INSTITUTE FOR MACHINE LEARNING AND ANALYTICS

Wrap up – Day I

Janis Keuper

Introduction to ML

INSTITUTE FOR MACHINE LEARNING AND ANALYTICS

Basic Types of Machine Learning Algorithms

Supervised Learning

Unsupervised Learning

Reinforcement Learning

- Labeled data
- Direct and quantitative evaluation

- Learn model from „ground truth“ examples
- Predict unseen examples

Introduction to ML

INSTITUTE FOR MACHINE LEARNING AND ANALYTICS

Basic Types of Machine Learning Algorithms

Supervised Learning

Unsupervised Learning

Reinforcement Learning

- NO Labeled data
- NO Direct and quantitative evaluation

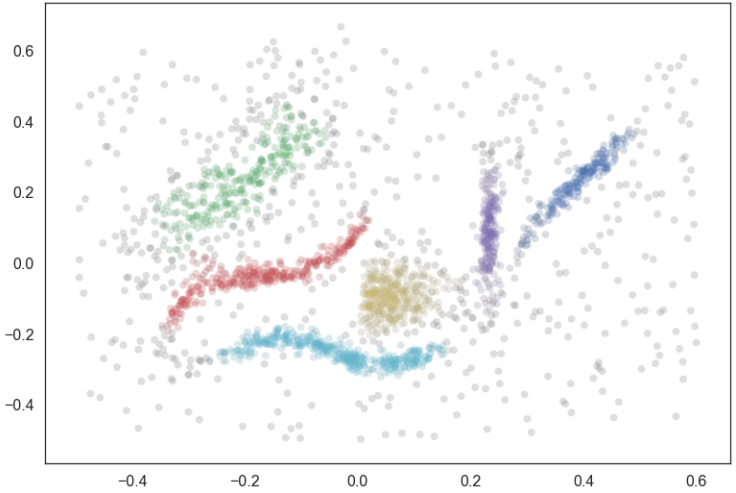
- Explore structure of data

Clustering Algorithms

INSTITUTE FOR MACHINE LEARNING AND ANALYTICS

Introduction

Cluster analysis or **clustering** is the task of **grouping a set of objects** in such a way that objects in the same group (called a cluster) are **more similar** (in some sense) to each other than to those in other groups (clusters). [Wikipedia]



Example 2d data set

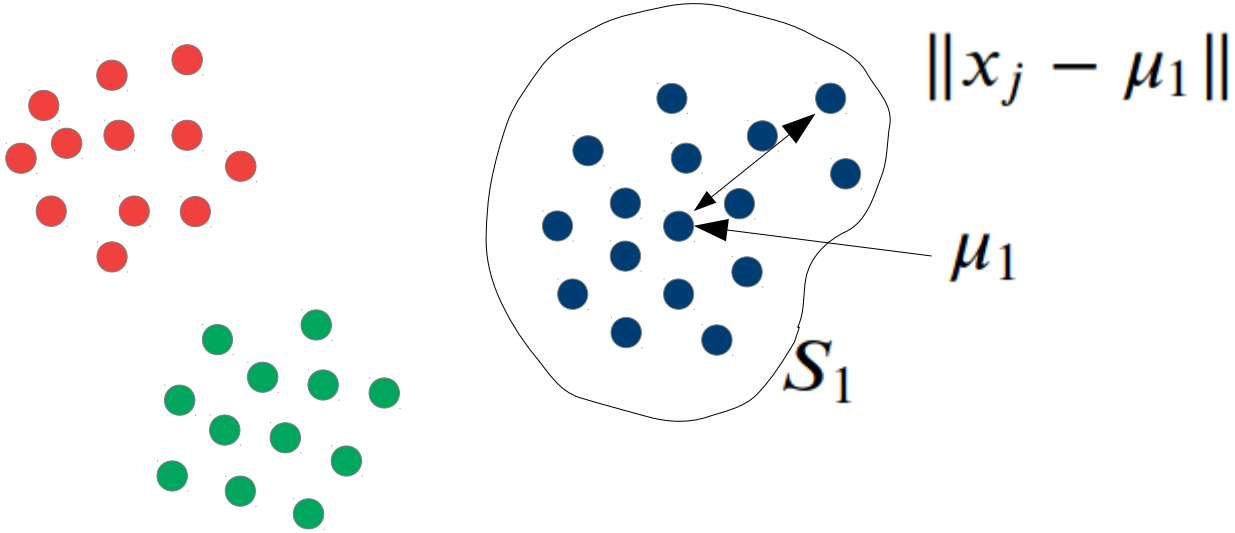
Clustering Algorithms: K-Means

INSTITUTE FOR MACHINE LEARNING AND ANALYTICS

Intuition:

