Introduction to unsupervised learning

OMRI MENDELS, NDG, INTEL

omri.mendels@intel.com

Unsupervised learning
Overview

Learning without a teacher

Generally, given a collection of observations $X_1, \dots, X_n$, sampled from a distribution $p(X)$, describe the properties of $p(X)$

Usually refers to:
- Clustering
- Dimensionality reduction such as principal components analysis and multidimensional scaling
- Association rules discovery

Unsupervised learning
Motivation

Targets may be hard to obtain / boring to generate

- Because they're expensive
- Because they require a lot of work to produce

Targets may be unknown

- Which users behave similarly
- The places a future user would visit
- Topics in an unseen corpus

Clustering
Definitions

The process of **finding groups** in data.

The process of dividing the data into groups, where **points within each group are close** (or similar) to each other.

The process of dividing the data into groups, where points within each group are close (or similar) to each other, and where **points of different groups are far** (or dissimilar) from each other.

The process of **dividing the feature space into regions with relatively high density of points**, separated by regions with relatively low density of points.

Clustering

Taxonomy of clustering methods

Hierarchical vs. **Partitional** Methods

Agglomerative vs. **Divisive** Methods (bottom up vs. top down)

Hard vs. **fuzzy** clusters

- Allowing samples to belong to more than one cluster

Deterministic vs. **probabilistic** clusters

- Estimate the probability of a sample belonging to a cluster

Deterministic vs. **stochastic** algorithms

Incremental vs. **non-incremental** (online vs. offline)

K-means clustering

<u>Problem setting:</u>

We know (or assume) how many clusters there are

We don't know which point belongs to which cluster

Two examples:

https://www.youtube.com/watch?v=BVFG7fd1H30

http://shabal.in/visuals/kmeans/3.html

K-Means clustering
Combinatorial Approach

In how many ways can we assign *K* labels to *N* observations?

For each such possibility, we can compute a cost. Pick up the assignment with best cost.

Number of possible assignments

$$S(N, K) = \frac{1}{K!} \sum_{k=1}^{K} (-1)^{K-k} \binom{K}{k} k^N$$

$$S(10, 4) = 34105, \text{ but } S(19, 4) \simeq 10^{10}$$

NP-Hard!

K-means clustering

Algorithm

1. Initialization: Perform a random selection of centroids
2. For each sample
   1. Find the closest centroid
   2. Assign the sample to the corresponding cluster
3. Recompute the centroid of that cluster
4. Repeat step 2 until a convergence criterion is met or after MAX_ITERATIONS

K-means clustering

Main assumptions:

1. Each cluster is spherical
2. The data can be naturally clustered into distinct k clusters
3. No noise in data

Things to consider:

1. K – the number of clusters. How to choose? A large number vs. small
2. Distance measure (usually squared Euclidean distance) – is this always the right choice?
3. Initial cluster positions
4. Curse of dimensionality
5. Is the solution optimal?

