

CLUSTERING

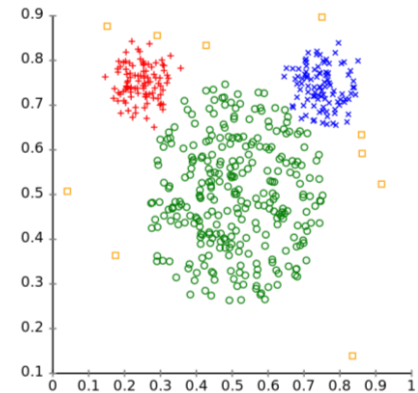


WHAT IS CLUSTERING

Unsupervised Learning (no labels/responses)

- **Clustering**
- Subspace Learning
- Manifold Learning

Finding structure in data.



Clustering Problem.

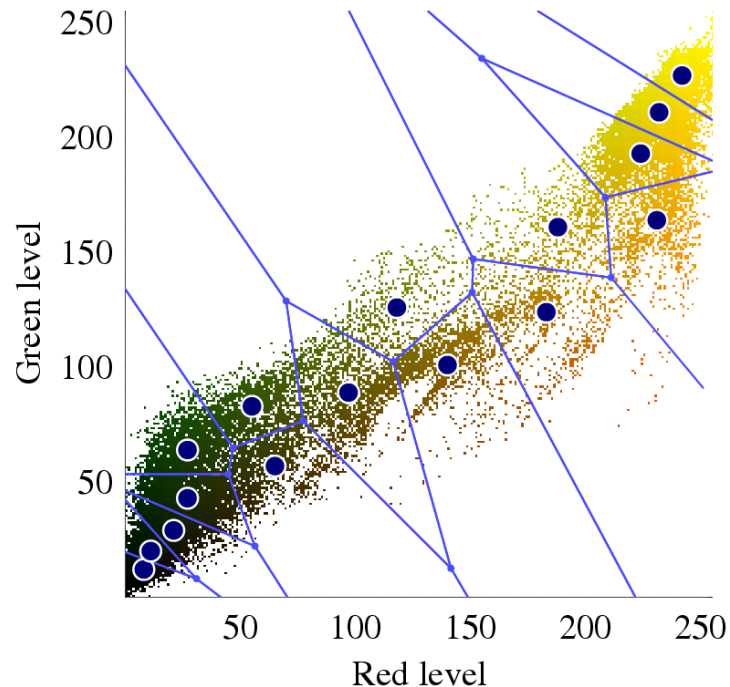
Input. Training data $\mathcal{S}_n = \{x^{(1)}, x^{(2)}, \dots, x^{(n)}\}$, each $x^{(i)} \in \mathbb{R}^d$.
Integer k

Output. Clusters $\mathcal{C}_1, \mathcal{C}_2, \dots, \mathcal{C}_k \subset \{1, 2, \dots, n\}$ such that every data point is in one and only one cluster.



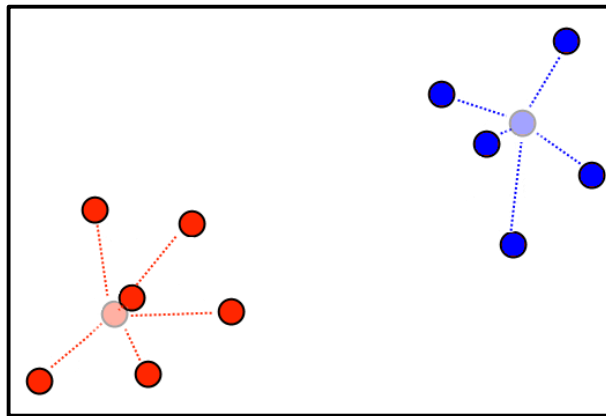
USES OF CLUSTERING

- Improve classification/regression (semi-supervised learning)
- Data compression

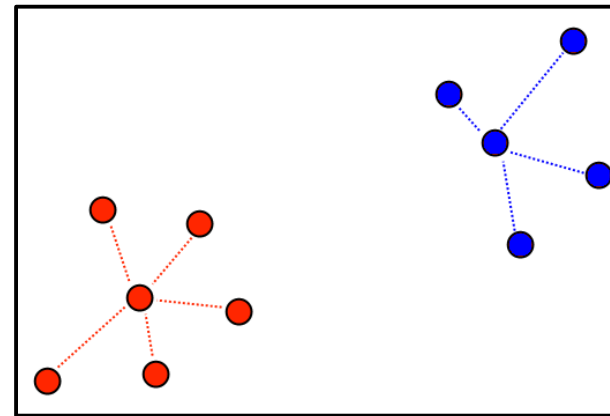


HOW TO SPECIFY A CLUSTER

1. By listing all its elements
2. Using a representative
 - a. A point in center of cluster (centroid)
 - b. A point in the training data (exemplar)



centroid



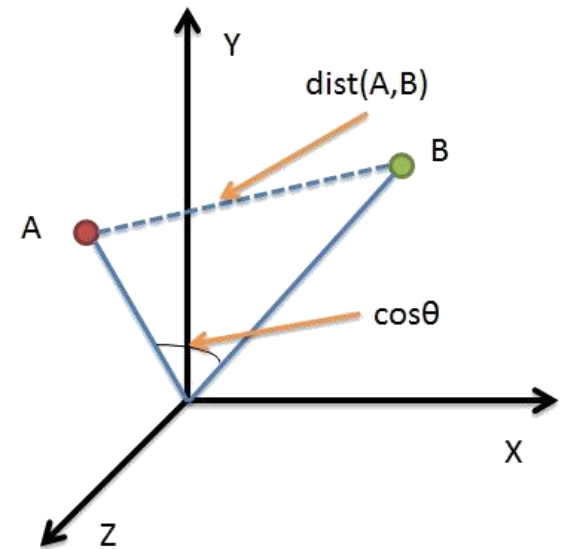
exemplar



LOSS FUNCTION

Distance metrics.

A measure of how similar or how close two data points are. Nearby points are more likely they belong to the same cluster.



- Cosine Similarity $\cos(x, x') = \frac{x \cdot x'}{\|x\| \|x'\|}$
- Euclidean Distance $\text{dist}(x, x') = \|x - x'\|^2$



OBJECTIVE FUNCTION

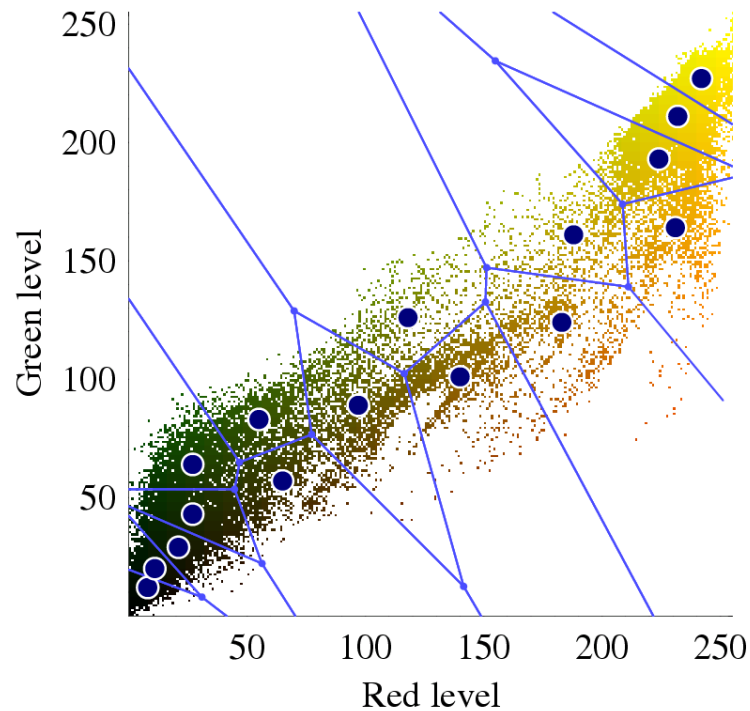
Cost of Clustering.

$$\text{cost}(\mathcal{C}_1, \dots, \mathcal{C}_k, z^{(1)}, \dots, z^{(k)}) = \sum_{j=1}^k \sum_{i \in \mathcal{C}_j} \|x^{(i)} - z^{(j)}\|^2$$

- Function of clusters $\mathcal{C}_1, \dots, \mathcal{C}_k$ and representatives $z^{(1)}, \dots, z^{(k)}$
- In some cases, it is enough to keep track of representatives.