$$\arg\min[S] \sum_{i=1}^{k} \sum_{x \in S_i} \|x - \mu_i\|$$

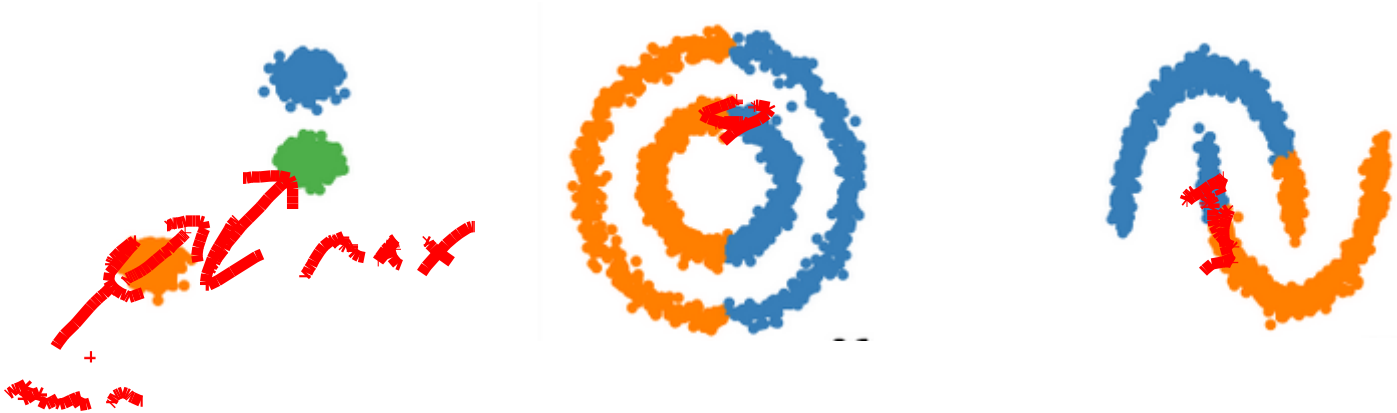
Clustering
for k=3



Clustering Algorithms: K-Means

INSTITUTE FOR MACHINE LEARNING AND ANALYTICS

Evaluation:



More practical examples in the Lab session…. Now!

