CS 199 ACC
More Spark

Prof. Robert J. Brunner

Ty Trauger
Osmar Coronel

MP 4

How was it?

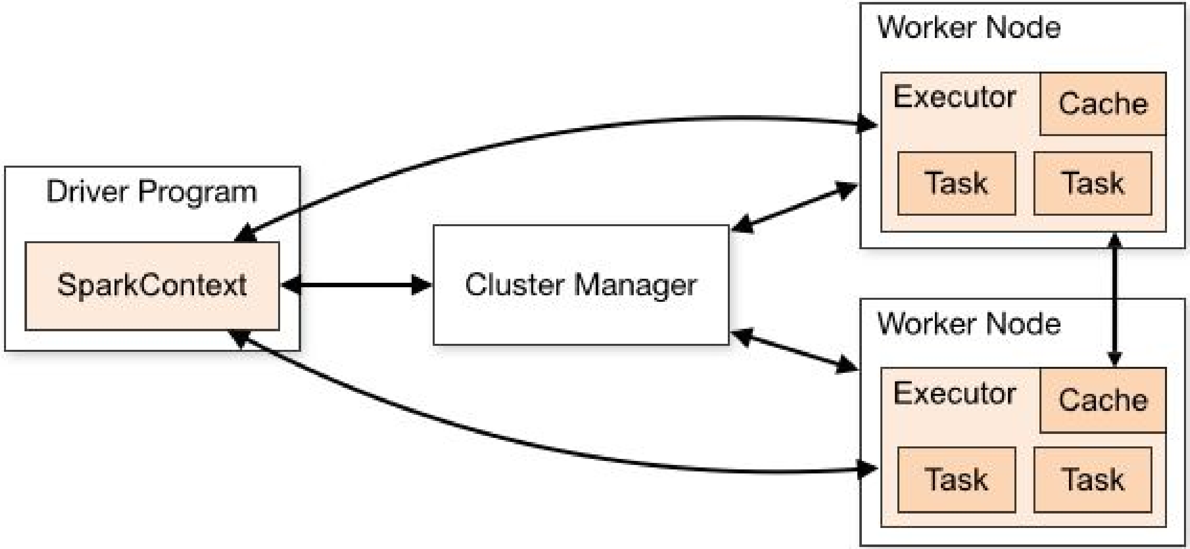
This Week

- Distributed Computation
- Apache Spark Adv.
- Some Hardware

Spark - Advanced

Behind the Scenes of Spark

- Driver
  - Manager of the executors
  - Only one
  - What is actually created when you run ‘spark-submit’
- Executor
  - Manager of the tasks
  - Executors take up cores
- Task
  - Runs a function
  - Think of as a thread



Auto-parallelization

- Why can Spark auto-parallelize your code, but GCC or LLVM cannot?

Auto-parallelization

- Why can Spark auto-parallelize your code, but GCC or LLVM cannot?
  - Immutability!
  - Due to LLVM or GCC not requiring that the data are immutable, they cannot predict what will necessarily happen next.
    - Think of a loop

Why map reduce?

- Again, for loops are mutable. For example:

```
arr = [0,1,2,3]

for i in arr:

    arr[i] = i**2
```

