Enough Machine Learning To Make Hacker News Readable Again

Ned Jackson Lovely

@nedjl / njl@njl.us

slides @ www.njl.us

A Simple, Achievable Project

A personalized filter for Hacker News

I Can Machine Learn and You Can Too!

Machine learning is just apply-ing statistics to big piles of data

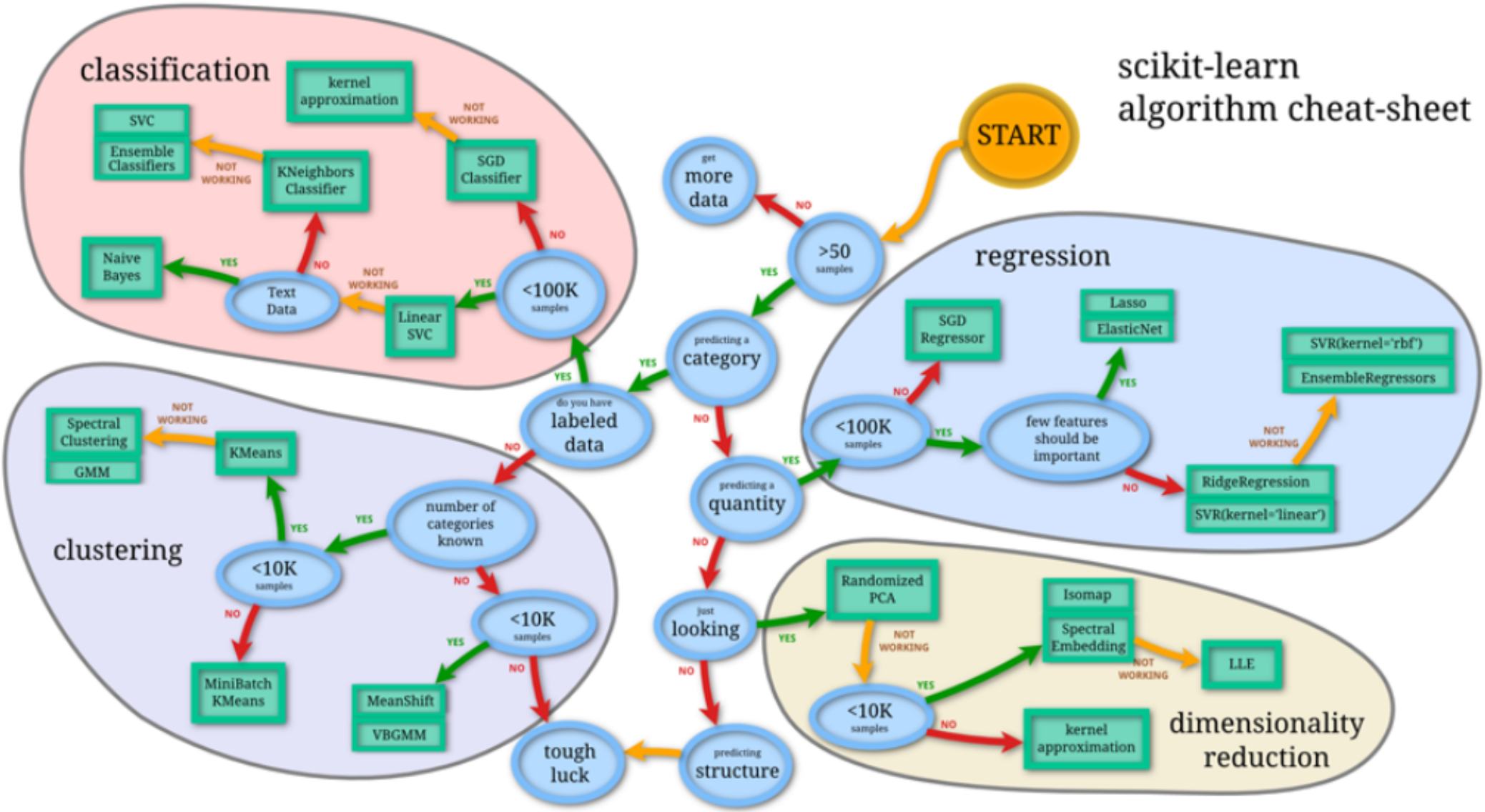
Get Data

Engineer the Data

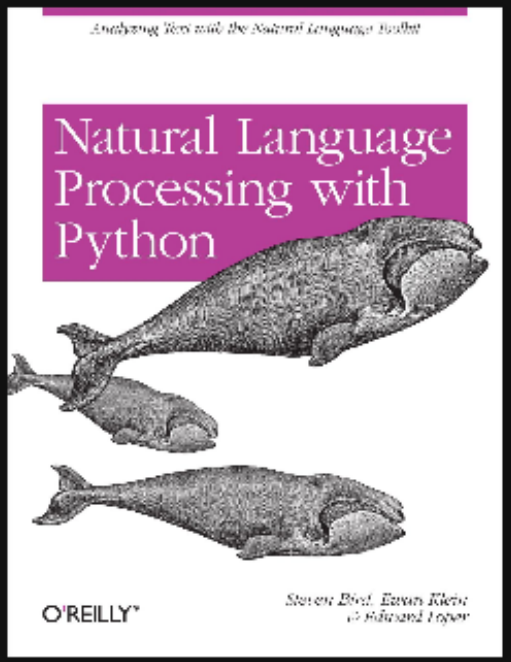
Train and Tune Models

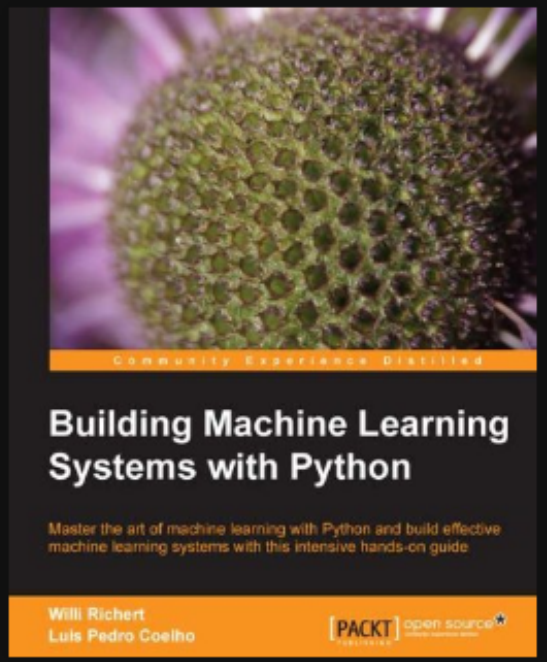
Apply Model

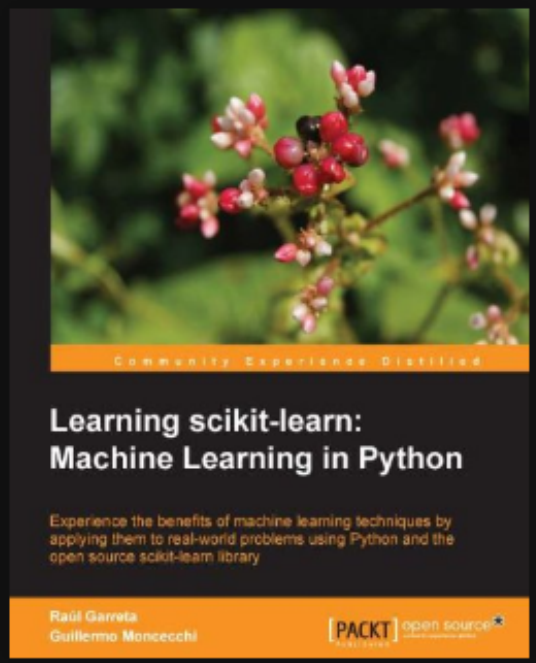
The scikit-learn Docs are Fantastic

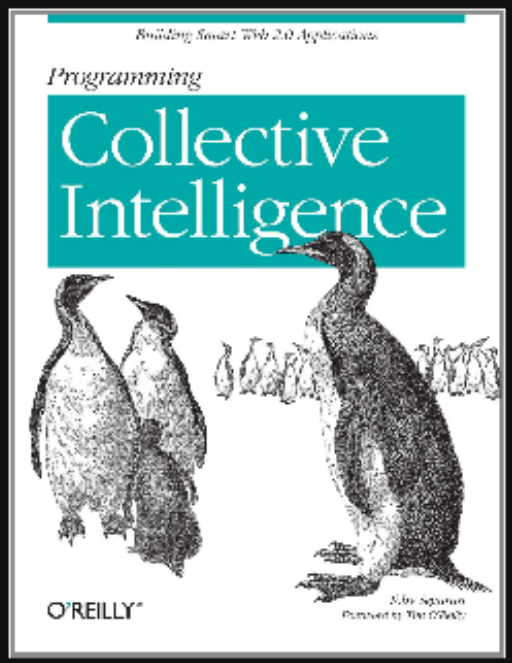

scikit-learn
algorithm cheat-sheet
START
get more data
NO
>50 samples
YES
predicting a category
YES
do you have labeled data
YES
<100K samples
NO
Linear SVC
NOT WORKING
Text Data
NO
Naive Bayes
YES
KNeighbors Classifier
NOT WORKING
Ensemble Classifiers
SVC
NOT WORKING
SGD Classifier
kernel approximation
NO
classification
NO
predicting a quantity
YES
<100K samples
NO
SGD Regressor
YES
few features should be important
YES
Lasso
ElasticNet
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
MiniBatch KMeans
YES
MeanShift
VBGMM
YES
<10K samples
YES
KMeans
NOT WORKING
Spectral Clustering
GMM
clustering

