Lecture 09. Clustering analysis and K-means

Xin Chen

These slides are based on slides from Mahdi Roozbahani

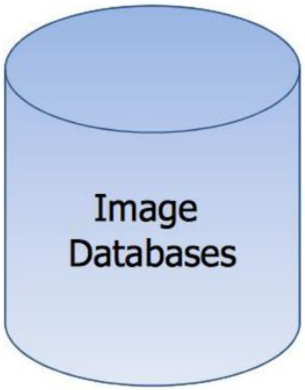
Logistics

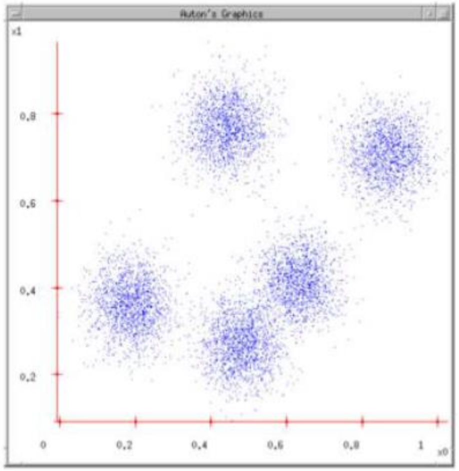
- Project proposal
  - Background & motivation
    - Why do people care?
    - More importantly, what is the existing approaches? How do you understand them?
  - Objectives:
    - Something based on the background information, what is missing? What is more important? What is your new angle?
- Writing a report/proposal
  - For any figures, plots and sentences that is not “yours”, you need to clearly cite where and who it come from.
  - Do not look like this. (If you submit it to somewhere like a conference, this would be a serious issue)
    - There is a dataset: A.
    - There is a paper/link B, which is about the topic.
    - The figure says C.
  - Everything on the report is your understanding. (How is this material you cite related to yours?)
- In general, remember to start from “small” and “solid”.

