Welcome to CS-304

A Data Mining Exploration

Objectives

- Exploring the facets of machine learning
- Data scientist check list
- Practical applications of converse inductive integrals in the context of epsilon

A Review

$$S_i^{(t)} = \{x_p : \|x_p - m_i^{(t)}\|^2 \le \|x_p - m_j^{(t)}\|^2 \ \forall\ 1 \le j \le k\},$$

$$m_i^{(t+1)} = \frac{1}{|S_i^{(t)}|} \sum_{x_j \in S_i^{(t)}} x_j$$

?????

Some Background

Introduction to Machine Learning

Introduction to ~~Machine~~ ~~Learning~~

Introduction to Data Mining

Depth versus Breadth

Types of problems with machine learning answers

Supervised versus Unsupervised

THUNDERBOLT
LABS

@bryanl

@thunderboltlabs

Required Knowledge

Math

Papers

Persistence

Regression

Linear regression

$$y = \alpha + \beta x,$$

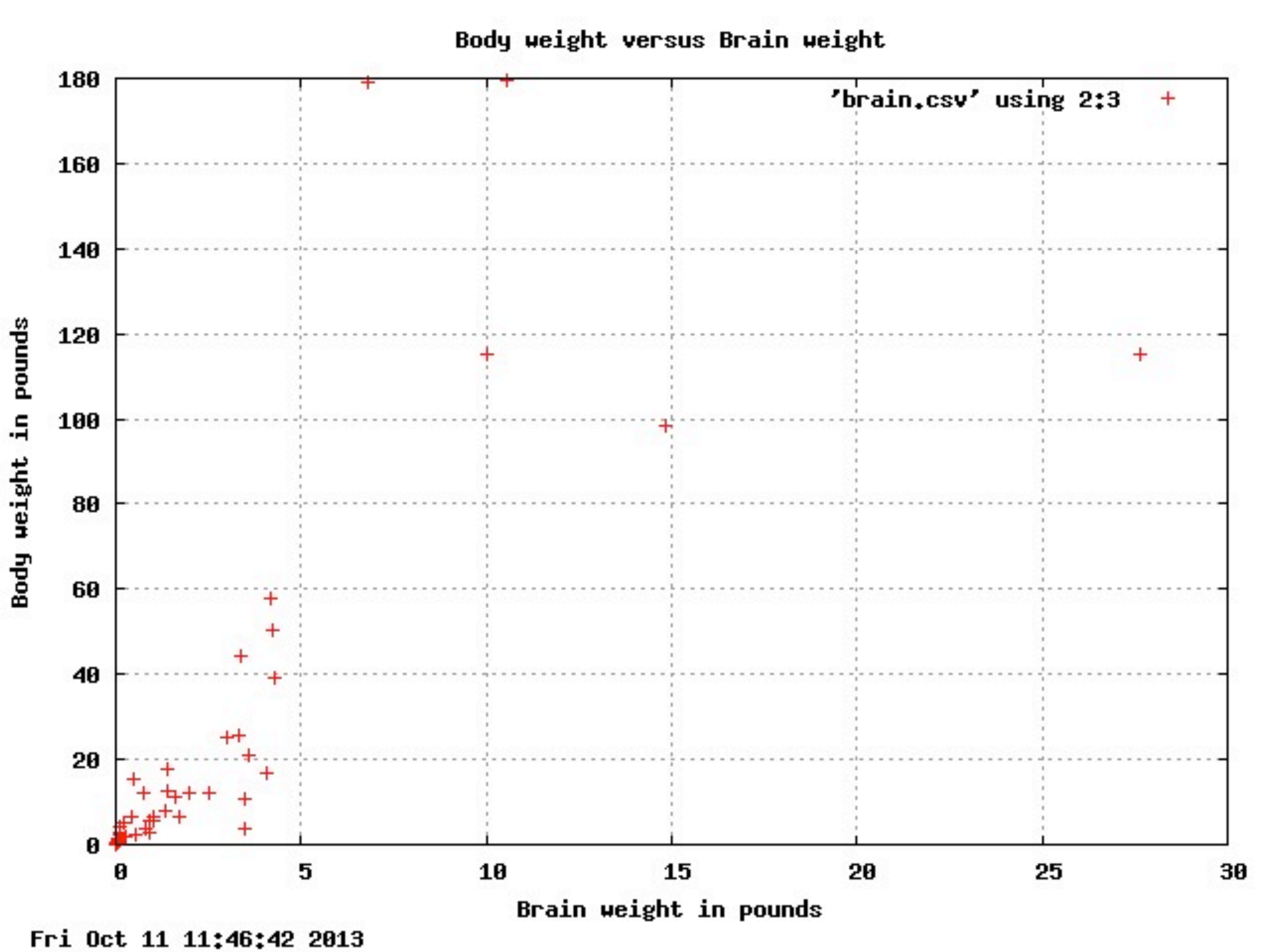
$$\text{Find } \min_{\alpha,\,\beta} Q(\alpha, \beta), \text{ where } Q(\alpha, \beta) = \sum_{i=1}^{n} \hat{\varepsilon}_i^{\,2} = \sum_{i=1}^{n} (y_i - \alpha - \beta x_i)^2$$

