EECS 510: SOCIAL MEDIA MINING
SPRING 2015

Data Mining Essentials 2:
Data Mining in Practice, with Python

Rosanne Liu
rosanne.liu@northwestern.edu
Outline

• Why Python?
• Intro to Python
• Intro to Scikit-Learn
• Unsupervised Learning
  – Demo on PCA, K-Means
• Supervised Learning
  – Demo on Linear Regression, Logistic Regression
Outline

• Why Python?
• Intro to Python
• Intro to Scikit-Learn
• Unsupervised Learning
  – Demo on PCA, K-Means
• Supervised Learning
  – Demo on Linear Regression, Logistic Regression
### Why Python?

What programming language do you use for data mining?

Source from: http://www.kdnuggets.com/polls/index.html
How much is your salary as analytics, data mining, data science professionals?

Source from: http://www.kdnuggets.com/polls/index.html
Should data scientist / data miners be responsible for their predictions?

<table>
<thead>
<tr>
<th>Should data scientists / data miners be responsible for their predictions? [242 voters]</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>No, they should not be responsible (108)</td>
<td>45%</td>
</tr>
<tr>
<td>Not sure (32)</td>
<td>13%</td>
</tr>
<tr>
<td>They can be held financially responsible, but if they also benefit from correct predictions (89)</td>
<td>37%</td>
</tr>
<tr>
<td>They can be held criminally responsible for wrong predictions (13)</td>
<td>5%</td>
</tr>
</tbody>
</table>

Source from: http://www.kdnuggets.com/polls/index.html
Why Python?

Think about the scientist’s needs:

- Get data (simulation, experiment control)
- Manipulate and process data.
- Visualize results... to understand what we are doing!
- Communicate results: produce figures for reports or publications, write presentations.
Why Python?

- Easy
  - Easy to learn, easily readable
  - Scientists first, programmers second
- Efficient
  - Managing memory is easy – if you just don’t care
- A single Language for everything
  - Avoid learning a new software for each new problem
More to Take Away

• Free distribution from http://www.python.org

• Known for it’s “batteries included” philosophy
  Similar to R, Python has a fantastic community around it and, luckily for you, this community can write

• Two popular versions, 2.7 or 3.x

• A single-click installer: Enthought Canopy

• Prepare yourself for code indentation heaven

```python
from math import sqrt
n = input("Maximal Number? ")
n = int(n)+1
for a in range(1,n):
    for b in range(a,n):
        c_square = a**2 + b**2
        c = int(sqrt(c_square))
        if ((c_square - c**2) == 0):
            print(a, b, c)
```
All the Good Modules

- **numpy, scipy**: basics for almost everything
- **Matplotlib**, a Python 2D plotting library [http://matplotlib.org](http://matplotlib.org)
- **NLTK**, Natural Language Toolkit [http://www.nltk.org](http://www.nltk.org)
- **Pandas**, Python Data Analysis Library [http://pandas.pydata.org](http://pandas.pydata.org)
- **mrjob**, route to writing MapReduce jobs [https://pythonhosted.org/mrjob/](https://pythonhosted.org/mrjob/)
- **IPython**, Interactive console with IDE-like features [http://ipython.org](http://ipython.org)
- **Theano/Pylearn2**, deep learning
  - [http://deeplearning.net/software/theano/](http://deeplearning.net/software/theano/)
  - [http://deeplearning.net/software/pylearn2/](http://deeplearning.net/software/pylearn2/)
- **More**: mlyp, PyBrain, Orange, Scrapy, ...
Outline

• Why Python?
• **Intro to Python**
• Intro to Scikit-Learn
• Unsupervised Learning
  – Demo on PCA, K-Means
• Supervised Learning
  – Demo on Linear Regression, Logistic Regression
The Use of Python: Simple demos

0 – Python Intro.ipynb
Outline

• Why Python?
• Intro to Python
• **Intro to Scikit-Learn**
  • Unsupervised Learning
    – Demo on PCA, K-Means
• Supervised Learning
  – Demo on Linear Regression, Logistic Regression
What is Scikit-learn

