CS 398 ACC
Spark MLlib

Prof. Robert J. Brunner

Ben Congdon
Tyler Kim

MP5

How’s it going?

Final Office Hours: After this lecture // Tomorrow 4-6pm

- **Please avoid Low-Effort/Private Piazza post**

Final Autograder run:

- Tonight ~9pm
- Tomorrow ~3pm

- Due tomorrow at 11:59 pm.
- Latest Commit to the repo at the time will be graded.
- Last Office Hours today after the lecture until 7pm.

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

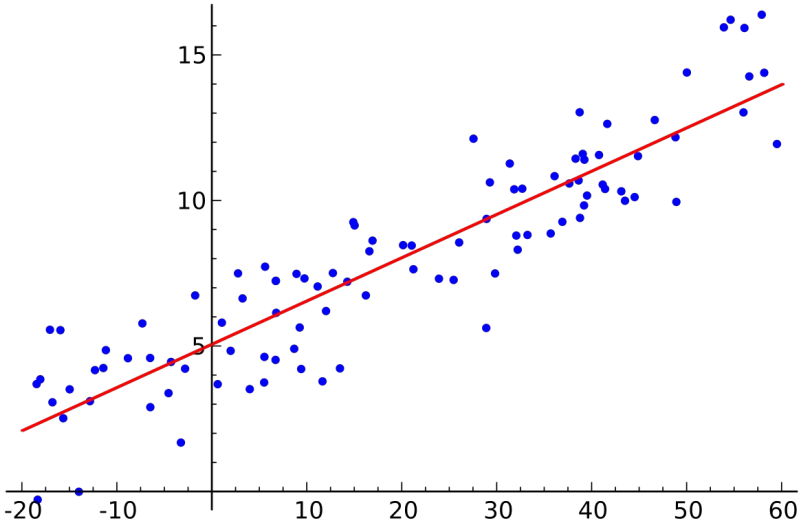
We will be covering three **broad** types of techniques:

