Open in Colab

T81-558: Applications of Deep Neural Networks

Module 8: Kaggle Data Sets

- Instructor: Jeff Heaton, McKelvey School of Engineering, Washington University in St. Louis
- For more information visit the class website.

Module 8 Material

- **Part 8.1: Introduction to Kaggle** [Video] [Notebook]
- Part 8.2: Building Ensembles with Scikit-Learn and Keras [Video] [Notebook]
- Part 8.3: How Should you Architect Your Keras Neural Network: Hyperparameters [Video] [Notebook]
- Part 8.4: Bayesian Hyperparameter Optimization for Keras [Video] [Notebook]
- Part 8.5: Current Semester's Kaggle [Video] [Notebook]

Part 8.1: Introduction to Kaggle

Kaggle runs competitions where data scientists compete to provide the best model to fit the data. A simple project to get started with Kaggle is the Titanic data set. Most Kaggle competitions end on a specific date. Website organizers have scheduled the Titanic competition to end on December 31, 20xx (with the year usually rolling forward). However, they have already extended the deadline several times, and an extension beyond 2014 is also possible. Second, the Titanic data set is considered a tutorial data set. There is no prize, and your score in the competition does not count towards becoming a Kaggle Master.

Kaggle Ranks

You achieve Kaggle ranks by earning gold, silver, and bronze medals.

- Kaggle Top Users
- Current Top Kaggle User's Profile Page
- Jeff Heaton's (your instructor) Kaggle Profile

- Current Kaggle Ranking System

Typical Kaggle Competition

A typical Kaggle competition will have several components. Consider the Titanic tutorial:

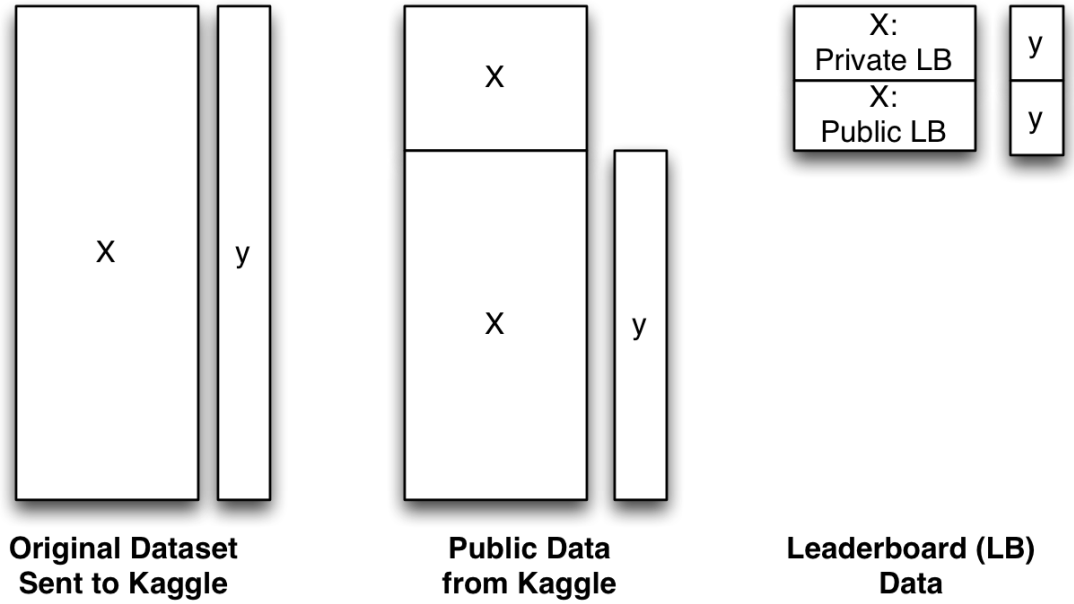
- Competition Summary Page
- Data Page
- Evaluation Description Page
- Leaderboard

How Kaggle Competition Scoring

Kaggle is provided with a data set by the competition sponsor, as seen in Figure 8.SCORE. Kaggle divides this data set as follows:

- **Complete Data Set** - This is the complete data set.
  - **Training Data Set** - This dataset provides both the inputs and the outcomes for the training portion of the data set.
  - **Test Data Set** - This dataset provides the complete test data; however, it does not give the outcomes. Your submission file should contain the predicted results for this data set.
    - **Public Leaderboard** - Kaggle does not tell you what part of the test data set contributes to the public leaderboard. Your public score is calculated based on this part of the data set.
    - **Private Leaderboard** - Likewise, Kaggle does not tell you what part of the test data set contributes to the public leaderboard. Your final score/rank is calculated based on this part. You do not see your private leaderboard score until the end.

Figure 8.SCORE: How Kaggle Competition Scoring



Preparing a Kaggle Submission

You do not submit the code to your solution to Kaggle. For competitions, you are scored entirely on the accuracy of your submission file. A Kaggle submission file is always a CSV file that contains the **Id** of the row you are predicting and the answer. For the titanic competition, a submission file looks something like this:

```
PassengerId,Survived
892,0
893,1
894,1
895,0
896,0
897,1
...
```

The above file states the prediction for each of the various passengers. You should only predict on ID's that are in the test file. Likewise, you should render a prediction for every row in the test file. Some competitions will have different formats for their answers. For example, a multi-classification will usually have a column for each class and your predictions for each class.

Select Kaggle Competitions

There have been many exciting competitions on Kaggle; these are some of my favorites. Some select predictive modeling competitions which use tabular data include:

- Otto Group Product Classification Challenge
- Galaxy Zoo - The Galaxy Challenge
- Practice Fusion Diabetes Classification
- Predicting a Biological Response

Many Kaggle competitions include computer vision datasets, such as:

- Diabetic Retinopathy Detection
- Cats vs Dogs
- State Farm Distracted Driver Detection

Module 8 Assignment

You can find the first assignment here: assignment 8

In [ ]: