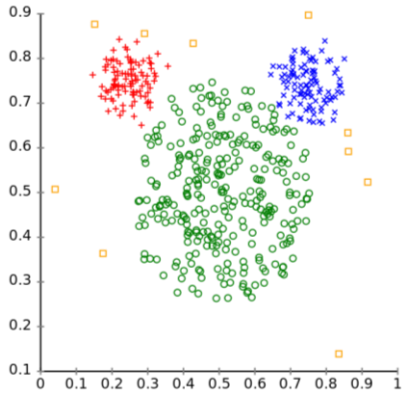
01.112/50.007 Machine Learning

Lecture 5

K-Means Clustering

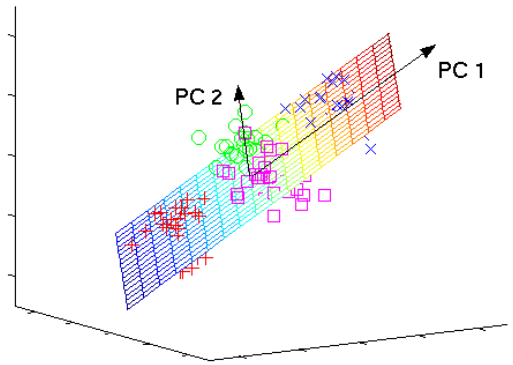
Unsupervised Learning

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



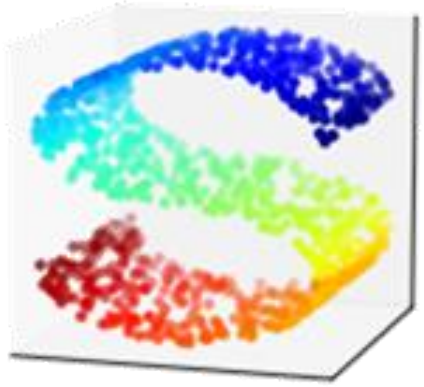
Clustering

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



Subspace Learning

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

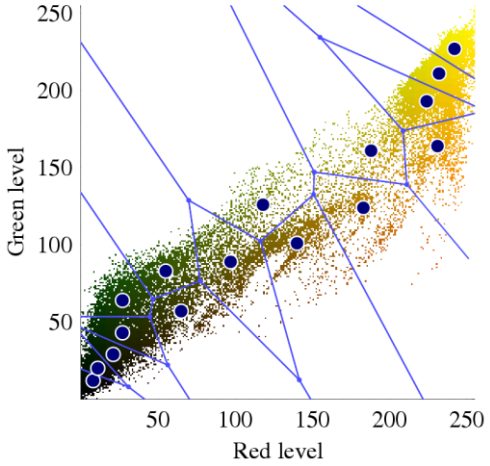


Manifold Learning

Uses of Unsupervised Learning

- Data compression

RGB
(12,165,73)
(11,167,79)
:



Labels
3
43
:

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

Uses of Unsupervised Learning

Improve classification/regression (semi-supervised learning)

1. From *unlabeled data,* learn good features $T: \mathbb{R}^d \to \mathbb{R}^m$.
2. To *labeled data*, apply transformation $T: \mathbb{R}^d \to \mathbb{R}^m$.

$$\left(T\left(x^{(1)}\right), y^{(1)}\right), \ldots, \left(T\left(x^{(n)}\right), y^{(n)}\right)$$

3. Perform classification/regression on transformed data.

RoadMap

Clustering

- K-Means Algorithm

Dim Reduction with Complete Features

Unsupervised

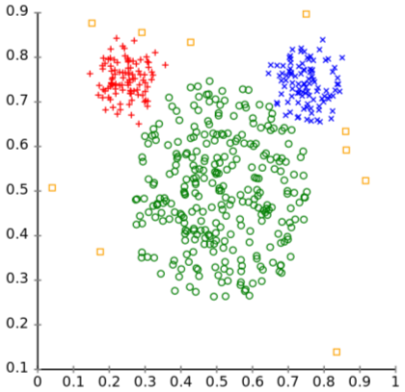
Recommender Systems
Collaborative Filtering
Missing Data Prediction

- K-Nearest Neighbors
- Matrix Factorization

Dim Reduction with Incomplete Features

Supervised or Unsupervised?