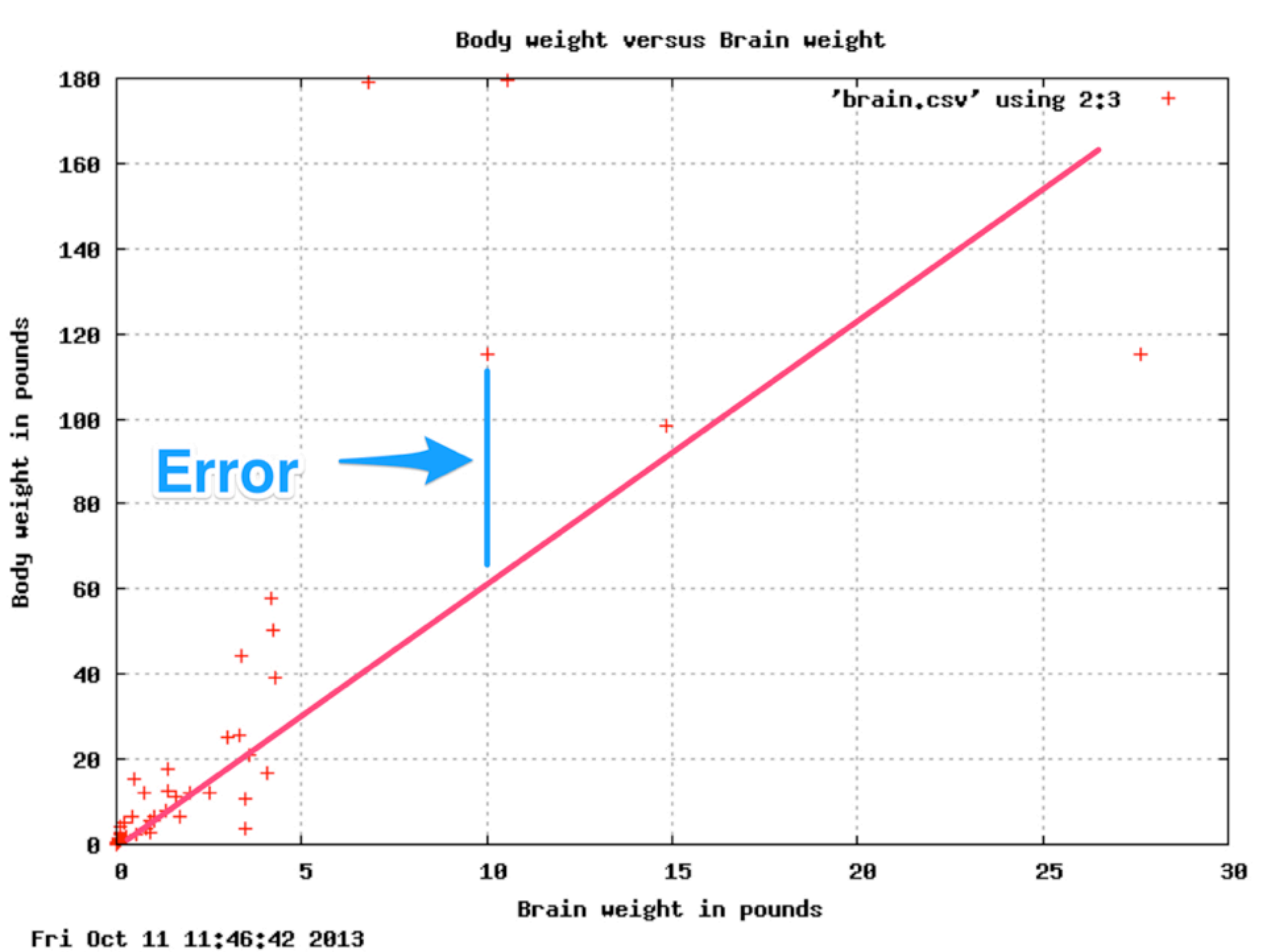
$$y = mx + b$$

$$y = mx + \alpha$$

$$y = \beta\chi + \alpha$$

```
3,1.350,8.100
4,465.000,423.000
5,36.330,119.500
6,27.660,115.000
7,14.830,98.200
8,1.040,5.500
9,4.190,58.000
10,0.425,6.400
11,0.101,4.000
12,0.920,5.700
13,1.000,6.600
14,0.005,0.140
15,0.060,1.000
16,3.500,10.800
17,2.000,12.300
18,1.700,6.300
```


Body weight versus Brain weight
'brain.csv' using 2:3
Body weight in pounds
Brain weight in pounds
180
160
140
120
100
80
60
40
20
0
0
5
10
15
20
25
30
Fri Oct 11 11:46:42 2013


Body weight versus Brain weight
'brain.csv' using 2:3
Body weight in pounds
Brain weight in pounds
0
20
40
60
80
100
120
140
160
180
0
5
10
15
20
25
30
Error
Fri Oct 11 11:46:42 2013

$$\mathrm{error}=(y_i-\alpha-\beta x_i)$$

$$Q=\Sigma(\mathrm{error})^2$$

Classification

How do we classify?

