Tuesday: Linear Regression Optimization - Cross Validation

Agenda

5 min: Overview

45 min: Suggested Readings

60 min: Exercise

Specific Learning Outcomes

- I can use cross-validation methods to assess the quality of their models.
- I can understand how to apply cross-validation, irrespective of the model of choice
- I can apply variations of k-folds cross-validation, namely, Leave one out cross-validation, and repeated k-folds.

Main Resources

As mentioned yesterday when talking about training and test datasets, we have to be careful of issues of sampling and biases in the data. We will always look for larger datasets that are more diverse to strengthen our models, but there are some techniques that can help us make the most of the data we have. The first such technique is called **k-fold** cross-validation.

Whereas we typically simply split our data into a training set and a test set, k-fold cross validation will split our data into **n** separate chunks called folds. Typically, we use 5 or 10 folds. This helps us identify potential biases and challenges by using each fold in turn as the test set.

For example, if we have 5 folds, we will train 5 models. The first model will use fold 1 as the test set and folds 2-5 as the training set. Model 2 will use fold 2 as its test set, and the remaining 4 folds as training, etc.

We then assess the accuracy of each model, looking for outliers. If the first 4 models were ~85% accurate, but model 5 was only 50% accurate, that tells us there is something special about that particular split of test and training data. Typically, we will want to see all models performing roughly the same, which will give us a sense of **the average accuracy of a model trained using the data available to us.**

Bear in mind that k-fold cross-validation can apply to **any and all supervised learning algorithms,** not only linear regression. We introduce it this early because it is a very powerful tool to add to your arsenal. The exercises below will run you through how to implement k-fold cross-validation and its variations.

Suggested Readings

- K-fold Introduction. [**link** (https://machinelearningmastery.com/k-fold-cross-validation/)]
- Advanced: Speeding up cross-validation. [**link** (https://www.thekerneltrip.com/machine/learning/speeding-up-cross-validation/)]

Exercise

- Cross-validation practice and challenges. [**link** (https://colab.research.google.com/drive/1fib16iCJrcTRpypNZoXNAg3j1UXElRMh?usp=sharing)]

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