GENERAL ASSEMBLY

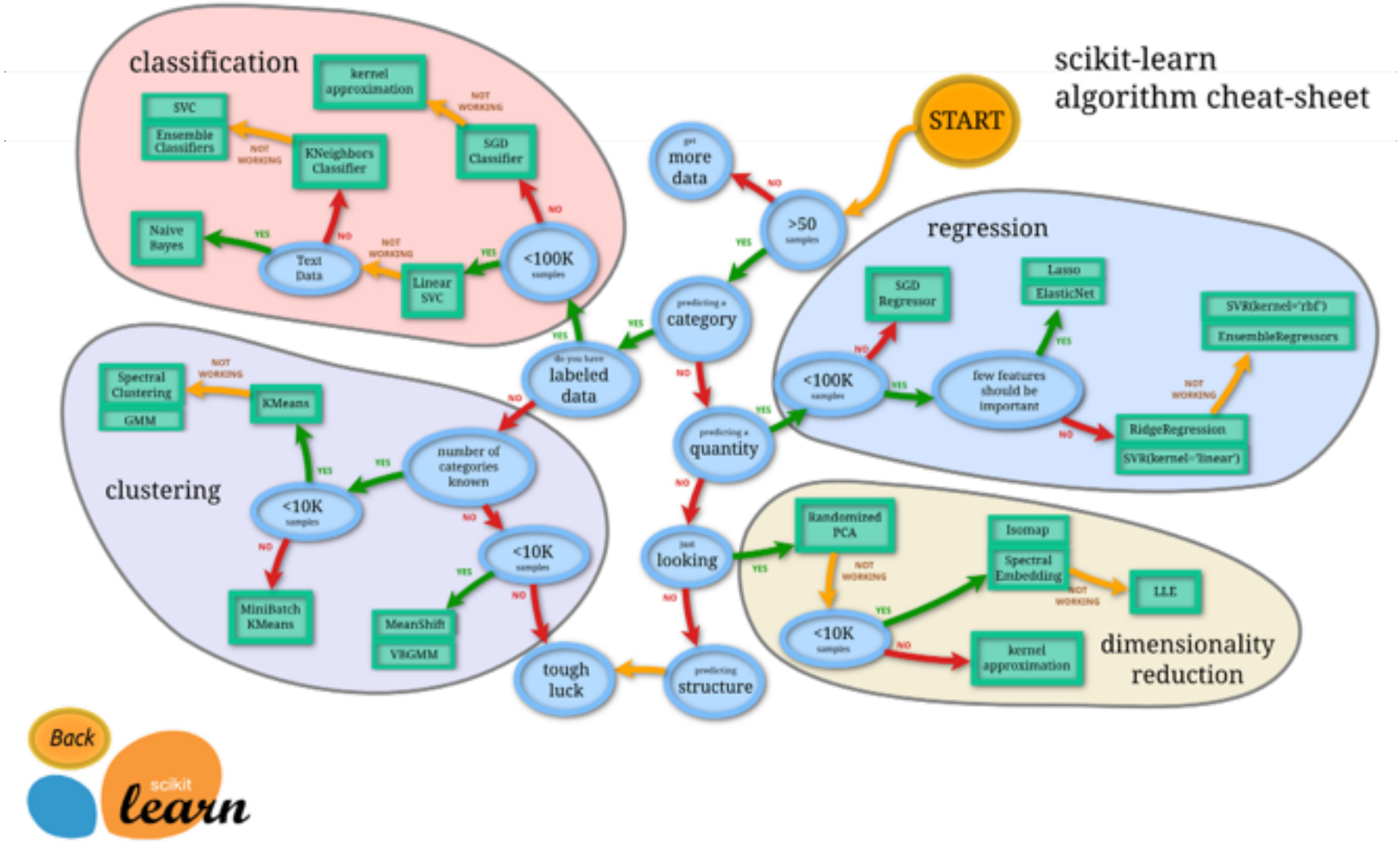
DATA SCIENCE
11 WEEK PART TIME COURSE

Week 5 – Dimensionality Reduction
Wednesday 27th April 2016

1. What is Dimensionality Reduction?
2. Why reduce dimensions?
3. What is Principal Component Analysis?
4. Lab
5. Real World Example
6. Discussion

