

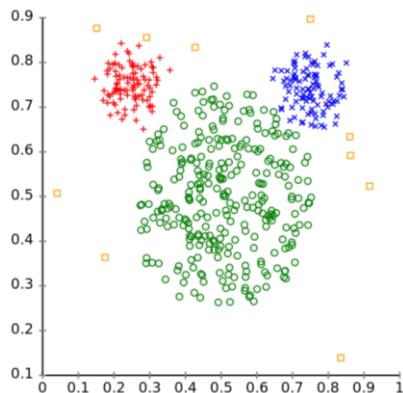


CLUSTERING



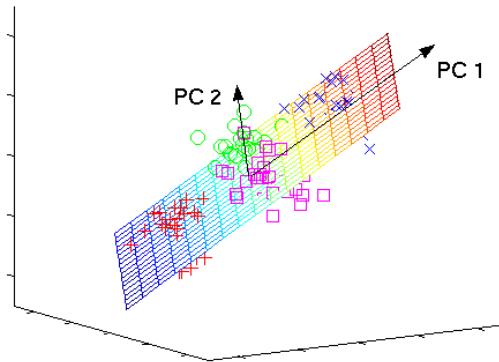
UNSUPERVISED LEARNING

- No labels/responses. Finding structure in data.
- Dimensionality Reduction.



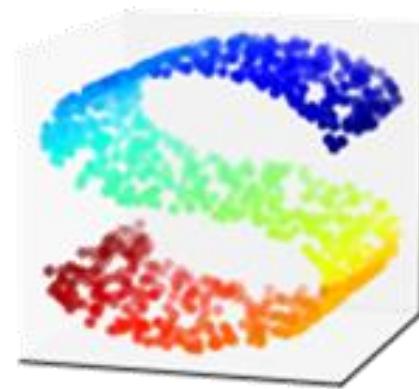
Clustering

$$T: \mathbb{R}^d \rightarrow \{1, 2, \dots, k\}$$



Subspace Learning

$$T: \mathbb{R}^d \rightarrow \mathbb{R}^m$$

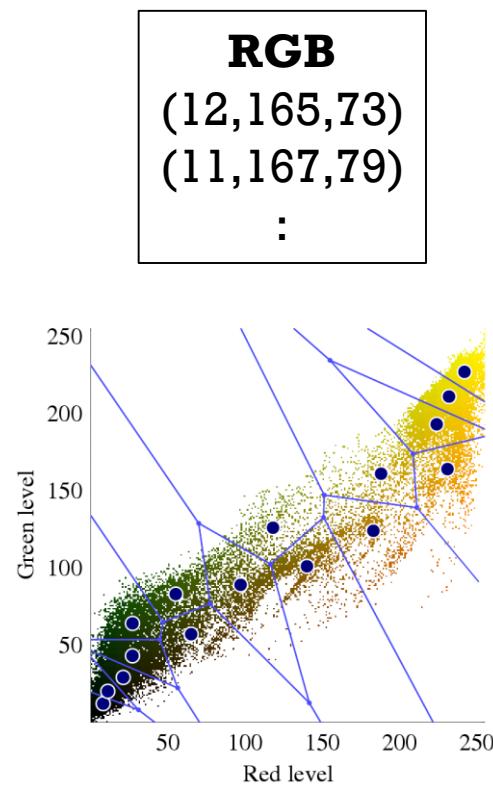


Manifold Learning



USES OF UNSUPERVISED LEARNING

- Data compression



Labels

3
43
:

Dictionary

1 ~ (10, 160, 70)
2 ~ (40, 240, 20)
:



Image compression

$K = 2$



$K = 3$



$K = 10$



Original image



USES OF UNSUPERVISED LEARNING

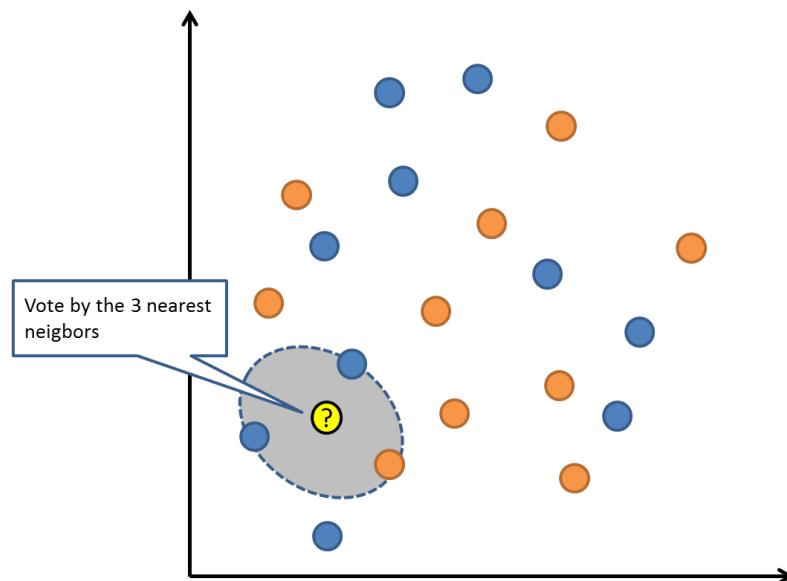
- Improve classification/regression (semi-supervised learning)
1. From *unlabeled data*, learn a good features $T: \mathbb{R}^d \rightarrow \mathbb{R}^m$.
 2. To *labeled data*, apply transformation $T: \mathbb{R}^d \rightarrow \mathbb{R}^m$.
$$(T(x^{(1)}), y^{(1)}), \dots, (T(x^{(n)}), y^{(n)})$$
 3. Perform classification/regression on transformed data.



K-NEAREST NEIGHBORS

Main Idea.

- Find a few users b_1, \dots, b_k (neighbors) that are similar to user a
- Use information from users b_1, \dots, b_k to predict ratings of user a



K-means vs K-nearest neighbours

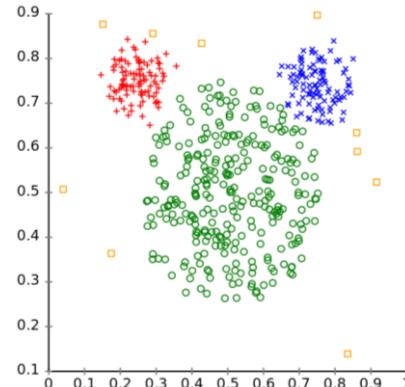
- K-means clustering is an unsupervised learning method, whereas KNN is a supervised learning technique.
- The 'K' in K-means refers to the number of clusters the algorithm is trying to learn from the data. The 'K' in KNN is the number of nearest neighbours used to classify a test sample.
- K-means has a training phase whereas KNN does not.

WHAT IS CLUSTERING

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.



Some clusters
could be empty!

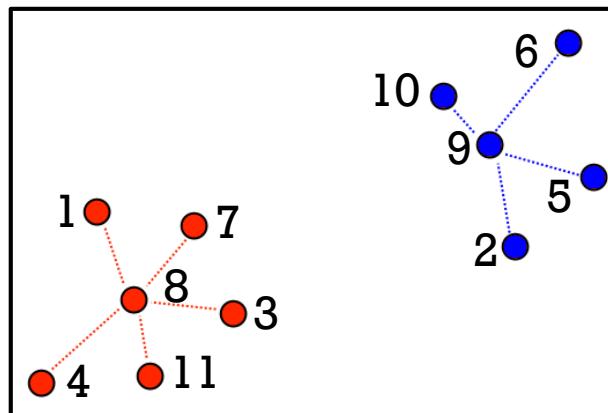


HOW TO SPECIFY A CLUSTER

- By listing all its elements

$$\mathcal{C}_1 = \{1,3,4,7,8,11\}$$

$$\mathcal{C}_2 = \{2,5,6,9,10\}$$



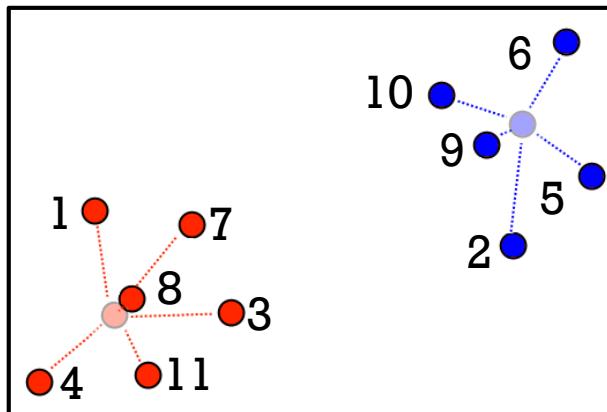
HOW TO SPECIFY A CLUSTER

Each point $x^{(i)}$ will be assigned the closest representative.

