Python Data Products

Course 4: Implementing and Deploying data-driven predictive systems

Lecture: Description of Capstone tasks

Learning objectives

In this lecture we will...

Describe the various components and tasks of the capstone project

Capstone tasks

The Capstone project requires you to:

- 1. Harness your knowledge of machine learning, evaluation, and feature design
- 2. Implement four practical tasks on a realworld dataset

Capstone tasks

The dataset you are given is a relatively large set of Amazon **musical instrument** reviews

- The code stub already separates the data into "training" and "test portions"
- You have to implement techniques that lead to good performance on the test set, but which are based only on data from the training set
- In other words, you have to implement your own training/validation/test pipeline

Capstone tasks

The four tasks consist of the following:

- 1. A basic **data processing** task
- 2. A **Classification** task
- 3. A **Regression** task
- 4. A Recommender Systems task

Task 1: Data processing

For the **data processing task**, you must implement a variety of simple functions to compute basic statistics of the data. E.g.:

- How many unique users are there in the dataset?
- What is the average rating?
- What fraction of reviews are verified?
- For each of these tasks you must fill in a function stub

Task 2: Classification

For the classification task, you must predict whether an Amazon review corresponds to a verified purchase.

- This is a binary classification task
- The main challenge in this task is that the data are imbalanced i.e., most reviews correspond to verified purchases
- Thus we will use a balanced evaluation metric (the BER)
- So, you must select classification techniques that lead to good performance on this evaluation metric
- Your method should beat a simple baseline that performs logistic regression based on the length and rating of the review

Task 3: Regression

For the **regression task**, you must use word features (and other features) to perform **sentiment analysis**

- This is a regression task (rating prediction)
- The main challenge in this task is to properly avoid overfitting, since you will be using high-dimensional features
- To do this, you will have to carefully implement a train/validation/test pipeline
- You will also have to carefully engineer your features to consider the different choices when pre-processing text
- You should beat a simple baseline that considers the 100 most popular words only

Task 4: Recommendation

For the **recommendation task**, you must predict ratings that users will give to items

- This is a recommender systems task (predict a rating given a user and an item)
- The main challenge in this task is to correctly implement a complex model
- Again you will have to be careful about overfitting, as well as initialization
- Your solution should outperform a simple bias-only model

Capstone project: evaluation

For all tasks, your goal is to beat the baselines given **on the test set**

- Beating these solutions on the training set should be easy you can make small modifications to the existing techniques.
- However these performance gains may not translate well to the test set unless you are careful about **overfitting** and correctly implementing a validation pipeline
- To do so will require leveraging several ideas from throughout this Specialization

Summary of concepts

Introduced the Capstone project for Course 4