DATA SCIENCE PART TIME COURSE

DIMENSIONALITY REDUCTION


scikit-learn
algorithm cheat-sheet
START
get more data
NO
>50 samples
YES
classification
kernel approximation
NOT WORKING
SVC
Ensemble Classifiers
NOT WORKING
KNeighbors Classifier
SGD Classifier
NO
Naive Bayes
YES
Text Data
NO
NOT WORKING
Linear SVC
YES
<100K samples
NO
regression
SGD Regressor
Lasso
ElasticNet
SVR(kernel='rbf')
EnsembleRegressors
<100K samples
NO
YES
few features should be important
YES
NOT WORKING
RidgeRegression
NO
SVR(kernel='linear')
YES
predicting a category
YES
do you have labeled data
NO
predicting a quantity
YES
NO
clustering
Spectral Clustering
GMM
NOT WORKING
KMeans
YES
number of categories known
YES
<10K samples
NO
NO
<10K samples
YES
MiniBatch KMeans
MeanShift
VBGMM
NO
just looking
YES
Randomized PCA
NOT WORKING
Isomap
Spectral Embedding
NOT WORKING
LLE
YES
<10K samples
NO
kernel approximation
dimensionality reduction
NO
predicting structure
tough luck
Back
scikit learn

A set of techniques for reducing the size (in terms of features, records, and/or bytes) of the dataset under examination.

In general, the idea is to regard the dataset as a matrix and to decompose the matrix into simpler, meaningful pieces.

- The number of features in our dataset can be difficult to manage, or even misleading (eg, if the relationships are actually simpler than they appear).

- Ideally, we would like to eliminate redundancy and consolidate the number of variables we're looking at.

- The complexity that comes with a large number of features is due in part to the curse of dimensionality.

- The complexity that comes with a large number of features is due in part to the curse of dimensionality.

- Namely, the sample size needed to accurately estimate a random variable taking values in a d-dimensional feature space grows exponentially with d (almost).

- (More precisely, the sample size grows exponentially with $l \leq d$, the dimension of the manifold embedded in the feature space).

- We'd like to analyze the data using the most meaningful basis (or coordinates) possible.

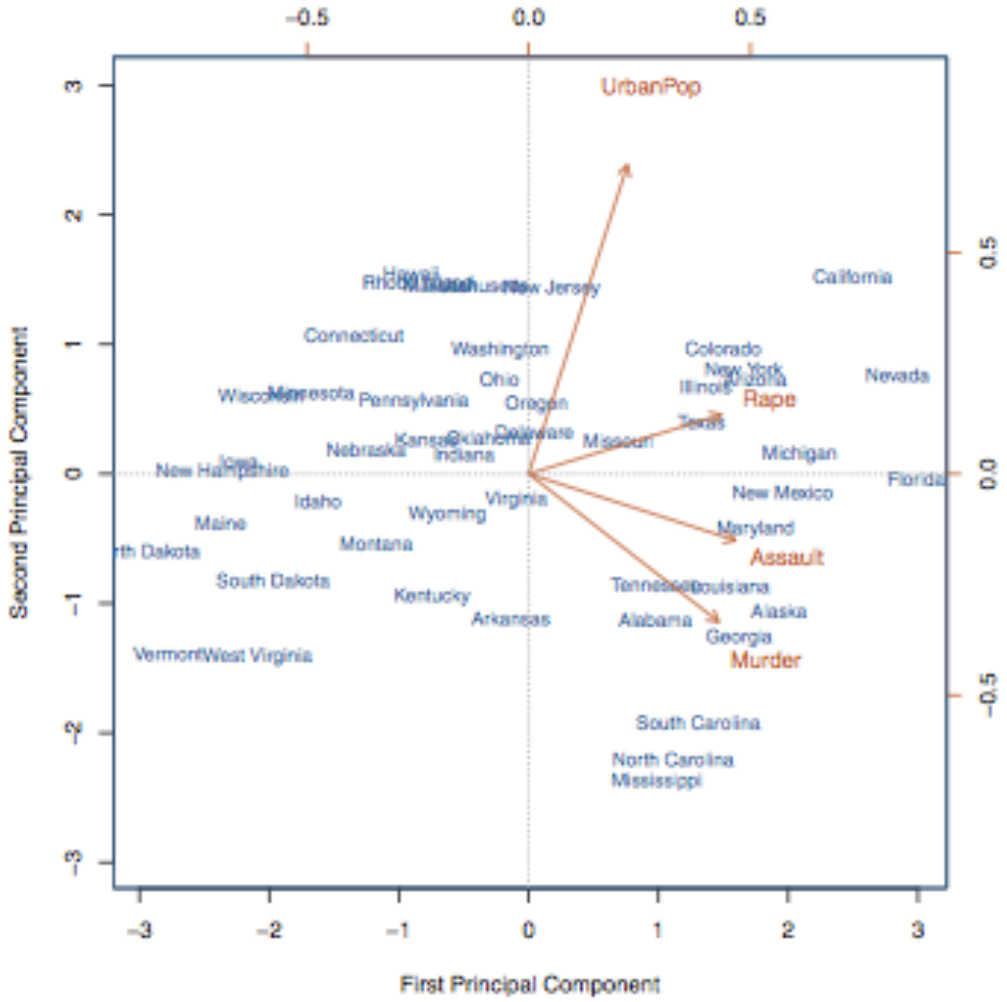
- More precisely: given an n x d matrix X (encoding n observations of a d-dimensional random variable), we want to find a k-dimensional representation of X ($k < d$) that captures the information in the original data, according to some criterion.

