: function to learn the multivariate linear regression
    - Predict: function to predict the input data
    - computeCost: calculate the cost function
    - gradientDescent: optimizes model parameters via gradient descent algorithm.
- test_linreg_univariate.py – module to test the univariate linear regression
  - plotData1D: scatter plots the univariate data
  - plotRegLine1D: given the trained model parameters, draws the regression line on the plot from the previous function.
  - visualizeObjective: plots the surface and the contour graphs of the cost function (only for demonstration purposes).
- test_linreg_multivariate.py

Datasets

- univariateData.dat – for univariate regression
- multivariateDat.dat – for multi-variate regression

Implementation

Just complete the *TODO* parts of the LinearRegression class!

Linear Regression

The $h(x)$ is the linear regression function and the θ is the weights matrix in this equation.

$$h(\theta) = \theta^T x$$

The learning process in linear regression is done by the finding the right weights such that it minimizes the loss function as:

$$\hat{\theta} = \min\ j(\theta)$$

$$j(\theta) = \frac{1}{2n}\sum_{i=1}^{n}\left(h(x^{(i)}) - y^{(i)}\right)^2$$

Or $j(\theta) = \frac{1}{2n}(X\theta - y)^T(X\theta - y)$

Where $X \in R^{n\times d}$and $y \in R^n$. This can be done via the gradient descent algorithm.

Gradient descent algorithm

The learning rule for this algorithm is:

$$\theta_{i+1} \leftarrow \theta_i - \alpha \frac{1}{2n}(X\theta - y)^T(X\theta - y)$$

Where α is the learning rate.

After implementing the linear regression, fit the model with both multi/uni-variate datasets and report the loss value. You can also include the plots obtained from test codes in your report.

Evaluating the model on the unseen data

Use your regression model and predict the test dataset, holdout.npz, and report the RMSE error between predicted and true values.

Part 2

Supervised Learning – Discrete Variable Prediction

In this section you will perform methods listed below for predicting discrete variables (a.k.a. classification).

- Logistic Regression Classifier
- Support Vector Machine (SVM) Classifier
- K-nearest neighbor (KNN) Classifier
- Decision Tree Classifier
- Random Forest Classifier
- Naive Bayes

Dataset and Tasks

The dataset to perform classification is the TalkingData's ad-tracking fraud dataset[1] and the task is to predict whether a user will download an app after clicking a mobile app advertisement. You should train your model on the train dataset and report your models accuracy on both test set and the train set. Alternatively, you can use a lighter version of this dataset[2] which offers additional features in parquet format[3]. In this case you should keep 0.2 of the dataset as the test set and fit the model on the rest and report the accuracy of your model on both of these sets. In addition, make sure to have the AUC curve, confusion matrix, and the decision boundary[4] of your predictions in your report.

Note: Bear in mind that preprocessing steps are **mandatory** for this task, e.g., **feature engineering**, and **feature selection**. Don't limit yourself only with the aforementioned methods, based on the quality of your work, extra scores may be granted for observing and testing other classification algorithms. Once again, we emphasize the report; it should contain all your questions and your innovative findings. Use figures, pictures, and tables, and DO NOT PUT ANY CODE IN THE REPORT**.**

[1] https://www.kaggle.com/c/talkingdata-adtracking-fraud-detection/data
[2] https://www.kaggle.com/matleonard/feature-engineering-data
[3] Readable by pandas framework
[4] Since this dataset is multivariate, pick two features that are the most determinant in your predictions. These two features need to be selected based on each of classifiers. It is recommended to use a smaller subset of your dataset for saving time and resources.