CIS192 Python Programming

Machine Learning in Python

Robert Rand

University of Pennsylvania

October 22, 2015

Outline

Installation

Recommended:

```
pip install -U numpy scipy matplotlib ipython[notebook] scikit-learn
```

You should install pip first if you don't have it.

Packages like Anaconda and Canopy also include the relevant libraries.

Numpy and Scipy

- Libraries for sophisticated mathematics and mathematical computing in Python.
- Include libraries for linear algebra.
- Optimized for efficient machine learning.

Matplotlib

- Library for plotting datasets.
- Use it to look at data before attempting machine learning techniques.
- Interfaces well with the IPython (an alternative shell for Python development).

Scikit Learn

- A machine learning library for Python.
- Uses numpy and scipy.
- Comes with a broad array of built in machine learning algorithms

Outline

Reducing People

Recall our PAC learning example from last class.

We were training on people, but we only recorded their height and their weight as relevant variables.

The problem of *data representation* is a core part of machine learning.

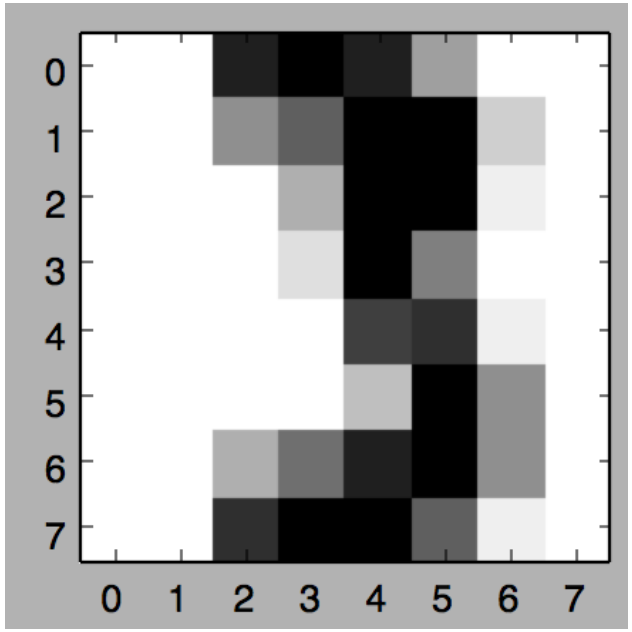
The Iris Dataset

The well-known Iris dataset (https://archive.ics.uci.edu/ml/datasets/Iris) models Irises as four simple numbers:

1. sepal length in cm
2. sepal width in cm
3. petal length in cm
4. petal width in cm

The Digits Dataset

We can represent hand-drawn digits by a simple 8 x 8 matrix with each entry corresponding to the shading of a given part of the box:



Bags of Words

We often represent text as a “bag of words”. That is, we record the number of times each word appears in document in an array. This allows us to compare the arrays against each other using standard methods.

Tf-Idf

The *term frequency-inverse document frequency* transformer also uses bag of words representation, but scales words according to how common they are in the broader corpus, preventing too much matching on “and”, “the” etc.

Outline

Coats and Bars

Suppose you didn't know the meaning of "boat" or "car" but you had a stack of photographs of boats and cars.

Could you somehow sort them into two stacks, which corresponded to boats and cars?

K-Means

- `cluster.KMeans()`
  - Parameter: `n_clusters` - specifies the number of clusters desired.
- Randomly assign initial position for each cluster.
- Repeat until stable:
  - Assign every point to its closest cluster c_i.
  - Move c_i to the center of points that are assigned to it.

  Every point is labeled with its cluster.

Other Models

- Gaussian Mixture Models general K-Means - K-means assumes clusters look like circle, where GMM can handle arbitrary elliptic clusters.
- Affinity Propagation ((cluster.AffinityPropagation()) identifies *exemplars* - points that can stand in for a given cluster. It may vary the number of clusters depending on the exemplars it finds.