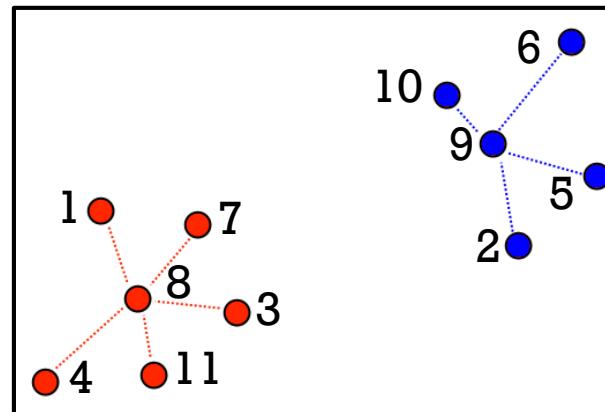
- Using a representative
 - a. A point in center of cluster (centroid)
 - b. A point in the training data (exemplar)

$$z^{(1)} = \begin{pmatrix} 1 \\ 1 \end{pmatrix}, z^{(2)} = \begin{pmatrix} 5 \\ 4 \end{pmatrix}$$

$$z^{(1)} = 8, z^{(2)} = 9$$



centroid

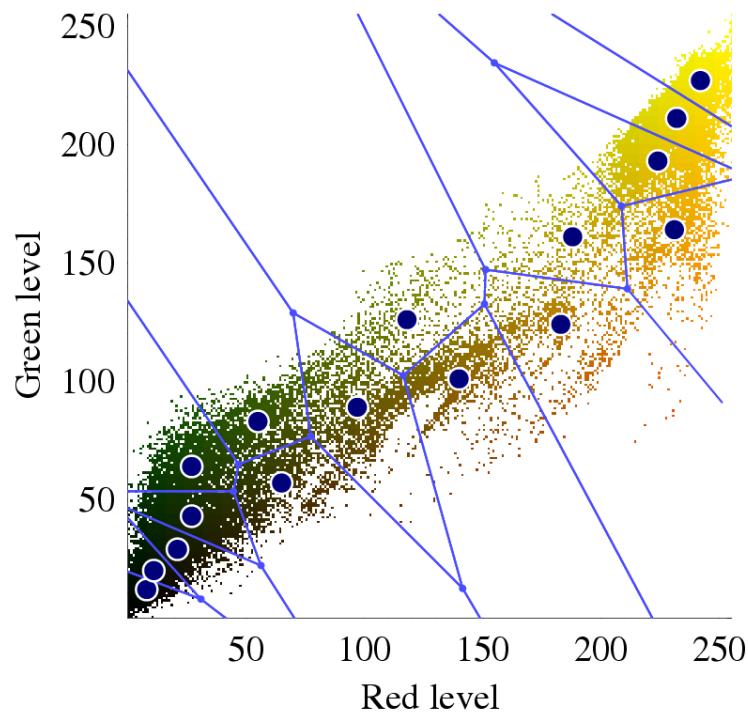


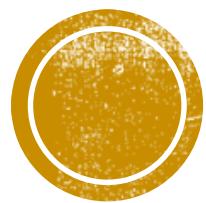
exemplar



VORONOI DIAGRAM

We can partition all the points in the space into regions, according to their closest representative.