Clustering Algorithms: DBSCAN

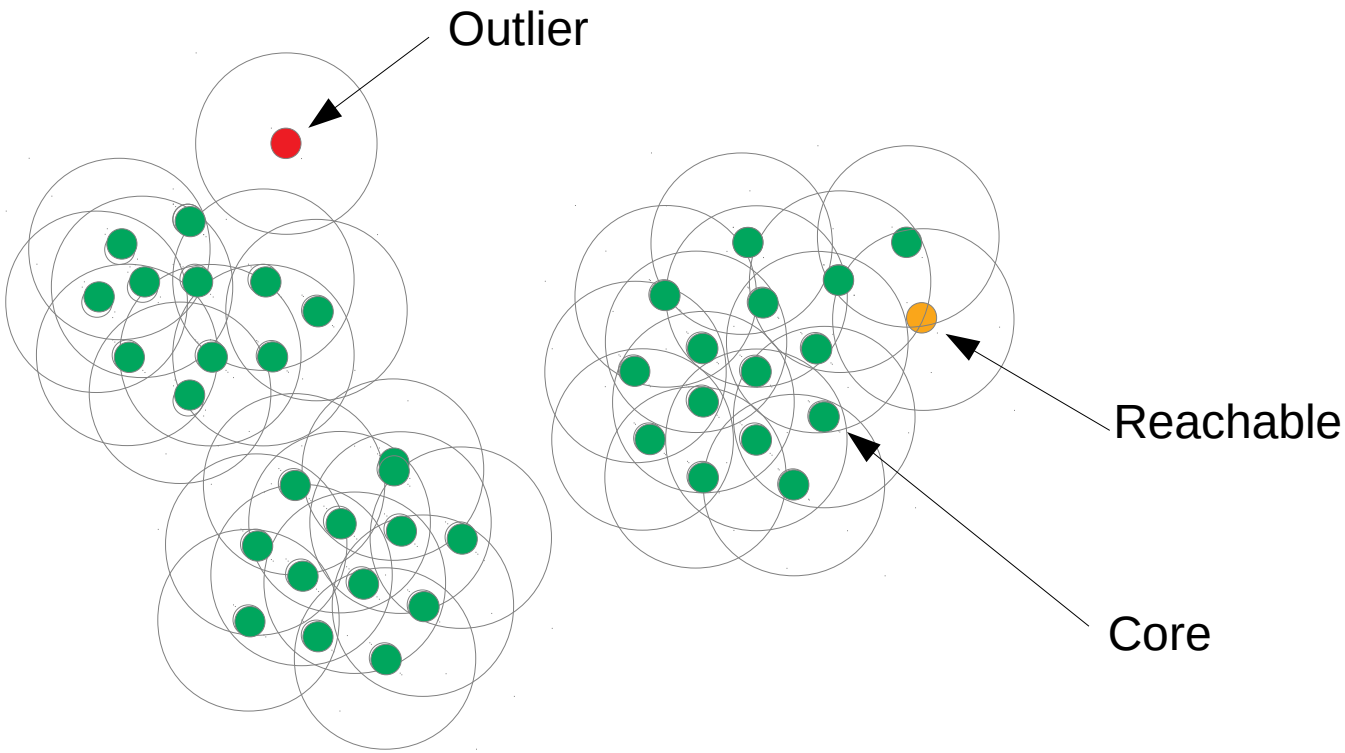
INSTITUTE FOR MACHINE LEARNING AND ANALYTICS

Core – Reachable - Outlier:

Init:

ε

n_ε = 3



Clustering Algorithms: DBSCAN

INSTITUTE FOR MACHINE LEARNING AND ANALYTICS

Evaluation:

More practical examples in the Lab session…. Now!

Clustering Algorithms: Evaluation

INSTITUTE FOR MACHINE LEARNING AND ANALYTICS

How to evaluate clustering:

- Visually → use dimension reduction techniques to visualize 2d or 3d

Clustering Algorithms: Evaluation

INSTITUTE FOR MACHINE LEARNING AND ANALYTICS

How to evaluate clustering:

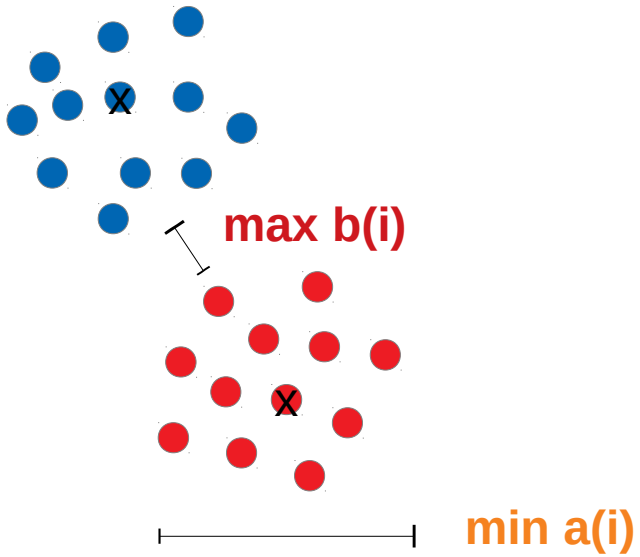
- Visually → use dimension reduction techniques to visualize 2d or 3d
- **Quantitative quality measures (what is a good cluster?)**
  - **Low intra cluster variance**

$$a(i) = \frac{1}{|C_i| - 1} \sum_{j \in C_i, i \neq j} d(i,j)$$

  - **High extra cluster variance**

$$b(i) = \min_{i \neq j} \frac{1}{|C_j|} \sum_{j \in C_j} d(i,j)$$

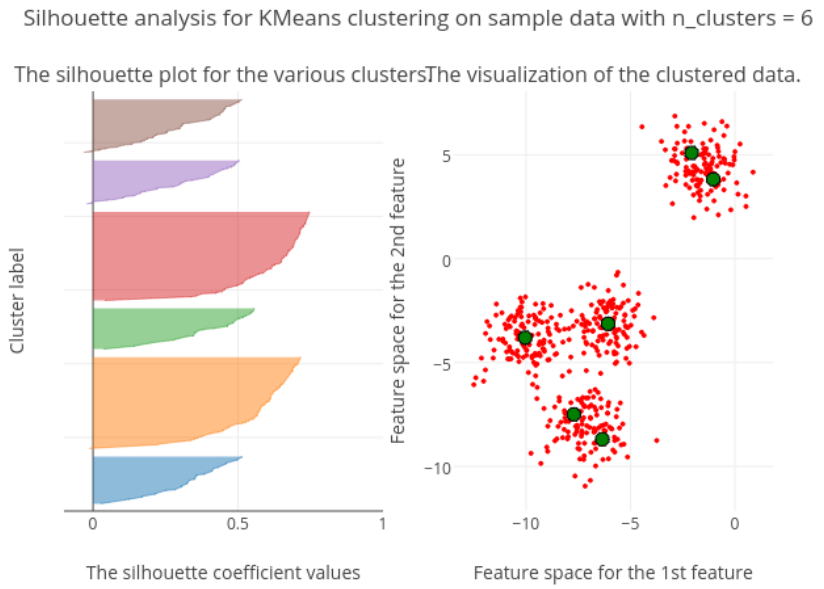
For each data point $i \in C_i$ (data point i in the cluster C_i)