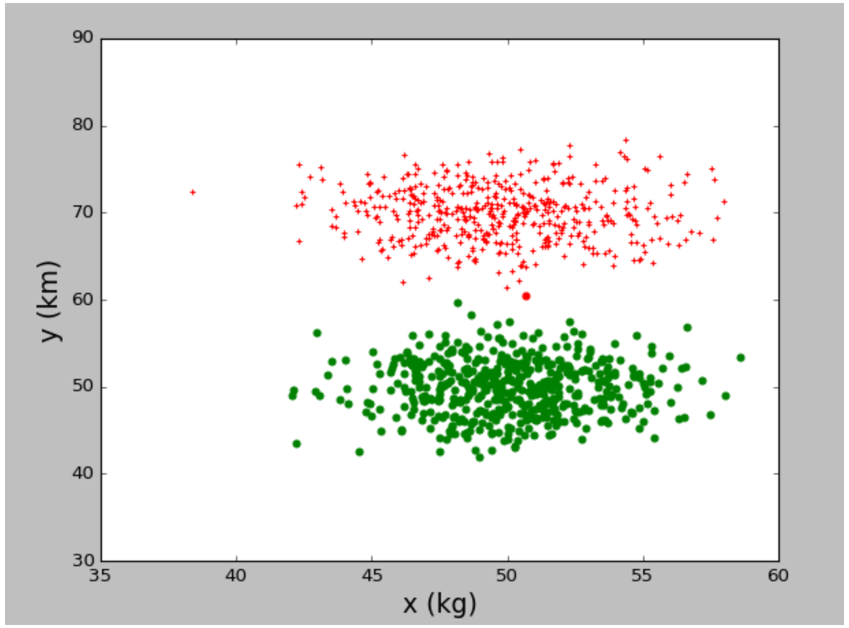
K-means clustering
Distance metric

K-means uses squared Euclidean distance. It’s optimization function is

$$\underset{S}{\arg\min} \sum_{i=1}^{k} \sum_{x \in S_i} \|x - \mu_i\|^2$$

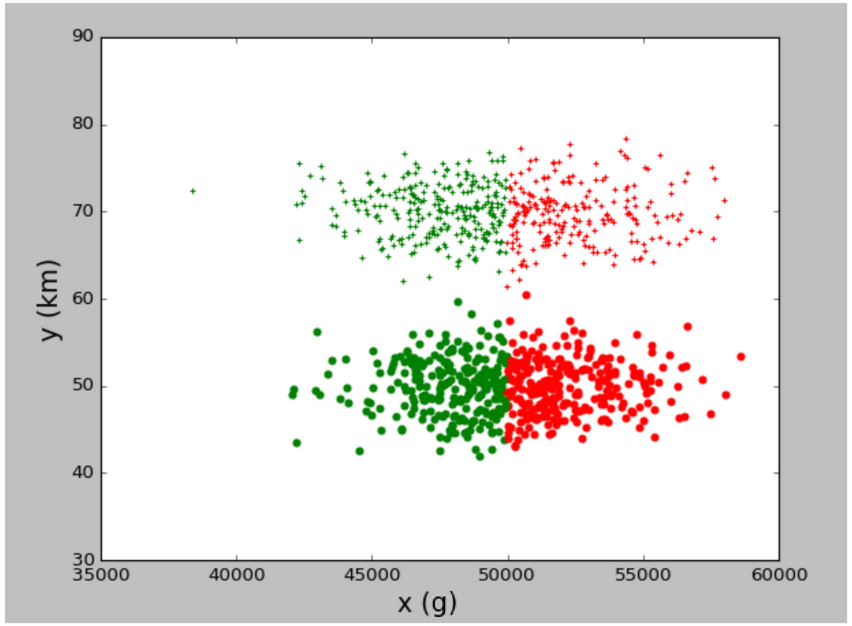
Where S is the set of samples, S_i is the set of samples currently associated with cluster i and x is a d dimensional vector

K-means clustering
Distance metric



Units = kg, perfect clusters

Vs.



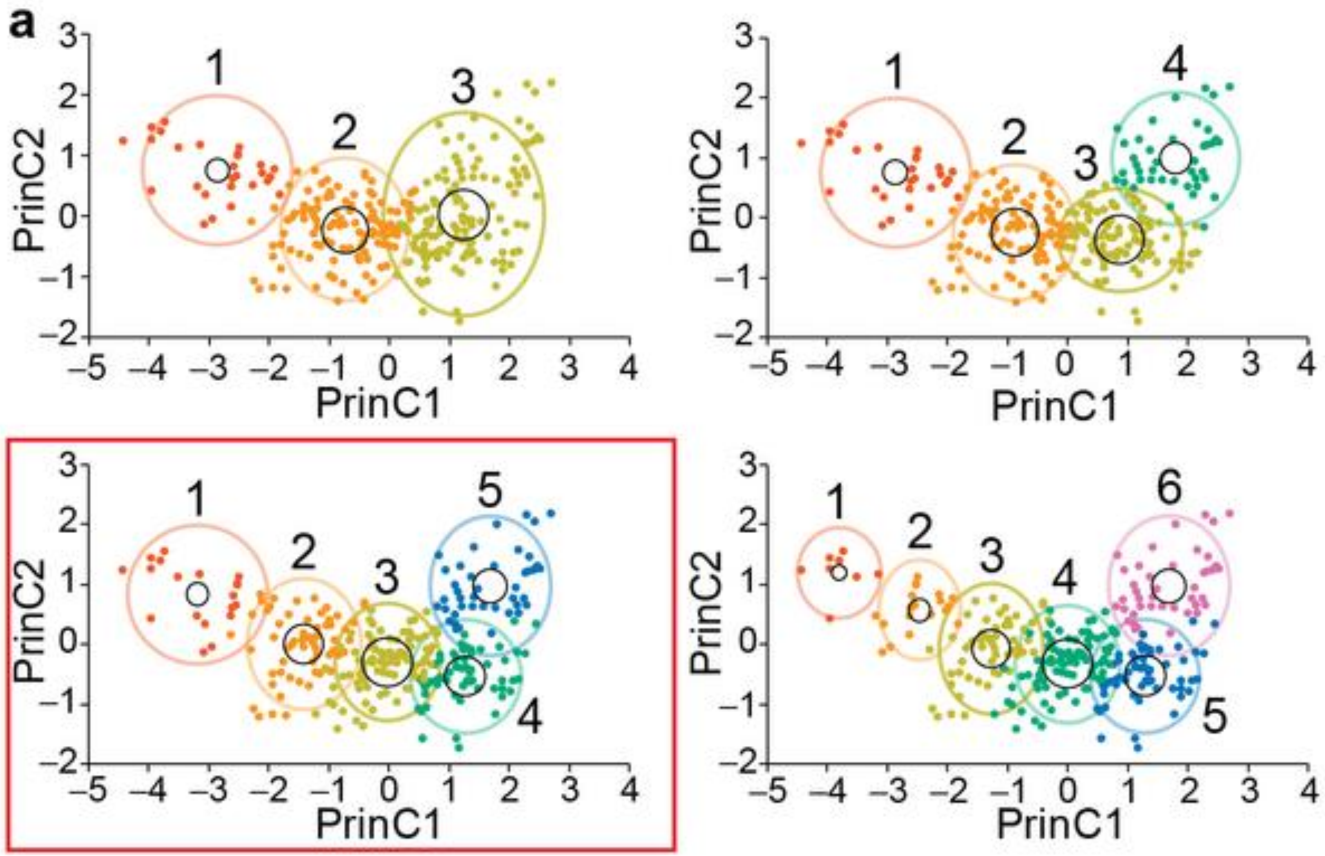
Units = g, clustering fails

L2 norm gives equal weight to each dimension

K-Means clustering
Choosing K

Elbow method

X-means

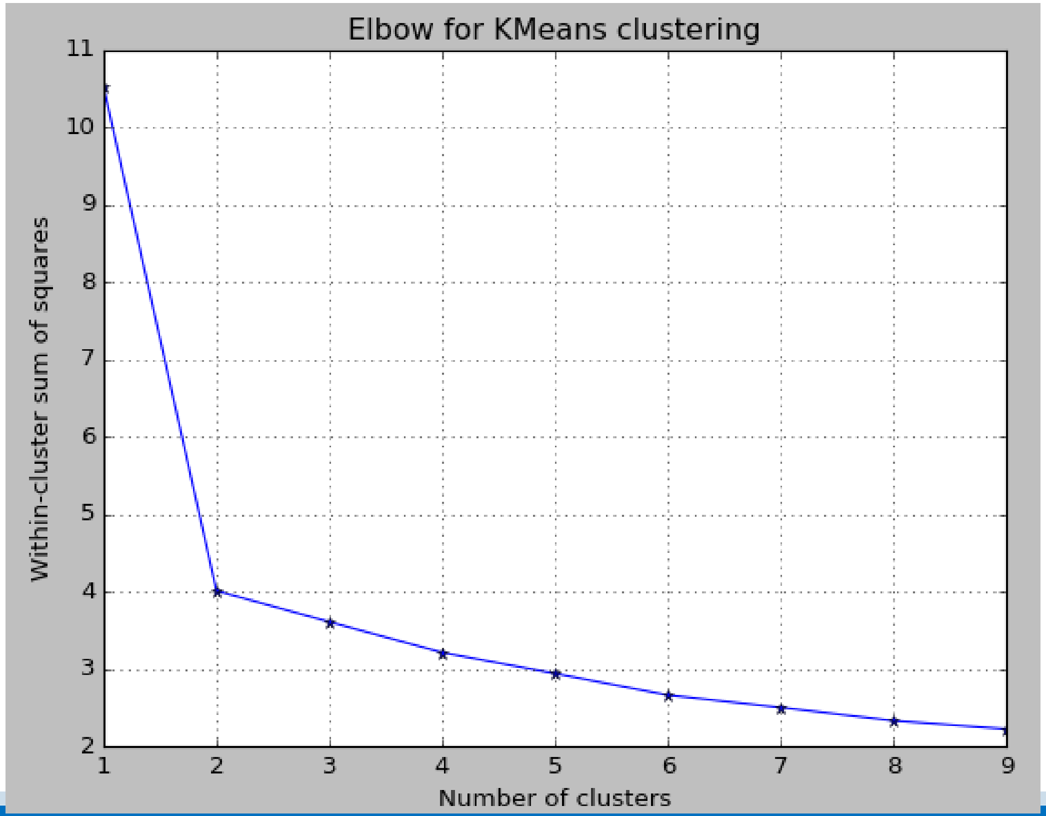


K-Means clustering
Elbow method

Look at the error (specifically sum of squared errors) to understand how compact clusters are.