DATA SCIENCE PART TIME COURSE

PRINCIPAL COMPONENTS

'It finds a low-dimensional representation of a data set that contains as much as possible of the variation. The idea is that each of the n observations lives in p-dimensional space, but not all of these dimensions are equally interesting. PCA seeks a small number of dimensions that are as interesting as possible, where the concept of interesting is measured by the amount that the observations vary along each dimension. Each of the dimensions found by PCA is a linear combination of the p features.'

- Introduction to Statistical Learning


UrbanPop
Rape
Assault
Murder
California
Nevada
Colorado
New York
Illinois
Texas
Michigan
Florida
New Mexico
Maryland
Alaska
Georgia
Louisiana
Alabama
Tennessee
Arkansas
Kentucky
South Carolina
North Carolina
Mississippi
Virginia
Missouri
Indiana
Nebraska
Kansas
Pennsylvania
Ohio
Oregon
Washington
Connecticut
New Jersey
Wyoming
Idaho
Maine
Montana
South Dakota
Vermont
West Virginia
New Hampshire
Second Principal Component
First Principal Component

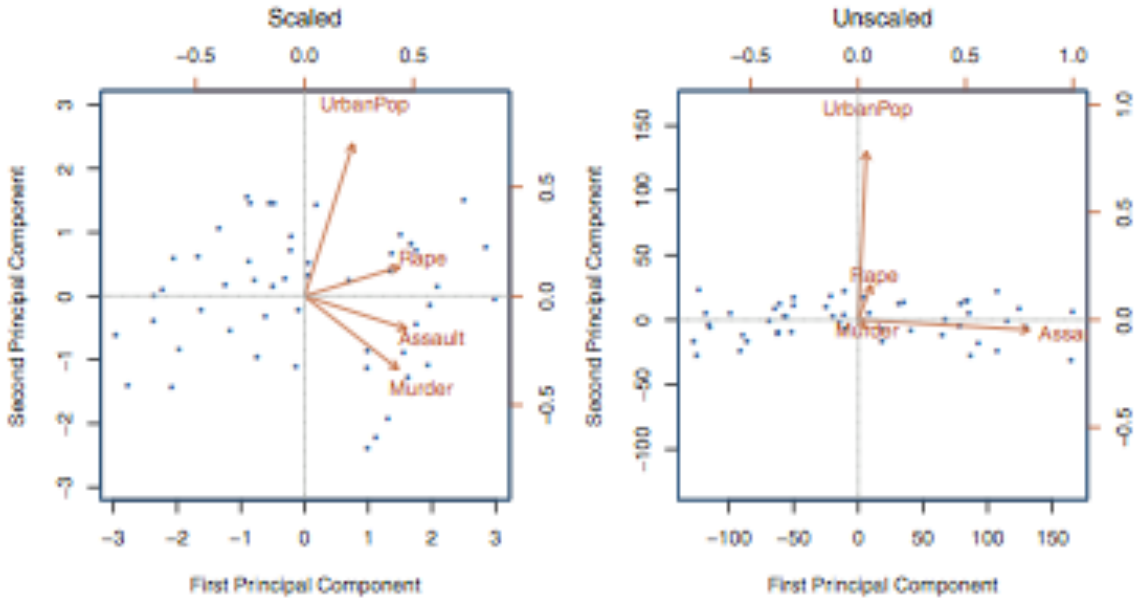
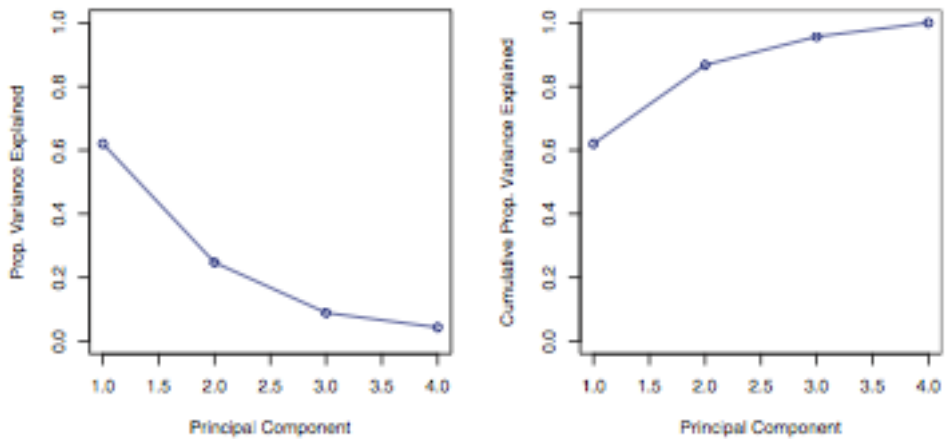