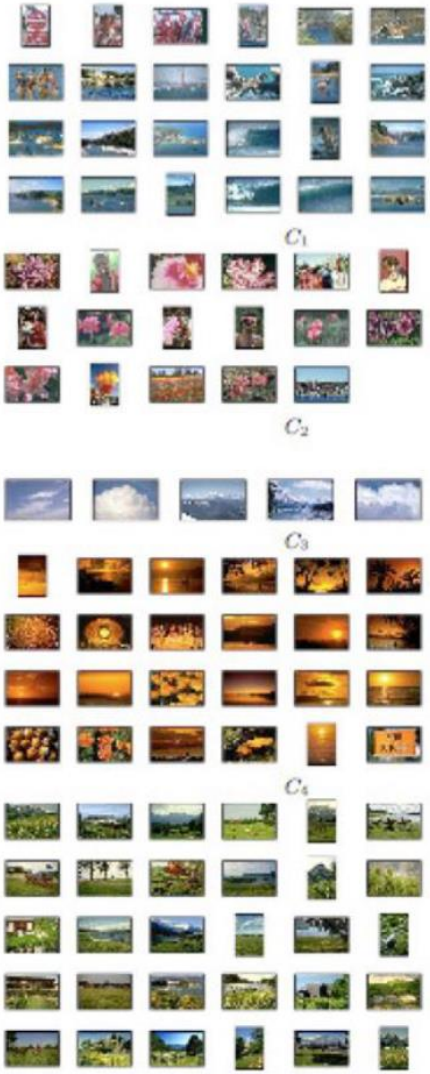
Outline

- Clustering
- Distance function
- K-means algorithm
- Analysis of K-means

Clustering images







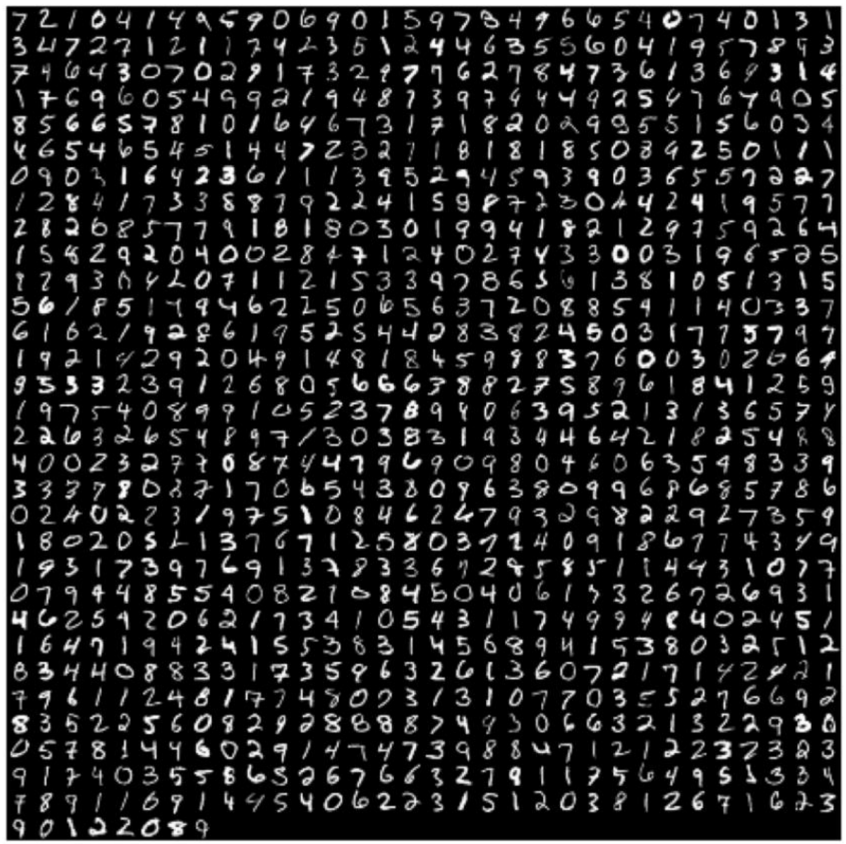
Goal of clustering:
Divide objects into groups and objects within a group are more similar than those outside the group.

Clustering other objects

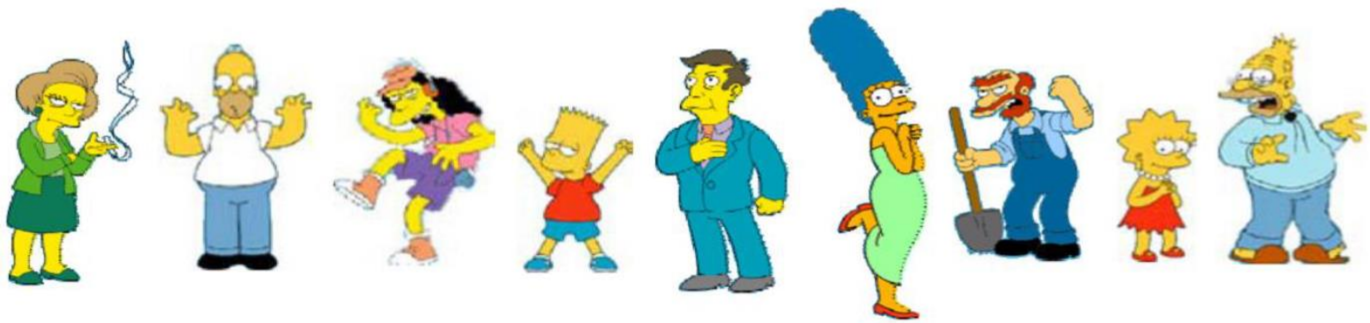
Piotr Belarusian
Pyotr Azerbaijani
Petros Greek
Pietro Italian
Pedro Portuguese
Pierre French
Piero Italian
Peter Dutch
Peder Danish
Peka Couldn't find it – Finish?
Peadar Irish

