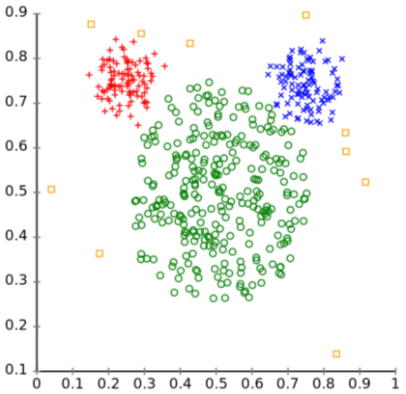
CLUSTERING

WHAT IS CLUSTERING

Unsupervised Learning (no labels/responses)

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

Finding structure in data.



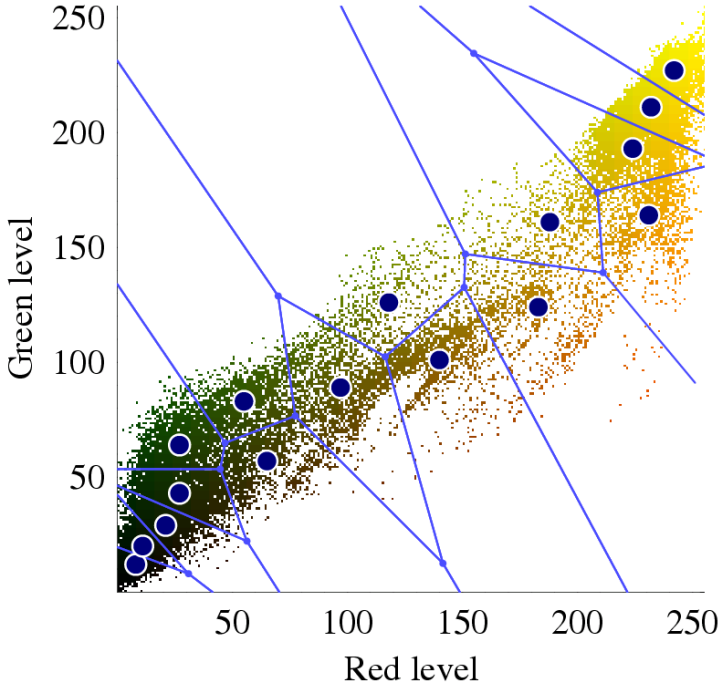
Clustering Problem.

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

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

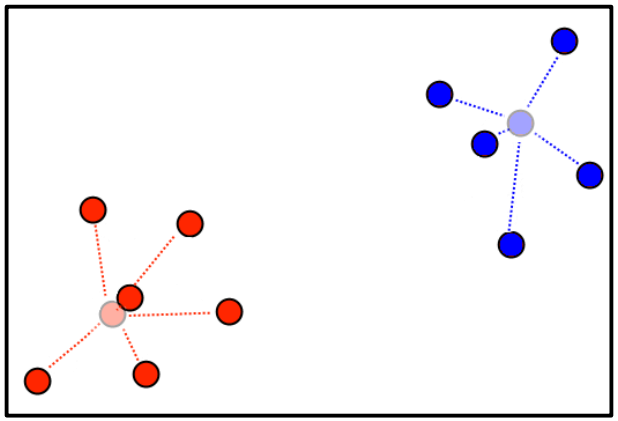
USES OF CLUSTERING

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

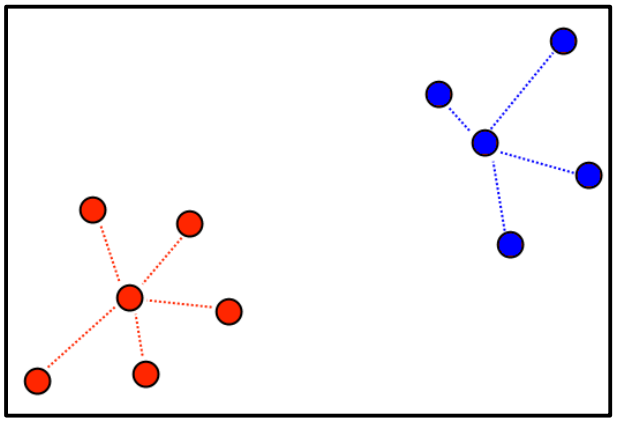


HOW TO SPECIFY A CLUSTER

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)

