Homework 2 Programming Assignment

10-605/10-805: Machine Learning with Large Datasets

Due Tuesday, September 29th at 1:30:00 PM Eastern Time

1 Introduction

This assignment involves understanding the basics of building machine learning pipelines and techniques in training linear regression models. The assignment will also involve using principal component analysis (PCA) and feature-based aggregation for exploratory data analysis.

This assignment consists of two major parts. The first part is to train a linear regression model to predict the release year of a song given a set of audio features. The second part of the assignment is to apply PCA to a light-sheet imaging dataset and complete a feature-based aggregation to find a moving visual pattern of neural activity.

2 Logistics

We provide the code template for this assignment in *two* Jupyter notebooks. What you need to do is to follow the instructions in the notebooks 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 notebooks `'hw2_part1.ipynb'` and `'hw2_part2.ipynb'` after downloading and unzipping `hw2.zip` at https://github.com/10605/released-hws/archive/f20-hw2.zip.

Next, import the notebooks into your Databricks account, 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. You can individually submit a notebook for debugging but **make sure to submit both notebooks for your final submission to receive full credit.**

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 homework files and fill in your answers again and resubmit.

2.3 Submission

1. Export both solution notebooks as IPython notebook files on Databricks via `File -> Export -> IPython Notebook`
2. Submit both completed notebooks via Gradescope (you can select both notebooks when uploading your solutions).

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 September 4th. 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. Create a cluster: `Clusters -> Create Cluster`. You can use any cluster name as you like. When configuring your cluster, make sure to choose **runtime version** `6.6`. Note: It may take a while to launch the cluster, please wait for its status to turn to '`active`' before start running.
4. Installing third-party packages that will be used in the homework on Databricks: `Clusters -> Cluster name -> Libraries -> Install New`. Then select `PyPI`, enter the package name as `nose`. Finally click `Install` to install it.
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. Make sure to install nose.

4 Linear Regression on the Million Song Dataset

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.

In this section, you will be implementing a common supervised learning pipeline, using a subset of the Million Song Dataset from the UCI Machine Learning Repository, to train a linear regression model to predict the release year of a song given a set of audio features.

In this part, we will cover

- Part 1: Reading and parsing the Million Song dataset
- Part 2: Creating and evaluating a baseline model
- Part 3: Training (via gradient descent) and evaluating a linear regression model
- Part 4: Training using SparkML and tune hyperparameters via grid search
- Part 5: Adding interactions between features

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

5 Principal Component Analysis (PCA)

This section delves into exploratory analysis of neuroscience data, specifically using principal component analysis (PCA) and feature-based aggregation. We will use a dataset of light-sheet imaging recorded by the Ahrens Lab at Janelia Research Campus.

Our dataset is generated by studying the movement of a larval zebrafish, an animal that is especially useful in neuroscience because it is transparent, making it possible to record activity over its entire brain using a technique called light-sheet microscopy. Specifically, we'll work with time-varying images containing patterns of the zebrafish's neural activity as it is presented with a moving visual pattern. Different stimuli induce different patterns across the brain, and we can use exploratory analyses to identify these patterns.

In this section you will learn about PCA, and then compare and contrast different exploratory analyses of the same data set to identify which neural patterns they best highlight.

In this part, we will cover:

- Part 1: Working through the steps of PCA on a sample dataset
- Part 2: Writing a PCA function and evaluating PCA on sample datasets
- Part 3: Parsing, inspecting, and preprocessing neuroscience data then perform PCA
- Part 4: Feature-based aggregation and PCA

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