Supervised vs Unsupervised


Natural Language Processing with Python
O'REILLY


Community Experience Distilled
Building Machine Learning Systems with Python
Master the art of machine learning with Python and build effective machine learning systems with this intensive hands-on guide
Willi Richert
Luis Pedro Coelho
PACKT open source


Community Experience Distilled
Learning scikit-learn: Machine Learning in Python
Experience the benefits of machine learning techniques by applying them to real-world problems using Python and the open source scikit-learn library
Raúl Garreta
Guillermo Moncecchi
PACKT open source


Programming
Collective Intelligence
O'REILLY

scikit-learn Patterns

Parallel Arrays

X, y

Set aside a validation set

```
1 | from sklearn.cross_validation import train_test_split
2 | X, X_val, y, y_val = train_test_split(X_full, y_full)
```

How Do You Work?

Build Pipelines
Optimize Parameters

Pipeline

A sequence of operations on data

```
1  from sklearn.pipeline import Pipeline
2  from sklearn.feature_extraction import text
3  from sklearn.svm import LinearSVC
4
5  p = Pipeline([ ('hv', text.HashingVectorizer()),
6                 ('svm', LinearSVC()) ])
```

Hyper-Parame-ters

Tune the Magic

```
from sklearn.grid_search import GridSearchCV
params = { 'svm__C':[0.5, 1.0, 2.0, 4.0],
           'svm__loss':['l1', 'l2'],
           'hv__ngram_range':[ (1,1), (1,2)],}
gs = GridSearchCV(p, params, verbose=2, n_jobs=-1)
gs = gs.fit(X,y)
```

Functions You'll See a Lot Of

`transform()`

`fit()`

`predict()`

transform(X [,y])

```
    hv = HashingVectorizer()
    hv.transform(['Simple is better than complex.',
                  'Sparse is better than dense.'])
<2x1048576 sparse matrix of type '<type 'numpy.float64'>'
    with 10 stored elements ...
```

fit(X,y)

```
1  svm = LinearSVC()
2  svm = svm.fit(X, y)
```

predict(X)

```
1 | y_predictions = svm.predict(X_new)
```

Dealing with the Super-Messy Real World

Get the Data

requests & lxml

	1852	73.0860299921%
	677	26.7166535122%
	5	0.197316495659%
	2534	

Bring The Pain			
rd data from Nieman-Marcus.	Framed	Y	Good
In 5 out of 6 stats for protecting your privacy	Framed	Y	Good
Course	Framed	Y	Good
neurship: How to Avoid Becoming a Stressed Out Loner	Framed	Y	Good
roducer Multi-consumer Queue on Ring Buffer	Framed	Y	Good
olution claimed for Navier-Stokes equations, one of the millennium problems	Framed	Y	Goo
he Day We Fight Back: MIT website hacked by Anonymous	Framed	Y	Goo
nonymous hacks MIT website on anniversary of Aaron Swartz suicide	Framed	Y	Goo

Classify Dreck and Non-Dreck

Title, URL, Submitter, Content of Link, Rank, Votes, Comments, Time of Day, Dreck or Not

Messy Data to Normalized Numpy Arrays

Bag Of Words

n-grams

Normalization

Stop Words

TF-IDF

Bag of Words

**"Time Flies Like An Arrow,
Fruit Flies Like Bananas"**

Flies	2
Like	2
Time	1
An	1
Arrow	1
Fruit	1
Bananas	1
Bitcoin	0
...	

n-grams

"Time Flies Like An Arrow,
Fruit Flies Like Bananas"

Time Flies
Flies Like
Like An
An Arrow
Arrow Fruit
...