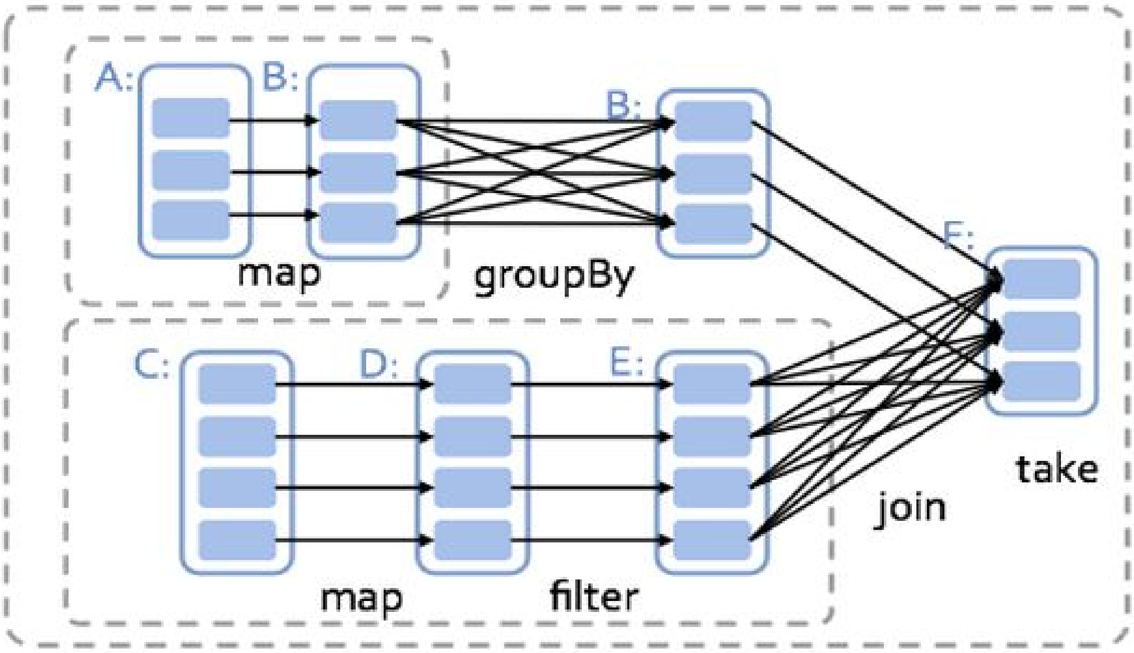
- This changes arr!

Why map reduce cont.

- A for loop can almost always be turned into a map
- Thus, by forcing a map reduce paradigm, you can force immutability while still allowing all of the things that can be done with a loop.

Spark Compiler

- Spark creates a graph of operations
- Divides operations into stages
- Each stage ends when a reduce/shuffle happens
- Auto-parallelization!



Distributed Machine Learning

The Options

mahout

Spark MLlib

TensorFlow

Machine Learning on Spark (MLlib)

- MLlib allows for distributed machine learning on very large datasets.
- Built on top of Spark so you can use it easily within Spark
- Designed to be similar in use to NumPy
- Can interoperate with NumPy and SciPy
- As of now, can only use RDD’s
    - no dataframes :(

Machine Learning Basics

What comes first?

Machine Learning Basics

What comes first?

Data, sparse and labeled

Machine Learning Basics

What comes first?

Data, sparse and labeled

How is the data represented?

Machine Learning Basics

What comes first?

Data, sparse and labeled

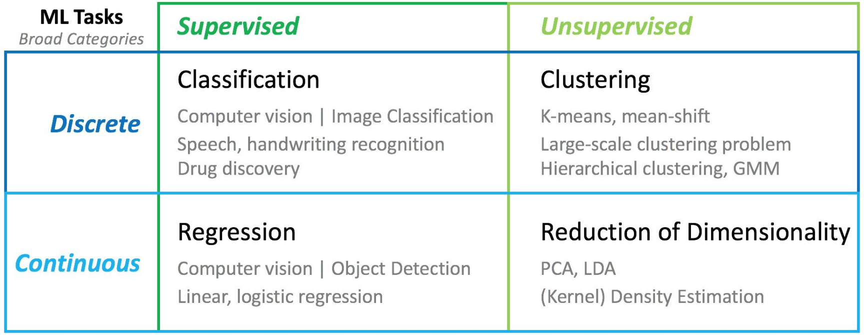
How is the data represented?

Continuous or Discrete? Supervised or Unsupervised?

