Machine Learning, with scikit-learn

Jeffrey Skonhovd

Georgia Institute of technology

$\mathrm{DATE}_O F_M Y_T ALK$

Outline

Who am I?

- Jeffrey Skonhovd
- Works at FTN Financial
- Twitter: @jskonhovd
- Github: jskonhovd

Overview

- What is Machine Learning?
  - Machine Learning is the study of computer algorithms that improve automatically through experience.
- How should I go about learning Machine Learning?
  - MOOCs
  - Don't get caught up in the implementations.
- Tools
  - WEKA
  - scikit-learn

Types

- Supervised Learning
- Supervised Learning is . . .
- Unsupervised Learning
- Unsupervised Learning is . . .
- Reenforcement Learning
- Reenforcement Learning is . . .

Some Boring, but important Definitions.

- Inductive Bias
- The inductive bias of a learning algorithm is the set of assumptions that the learner uses to predict outputs given inputs that it has not encountered.
- Occam's Razor assumes that the hypotheses with the fewest assumptions should be selected.
- Cross-validation
- The basic idea of Cross-validation to leave out some of the data when fitting the model.

Scikit-learn

- Scikit-learn is a set of simple and efficient tools for data mining and data analysis.
- Uses Python!!!
- `http://scikit-learn.org/`

Decision Trees

- Decision Tree learning is a method for approximating discrete-valued target functions, in which the learned function is represented a decision tree.
- Maximize Information Gain
- Information Gain measures how well a given attribute separates the training examples according to their target classifcation.

Decision Trees: Example


```
import numpy as np
import pylab as pl
from sklearn.datasets import load_iris
from sklearn.tree import DecisionTreeClassifier
# Parameters
# Load data
iris = load_iris()
clf = DecisionTreeClassifier()
X = iris.data[:, [1, 2]]
y = iris.target
clf = clf.fit(X, y)
plotCustom(X, y, [1, 2], clf)‘
```

kNN: Example


```
from sklearn import neighbors
import numpy as np
import pylab as pl
from sklearn import cross_validation
from sklearn.datasets import load_iris

iris = load_iris()

X = iris.data[:, [1, 2]]
y = iris.target

clf = neighbors.KNeighborsClassifier(3, 'distance')

plotCustom(X, y, [1,2], clf)
```

SVM


```
from sklearn import svm
import numpy as np
import pylab as pl
from sklearn import cross_validation
from sklearn.datasets import load_iris

iris = load_iris()

X = iris.data[:, [1, 2]]
y = iris.target
C = 1.0
clf = svm.SVC(kernel='linear', C=C)
rbf_svc = svm.SVC(kernel='rbf', gamma=0.7, C=C)
poly_svc = svm.SVC(kernel='poly', degree=3, C=C)
lin_svc = svm.LinearSVC(C=C)
clf.fit(X, y)
```

kMeans


```
from time import time
import numpy as np
import pylab as pl
from sklearn.cluster import KMeans
from sklearn.datasets import load_digits
from sklearn.decomposition import PCA
from sklearn.preprocessing import scale
from sklearn.datasets import load_iris

iris = load_iris()

X = iris.data[:, [2, 3]]
y = iris.target
n_digits = len(np.unique(y))
kmeans = KMeans(init='k-means++', n_clusters=n_digits
kmeans.fit(X)
```

- Resources