What Is Clustering



Clustering Problem.

Input.

Training data $S_n = \{x^{(i)}; i = 1, 2, \ldots, n\}$, 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.

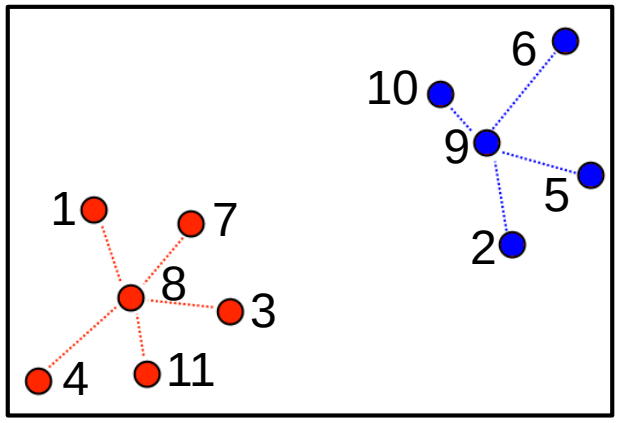
Some clusters could be empty!

How to Specify a Cluster

- By listing all its elements

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

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

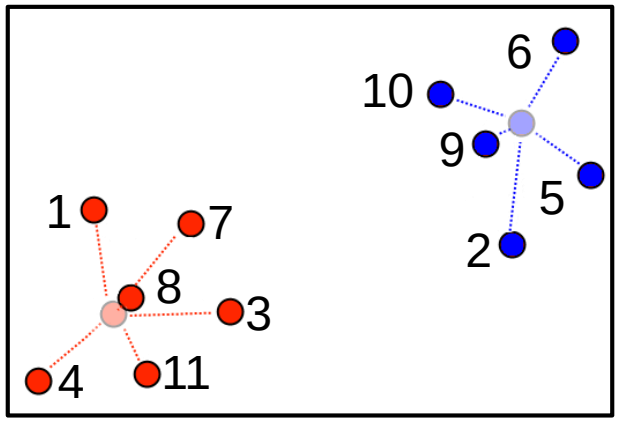


How to Specify a Cluster

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

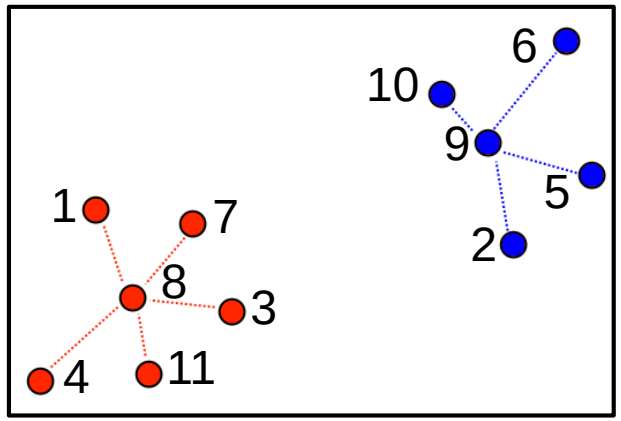
Each point $x^{(i)}$ will be assigned the closest representative.