TRAINING LOSS

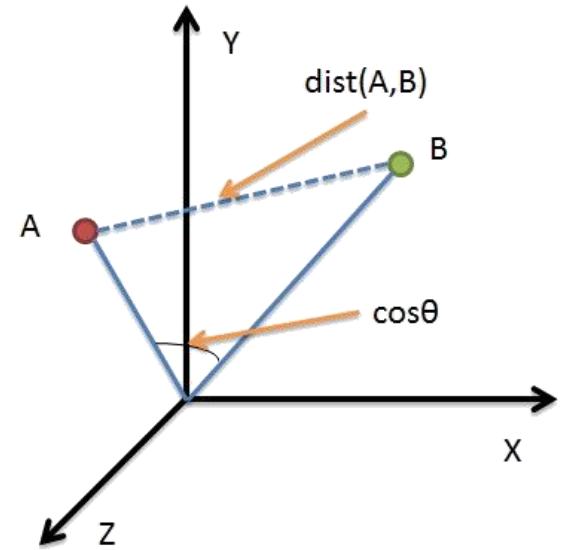


DISTANCE METRICS

(sometimes called *loss functions*)

A measure of how close two data points are.
Nearby points (i.e. distance is *small*) are
more likely they belong to the same cluster.

- Euclidean Distance $\text{dist}(x, y) = \|x - y\|^2$

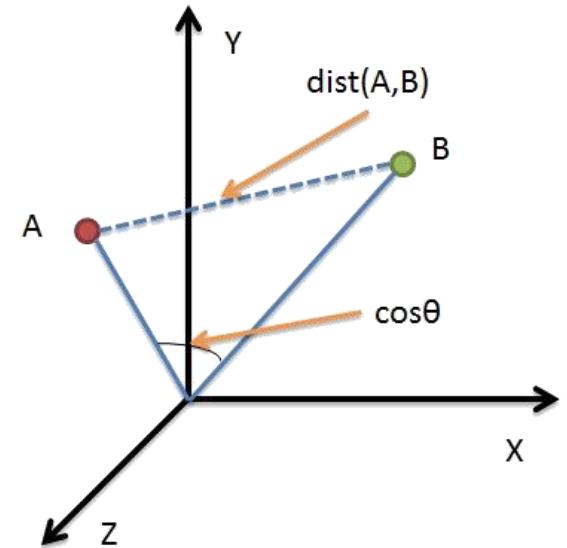


SIMILARITY FUNCTIONS

(sometimes called *kernels, correlation*)

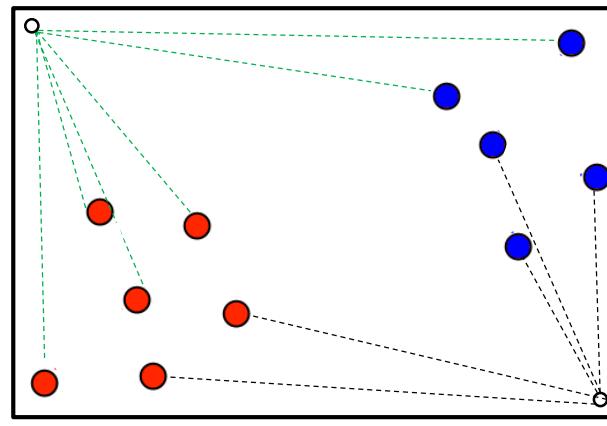
A measure of how alike two data points are.
Similar points (i.e. similarity is **large**) are
more likely they belong to the same cluster.

- Cosine Similarity $\cos(x, y) = \frac{x^T y}{\|x\| \|y\|}$



TRAINING LOSS

Sum of squared distances to closest representative.



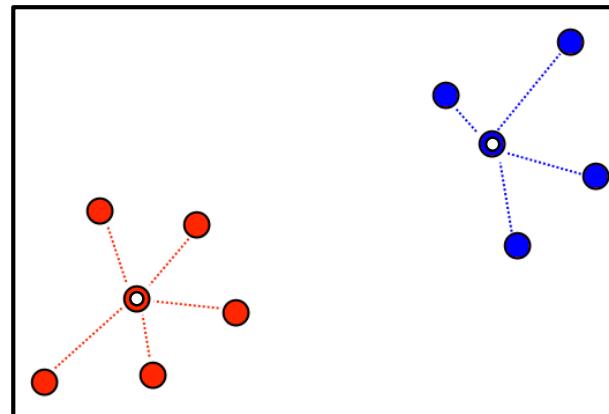
$$\text{loss} \approx 11 \times (1)^2 = 11$$

assume length of each
edge is about 1



TRAINING LOSS

Sum of squared distances to closest representative.



$$\text{loss} \approx 9 \times (0.1)^2 = 0.09$$

assume length of each
edge is about 0.1



TRAINING LOSS

Optimizing over representatives.