Linguistic Similarity

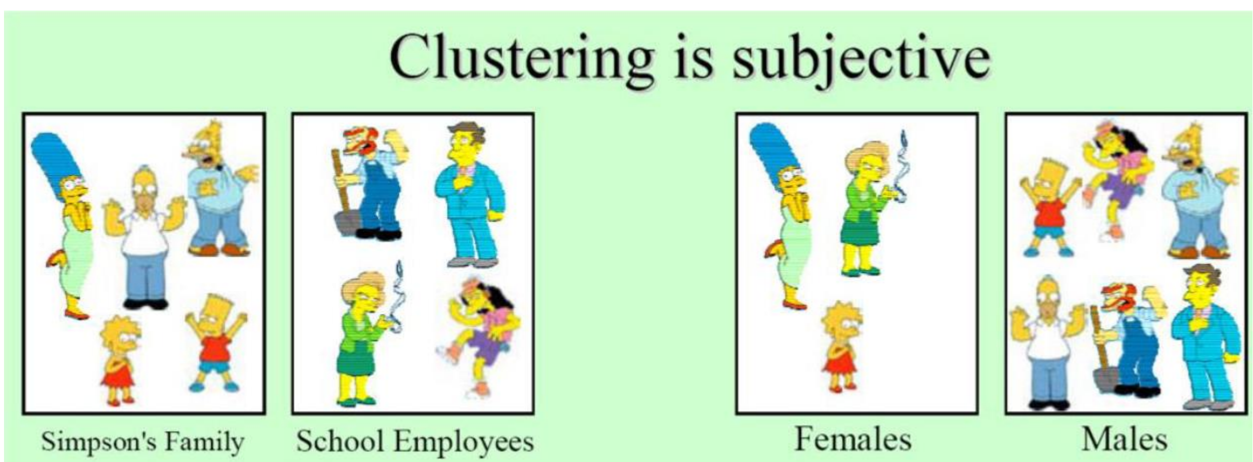
Clustering hand digits



Clustering is subjective



What is considered similar/dissimilar?



What is clustering in general?

- First we need to pick similarity/dissimilarity function?

- The algorithm figures out the grouping of objects based on the chosen dissimilarity/dissimilarity function:
  - Points within a cluster is similar
  - Points across cluster are not similar

- Issues for clustering:
  - How to represent objects? (vector space? Normalization)
  - What is similarity/dissimilarity function?
  - What are the algorithm steps?

Outline

- Clustering
- Distance function
- K-means algorithm
- Analysis of K-means

Properties of similarity function

- Desired properties of dissimilarity function
  - Symmetry: $d(x, y) = d(y, x)$
    - Otherwise you can claim "Alex looks like Bob, but Bob looks nothing like Alex.
  - Positive separability: $d(x, y) = 0, \text{if and only if } \mathrm{x} = \mathrm{y}$
    - Otherwise there are objects that are different, but you cannot tell apart
  - Triangular inequality: $d(x, y) \leq d(x, z) + d(z, y)$
    - Otherwise you can claim "Alex is very like Bob, and Alex is very like Carl, but Bob is very unlike Carl.

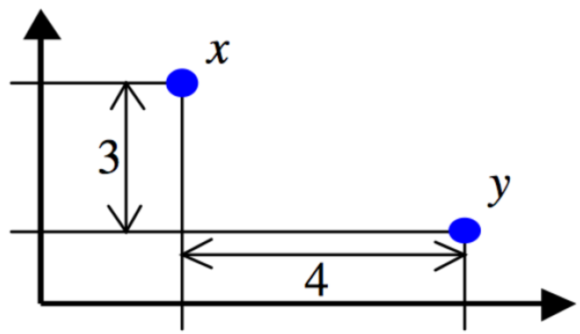
Distance functions for vectors

- Suppose two data points, both in R^d
  - $x = (x_1, x_2, \dots, x_d)^T$
  - $y = (y_1, y_2, \dots, y_d)^T$

- Euclidian distance: $d(x, y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$

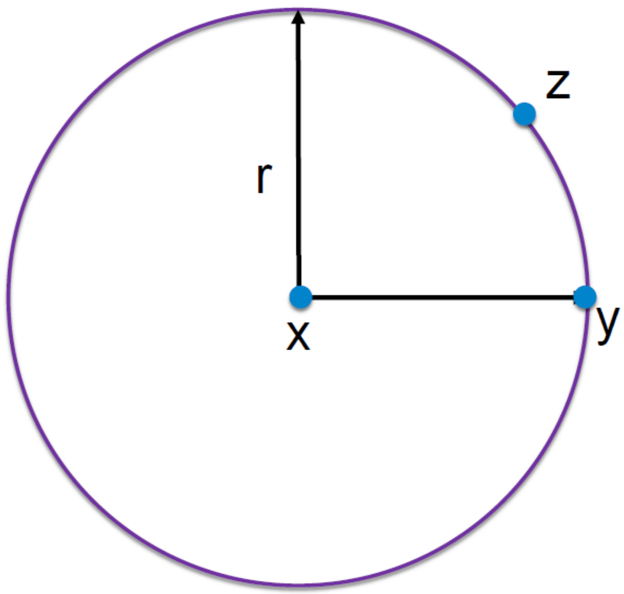
- Minkowski distance: $d(x, y) = \sqrt[p]{\sum_{i=1}^{n} (x_i - y_i)^p}$
  - Manhattan distance: $\mathrm{p} = 1, d(x, y) = \sum_{i=1}^{d} |x_i - y_i|$
  - "inf"-distance: $\mathrm{p} = \infty, d(x, y) = \max(|x_i - y_i|)$

