Monday: Crafting Training Sets

Agenda

5 min: Overview

55 min: Suggested Readings

60 min: Exercise

Specific Learning Outcomes

- I can understand the need for training and validation data sets
- I can understand the ethical implications and challenges that come from training models
- I can leverage EDA techniques in creating good data sets.
- I can use relevant sklearn functions to create good train/test data sets.

Overview

As mentioned in the intro, the first step in our predictive analytics work is to identify our test and training data sets. In

this section, we will define key concepts, and run you through a few exercises on how to use sklearn to achieve this.

Main Resources

Understanding your variables:

First, you must analyze your variables, and determine which variable you want your model to **predict - we will refer to it as the dependent variable.**

Next, you must establish which other variables will help you predict your dependent variable. These will be referred to as **independent variables.**

It is important to perform exploratory data analysis to identify if there is a **relationship between your dependent and independent variables.** This does not mean that your independent variable **causes** the dependent one, just that they are connected.

For example, if we have a dataset on students, we may find variables such as student height, mock exam results, and national exam results. Plotting mock exam results against national exam results, you will see them to roughly take the shape of a line, which makes intuitive sense: Students who do poorly in the mock are likely not to be ready for the national exam, and vice versa.

Plotting height against national exam results will probably lead to a much more scattered plot, indicating that there isn't a strong relationship between height and academic performance.

Therefore, as we create our training and testing set to predict national exam performance, we will want to include mock exam performance, but not height.

Why do we need two sets?

This is where the machine learning actually happens: The training set includes data on your dependent variable, alongside all independent variables you choose to include. Your supervised learning algorithm will then go through this data set and for a given row try to **predict** what the dependent variable should be given the independent ones, then adjust its understanding of the process based on how good its **prediction** was. Over time, your algorithm will get really good at recognizing the patterns in your data set.

Why do we need the **test** set then? Well, the **test** set is not used for training, but to validate how good the model you've created is at predicting the desired dependent variable.

Later this week, we will explore **ethical considerations** when creating train and test datasets. Remember this though: Your predictive model is only as good as the data you've used to train it. There have been many challenges with training, the reading

and exercises below will run you through ways to deal with them.

Suggested Readings

- How do you know you have enough data? [link (https://towardsdatascience.com/how-do-you-know-you-have-enough-training-data-ad9b1fd679ee)]
- Google's image recognition fiasco - bring diversity to your data! [link (https://www.theverge.com/2018/1/12/16882408/google-racist-gorillas-photo-recognition-algorithm-ai)]

Exercise

- Crafting good training sets in python. [Link (https://colab.research.google.com/drive/1NToj9e-YwP7fakXQq51lyV0d8uK9hpeg)]

"Machine learning will automate jobs that most people thought could only be done by people." ~ Dave Waters