Principles of Computer Science II

Working with Data Sets

Ioannis Chatzigiannakis

Sapienza University of Rome

Lecture 15

K-means Algorithm

- Developed and published in Applied Statistics by Hartigan and Wong, 1979.
- Many variations have been proposed since then.
- Standard/core function of R, Python, Matlab, ...
- Assumes Euclidean space/distance

The aim of the K-means algorithm is to divide M points in N dimensions into k clusters so that the within-cluster sum of squares is minimized.

$$\min_{C_1,\ldots,C_K} \sum_{k=1}^{k} \frac{1}{|C_k|} \sum_{i,i' \in C_k} \sum_{j=1}^{p} (x_{ij} - x_{i'j})^2$$

Cluster Initialization

- Start by picking k, the number of clusters
- Initialize clusters by picking one point per cluster

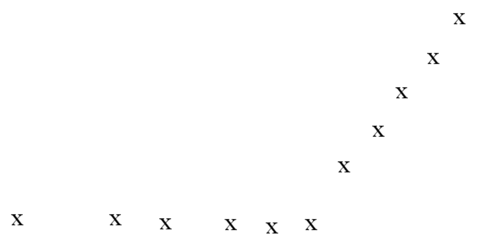
Example: Pick one point at random, then $k-1$ other points, each as far away as possible from the previous points

Populating Clusters

1. For each point, place it in the cluster whose current centroid it is nearest
2. After all points are assigned, update the locations of centroids of the k clusters
3. Reassign all points to their closest centroid
   - Sometimes moves points between clusters
4. Repeat 2 and 3 until convergence