Example



- Euclidian distance: $\sqrt{4^2 + 3^2} = 5$
- Manhattan distance: $4 + 3 = 7$
- "inf"-distance: $max\{4,3\} = 4$

Some problems with Euclidean distance



Hamming Distance

- Manhattan distance is also called Hamming distance when all features are binary
  - Count the number of difference between two binary vectors
  - Example,$x, y \in \{0, 1\}^{17}$

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
x	0	1	1	0	0	1	0	0	1	0	0	1	1	1	0	0	1
y	0	1	1	1	0	0	0	0	1	1	1	1	1	1	0	1	1

$$d(x, y) = 5$$

Edit distance

- Transform one of the objects into the other, and measure how much effort it takes

x	I	N	T	E	*	N	T	I	O	N
	|	|	|	|	|	|	|	|	|	|
y	*	E	X	E	C	U	T	I	O	N
	d	s	s		i	s				

d: deletion (cost 5)
s: substitution (cost 1)
i: insertion (cost 2)

$$d(x, y) = 5 * 1 + 1 * 3 + 2 * 1 = 10$$

Outline

- Clustering
- Distance function
- K-means algorithm
- Analysis of K-means

Results of K-means clustering

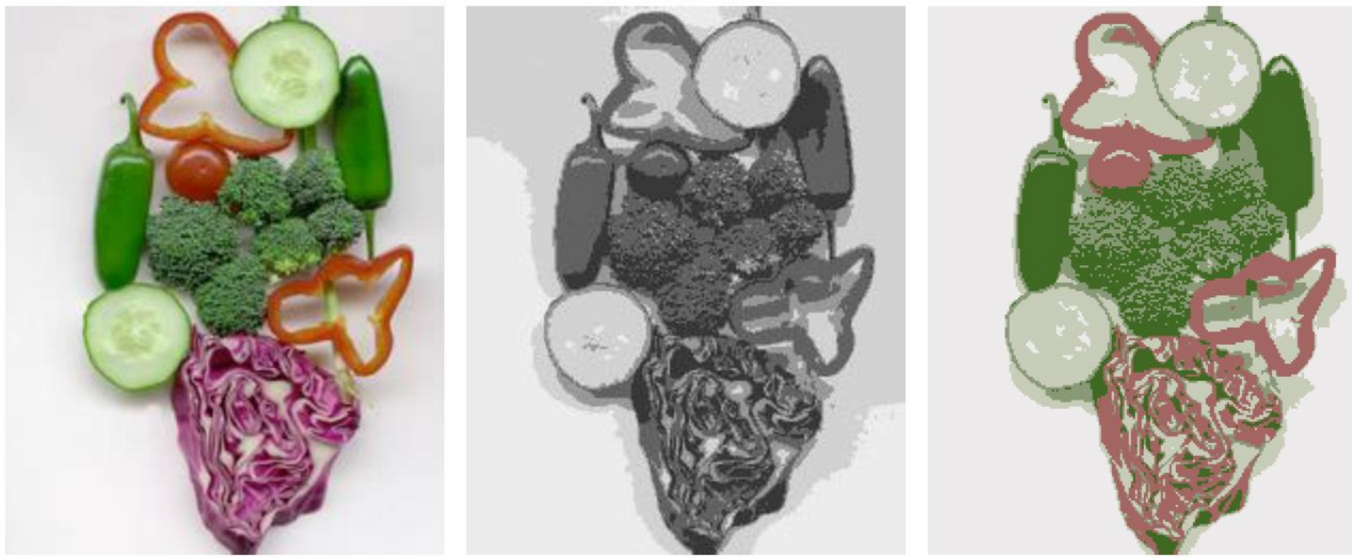
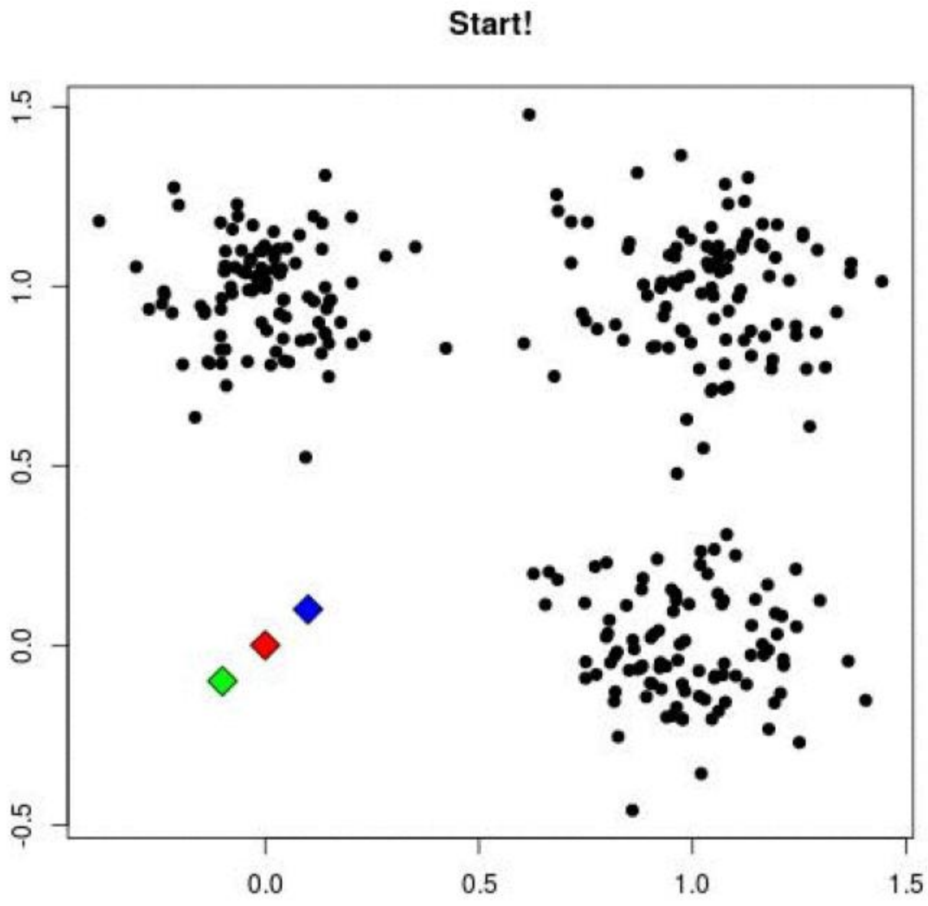


Image Clusters on intensity Clusters on color

Clustering using intensity only and color only

* Pictures from Mean Shift: A Robust Approach toward Feature Space Analysis, by D. Comaniciu and P. Meer http://www.caip.rutgers.edu/~comanici/MSPAMI/msPamiResults.html

K-Means algorithm



[Visualizing K-Means Clustering]

K-means algorithm

- Initialize k cluster centers, $\{c^1, c^2, \ldots, c^k\}$, randomly
- Do
  - Decide the cluster memberships of each data point, x^i, by assigning it to the nearest cluster center (cluster assignment)