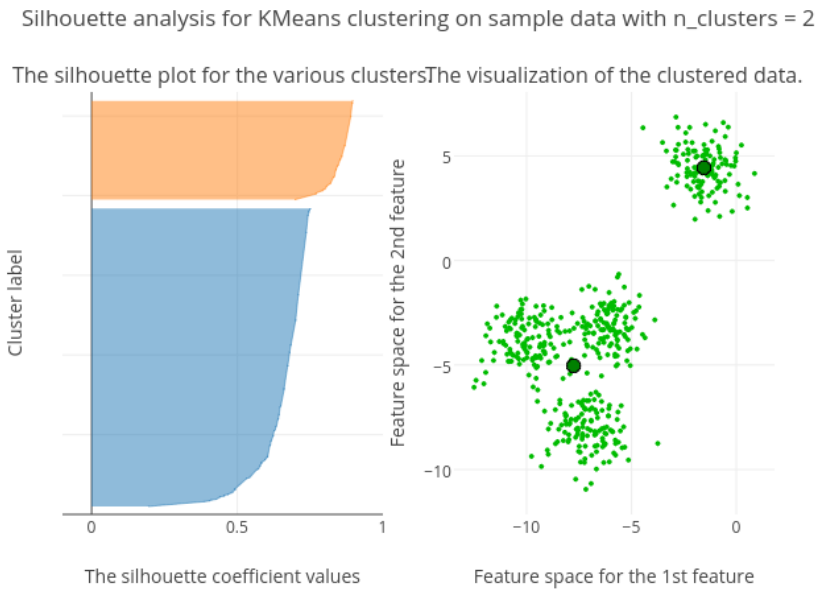


Clustering Algorithms: Evaluation

INSTITUTE FOR MACHINE LEARNING AND ANALYTICS

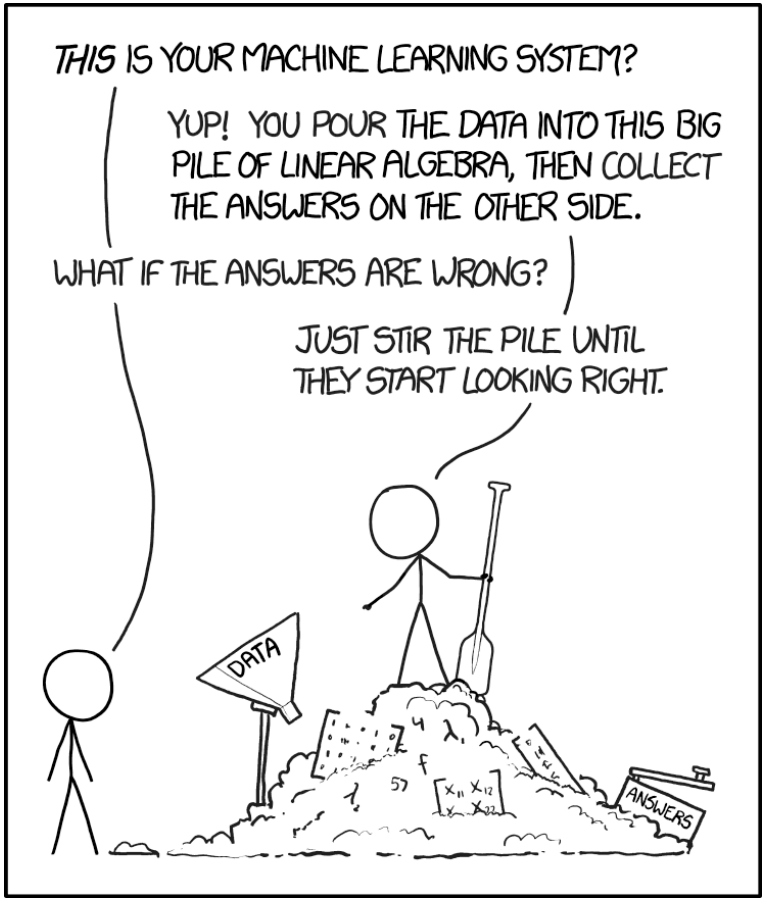
Silhouette Diagrams: finding the best number of clusters





