# Writing Smart Programs

Anoop Thomas Mathew (atmb4u)
Profoundis Inc.

# History

then:

low level ⇝ high level

now:

code driven ⇝ data driven

# What is Smart

everyone makes mistakes

- looking at past to avoid future mistakes

everyone misses what comes next

- predicting what to expect

clusters of items do exist

- automatically group things based on similarity

# Jargon Buster

**Dataset**

**Data Cleaning**

**Dimension**

**Model**

**Training**

**Parameters**

**Dimensionality Reduction**

**Accuracy**

**Overfitting**

**Underfitting**

**Testing**

**Domain**

# Parameter Optimization

Heuristic / Statistical / Machine Learning

- **“know parameters well”**

Eg: stock market prediction disaster; no. of lawyers vs. no. of suicides

# Supervised vs Unsupervised

we know what we want

*vs.*

find what's interesting

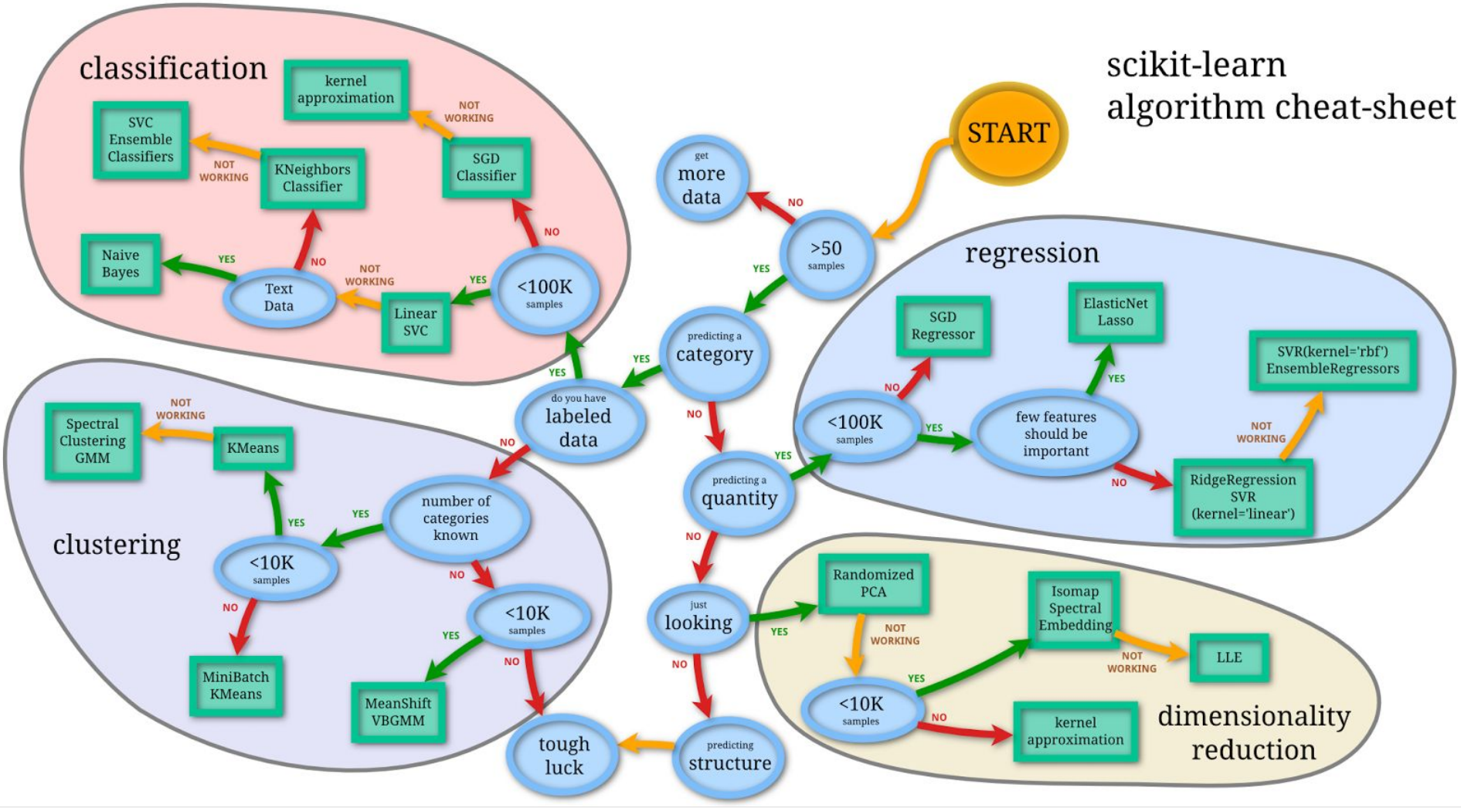
NB: training data, accuracy, semi-supervised

# Classification / Regression / Clustering

- ★ rain prediction
- ★ digit recognition
- ★ customer segmentation
- ★ time-series prediction
- ★ spam filtering

Algorithm Examples

- ★ K-means
- ★ SVR
- ★ SVC
- ★ Naive Bayes
- ★ Random Forest Decision Tree


scikit-learn
algorithm cheat-sheet
START
get more data
NO
>50 samples
YES
predicting a category
YES
do you have labeled data
YES
<100K samples
NO
SVC Ensemble Classifiers
NOT WORKING
KNeighbors Classifier
NOT WORKING
kernel approximation
NOT WORKING
SGD Classifier
NO
Naive Bayes
YES
Text Data
NO
NOT WORKING
Linear SVC
YES
classification
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
tough luck
NO
number of categories known
YES
<10K samples
NO
KMeans
YES
Spectral Clustering GMM
NOT WORKING
MiniBatch KMeans
<10K samples
YES
MeanShift VBGMM
NO
clustering

# The ML Process

**plan** ↠ **collect** ↠ **execute** ↠ **test**

| | | | | |
|---|---|---|---|---|
| time: | 50% | 30% | 5-10% | 15-20% |

# DEMO 1

## SHOPPING PREDICT

(https://github.com/atmb4u/smarter-apps-2016)

# DEMO 2

## Support Vector Machine

(https://github.com/atmb4u/smarter-apps-2016)

# Dummy Tasks

- ★ user auto-login redirect
- ★ predict if a user will convert to paid

# Thank You

@atmb4u