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



centroid

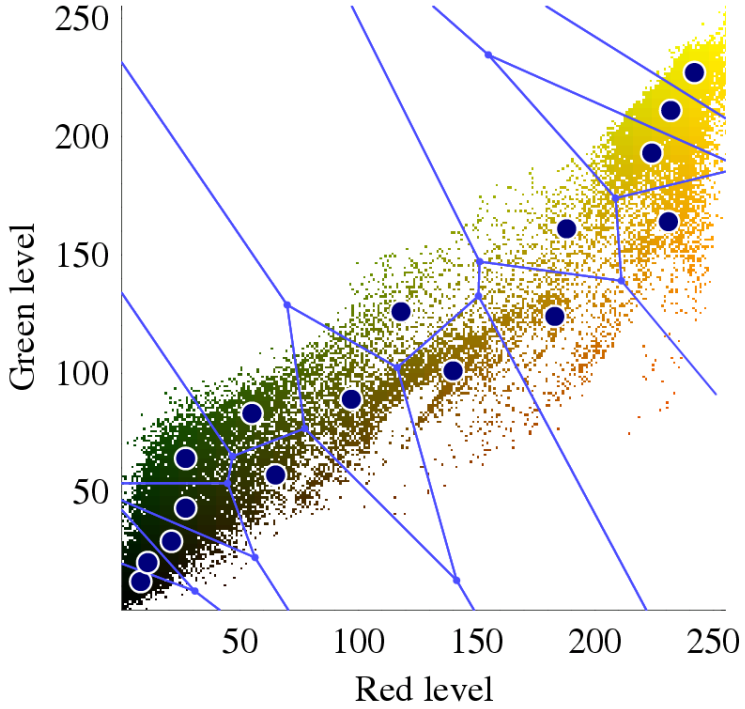
$z^{(1)} = 8, z^{(2)} = 9$



exemplar

Voronoi Diagram

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

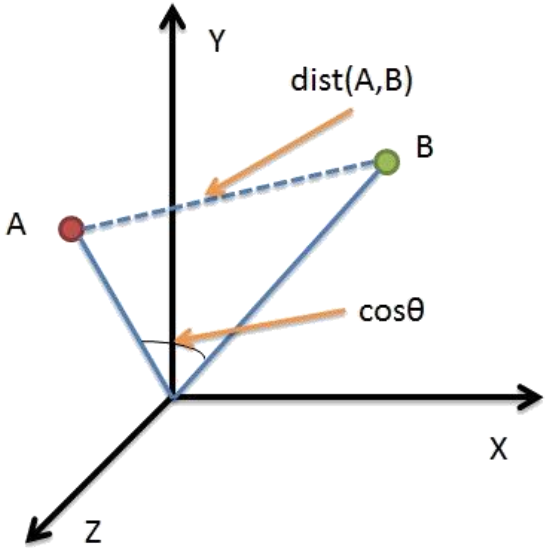


Training Loss

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 that they belong to the same cluster.