$$SSE = \sum_{i=1}^{k} \sum_{x \in c_i} (x - m_i)^2$$

m_i = centroid of cluster i



K-Means Clustering

X-means algorithm

Input: data D, $K = \{k_j, \ldots, km\}$

1. Run k-means
2. For each cluster, re run k means with k=2 on each cluster found in 1
3. Evaluate k=2 vs k=1 and decide if to split or not (using BIC or other criteria)

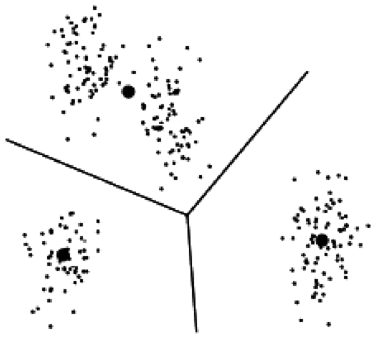
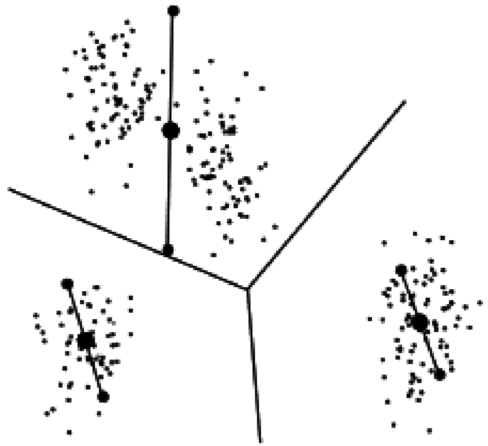


Figure 1. The result of running K-means with three centroids.



K-Means Clustering

X-means algorithm

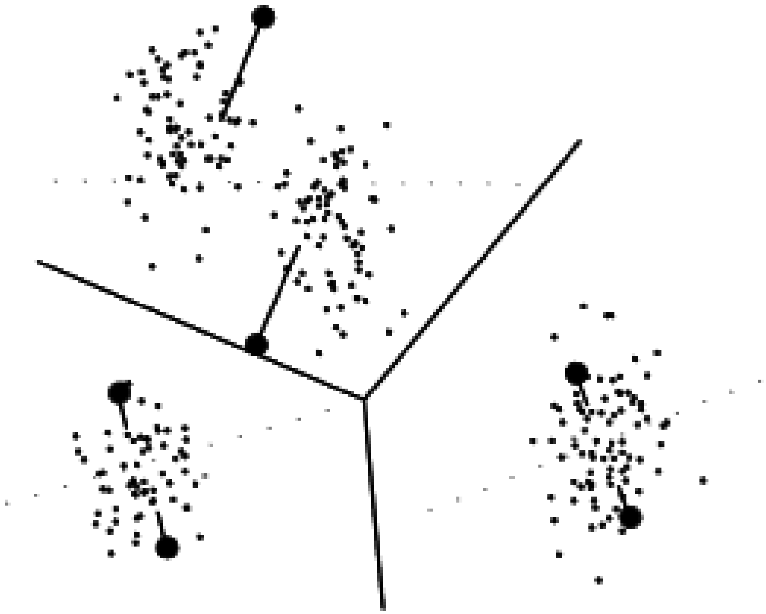


Figure 3: The first step of parallel local 2-means. The line coming out of each centroid shows where it moves to.