- 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

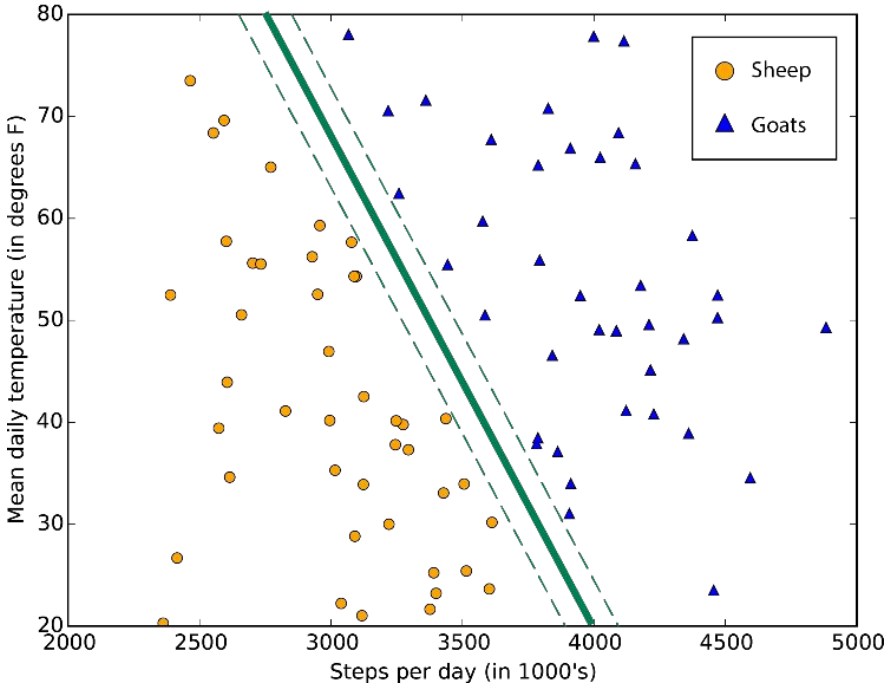
Regression

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



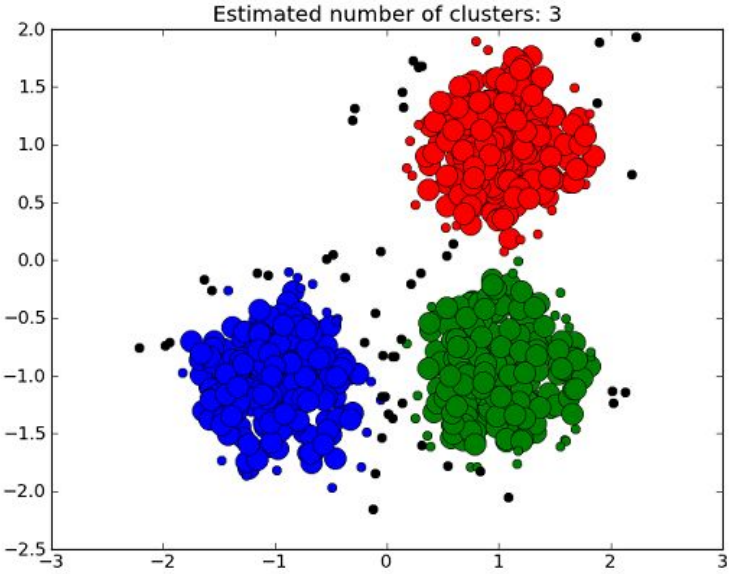
Classifiers

- 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

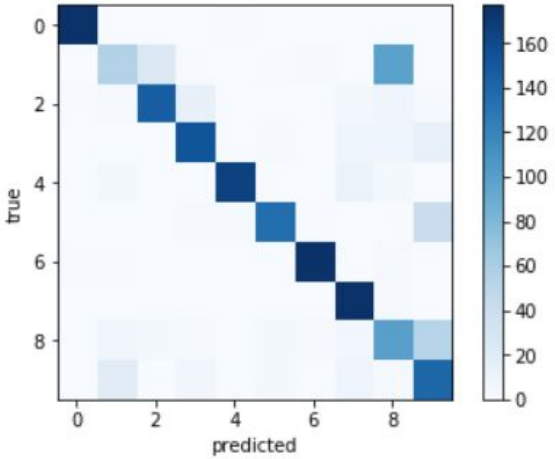


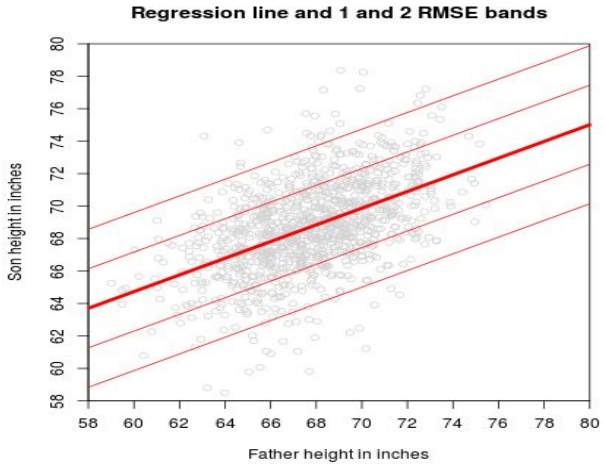
How Do I Know If My Model Is Any Good?

- 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...