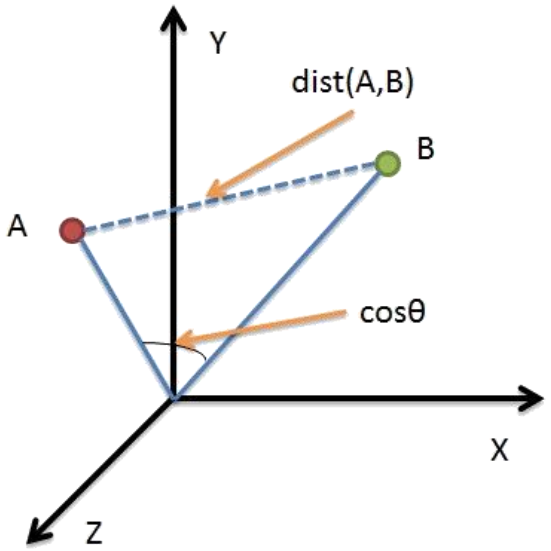


- Cosine Similarity $\cos(x, y) = \frac{x^{\top} y}{\|x\| \, \|y\|}$

Distance Metrics

(sometimes called *loss functions)*

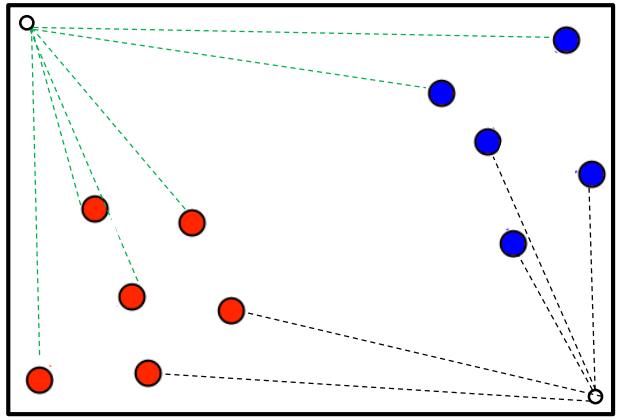
A measure of how close two data points are. Nearby points (i.e. distance is *small)* are more likely as they belong to the same cluster.



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

Training Loss

Sum of squared distances to closest representative.

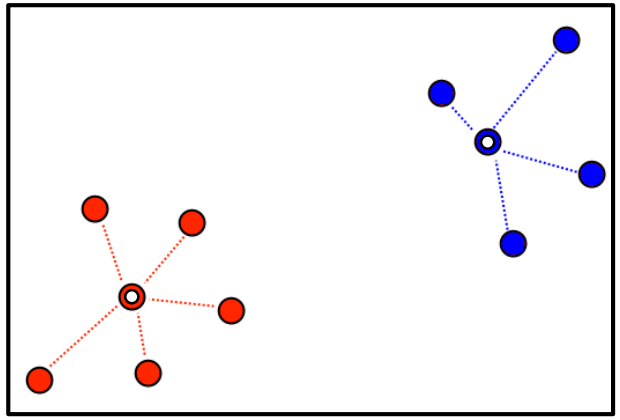


$$\text{loss} \approx 11 \times (1)^2 = 11$$

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 we use a similarity function instead?

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

Training Loss

Optimizing over **clusters**.

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

Training loss

Optimizing both **clusters** and **representatives**.

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

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

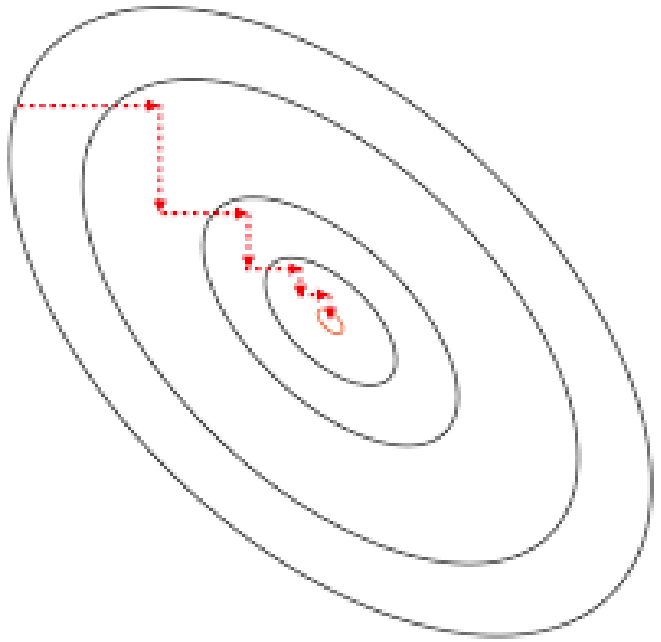
These representatives need not be the centroids of the clusters.