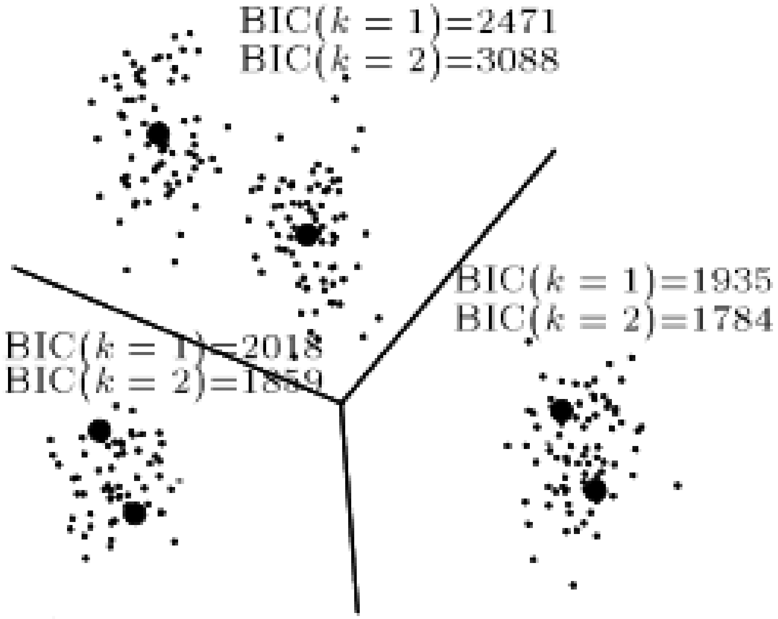


Figure 4: The result after all parallel 2-means have terminated.

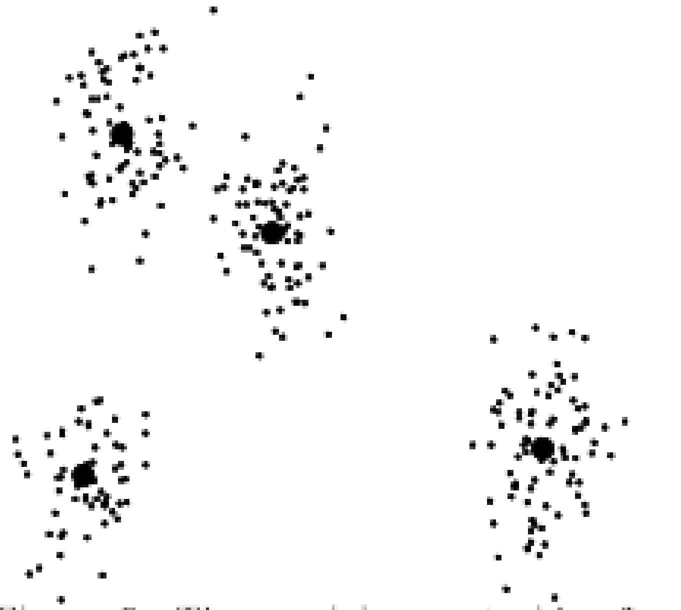
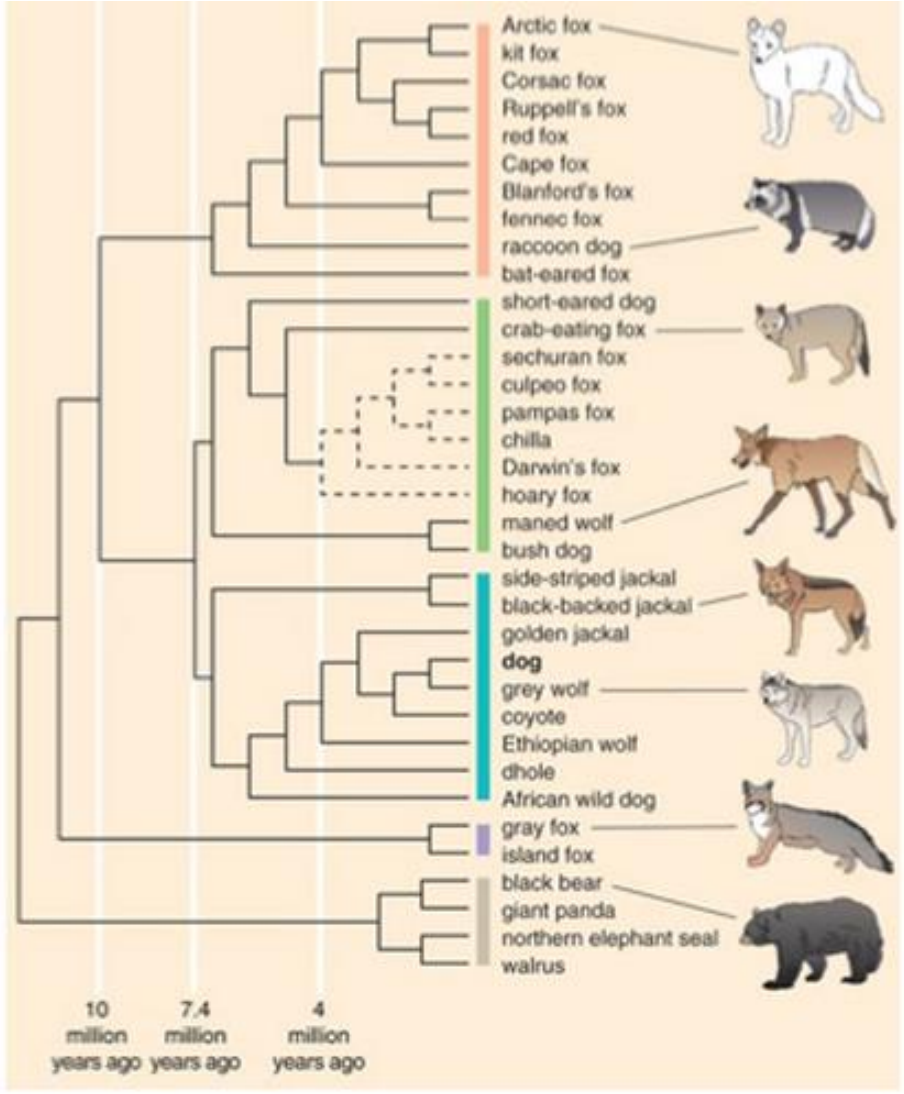


