Data Science boot camp @NYU-CUSP

With Eric Schles

The abc’s of data science

It’s all about the data:

1. Getting the data
2. cleaning the data
3. analyzing the data
4. moving the data around (ETL)

Where does this data come from?

from the internet

from the database

from a file

Requests

http://docs.python-requests.org/en/latest/

--How to download html locally.

[demo here]

lxml

http://lxml.de/

--How to parse the data

http://thomaslevine.com/dada/web-sites-to-data-tables-in-depth/

--best intro to web scraping EVER

[demo here]

Web scraping review

http://www.baseball-reference.com/

--download and write the american league and national league tables to csv files

http://hoopdata.com/teamff.aspx

--download and write the table found on this page to a csv file

pandas

Pandas is an easy, versatile package for taking the best of R and Python and putting them together.

Deal with R style dataframes.

[demo]

Feature

Often, A feature is about taking a piece of human readable data or unstructured data and turning it into something that can be processed mathematically

Label

A label is the y values that the features from your data map to.

Naive Bayesian Classifier

$$p(C|F_1,\dots,F_n) = \frac{1}{Z} p(C) \prod_{i=1}^{n} p(F_i|C)$$

This classifier treats of the F[i]’s as independent and returns a probability that a given set of features for an object implies the object has a certain label.

[demo]

example

It’s about taking things like human language and turning them into mathematical equations and then deriving meaning from those equations.

[demo here]

modeling

Linear regression:

$$y_i = \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \varepsilon_i$$

The y[i]’s are linearly related to the vector of x[i]’s

Where x[i] is of length p.

[demo]

example

How much should you expect to spend on groceries next year?

Questions for the class:

what are some meaningful variables you might care about?

[demo]

Analysis

Pushing the data to a model

Determining which features of the data translate to labels (and how).

creating a visualization of the data

Analysis exercise

Take the potential features that you saved to a CSV and try to do some analysis. Specifically:

Use this data set: http://www.baseball-reference.com/teams/KCR/2014.shtml

To try and find the most valuable player on the team.

things to keep in mind

What are the features you want to use?

What are the labels you want to use?

What features don't really matter?

generalizing

Look at all the sports teams and try to determine how many good players there are on each team.

How does the number of good players affect the teams overall ranking?

Number of wins?

Is it better to have a few really good players or a lot of better than average players?