FIGURE 10.3. *Two principal component biplots for the* `USArrests` *data.* Left: *the same as Figure 10.1, with the variables scaled to have unit standard deviations.* Right: *principal components using unscaled data.* `Assault` *has by far the largest loading on the first principal component because it has the highest variance among the four variables. In general, scaling the variables to have standard deviation one is recommended.*


Prop. Variance Explained
0.0
0.2
0.4
0.6
0.8
1.0
1.0
1.5
2.0
2.5
3.0
3.5
4.0
Principal Component
Cumulative Prop. Variance Explained
0.0
0.2
0.4
0.6
0.8
1.0
1.0
1.5
2.0
2.5
3.0
3.5
4.0
Principal Component

LAB

git remote -v

git remote add upstream https://github.com/ihansel/SYD_DAT_3.git

git remote -v

git fetch upstream

git checkout master

git merge upstream/master

OR git reset –hard upstream/master

DISCUSSION TIME

- Real Problems
- Homework & Readings
- Project & Hackday
- (Another) Survey

DATA SCIENCE – Week 6 Day 1

REAL PROBLEMS

DATA SCIENCE – Week 6 Day 1

HOMEWORK

Two Parts

- Part 1: Pick one data science topic that interests you, could be an article, or how a Kaggle competition was won, or an article incorporating data journalism. Write a 5 minute presentation with 5 slides. You should explain, what was done, the motivation behind it, the results and what you can think of that would improve the project.
- Part 2: Project work

Readings

Read the following

- Chapter 8 of Introduction to Statistical Learning, Tree Based Methods (30 pages)