Performance Metrics

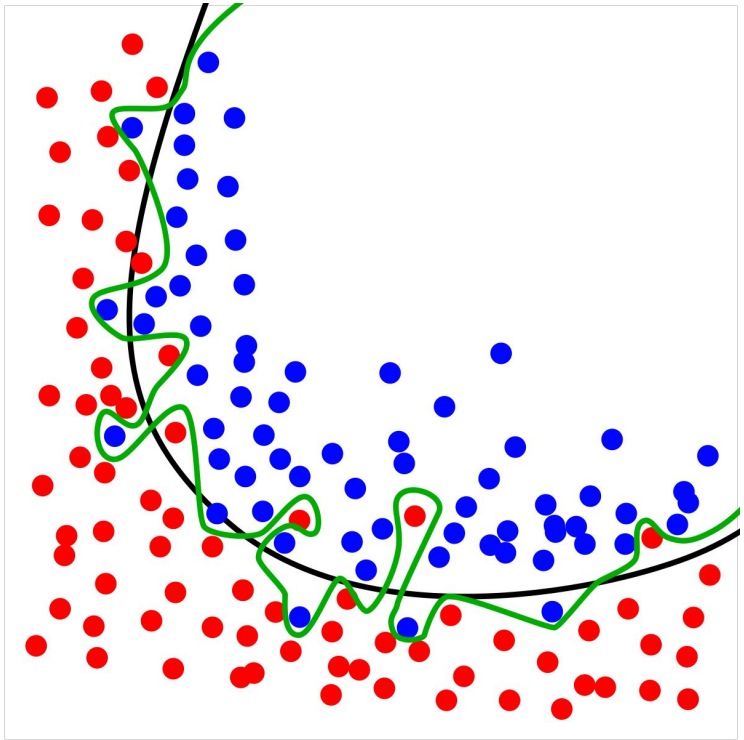
- Confusion Matrix
  - Useful for Classification
- RMSE - Root Mean Square Error
  - Useful for Regression





Overfitting

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



Improve Models with Data

- **Get More Data**
  - Invent, Simulate, Resample...
- **Transform Data**
  - Reshape the distribution, Rescale the data...
- **Feature Engineering**
  - Create and add new features
- **Clean Data**
  - Missing data handling, Reduce Noise...

Improving Models

Feature Selection

- Selecting features to improve the prediction model
    - Use when there are a lot of features (noise) and not enough data points
    - Sometimes adding more feature can also improve the model as it decrease bias.

To

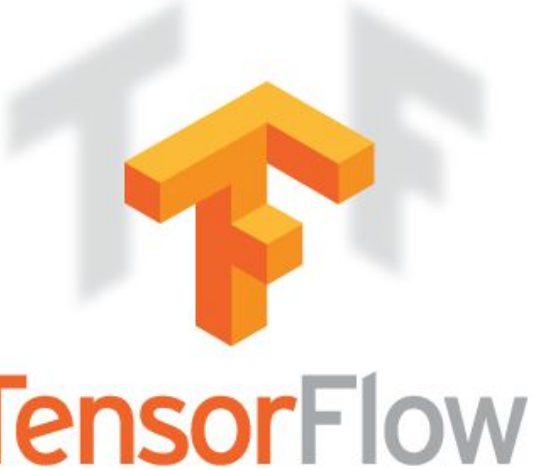
- **Reduce Overfitting**
- **Improve Accuracy**
- **Reduce overall Training**

Distributed Machine Learning

The Options

Apache Singa

MLlib



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

Machine Learning on Spark (MLlib)

- Can use RDDs or DataFrames
  - Unfortunately, they have slightly different feature sets...

- RDD API:
  - `pyspark.mllib.*`
  - Original API, now in “Maintenance Mode”

- DataFrame API:
  - `pyspark.ml.*`
  - Primary API for MLlib for Spark 2.0+
  - Support for ML “pipelines”
    - Less “glue” code necessary

When to use MLlib?

- When your data is LARGE
- To work with the Spark Ecosystem
- Real-Time Machine Learning (with Spark Streaming)

Wednesday

Spark MLlib Demo + Office Hours

MP 6

Due in next next **Tuesday, March 13th (you have 2 weeks)** at 11:59pm

Topic: “Spark MLlib”

> Check Piazza for Q&A and Announcements