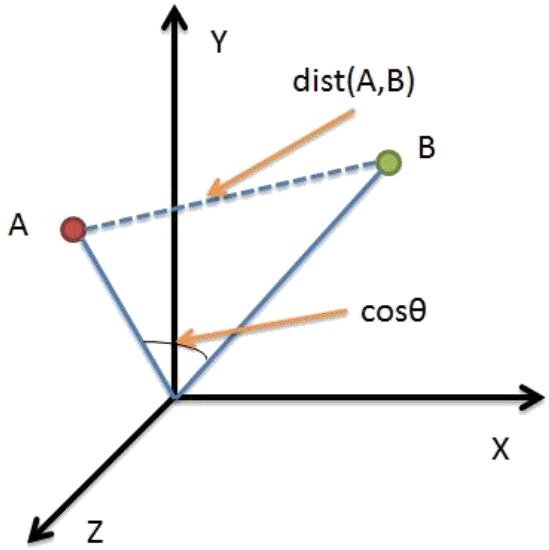


centroid



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 $\text{dist}(x, x') = \|x - x'\|^2$

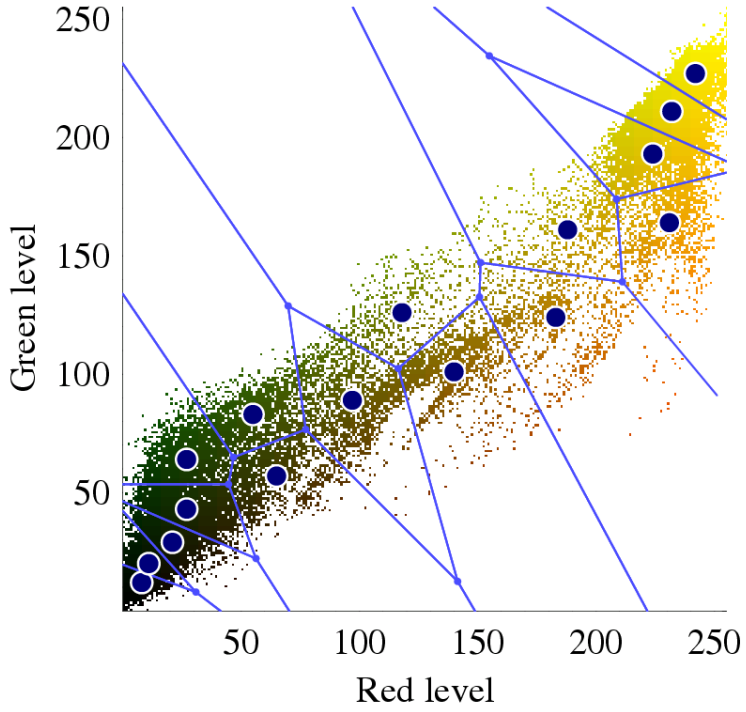
OBJECTIVE FUNCTION

Cost of Clustering.

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

- 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.