Figure 5: The surviving centroids after all the local model scoring tests.

Hierarchical Clustering

Bottom up approach

- Start where every sample is a cluster
- Group pairs of clusters each time
- Create a new cluster out of the grouped pair
- Iterate until all data is grouped in a single cluster



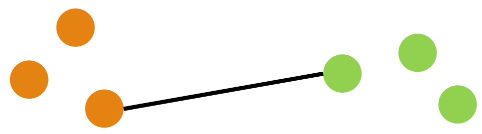
http://www.cs.princeton.edu/courses/archive/spr08/cos424/slides/clustering-2.pdf

Hierarchical Clustering

Grouping pairs of clusters

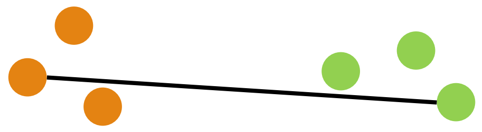
Nearest neighbor or **Single-link**

- Distance between closest elements in clusters
- $D(c_1, c_2) = \min(D(x_1, x_2))$



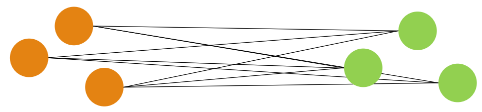
Complete link

- Distance between farthest elements in clusters
- $D(c_1, c_2) = \max(D(x_1, x_2))$



Average link

- Average of all pairwise distances
- $D(c_1, c_2) = \frac{1}{|c_1|}\frac{1}{|c_2|}\sum_{x \in c_1}\sum_{x \in c_2} D(x_1, x_2)$

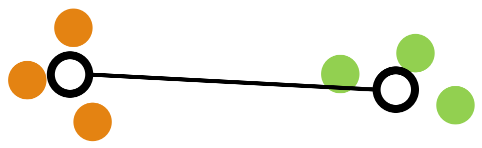


Hierarchical Clustering

Grouping pairs of clusters

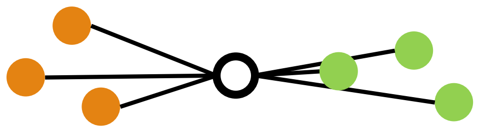
Centroids

- Distance between centroids (means) of two clusters
- $D(C_1, C_2) = D\left(\frac{1}{|c_1|}\sum_{x \in c_1} \vec{x}\,, \frac{1}{|c_2|}\sum_{x \in c_2} \vec{x}\right)$

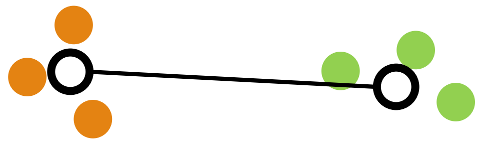


Ward's method

- How does the total distance from centroids (TD) change when joining two clusters
- $TD_{c_1 \cup c_2} = \sum_{x \in c_1 \cup c_2} D\left(x, \mu_{c_1 \cup c_2}\right)^2$



Vs.