How do I use a
similarity function
instead?

$$\mathcal{L}_{n,k}(z^{(1)}, \dots, z^{(k)}; \mathcal{S}_n) = \sum_{i=1}^n \min_{1 \leq j \leq k} \|x^{(i)} - z^{(j)}\|^2.$$



TRAINING LOSS

Optimizing over clusters.

$$\mathcal{L}_{n,k}(\mathcal{C}_1, \dots, \mathcal{C}_n; \mathcal{S}_n) = \sum_{j=1}^n \sum_{i \in \mathcal{C}_j} \left\| x^{(i)} - \frac{1}{|\mathcal{C}_j|} \sum_{i' \in \mathcal{C}_j} x^{(i')} \right\|^2.$$



TRAINING LOSS

Optimizing both clusters and representatives.

Instead of the distance metric, you can use the *negative similarity* function.

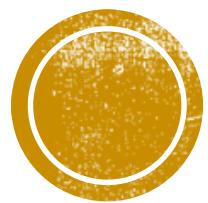
$$\mathcal{L}_{n,k}(\mathcal{C}_1, \dots, \mathcal{C}_k, z^{(1)}, \dots, z^{(k)}; \mathcal{S}_n) = \sum_{j=1}^k \sum_{i \in \mathcal{C}_j} \|x^{(i)} - z^{(j)}\|^2$$

These clusters need not consist of points closest to the representatives.

These representatives need not be the centroids of the clusters.



5 min break



K-MEANS



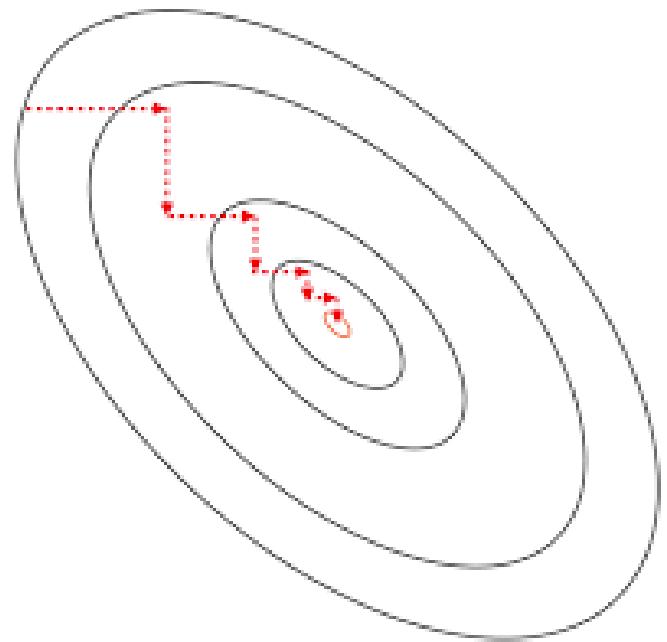
OPTIMIZATION ALGORITHM

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

Coordinate Descent (Gradient).

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 (Optimization).

Repeat until convergence:

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



OPTIMIZATION ALGORITHM

Coordinate Descent (Optimization)

Repeat until convergence:

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

$$\mathcal{L}_{n,k}(\mathcal{C}_1, \dots, \mathcal{C}_k, z^{(1)}, \dots, z^{(k)}; \mathcal{S}_n) = \sum_{j=1}^k \sum_{i \in \mathcal{C}_j} \|x^{(i)} - z^{(j)}\|^2$$