Machine Learning Techniques

We will be covering three broad types of techniques:

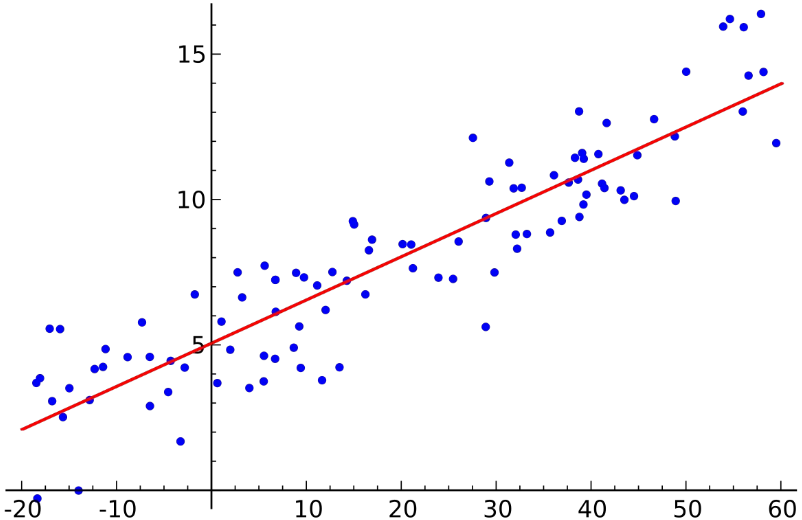
- Regression
  - Tries to predict an output given data (continuous)
- Classifiers
  - Takes data and try to assign it a label (discrete)
- Clustering
  - Don’t know labels or numbers.
  - Groups similar data points into a group (or ‘cluster’).



ML Tasks *Broad Categories*	***Supervised***	***Unsupervised***
Discrete	Classification Computer vision | Image Classification Speech, handwriting recognition Drug discovery	Clustering K-means, mean-shift Large-scale clustering problem Hierarchical clustering, GMM
Continuous	Regression Computer vision | Object Detection Linear, logistic regression	Reduction of Dimensionality PCA, LDA (Kernel) Density Estimation

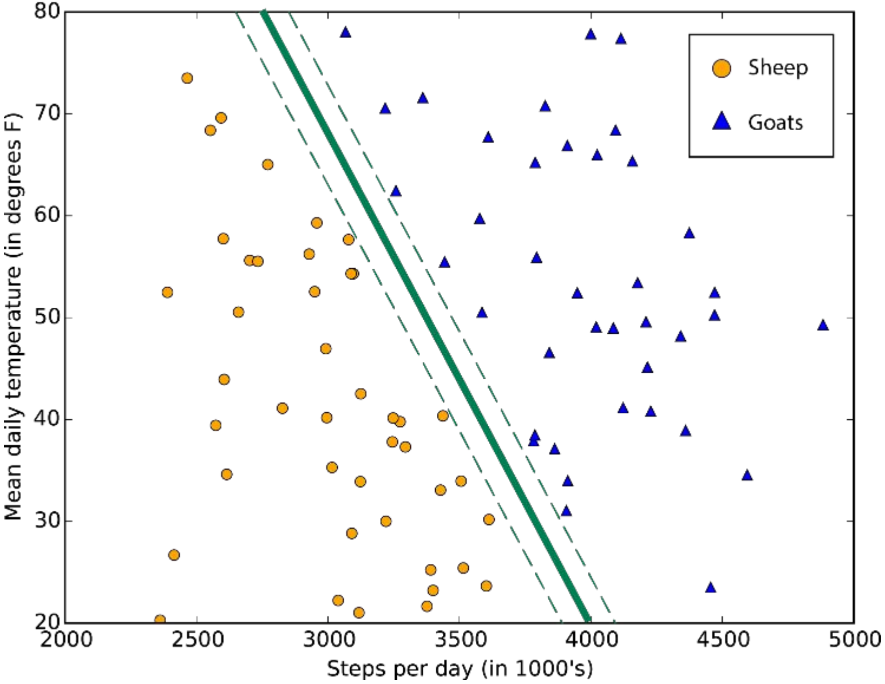
Regression

- Fits a function to your data.
  - For example, linear regression finds a line of best fit