[plots: https://plot.ly/scikit-learn/plot-kmeans-silhouette-analysis/]

Discussion

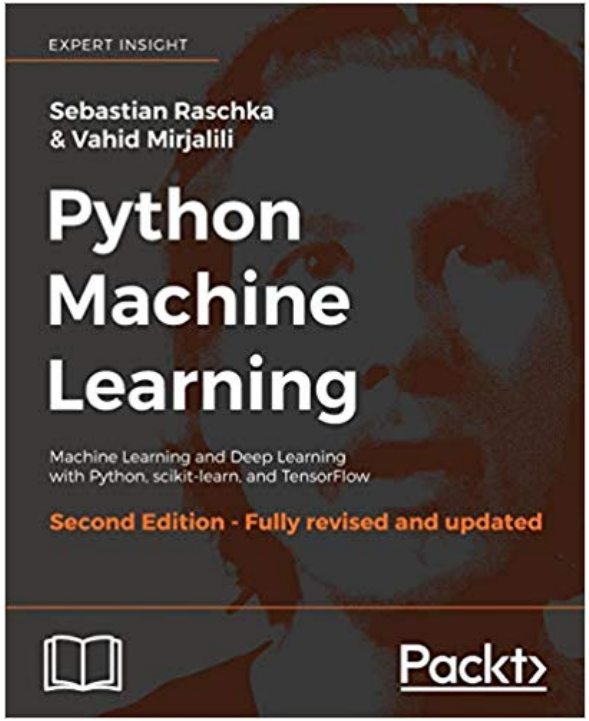


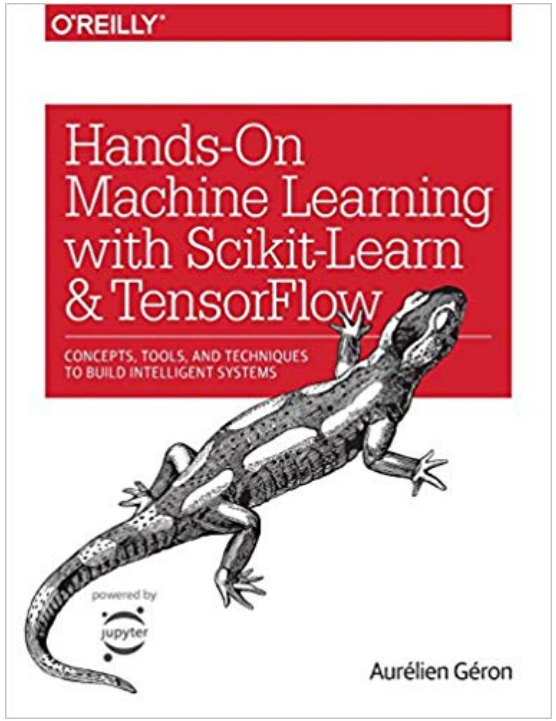
https://xkcd.com/1838/

Literature

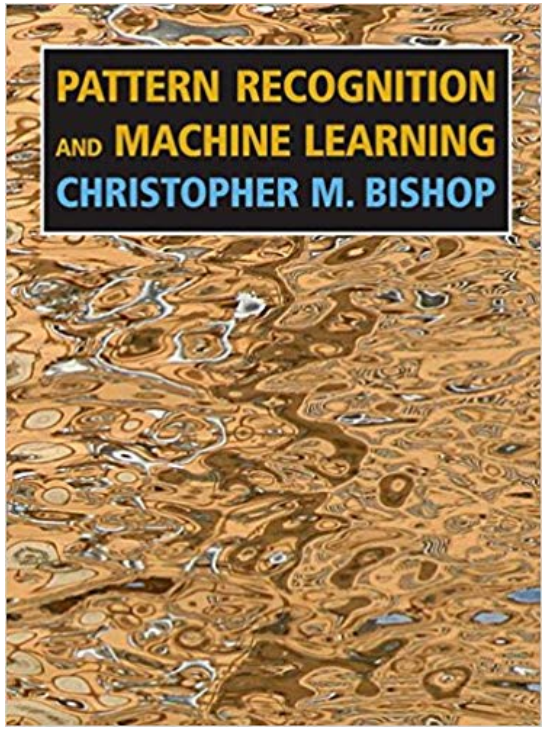
INSTITUTE FOR MACHINE LEARNING AND ANALYTICS

„Hands on“ books:





Literature



„complete“ theory

Literature

INSTITUTE FOR MACHINE LEARNING AND ANALYTICS

Learning by doing:

www.kaggle.com