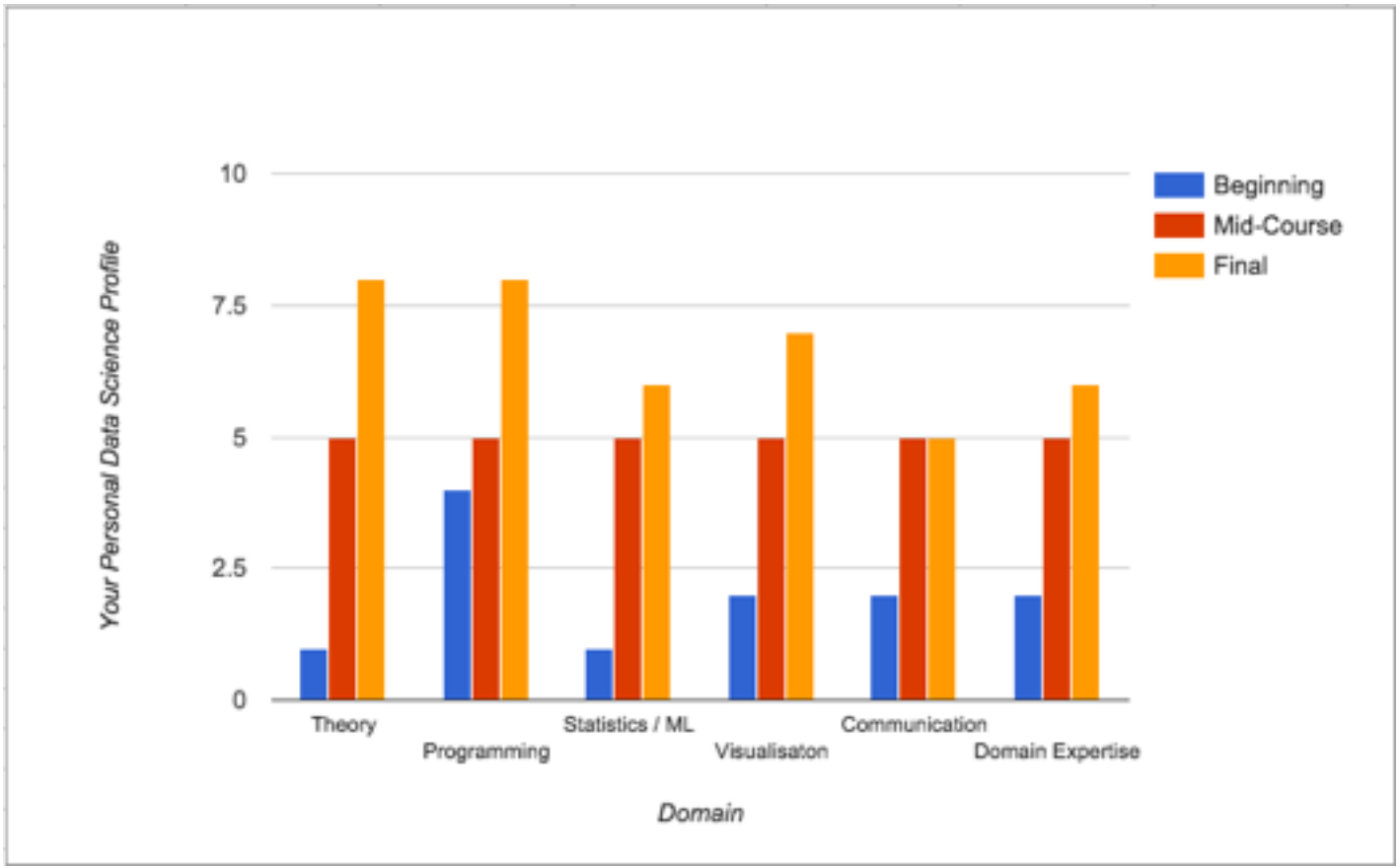
DATA SCIENCE – Week 6 Day 1

Project

- I will be checking progress over the weekend and making some recommendations and asking questions.
- Have most recent progress up on github by Saturday

Hackday

- Opportunity to work on project, cover any material you want to review, & maybe a few bonus topics ?????
- Who's (really) keen ?
- When suits everyone ?
- Saturday 7th May 2016


10
7.5
5
2.5
0
Your Personal Data Science Profile
Beginning
Mid-Course
Final
Theory
Programming
Statistics / ML
Visualisaton
Communication
Domain Expertise
Domain