Binary classification

Linear classification

Support Vector Machines

Decision trees

Clustering

Jaccard Coefficient

Group documents

Detect plaguirism

K-Means Clustering

Survey of the Ruby Landscape

AI4R

SciRuby

JRuby and Mahout

Rails on Ruby

Fast Math

Easy Plotting

Integrated Environment

Let me dance!

Moving Forward

Want to learn more?

Linear Algebra

Calculus

Coursera ML

Wikipedia

Now, if you want to get serious

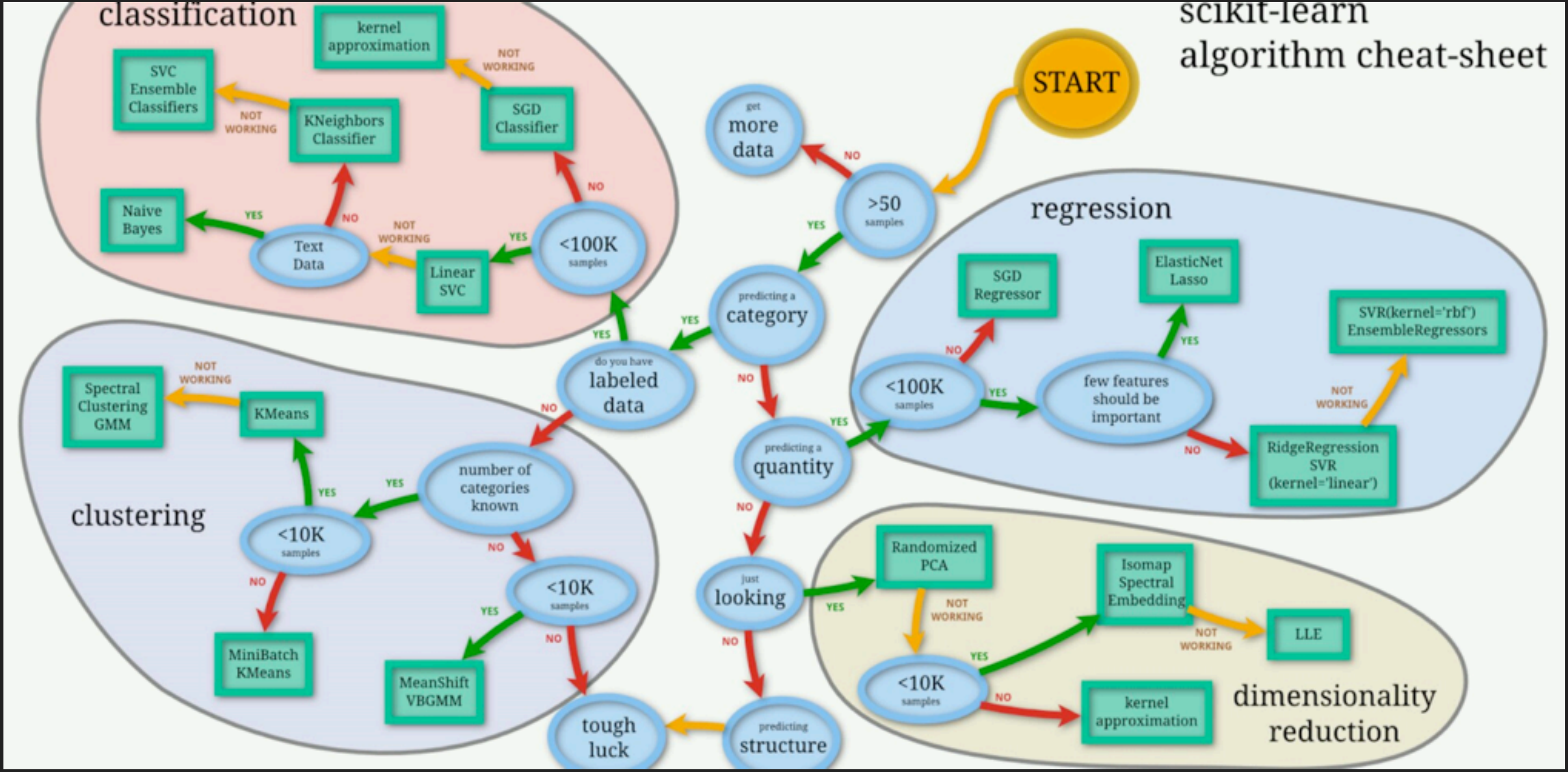
1. Find a dataset

2. Find another language

3. ...

4. Profit?

We haven’t event scratched the surface


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

BigML

Dundas

Kaggle

Python Land

Mahout

Shark with Spark