OPTIMIZATION ALGORITHM

1. Initialize centroids $z^{(1)}, \dots, z^{(k)}$ from the data.
2. Repeat until no further change in training loss:

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

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

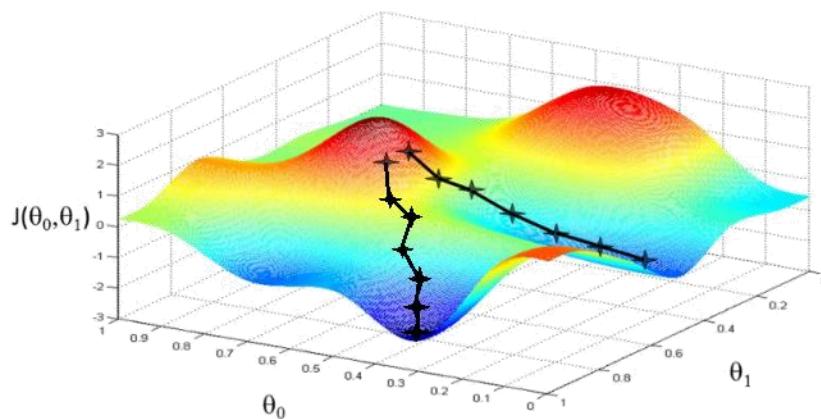
- b. For each $j \in \{1, \dots, k\}$,

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



CONVERGENCE

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



Challenge.

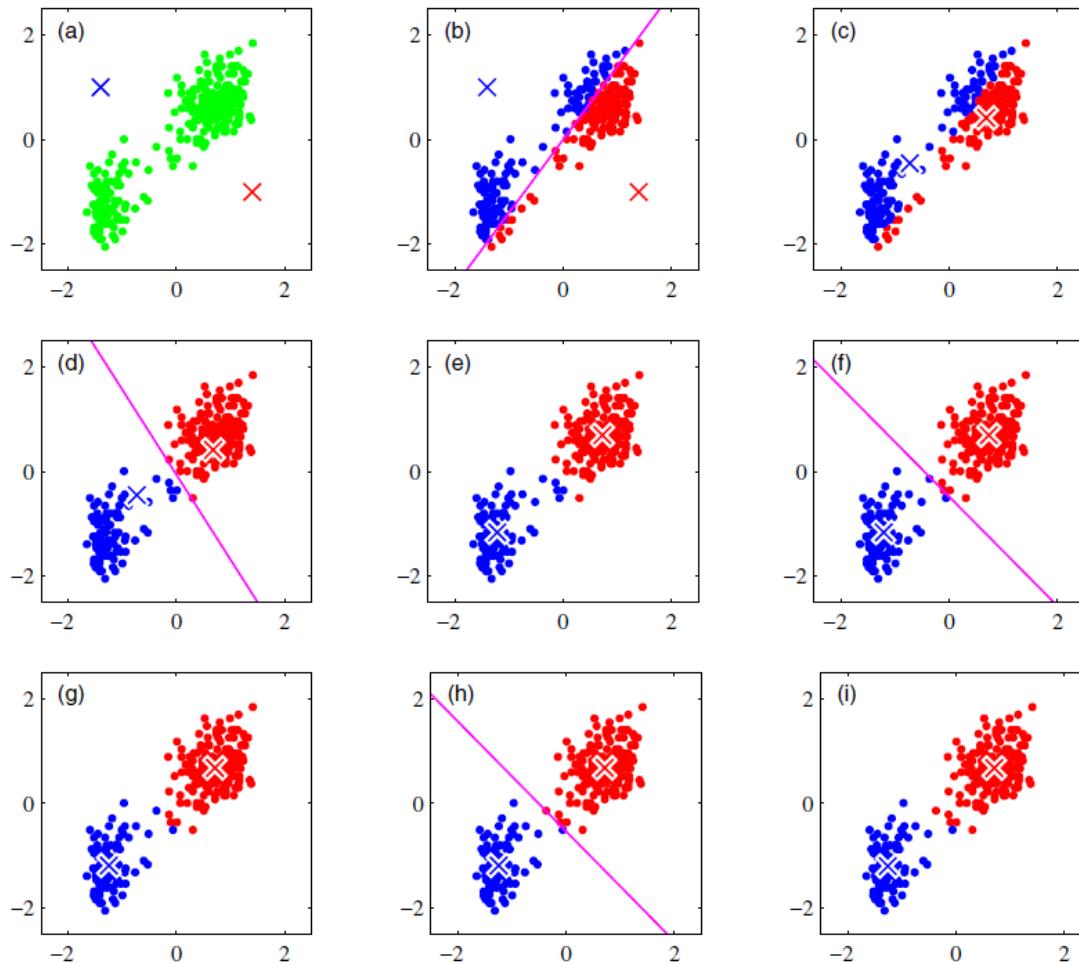
Why does the algorithm terminate in a finite number of steps?

