Introduction to Machine Learning

Homework 1 - Summer 2022

Homework Objectives

In this homework, you are to help a fictitious insurance company. Your task is to estimate medical cost of individuals based on their attributes such as age, bmi, etc. The insurance company will use your model to adjust the insurance fee for each individual separately. This homework will provide you with an opportunity to learn and apply data visualization, data preparation, and several traditional ML algorithms.

Step 1: Get the Data

Download **Medical Cost Personal Dataset** from Kaggle website. Kaggle is an online community platform for data scientists which allows users to collaborate, find and publish datasets, use GPU integrated notebooks, and compete to solve data science challenges. For easier setup, you are recommended to use Kaggle notebooks.

> Check out **this video** to learn more about Kaggle.

Step 2: Discover and Visualize the Data to Gain Insights

(a) Visualize the data. Try to provide high quality plots. You have to provide a description for each of the figures.

(b) Compute and visualize the correlation between every pair of attributes. Explain your observations.

(c) If possible, combine two or more attributes to create new attributes. Compute the correlation between the new attributes and the insurance cost.

> Check out **this video** to learn more about data visualization in python.
> Check out **this video** to learn more about data engineering.

Step 3: Prepare the Data for Machine Learning Algorithms

(a) Split the data into training and test sets.

(b) Check if the data contains missing values. If so, fill in the missing values in an appropriate manner.

(c) ML algorithms assume two nearby values are more similar than two distant values. Encode categorical attributes using one-hot encoding, or replace them with related numerical features instead.

(d) Scale the features.

> Check out **this article** to learn more about methods to fill missing values.
> Check out **this article** to learn more about encoding methods.
> Check out **this article** to learn more about feature scaling methods.

Step 4: Implement Linear Regression and Gradient Descent from Scratch

The next steps are to build and train ML models on the prepared data. In this step, you implement and train the linear regression model using gradient descent algorithm. You are **NOT** allowed to use machine learning libraries and frameworks such as scikit-learn. However, it is recommended to use fundamental packages like NumPy and Pandas.

(a) Recall the concepts and equations of linear regression and gradient descent.

(b) Implement and train your model on the training set. Do not regularize the model at this step. Visualize the error of your model during the training process.

(c) Report the error of your model on the test set.

(d) Include the regularization terms into your equations.

(e) Implement and train the regularized model on the training set. Visualize and compare the error of your model during the training process.

(f) Report and compare the error of the regularized model on the test set.

(g) Try to improve the results of your model. You may refer to course videos, textbooks, or internet to complete this task.

> Check out **this video** to learn more about linear regression and gradient descent.
> Check out **this article** to learn more about model regularization.

Step 5: Build and Compare Different ML Algorithms

In this step you build, evaluate, and compare different ML algorithms. These algorithms are SVM, KNN, Decision Tree, Random Forest, and Gradient Boosting regression. For each of these algorithms:

(a) Read their documentation page on the scikit-learn website. Try to understand all of the parameters and attributes.

(b) Build and train the algorithm on the prepared data. You may use grid search for parameter tuning.

(c) Report the error of your model on the test set. If possible, provide one or more visualizations.

> Check out **this video** to learn more about grid search.

Step 6: Check Out Other Notebooks

You can view many other notebooks for each dataset on the Kaggle website. Sort them by the number of votes and look at a number of them. Try to understand and learn interesting things they have done but do not copy any part of their codes. Do not forget to **give credit** where credit is due.

Step 7: Course Feedback

I would greatly appreciate your feedback on the coursework:

(i) How hard did you find this homework? How long did it take you to finish? Does that seem unreasonably difficult or time-consuming for an optional course?

(ii) Did you read through the textbook? If so, did you find it useful?

(iii) Is there anything in particular that needs to be improved? Is there anything in particular that you think was helpful?