Normalization

```
1  from nltk.stem.snowball import EnglishStemmer
2  stemmer = EnglishStemmer()
3  input='when he flies he likes to fly upon early flights'
4  print(' '.join(stemmer.stem(x) for x in input.split()))
5
6  "when he fli he like to fli upon earli flight"
```

Stop Words

```
1  from sklearn.feature_extraction import text
2
3  print(', '.join(x for x in
4                  list(text.ENGLISH_STOP_WORDS)[:8]))
5
6  all, six, less, being, indeed, over, move, anyway
```

TF-IDF

Term-Frequency, Inverse Document Frequency

```
1  from sklearn.feature_extraction import text
2
3  tfidf = text.TfidfVectorizer()
4  tfidf = tfidf.fit(X, y)
```

Engineering Features

Pulling Out the Relevant Text

```
from readability.readability import Document
from lxml.html import fragment_fromstring

def get_relevant(raw):
    try:
        cleaned = Document(raw).summary(True)
        et = fragment_fromstring(cleaned)
        return ' '.join(x.text for
                        x in et.getiterator()
                        if x.text)
    except Exception as e:
        print(e)
        return ''
```

Roll your own features

```
1  from sklearn import base
2  from sklearn.preprocessing import StandardScaler
3  from sklearn.pipeline import Pipeline
4
5  class FeatureLength(base.BaseEstimator,
6                      base.TransformerMixin):
7      def fit(self, X, y=None):
8          return self
9      def transform(self, X, y=None):
10         return [[len(x)] for x in X]
11
12 length_pipeline = Pipeline([ ('fl', FeatureLength()),
13                               ('ss', StandardScaler()),])
```

Combine Features

```
from sklearn.pipeline  import FeatureUnion
union = FeatureUnion([ ('tfidf', TfidfVectorizer()),
                       ('length', length_pipeline), ])
```

My Data Starts as Dictionaries

```
1  from sklearn import base
2
3  class DictFeatureExtractor(base.BaseEstimator,
4                             base.TransformerMixin):
5      def __init__(self, prop_name):
6          self.prop_name = prop_name
7      def fit(self, X, y=None):
8          return self
9      def transform(self, X, y=None):
10         return [x[self.prop_name] for x in X]
```

Maybe Hostnames Are Relevant

```
1  import urlparse
2
3  class HostnameExtractor(BaseEstimator, TransformerMixin):
4      def fit(self, X, y=None):
5          return self
6      def transform(self, X, y=None):
7          return [{'host':urlparse.urlparse(x).netloc}
8                      for x in X]
```

A Hostname Pipeline

```
1  from sklearn.feature_extraction import DictVectorizer
2
3  hosts = Pipeline([
4                ('dfe', DictFeatureExtractor('url')),
5                ('he', HostnameExtractor()),
6                ('dv', DictVectorizer()), ])
```

The Actual Application

Predict

```
1  from pickle import load
2  classifier = load(open('ridiculous_pipeline.pickle'))
3  classifier.predict(X_input_data)
```

1 [Full-disclosure] Administrivia: The End
171 points by Morgawr 2 hours ago
comments

ki triangle page to end most sierpinski triangle pages
56 points by deedee 2 hours ago
comments

3 Needy robotic toaster sells itself if neglected
20 points by Baustin 1 hour ago
comments

4 Lessons from a Silicon Valley job search
11 points by padolsey 41 minutes ago
comments

5 Nodemailer: Easy as cake e-mail sending from your Node.js applications
49 points by tilt 4 hours ago
comments

rom an IMAP server failure
28 points by Aoyagi 2 hours ago
comments

7 Results of operation against child exploitation enterprise on tor network
21 points by alternize 2 hours ago
comments

8 Why air travel hasn't gotten any faster since the 1960s
9 points by nmorell 1 hour ago
comments

9 What Happens to Older Developers?
192 points by dsirijus 4 hours ago
comments

10 Moto 360
587 points by uptown 21 hours ago
comments

11 Sony announces Project Morpheus, a virtual reality headset coming to PS4
176 points by nzonbi 12 hours ago
comments

12 The new Expats.Stackexchange Public Beta has launched
47 points by nsaparanoid 6 hours ago
comments

13 Storage Pod 4.0: Direct Wire Drives – Faster, Simpler and Less Expensive
8 points by nuriaion 1 hour ago
comments

14 Model View Whatever
3 points by khanlou 7 minutes ago
comments

onfiguration Patterns
9 points by frist45 2 hours ago
comments

JS: Divide and conquer complex web apps
44 points by porker 7 hours ago
comments

17 Android coming to wearables
374 points by deepblueocean 22 hours ago
comments

18 I don't want to be a Real Programmer
54 points by verybadalloc 6 hours ago
comments

hn.njl.us

If You Have The Data...

Unsupervised Learning

Cluster Hacker News Articles

Regression

Predict article scores

Machine Learning is Becoming Engineering

Go Engineer!

Thank You!

Ned Jackson Lovely

@nedjl / njl@njl.us

slides @ www.njl.us