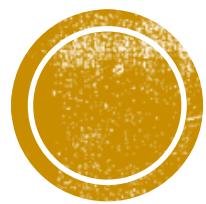
* not in syllabus

Repeat algorithm over many initial points, and pick the configuration with the smallest training loss.



Illustration





DISCUSSION



INITIALIZATION

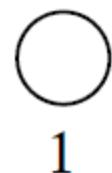
Optimization

- Empty clusters
 - Pick data points to initialize clusters
- Bad local minima
 - Initialize many times and pick solution with smallest training loss
 - Pick good starting positions



INITIALIZATION

Optimization



Starting position of centroids



Final position of centroids

Problem.

How to choose good starting positions?

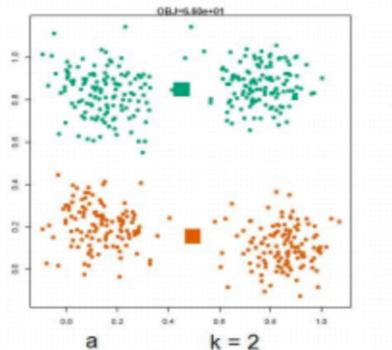
Solution.

Place them far apart with high probability.

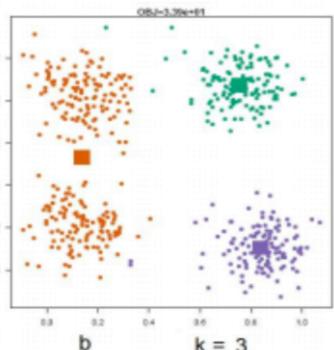


NUMBER OF CLUSTERS

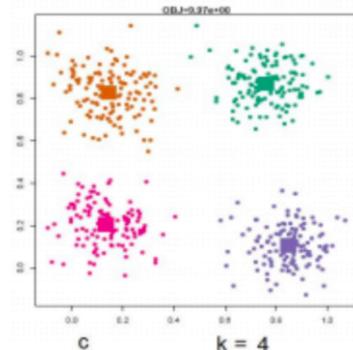
Generalization



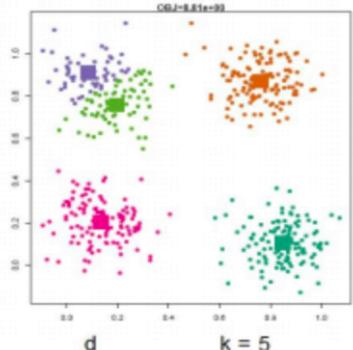
a $k = 2$



