Homework 2

17-400/17-700: Data Science and Machine Learning at Scale

Due Monday, October 5th at 11:59 PM

1 Introduction

This assignment builds upon the Spark knowledge you've gained in Homework 1 to build machine learning applications using the Spark MLlib library.

This assignment consists of two major sections. The first section is an end-to-end exercise of performing Extract-Transform-Load (ETL) and exploratory data analysis on a real-world dataset, and then applying several different machine learning algorithms to solve a supervised regression problem on the dataset. The second section involves training a linear regression model to predict the release year of a song given a set of audio features.

2 Logistics

As with Homework 1, we provide the code template for this assignment in a Jupyter notebook. What you need to do is to follow the instructions in the notebook and implement the missing parts marked with `'<FILL IN>'` or `' YOUR CODE HERE'`. Most of the `'<FILL IN>/YOUR CODE HERE'` sections can be implemented in just one or two lines of code.

2.1 Getting lab files

The code for this homework is in a single notebook file. You can obtain the notebook `'assignment_notebook.ipynb'` after downloading and unzipping `hw2.zip` at
https://github.com/17-700/released-hws-fa2020/raw/master/hw2/hw2.zip.

Next, as for Homework 1, import the notebook into your Databricks workspace. You can refer to the instructions of Homework 1 if you a refresher on how to set up your environment - the requirements are identical for this homework.

2.2 Preparing for submission

We provide several public tests via `assert` in the notebook. You may want to pass all those tests before submitting your homework.

In order to enable auto-grading, please do not change any function signatures (e.g., function name, parameters, etc) or delete any cells. If you do delete any of the provided cells (even if you re-add them), the autograder will fail to grade your homework. If you do this, you will

need to re-download the empty 'assignment_notebook.ipynb' file and fill in your answers again and resubmit.

Also be sure to comment out any Databricks-specific functions such as `dbutils` before submitting your notebook. The autograder environment runs outside of Databricks, and so these function calls will fail and cause subsequent statements in the same cell to fail. This in turn might lead to cascading errors in later cells and unintuitive errors in the autograder output.

2.3 Submission

1. Export your solution notebook as a IPython notebook file on Databricks via `File -> Export -> IPython Notebook`
2. Submit your solution via Gradescope (Please don't rename your notebook file).

3 Section I: Power Plant Machine Learning Pipeline Application

This section is an end-to-end exercise of performing Extract-Transform-Load (ETL), exploratory data analysis, and model development on a real-world example of predicted demand, actual demand, and available resources from the California power grid. Our goal is to accurately predict power output given a set of environmental readings from various sensors in a natural gas-fired power generation plant.

This section covers:

- Loading our data into a format we can query and use
- Exploring and visualizing the data
- Preparing the data for machine learning
- Modeling the data and making predictions
- Model tuning and evaluation

4 Section II: Linear Regression

This section covers a common supervised learning pipeline, using a subset of the Million Song Dataset from the UCI Machine Learning Repository. Our goal is to train a linear regression model to predict the release year of a song given a set of audio features.

This section covers:

- Reading and parsing the initial dataset through visualization of the features and labels
- Creating and evaluating a baseline model
- Training and evaluating a linear regression model
- Training using SparkML and grid search
- Adding interactions between features

See the notebook for detailed descriptions and instructions of each question.