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

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

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

K-Means Algorithm

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

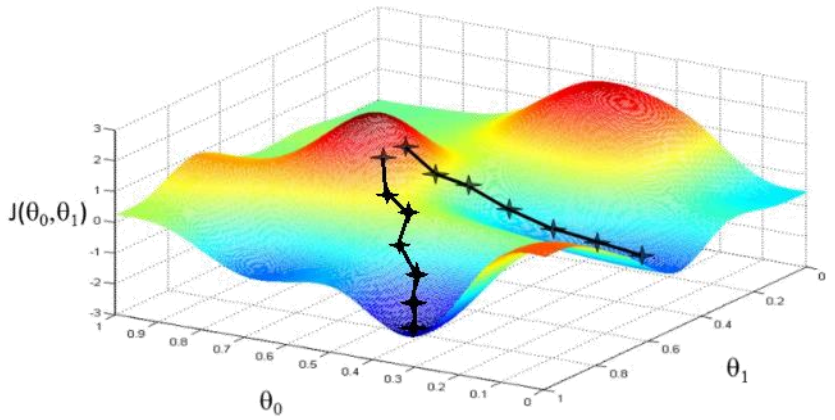
   a. For each $j \in \{1, \ldots, k\}$,
   $$\mathcal{C}_j = \{\, i \text{ such that } x^{(i)} \text{ is closest to } z^{(j)} \,\}.$$

   b. For each $j \in \{1, \ldots, k\}$,
   $$z^{(j)} = \frac{1}{|\mathcal{C}_j|} \sum_{i \in \mathcal{C}_j} x^{(i)} \quad \text{(cluster mean)}$$

Animation: https://towardsdatascience.com/k-means-clustering-introduction-to-machine-learning-algorithms-c96bf0d5d57a

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?

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

Convergence

The cost is lowered with each iteration

$$\text{cost}(C_1, \ldots, C_k, z^{(1)}, \ldots, z^{(k)})$$

Find new clusters for fixed centroids

$$\begin{aligned}\text{cost}(C_1, \ldots, C_k, z^{(1)}, \ldots, z^{(k)}) &\overset{(a)}{\geq} \min_{C_1,\ldots,C_k} \text{cost}(C_1, \ldots, C_k, z^{(1)}, \ldots, z^{(k)}) \\ &= \text{cost}(C'_1, \ldots, C'_k, z^{(1)}, \ldots, z^{(k)})\end{aligned}$$

Convergence

The cost is lowered with each iteration

$$\text{cost}(C_1, \ldots, C_k, z^{(1)}, \ldots$$

Find new clusters for fixed centroids

$$\text{cost}(C_1, \ldots, C_k, z^{(1)}, \ldots, z^{(k)}) \overset{(a)}{\geq} \min_{C_1,\ldots,C_k} \text{cost}(C_1, \ldots, C_k, z^{(1)}, \ldots, z^{(k)})$$
$$= \text{cost}(C'_1, \ldots, C'_k, z^{(1)}, \ldots, z^{(k)})$$

Either cluster points are re-assigned or remain in the same cluster as no other cluster can achieve lower cost for given centroids.

Convergence

The cost is lowered with each iteration

$$\text{cost}(C_1, \ldots, C_k, z^{(1)}, \ldots$$

Find new clusters for fixed centroids

$$\text{cost}(C_1, \ldots, C_k, z^{(1)}, \ldots, z^{(k)}) \overset{(a)}{\geq} \min_{C_1,\ldots,C_k} \text{cost}(C_1, \ldots, C_k, z^{(1)}, \ldots, z^{(k)})$$
$$= \text{cost}(C'_1, \ldots, C'_k, z^{(1)}, \ldots, z^{(k)})$$

Either cluster points are re-assigned or remain in the same cluster as no other cluster can achieve lower cost for given centroids.

Find new centroid for new clusters

$$\text{cost}(C'_1, \ldots, C'_k, z^{(1)}, \ldots, z^{(k)}) \overset{(b)}{\geq} \min_{z^{(1)},\ldots,z^{(k)}} \text{cost}(C'_1, \ldots, C'_k, z^{(1)}, \ldots, z^{(k)})$$
$$= \text{cost}(C'_1, \ldots, C'_k, z'^{(1)}, \ldots, z'^{(k)})$$

Convergence

The cost is lowered with each iteration

$$\text{cost}(C_1, \ldots, C_k, z^{(1)}, \ldots$$

Find new clusters for fixed centroids

$$\text{cost}(C_1, \ldots, C_k, z^{(1)}, \ldots, z^{(k)}) \overset{\text{(a)}}{\geq} \min_{C_1,\ldots,C_k} \text{cost}(C_1, \ldots, C_k$$
$$= \text{cost}(C'_1, \ldots, C'_k, z^{(1)}, \ldots$$