• A Python Machine Learning Library
• Focused on modeling data
• Developed by David Cournapeau as a Google summer of code project in 2007.
• First public release (v0.1 beta) published in late January 2010.
• Now has more than 30 active contributors and has had paid sponsorship from INRIA, Google, Tinyclues and the Python Software Foundation.
• The library is built upon the SciPy that must be installed before you can use scikit-learn.
Outline

• Why Python?
• Intro to Python
• Intro to Scikit-Learn

• **Unsupervised Learning**
  – Demo on PCA, K-Means

• Supervised Learning
  – Demo on Linear Regression, Logistic Regression
The use of Scikit-Learn: unsupervised learning demos
PCA Summary

- PCA projects to axis with greatest variance
- Often provides good first insight into dataset

\[
\begin{align*}
\bar{X} & \leftarrow X - \text{mean}(X) \quad \bar{X} \in \mathbb{R}^{n \times N} \\
W & \leftarrow \text{PCA}(\bar{X}, 2) \quad W \in \mathbb{R}^{N \times M} \\
X_{\text{PCA}} & \leftarrow \bar{X} \cdot W \quad X_{\text{PCA}} \in \mathbb{R}^{n \times M}
\end{align*}
\]

- Identify important variables in projection matrix \(W\):

\[
W = \begin{bmatrix}
0.36 & -0.08 & 0.85 & 0.35 \\
-0.65 & -0.72 & 0.17 & 0.07
\end{bmatrix}
\]
1 – PCA.ipynb
K-Means Algorithm

$k$-Means finds assignments $j$ and cluster centers $\mu$ by solving

$$\min_{\mu} \sum_{i=0}^{N} \min_{j} \|\mu_j - x_i\|^2$$

(The 1)

The algorithm is simple:

1. Set $\mu, j$ to a random value
2. Solve (1) for $j$
3. Solve (1) for $\mu$
4. If $j$ or $\mu$ changed significantly, go to step 2.
2 – k means.ipynb
Outline

• Why Python?
• Intro to Python
• Intro to Scikit-Learn
• Unsupervised Learning
  – Demo on PCA, K-Means

• **Supervised Learning**
  – Demo on Linear Regression, Logistic Regression, kNN
The use of Scikit-Learn: supervised learning demos
Linear Regression

\[ y = w_1 x_1 + b \]

\[ y = w_2 x_2 + w_1 x_1 + b \]

To find \( w \) and \( b \), minimize the error:

\[ E = \sum_{i=0}^{N} (y_i - (w_i x_i + b))^2 \]
3 – LinearRegression1.ipynb
3 – LinearRegression2.ipynb
Logistic Regression

For two classes $-1, +1$.
Decision boundary given by hyperplane.
Hyperplane defined by normal vector and offset:

$$y = \text{sign}(\langle w, x \rangle + b)$$

$$w \in \mathbb{R}^n, b \in \mathbb{R}$$
Logistic Regression

Relation to regression:

\[ p(y = +1 \mid x) = \text{logistic}(\langle w, x \rangle + b) \]

As probabilities are between 0 and 1, the logistic function squashes the regression result:

\[ p(y = +1 \mid x) > 0.5 \iff \langle w, x \rangle + b > 0 \]

Need to solve:

\[ \max_w \sum_{i=0}^{n} \log(p(Y = y_i \mid x_i)) \]
4 – LogisticRegression.ipynb
Nonlinear Problems

- Logistic regression works well if the data is linearly separable, but...
K Nearest Neighbors

- Classification: same setup as logistic regression.
- Very simple but powerful idea: Do as your neighbors do.
- For a new point $x$ look at the nearest (or the two nearest or three nearest, ...) point(s) in the training data for a label.
- Usual distance measure: Euclidean distance
Simple Algorithm

• Pick a $k$, for example $k = 3$.
• Want to classify new example $x$.
• Compute $d_i = d(x_i, x)$, i.e. $d(x_i, x) = ||x_i - x||$.
• Sort $d_i$, take $k$ smallest: $d_{i_0}, d_{i_1}, d_{i_2}$.
• Assign $y$ that appears most often among $y_{i_0}, y_{i_1}, y_{i_2}$.