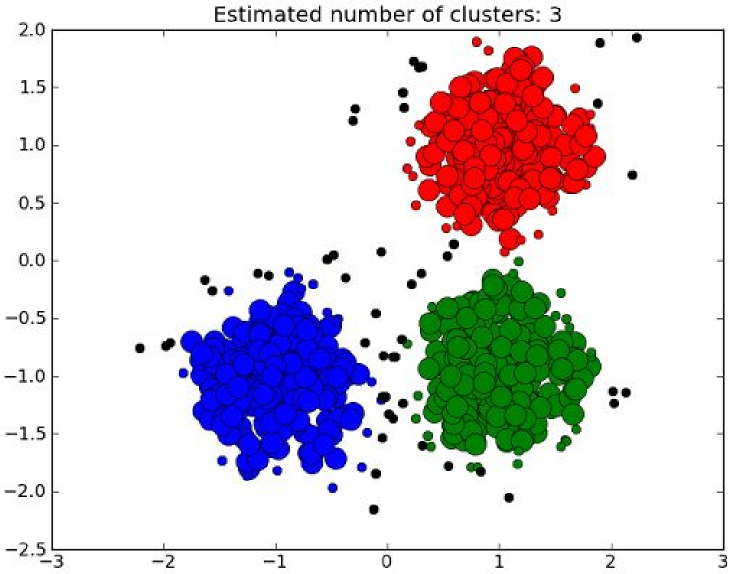
Classifiers

- Takes data and assigns them a label based on what it is ‘closest’ to.
- Supervised



Clustering

- Unsupervised; used when there are no labels
- The algorithm determines the clusters

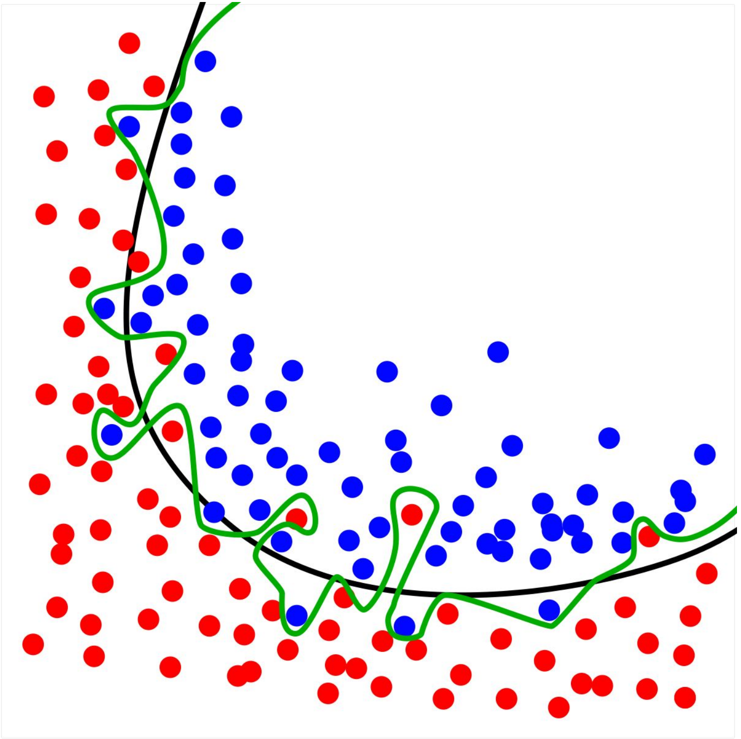


How Do I Know If My Model Is Any Good?