Find new centroid for new clusters

$$\text{cost}(C'_1, \ldots, C'_k, z^{(1)}, \ldots, z^{(k)}) \overset{\text{(b)}}{\geq} \min_{z^{(1)},\ldots,z^{(k)}} \text{cost}(C'_1, \ldots, C'_k, z^{(1)}, \ldots, z^{(k)})$$
$$= \text{cost}(C'_1, \ldots, C'_k, z'^{(1)}, \ldots, z'^{(k)})$$

Either cluster points are re-assigned or remain in the same cluster as no other cluster can achieve lower cost for given centroids.

Either the centroids remain same or get updated to minimize distance to their cluster points.

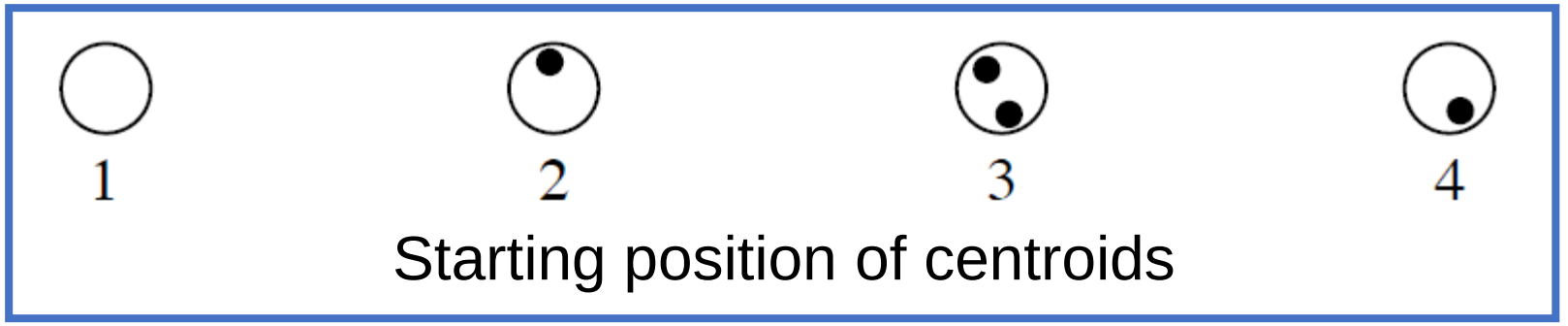
Discussion

Initialization

- 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

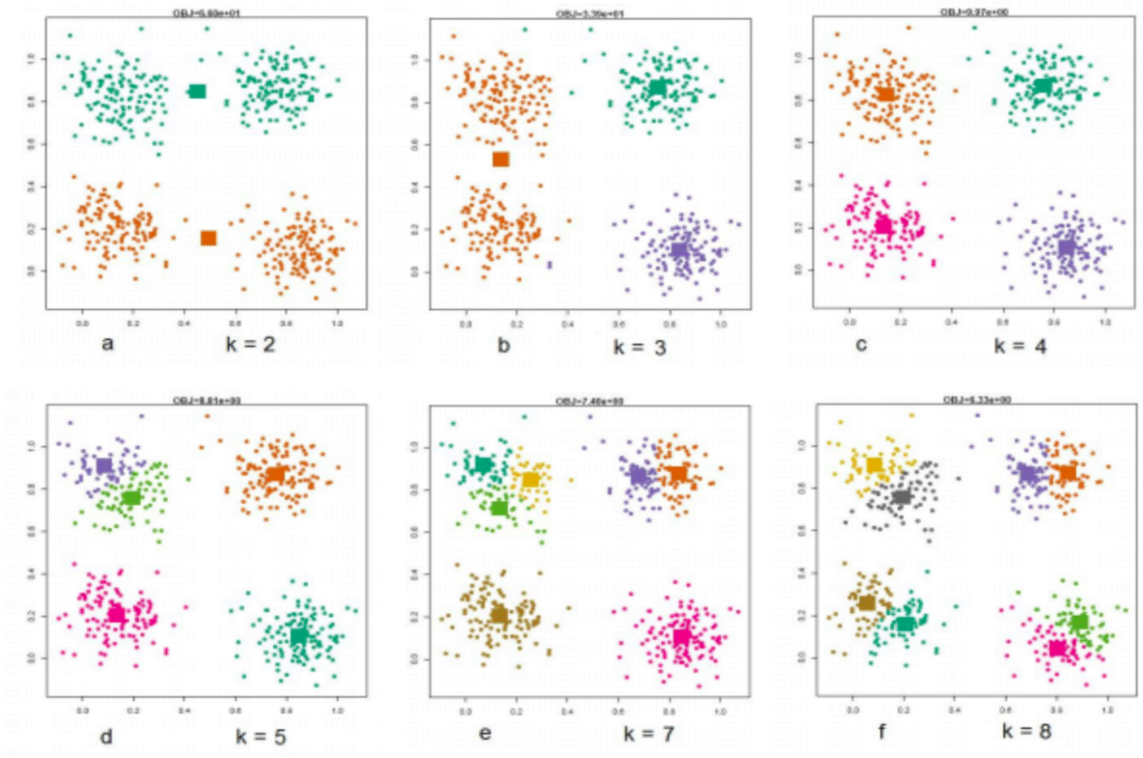
Optimization



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

Number of Clusters

Generalization