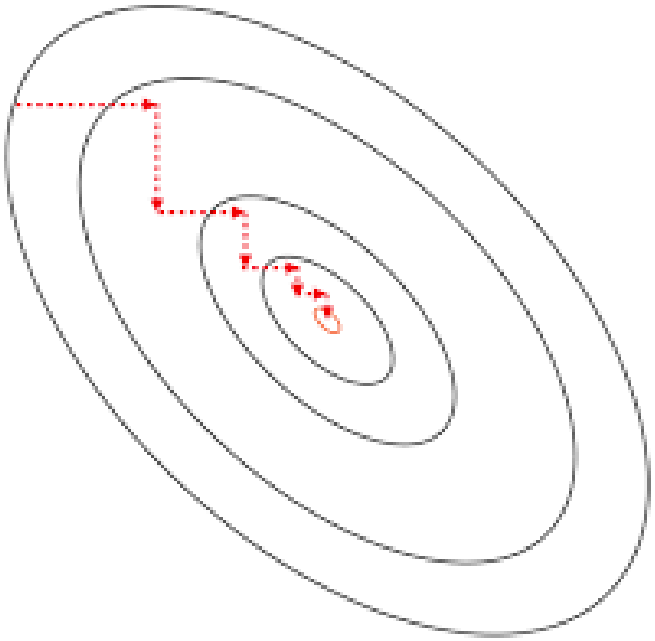
VORONOI DIAGRAM



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 $z^{(j)}$.

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

OPTIMIZATION ALGORITHM

Coordinate Descent.

Repeat until convergence:

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

OPTIMIZATION ALGORITHM

1. Initialize centroids $z^{(1)}, \ldots, z^{(k)}$.
2. Repeat until no further change in cost:

   a. For each $j \in \{1, \ldots, k\}$,

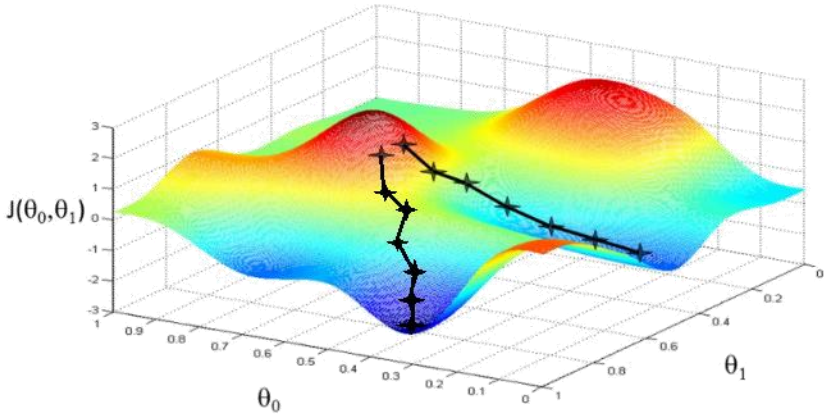
   $$\mathcal{C}_j = \{\, i \text{ such that } x^{(i)} \text{ is closest to } z^{(j)} \,\}.$$

   a. For each $j \in \{1, \ldots, k\}$,

   $$z^{(j)} = \frac{1}{|\mathcal{C}_j|} \sum_{i \in \mathcal{C}_j} x^{(i)} \quad \text{(cluster mean)}$$