VORONOI DIAGRAM



In some cases, it is enough to keep track of representatives.



OPTIMIZATION ALGORITHM

Goal. Minimize $\mathcal{L}(x, y)$.

Coordinate Descent (Method 1).

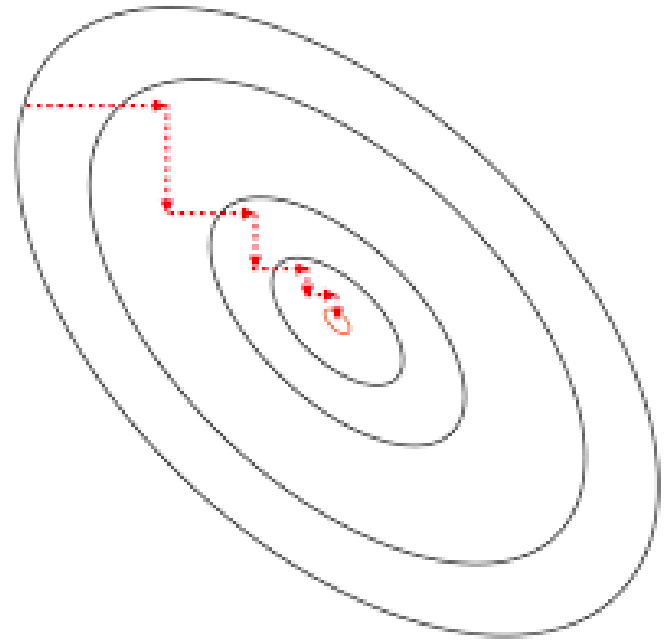
Repeat until convergence:

1. Move in direction of $\partial\mathcal{L}/\partial x$.
2. Move in direction of $\partial\mathcal{L}/\partial y$.

Coordinate Descent (Method 2).

Repeat until convergence:

1. Find optimal x while holding y constant.
2. Find optimal y while holding x constant.





K-MEANS



OBJECTIVE FUNCTION

Define cluster \mathcal{C}_j to be the set of points $x^{(i)}$ whose closest representative is $z^{(j)}$.

$$\text{cost}(z^{(1)}, \dots, z^{(k)}) = \sum_{i=1}^n \min_{1 \leq j \leq k} \|x^{(i)} - z^{(j)}\|^2.$$



OPTIMIZATION ALGORITHM

Coordinate Descent.

Repeat until convergence:

- Find best clusters given centroids
- Find best centroid given clusters



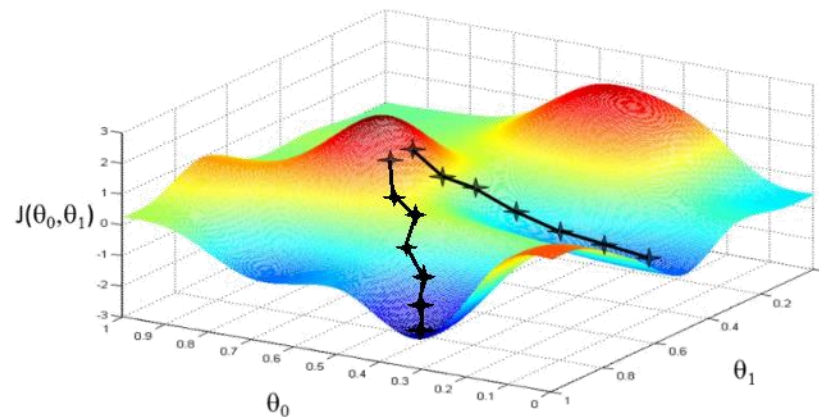
OPTIMIZATION ALGORITHM

1. Initialize centroids $z^{(1)}, \dots, z^{(k)}$.
2. Repeat until no further change in cost:
 - a. For each $j \in \{1, \dots, k\}$,
$$\mathcal{C}_j = \{ i \text{ such that } x^{(i)} \text{ is closest to } z^{(j)} \}.$$
 - a. For each $j \in \{1, \dots, k\}$,
$$z^{(j)} = \frac{1}{|\mathcal{C}_j|} \sum_{i \in \mathcal{C}_j} x^{(i)} \text{ (cluster mean)}$$



CONVERGENCE

- Cost always decreases in each step (coordinate descent).
- Converges to local minimum, not necessarily global minimum.



