Homework 2

10-405/10-605: Machine Learning with Large Datasets

Due Monday, February 17th at 9:00 AM

1 Introduction

This assignment involves understanding basics of building machine learning pipelines and techniques in training linear regression models.

This assignment consists of two major parts. The first part is to build a machine learning pipeline to apply several different machine learning algorithms to solve a supervised regression problem on a real-world data-set. The second part is to train a linear regression model to predict the release year of a song given a set of audio features.

2 Logistics

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

You can obtain the notebook '`assignment_notebook.ipynb`' after downloading and unzipping `hw2.zip` at https://github.com/10605/released-hws/raw/master/hw2/hw2.zip.

Next, import the notebook into your Databricks, which provides you a well-configured Spark environment and will definitely save your time (see the next section for details).

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.

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 Setting up environments on Databricks

We provide step-by-step instructions on how to configure your Databricks platform. We also introduced it in detail during the recitation on January 17. The recitations slides can be found here.

1. Sign up for the **Community Edition** of Databricks here: https://databricks.com/try-databricks.
2. Import the notebook file we provide on your homepage: `Workspace -> Users -> Import`
3. Installing third-party packages that will be used in the homework on Databricks: `Workspace -> Create -> Library`. Then select `PyPI`, enter the package name as `nose`. Finally click `Create` to install it. You may also want to check `install automatically on all clusters` option to install the PyPI package on all clusters you create.
4. Create a cluster: `Clusters -> Create Cluster`. You can use any cluster name as you like. When configuring your cluster, make sure to choose **the default Databricks runtime version and Python version** `Python 3`. Note: It may take a while to launch the cluster, please wait for its status to turn to '`active`' before start running.
5. You can start to play with the notebook now!

Note: Databricks Community Edition only allows you to launch one 'cluster'. If the current cluster is 'terminated', then you can either (1) delete it, and then create a new one, or (2) activate and attach to the existing cluster when running the notebook.

4 Power Plant Machine Learning Pipeline Application

This section is an end-to-end exercise of performing Extract-Transform-Load 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.

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. In this exercise, we are going to go through a general machine learning pipeline, including data preparation, data modeling, and tuning and evaluation.

5 Linear Regression

This section covers a common supervised learning pipeline, using a subset of the [Million Song Dataset](http://labrosa.ee.columbia.edu/millionsong/) from the [UCI Machine Learning Repository](https://archive.ics.uci.edu/ml/datasets/YearPredictionMSD). Our goal is to train a linear regression model to predict the release year of a song given a set of audio features.

We are going to create and evaluate a baseline model, train a linear regression model via self-implemented gradient descent, explore MLlib from pyspark to use stochastic gradient descent, and add interactions between features for better feature selection.

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