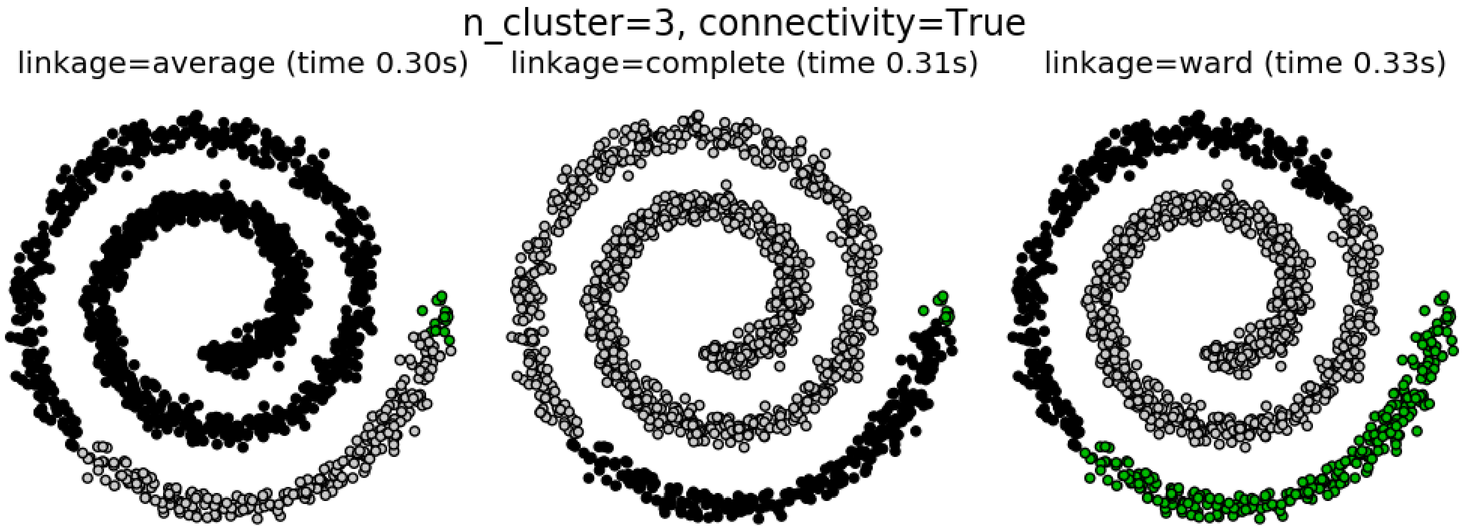
Hierarchical Clustering
Grouping pairs of clusters

If the clusters are natural (tight and roundish…), all of these methods produce similar results.

That's often not the case!

When to use which? What about outliers?

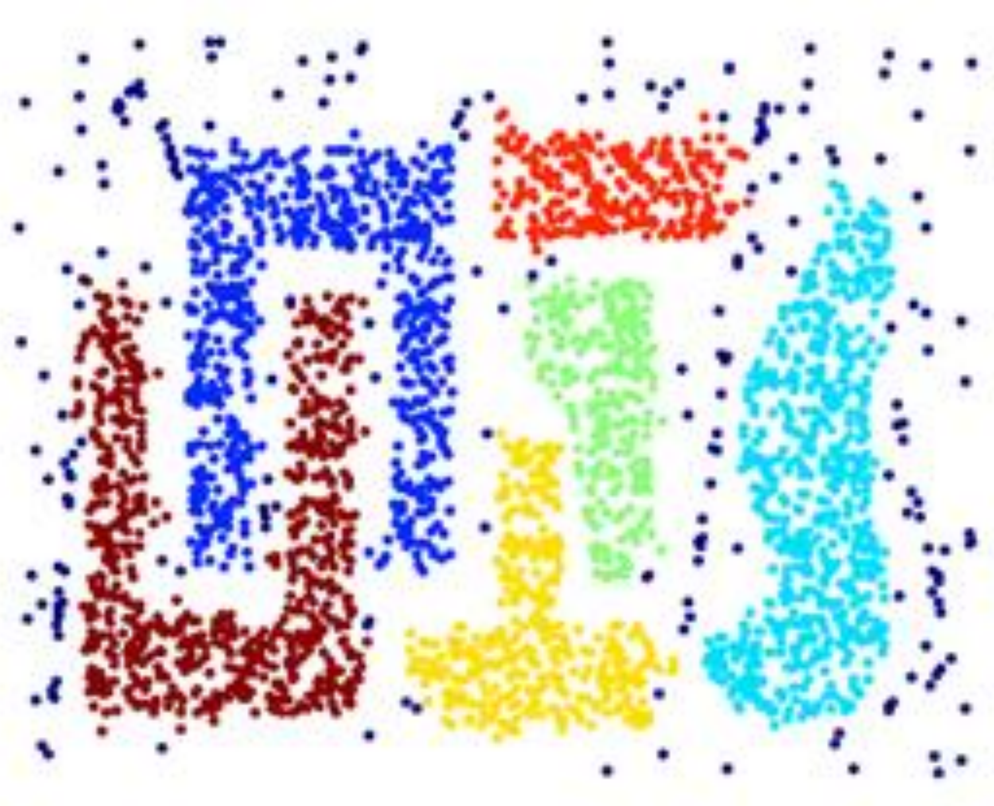
- Coordinates of all visits to my home
- Clustering users according to behavior



DBSCAN

Density based clustering

- Can cluster non convex shapes
- Doesn't necessarily cluster all points
- More on this tomorrow!