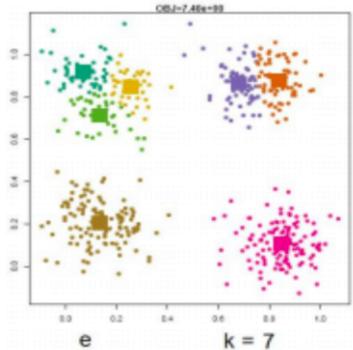
b $k = 3$



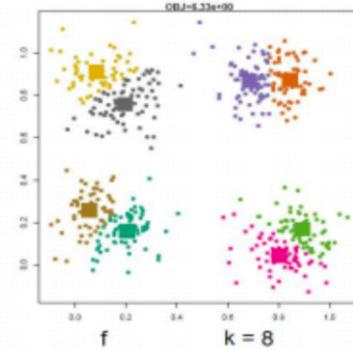
c $k = 4$



d $k = 5$



e $k = 7$



f $k = 8$



NUMBER OF CLUSTERS

Generalization

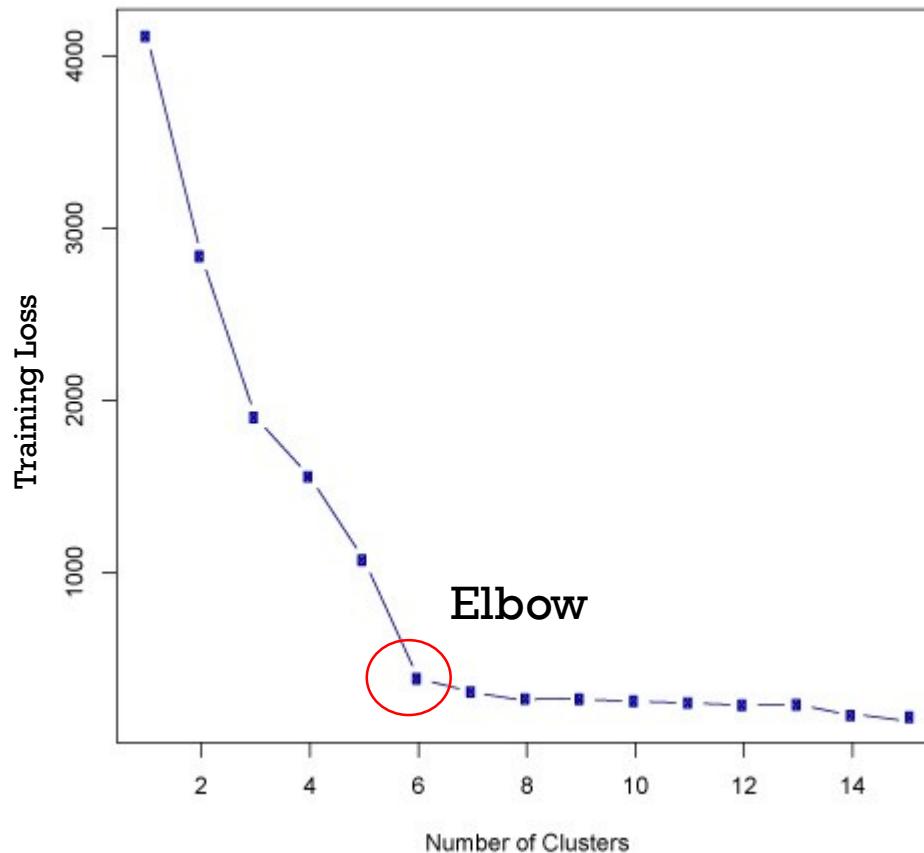
How do we choose k , the optimal number of clusters?

- Elbow method
 - Training Loss
 - Validation Loss
- Semi-supervised learning
 - Accuracy in supervised task



ELBOW METHOD

Generalization

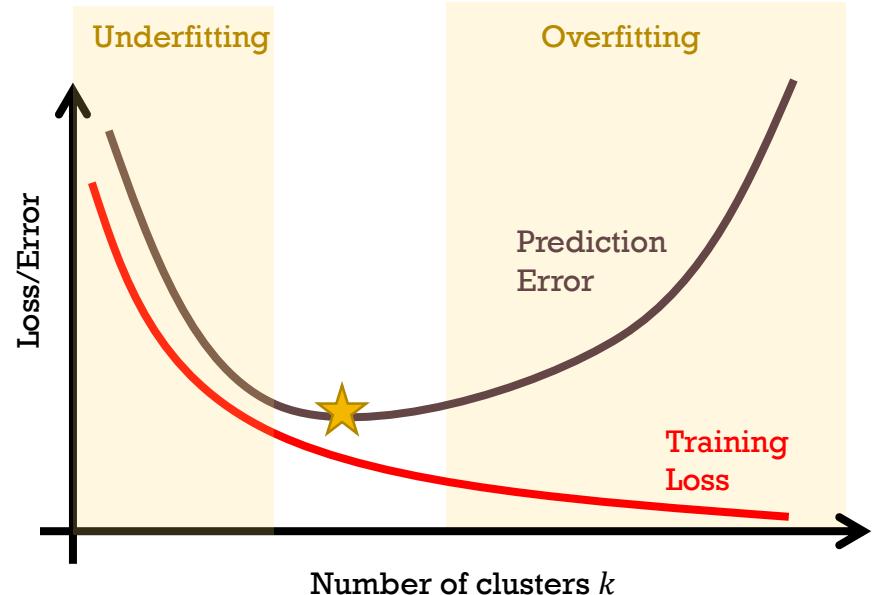


SEMI-SUPERVISED LEARNING

Generalization

Supervised task with small *labeled* data S'

- For each number of clusters k ,
 - Perform k -means on *unlabeled* data.
 - Transform S' using learned clusters e.g. compute distance to each centroid.
 - Use new features for supervised task, and compute prediction error.
- Pick k with smallest prediction error.



K-MEDOIDS

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](#)