- Check your data and clean it up!
    - Good models only come from good data
    - Don’t Overfit!!
- Metrics
    - Precision, accuracy, area under ROC, true positive rate, root mean squared error, etc…
    - Lots of them, but we won’t have you worry with most of them

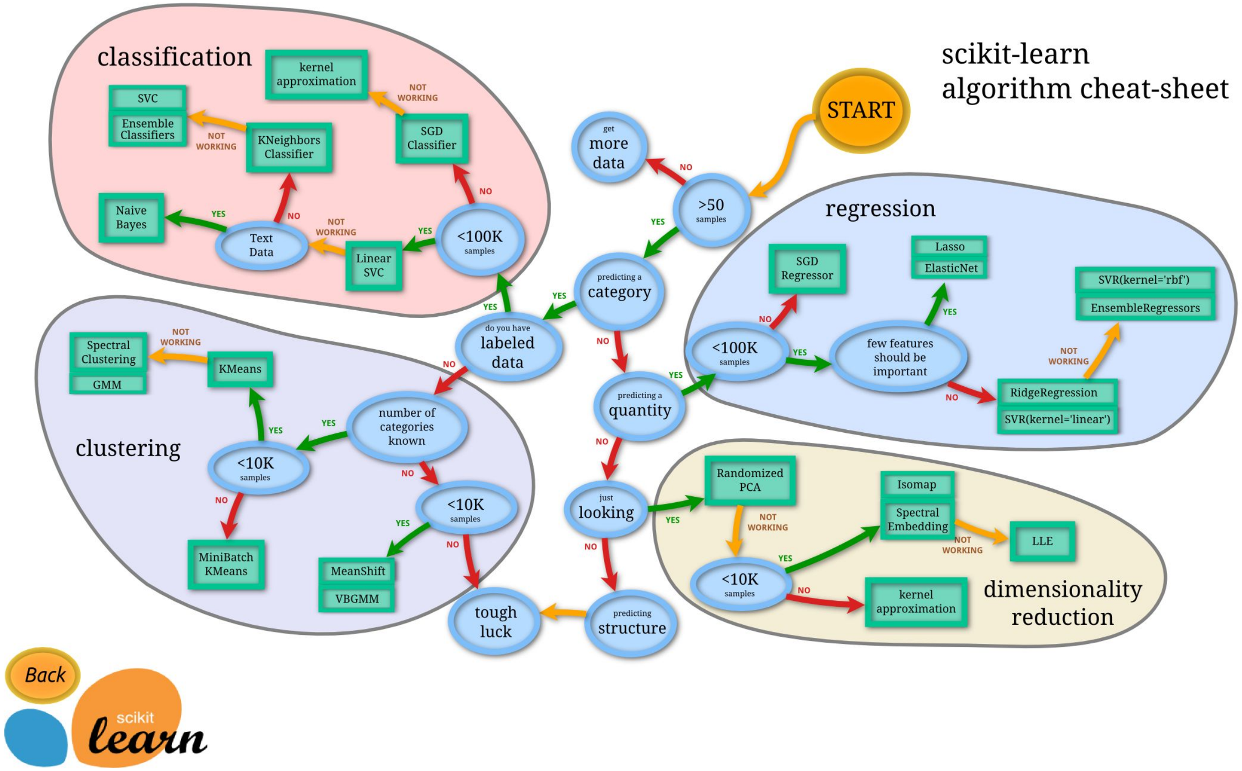
Overfitting

- When your model is too good
- Happens when your model ‘learns’ random noise in your training data.



Useful Cheat Sheet

Note: Since we are using MLlib and not scikit-learn, the values are off; for a more accurate value multiply by ~100. Also, MLlib does not support all of the algorithms



Demo

- Basic Linear Regression demo

When to use MLlib?

- When your data is LARGE.
- When your task is not GPU intensive
  - A lot of machine learning benefits from a single GPU than 100 CPUs.

MP 5

Due in one week (10/18) at 11:55pm

Start it early

Warning!

- We **don't guarantee** the cluster uptime
- Even though Hadoop is scalable, reliable, and fault tolerant (all those buzzwords), imagine what would happen if all thirty of you tried to log on to the cluster and submit a huge mapreduce job at the deadline
    - Either the cluster crashes or it runs at a snail's pace.
- As with course policy, if it's late it is late.

Attendance

bit.ly/199attendance2