Challenge.

Why does the algorithm terminate in a finite number of steps?



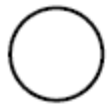


DISCUSSION



INITIALIZATION

Optimization



1



2



3

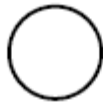


4

Starting position of centroids



1



2



3



4

Final position of centroids

Problem.

How to choose good starting positions?

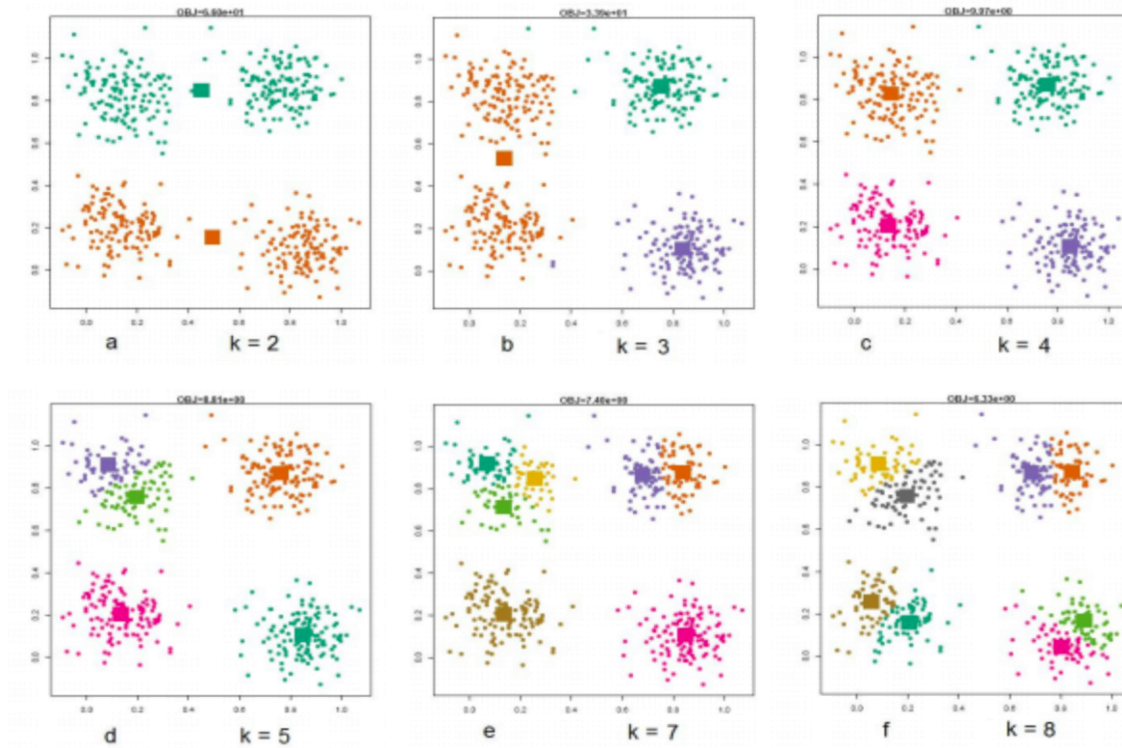
Solution.

Place them far apart with high probability.



NUMBER OF CLUSTERS

Generalization



Problem.

How do we choose k , the optimal number of clusters?

