Machine Learning with Spark

What is Spark?

- Data processing program
- NOT a language

- Can be used in Hadoop and with other platforms

- 100x faster than MapReduce

- Use with: Scala, Python, or Java

Components of Spark

- Spark 1.0 (Core)
- Spark Streaming: real-time
- Spark SQL
- Spark 2.0 (MLLib)
- GraphX: social networks

Spark Data Storage

- Resilient Distributed Datasets (RDDs)
  - Spark 1.0
  - Store data across the cluster
  - Slow!

- DataSets
  - Like an RDD, but faster

- DataFrames
  - Even faster
  - For relational data

Why Run Spark in Scala?

- Spark was written in Scala
- More up-to-date changes
- Better understanding and functionality
- Improved data efficiency

val ‘n var

val

- Value
- Cannot be changed
- More common

var

- Variable
- Changeable

Comments in Scala

// Commenting!

Feature Importance

- Features = Variables or Columns

- Which is the most important to the accuracy of the model?

- Weighting system

- The higher the better

What is a Hyperparameter?

- Components of a ML model

- By playing with them, you can get better model fit

Hyperparameters for Decision Trees

- Maximum Depth: Number of decisions it can make
- Maximum Bins: Number of decision rules it can have
- Impurity: How much category mix-up you'll allow
- Minimum Information Gain: Only keep things that will improve accuracy

What is a pipeline?

- Chaining operations together
- You don’t have to manually run code once created
- Does take computer time / processing power

Questions?