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

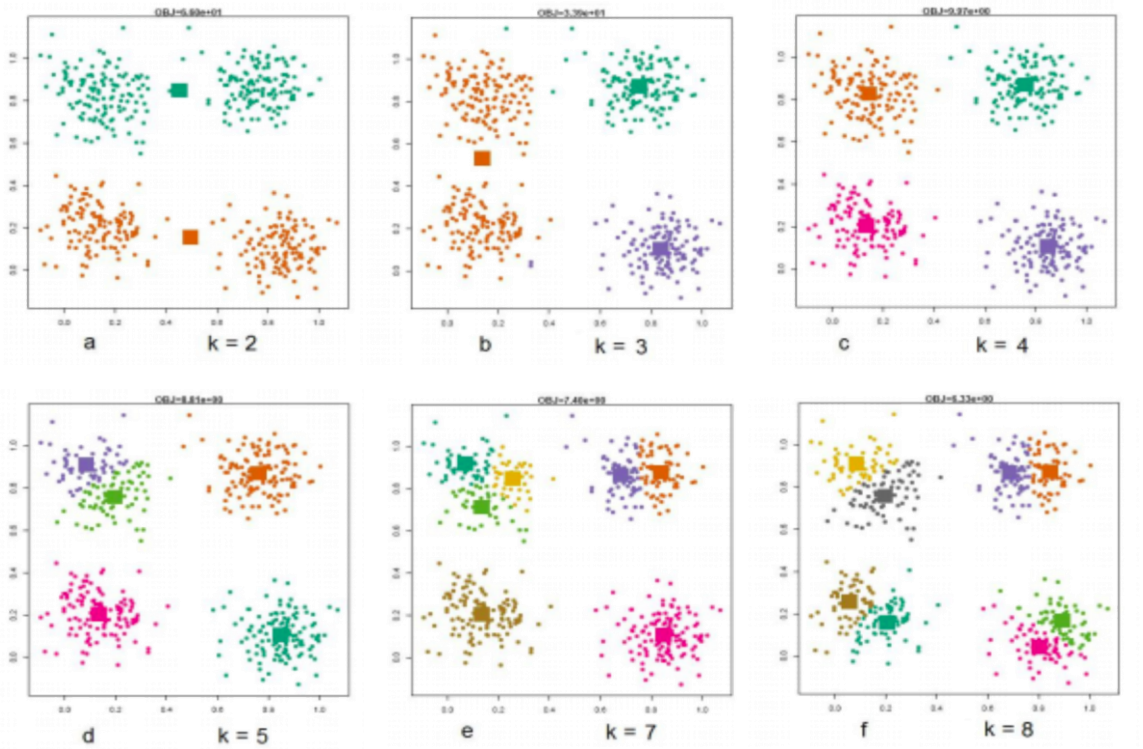
INITIALIZATION

Optimization

Problem. How to choose good starting positions?
Solution. Place them far apart with high probability.

NUMBER OF CLUSTERS

Generalization

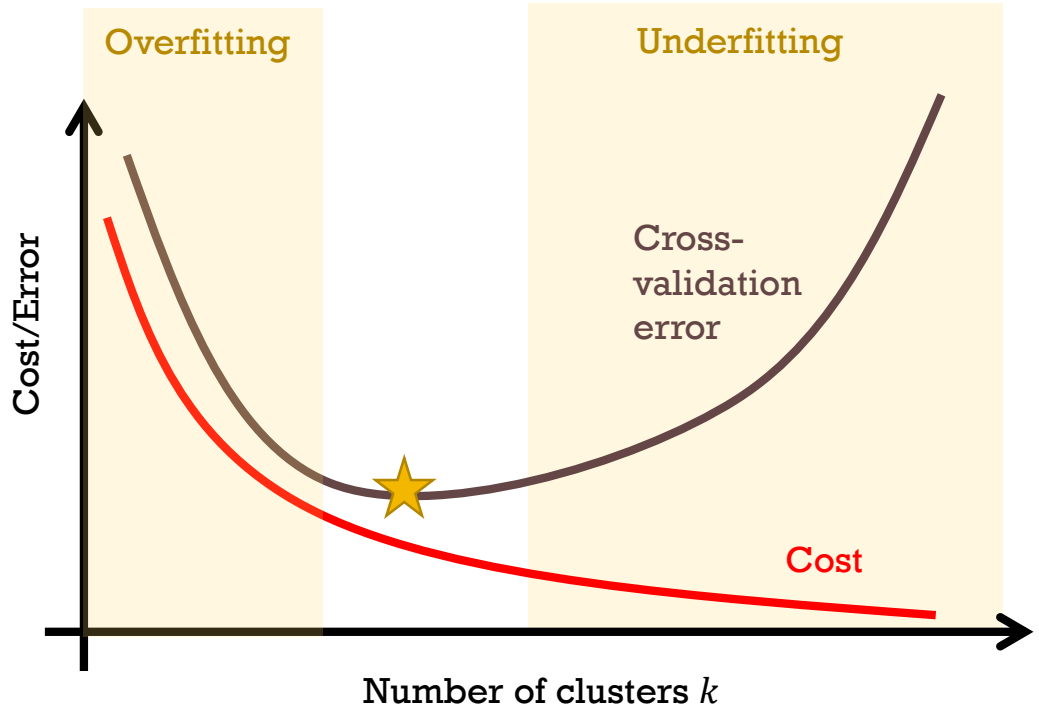


Problem. How do we choose k, the optimal number of clusters?