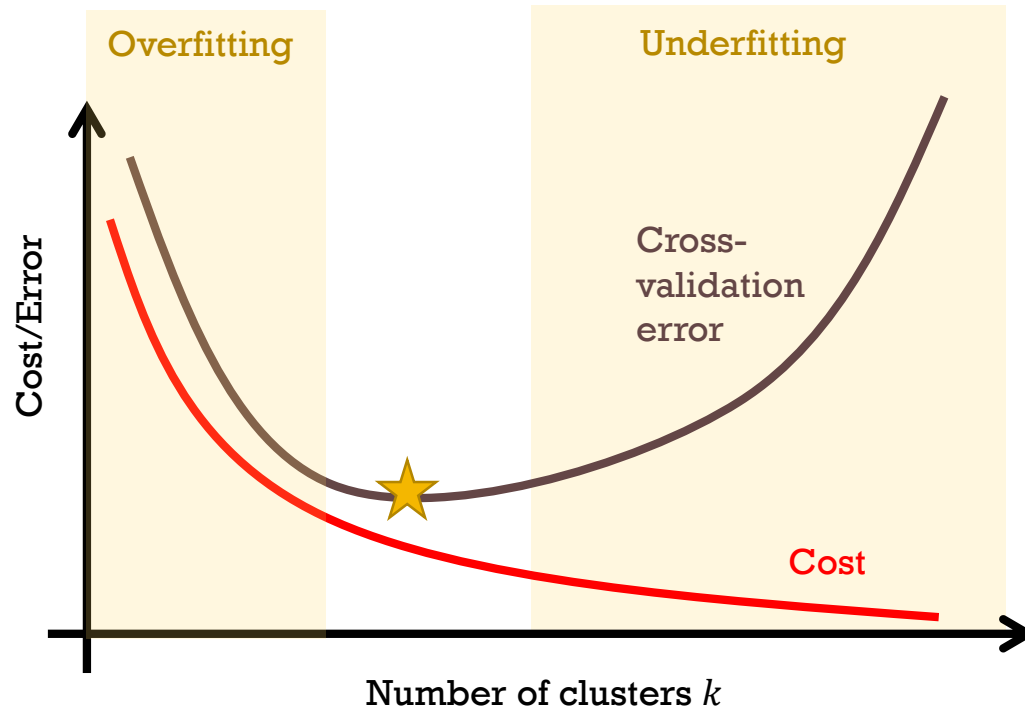
Solution.

Cross-validation.



SEMI-SUPERVISED LEARNING

Cross-validation error



K-MEDROIDS

Use exemplars
instead of centroids.

e.g. Google News.

Repeat until convergence:

- Find best clusters
given exemplars
- Find best exemplars
given clusters



People Are Drilling Headphone Jacks Into the **iPhone 7**

Fortune - 1 hour ago

He then takes the bit to the **iPhone 7** and drills a hole into the device. ... Instead, **Apple** shipped **iPhone 7** units with an adapter that lets users ...

iPhone 7 review: Not **Apple's** best

Expert Reviews - 2 hours ago

Please don't drill a headphone jack into your **iPhone 7**

BGR - 2 hours ago

Apple iPhone 7 Users: Please DO NOT Drill a 3.5mm Hole on it to ...

News18 - 7 hours ago

Video claiming drilling into **iPhone 7** will reveal hidden headphone ...

Highly Cited - **The Guardian** - 1 hour ago

Clueless **iPhone 7** owners tricked into DRILLING hole in their ...

Highly Cited - **The Sun** - 24 Sep 2016



Expert Reviews



BGR



The Guardian



News18



International ...



Herald Sun

[View all](#)



Pegatron CEO slams analysts, 'cautiously optimistic' about **Apple** ...

AppleInsider (press release) (blog) - 3 hours ago

The CEO of **Apple's** manufacturing partner Pegatron notes that the **iPhone 7** is exceeding estimates on the strength of the phone alone, and ...

Google Nexus 2016' Specs: Solution to **Apple iPhone 7** ...

University Herald - 3 hours ago

Apple Supplier Pegatron Hints of Higher **iPhone 7** Demand while ...

Patently Apple - 2 hours ago

iPhone 7 vs Samsung Galaxy S7: Which is the best smartphone to ...

Alphr - 5 hours ago

Samsung Galaxy Note 7 Explosions Boost **iPhone 7** Sales, Top ...

Softpedia News - 8 hours ago



University He...



Patently Apple



Alphr



Softpedia News



Expert Reviews

[View all](#)

SUMMARY

- Clustering
 - Distance Metric
 - Cost of Clustering
- Representatives
 - Centroids
 - Exemplars
 - Voronoi Diagrams
- Optimization
 - Coordinate Descent
 - Initialization
- Generalization
 - Number of Clusters
- Uses
 - Data Compression
 - Semi-Supervised Learning