Association rules mining

Association rules Mining
Problem setting

Originally created for shopping

Understanding buying patterns can help to increase sales

Data: list of transactions. Each shopper has a list of items he/she bought

Main objective: find pairs or (sets of) items that are bought together, and find items that influence on the purchase of other items

Association rules analysis is a technique to **uncover how items are associated** to each other.

Association rules Mining

Metrics

Two main metrics to measure association:

Support: The number of times a subset of items appear together

Confidence: How likely item Y is purchased when item X is purchased, expressed as {X -> Y}

Support {🍎} = $\frac{4}{8} = P(apple)$

Confidence {🍎 → 🍺} = $\frac{\text{Support \{🍎, 🍺\}}}{\text{Support \{🍎\}}}$ $\left(P(x|y) = \frac{p(x,y)}{p(y)} \right)$

Transaction 1	🍎 🍺 🍚 🍗
Transaction 2	🍎 🍺 🍚
Transaction 3	🍎 🍺
Transaction 4	🍎 🍐
Transaction 5	🍼 🍺 🍚 🍗
Transaction 6	🍼 🍺 🍚
Transaction 7	🍼 🍺
Transaction 8	🍼 🍐

Transaction	Support	Confidence
Canned Beer → Soda	1%	20%
Canned Beer → Berries	0.1%	1%
Canned Beer → Male Cosmetics	0.1%	1%

Association Rules Mining

Apriori Algorithm

Step 0. Start with itemsets containing just a single item, such as {apple} and {pear}.

Step 1. Determine the support for itemsets. Keep the itemsets that meet your minimum support threshold, and remove itemsets that do not.

Step 2. Using the itemsets you have kept from Step 1, generate all the possible itemset configurations.

Step 3. Repeat Steps 1 & 2 until there are no more new itemsets.

Association Rules Mining

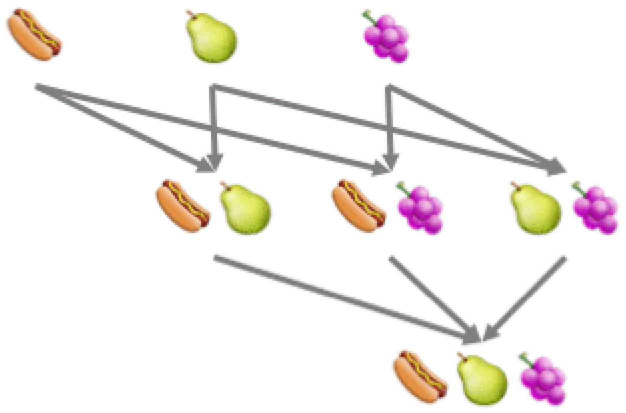
Apriori Algorithm - Example

0. Start with itemsets containing just a single item.

1. Determine the support for itemsets. Keep the itemsets that meet your minimum support threshold, and remove itemsets that do not.

2. Using the itemsets you have kept from Step 1, generate all the possible itemset configurations.

3. Repeat Steps 1 & 2 until there are no more new itemsets.



Wrapping up

Unsupervised learning

Used for EDA (Exploratory data analysis)

- Understanding data structure, distribution, dimensionality

Also used in live systems

- Visit/stay detection
- Pretraining for deep neural networks
- Image segmentation

For intelligent personalization:

- Association rules mining for predicting user actions given context
- Clustering for home/work detection
- Clustering of arrival and leave times for routine detection
- Clustering of users by behavioral traits

References

Introduction (http://www.uio.no/studier/emner/matnat/ifi/INF3490/h15/lectures/oneinone_ul_lecture.pdf)

Definitions (http://www.ee.columbia.edu/~vittorio/UnsupervisedLearning.pdf)

Apriori (http://www.kdnuggets.com/2016/04/association-rules-apriori-algorithm-tutorial.html)

X-means (http://www.aladdin.cs.cmu.edu/papers/pdfs/y2000/xmeans.pdf)

Hierarchical clustering (https://www.youtube.com/watch?v=VMyXc3SiEqs)