

Text Analysis with Python

Arieda Muço

Central European University

Spring 2019

Machine learning

Applying algorithms that iteratively learn from data.

- Used for: Fraud detection, pattern and image recognition, text sentiment analysis, email spam filtering, credit scoring, new pricing models, recommendation engines...

Types of learning

- Supervised Learning

- ▶ The program is “trained” on a pre-defined set of “training examples”, which facilitate its ability to reach an accurate conclusion when given new data
- ▶ In this case, we have a “target” or dependent variable
- ▶ We also have “labeled” data

- Unsupervised Learning

- ▶ Find patterns and relationships between the data
- ▶ No “target” or dependent variable
- ▶ We don’t have “labeled” data

- Deep Learning (Artificial Neural Networks)

Some references

- Books

- ▶ An Introduction to Statistical Learning
<http://www-bcf.usc.edu/~gareth/ISL/ISLR%20Sixth%20Printing.pdf>
- ▶ Introduction to Machine Learning
<http://robotics.stanford.edu/people/nilsson/MLBOOK.pdf?>
- ▶ Elements of Statistical Learning https://statweb.stanford.edu/~tibs/ElemStatLearn/printings/ESLII_print10.pdf?
- ▶ Machine Learning with Python by Sara Guido and Andreas Muller

- Andrew Ng

- ▶ Notes <http://cs229.stanford.edu/materials.html>
- ▶ Video <https://www.coursera.org/learn/machine-learning>

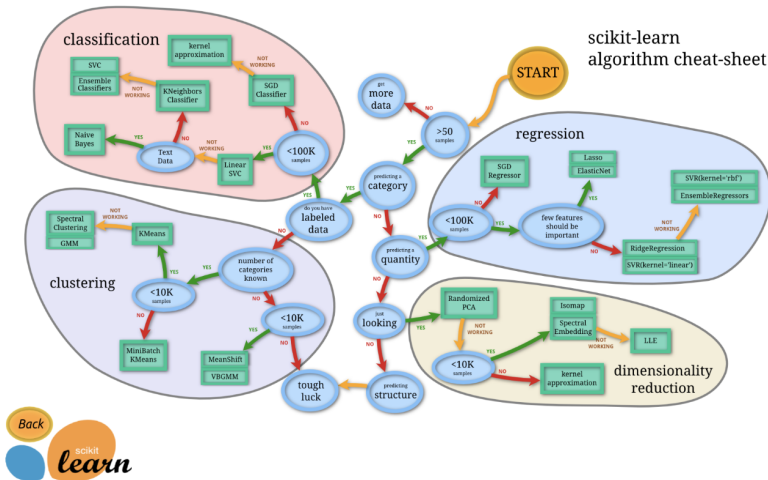
Python Libraries

We are mainly going to use the following libraries

- pip install scikit-learn or conda install scikit-learn (Machine Learning)
- pip install nltk or conda install nltk (Text Analysis)

The most used libraries

Algorithm Cheat Sheet



Learning Machine Learning

Machine Learning takes time to learn

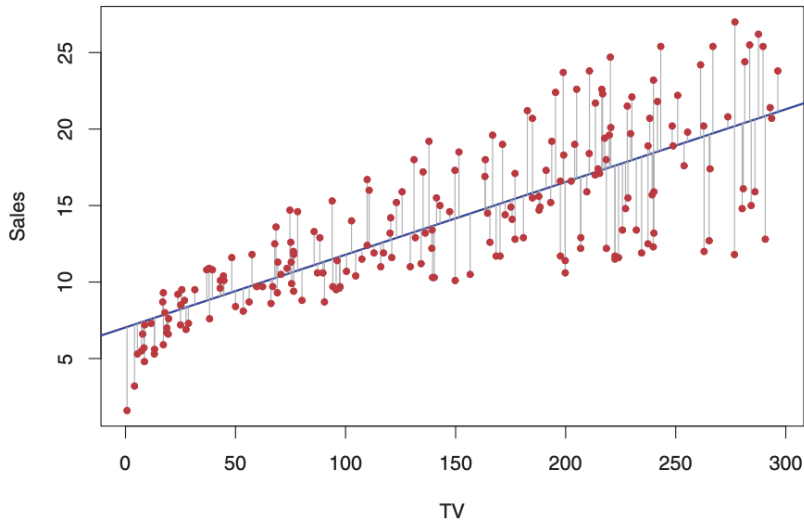
- Be patient with yourself
- Ask questions. I'm happy to answer questions to point you towards material where you can further deepen your understanding
- Most importantly: Google

Linear Regression

Scikit-learn regression model

- Train the regressor using the `fit()` method
- Predict new labels using the `predict()` method
- Using the housing price dataset

Linear Regression



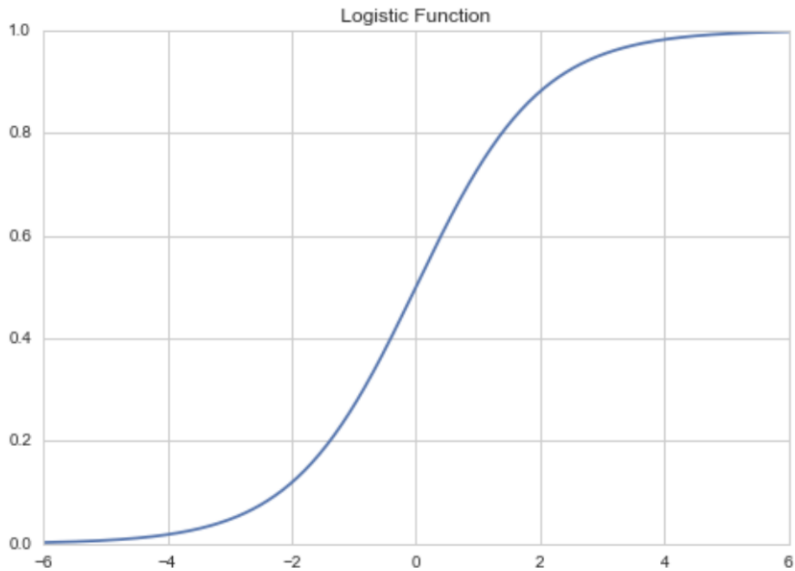
Let's try it out!!!

Logistic Regression

Perform a binary classification, so that we have two outputs, yes or not

- Emails: spam or not spam
- Credit: default or not

It is also possible to have more than two classes (Multinomial)



Let's try it out!!!

Naive Bayes

- Naive Bayes is one of the most practical machine learning algorithms
- It performs very well with text data
- It learns and predicts very fast and it does not require lots of storage
- It takes the name after Bayes as the Bayes theorem is applied. It's called "NAIVE" because all features are assumed to be independent of each other
 - ▶ This is rarely the case, however, the algorithm still returns very good accuracy in practice even when the independent assumption does not hold

Let's try it out!!!