Solution. Cross-validation.

SEMI-SUPERVISED LEARNING

Cross-validation error



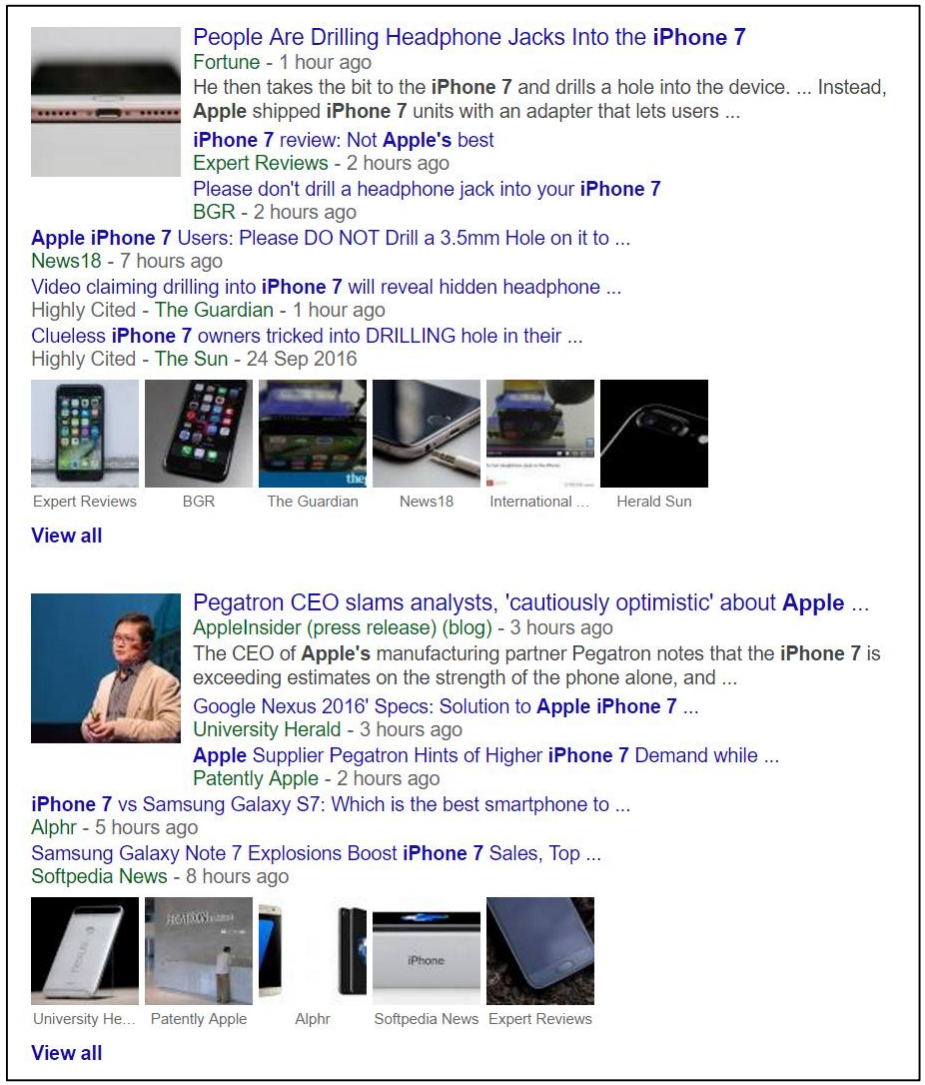
K-MEDROIDS

Use exemplars
instead of centroids.

e.g. Google News.

Repeat until convergence:

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



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