Number of Clusters

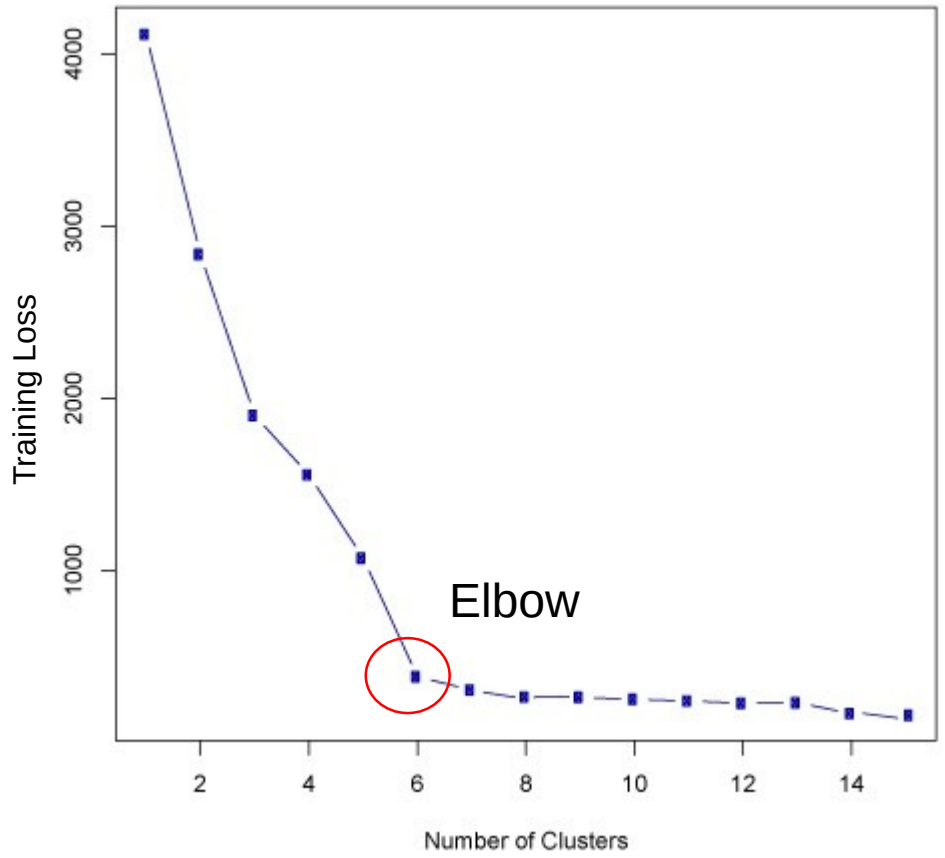
Generalization

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

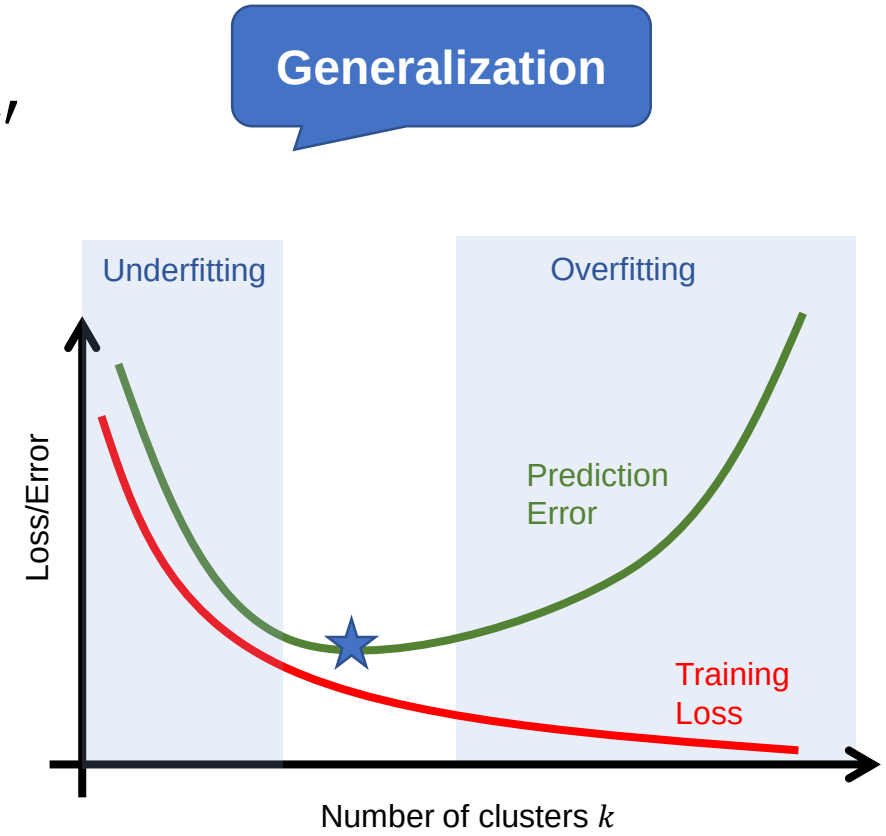
Generalization



Semi-Supervised Learning

Supervised task with small *labeled* data $\mathcal{S}'$

- For each number of clusters k,
  - Perform k–means on *unlabeled* data.
  - Transform $\mathcal{S}'$ 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.



Summary

- **Clustering**
  - Distance Metric
  - Similarity Function
  - Training Loss
- **Representatives**
  - Centroids
  - Exemplars
  - Voronoi Diagrams
- **k-Means Algorithm**
  - **Optimization**
    - Coordinate Descent
    - Initialization
  - **Generalization**
    - Number of Clusters
  - **Applications**
    - Dimensionality Reduction
    - Data Compression
    - Semi-Supervised Learning

Intended Learning Outcomes

Clustering

- Describe the differences between distance metrics and similarity functions. List examples of each of them.
- Write down the training loss using the Euclidean distance.
- Describe two ways of picking representatives for clusters. Explain how Voronoi diagrams are derived from the representatives.
- List two important applications of clustering, and how they are related to dimensionality reduction.

Intended Learning Outcomes

K-Means Algorithm

- Describe the k-means algorithm, and point out how it is based on coordinate descent.
- Explain why it is important to run the k-means algorithm several times at various starting points.
- Describe a procedure for estimating k, the number of clusters.