Intro to Kaggle Classroom

Vinay Bhaip

September 2019

1 Introduction

Kaggle Classroom is the platform we will be using this year to host machine learning competitions. To ease the learning curve and encourage as many people as possible to compete, this lecture will help set you up with how to use Kaggle Classroom.

2 Setting Up

The steps for getting set up are pretty simple, and we make it easy to start coding right away.

1. Create a Kaggle account if you don't already have one by clicking "sign up" in the top right.
2. Click on the competition link, which will be posted in the lecture table on this page. Just to make sure you're able to submit something for practice, we've created a sample Kaggle Classroom competition at https://www.kaggle.com/c/tj-ml-sample-competition.
3. Download the training and testing data.
4. Download any shell code from the website. For nearly all competitions, we'll provide you shell code so that you can focus primarily on making the model.

3 Creating a Model

After downloading the training and testing data, the next step is to actually write your algorithm and train it on the training data. The shell code we provide will handle reading in the training and testing data so all you have to do is make the model and run it on the testing data.

When you run it on the testing data, you'll need to create a submission file with the predictions in the format shown in the sample submission file. It's important to note here that the format has to match exactly, otherwise Kaggle

won't accept your file. Generally, we ask for the output file to be in the CSV format with the first column corresponding to the test data number and the remaining columns corresponding to your predictions.

4 Submitting Solutions

All you have to do is upload the CSV file you generated to Kaggle and you're done! You'll be able to submit multiple submission files if you want to try to improve your algorithm. After a few seconds, you'll see how you rank on the public leaderboard. It's important to note that the public leaderboard displays how well you performed on a subset of the testing data. The reason this is done is to prevent people from randomly changing the output file to maximize their score. This encourages people to create the best generalized algorithm. Once the competition ends, the private leaderboard will display the scores for the remaining testing data.

Your ranking for the public leaderboard and the private leaderboard might change, since your score is dependent on different data. While it's disappointing to see how you might jump from the top of the leaderboard to way below, it's important to understand that this happens all the time in real Kaggle competitions with thousands of dollars the line.