Convergence: Points do not move between clusters and centroids stabilize

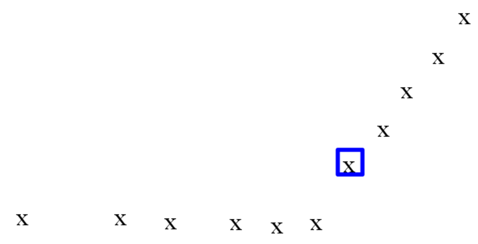
A Simple Example



x … data point
□ … centroid

Clusters after round 1

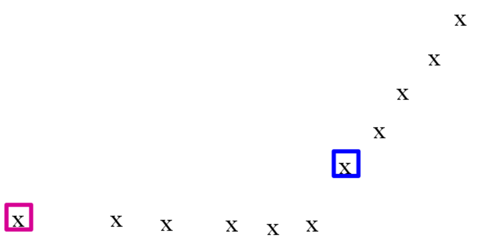
A Simple Example



x … data point
□ … centroid

Clusters after round 1

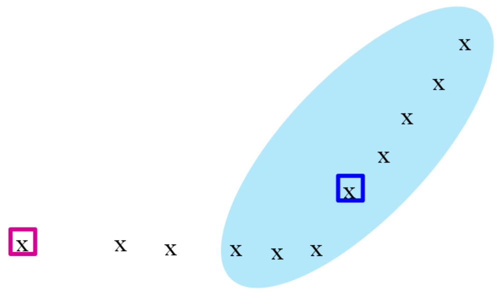
A Simple Example



x … data point
□ … centroid

Clusters after round 1

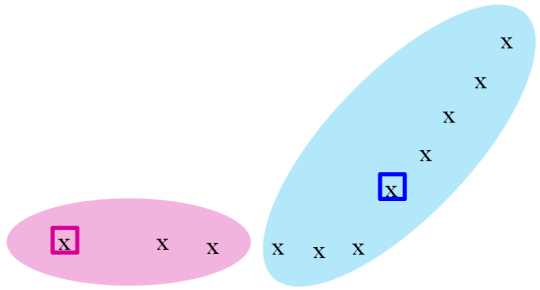
A Simple Example



x … data point
□ … centroid

Clusters after round 1

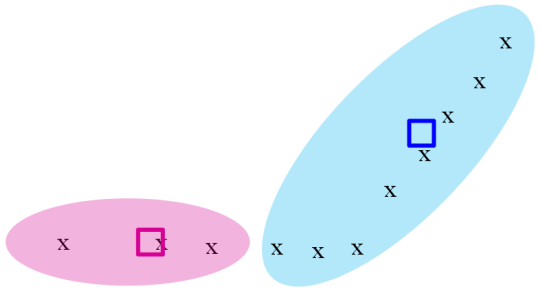
A Simple Example



x … data point
□ … centroid

Clusters after round 1

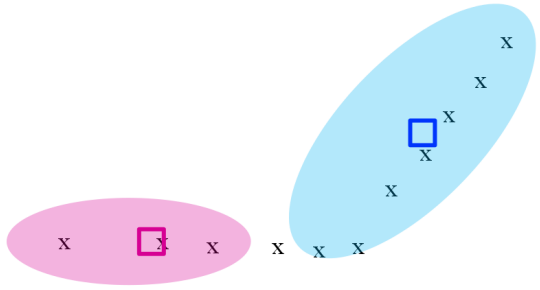
A Simple Example



x … data point
□ … centroid

Clusters after round 2

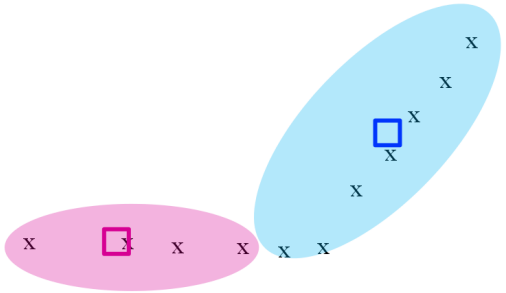
A Simple Example



x … data point
□ … centroid

Clusters after round 2

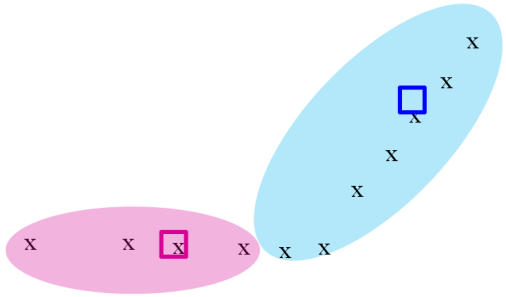
A Simple Example



x … data point
□ … centroid

Clusters after round 2

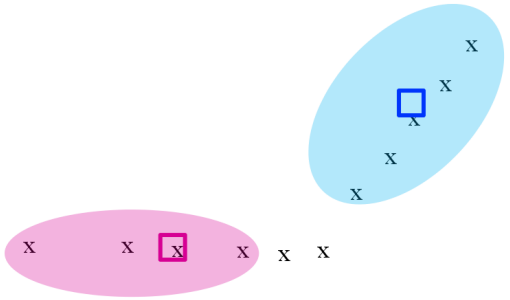
A Simple Example



x … data point
□ … centroid

Clusters at the end

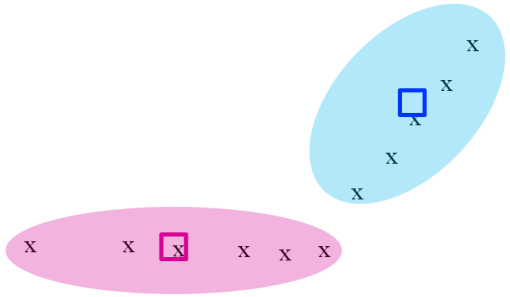
A Simple Example



x … data point
□ … centroid

Clusters at the end

A Simple Example

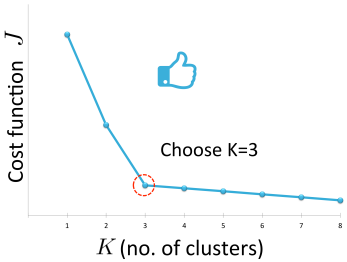


x … data point
□ … centroid

Clusters at the end

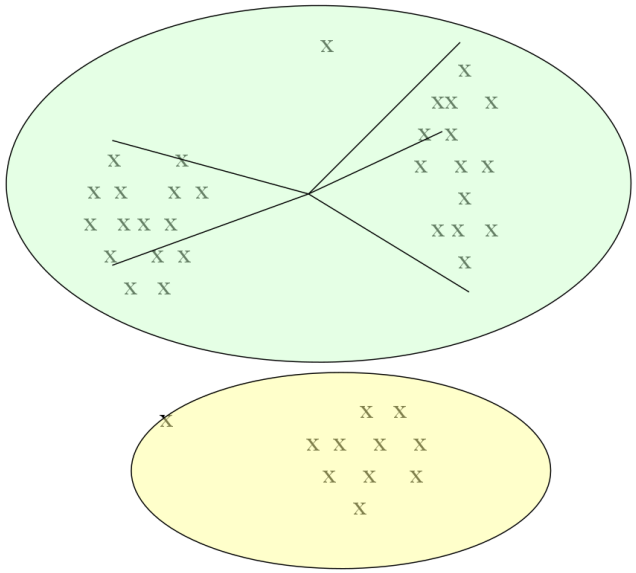
How to select k?

- We use the elbow method to determine the optimum number of clusters.
- Try different k, looking at the change in the average distance to centroid as k increases.
- Average falls rapidly until right k, then changes little.



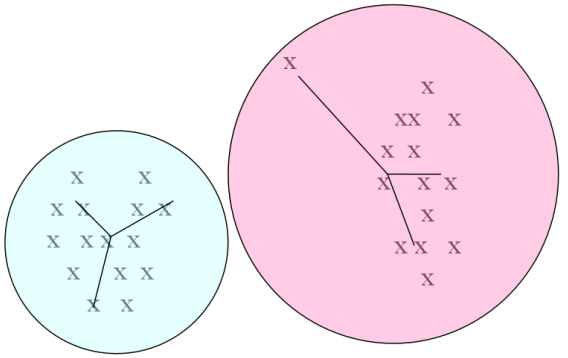
Selection of k – an example

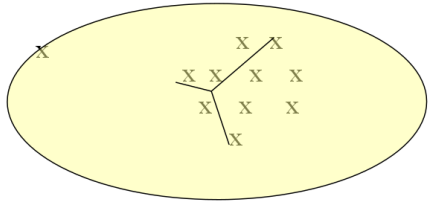
Too few;
many long
distances
to centroid.



Selection of k – an example

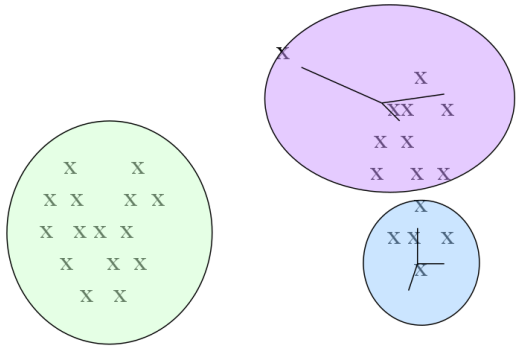
Just right;
distances
rather short.





Selection of k – an example

Too many;
little improvement
in average
distance.



One-dimensional clustering

```
1 from sklearn.cluster import KMeans
2 kmeans = KMeans(n_clusters=3, init='random')
3
4 kmeans.fit([row[0:1] for row in dataset])
5
6 c = kmeans.predict([row[0:1] for row in dataset])
```

Evaluating the outcome