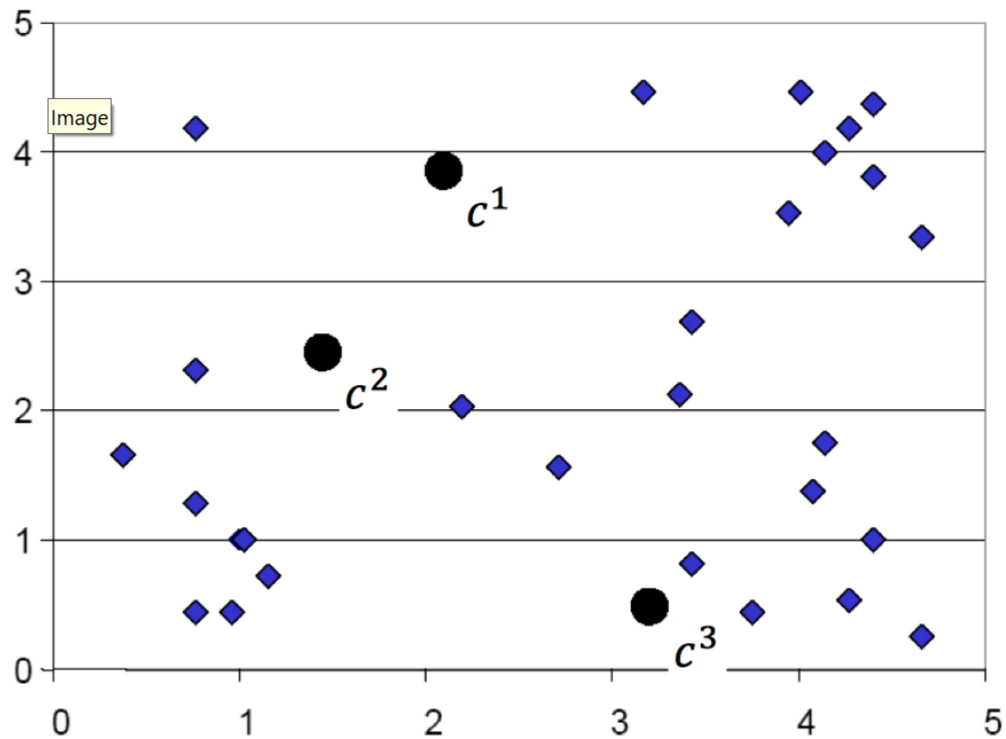
$$\pi(i) = argmin_{j=1,\ldots,k} \left\|x^i - c^j\right\|^2$$

  - Adjust the cluster centers (center adjustment)

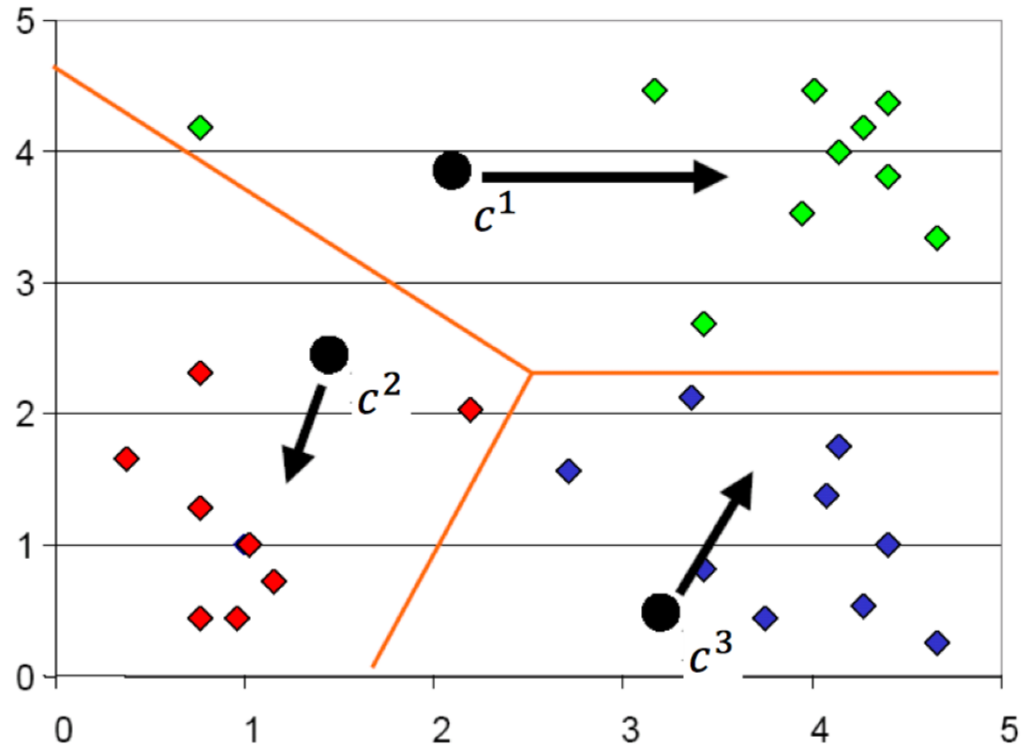
$$c^j = \frac{1}{|\{i: \pi(i) = j\}|} \sum_{i:\pi(i)=j} x^i$$

- While any cluster center has been changed

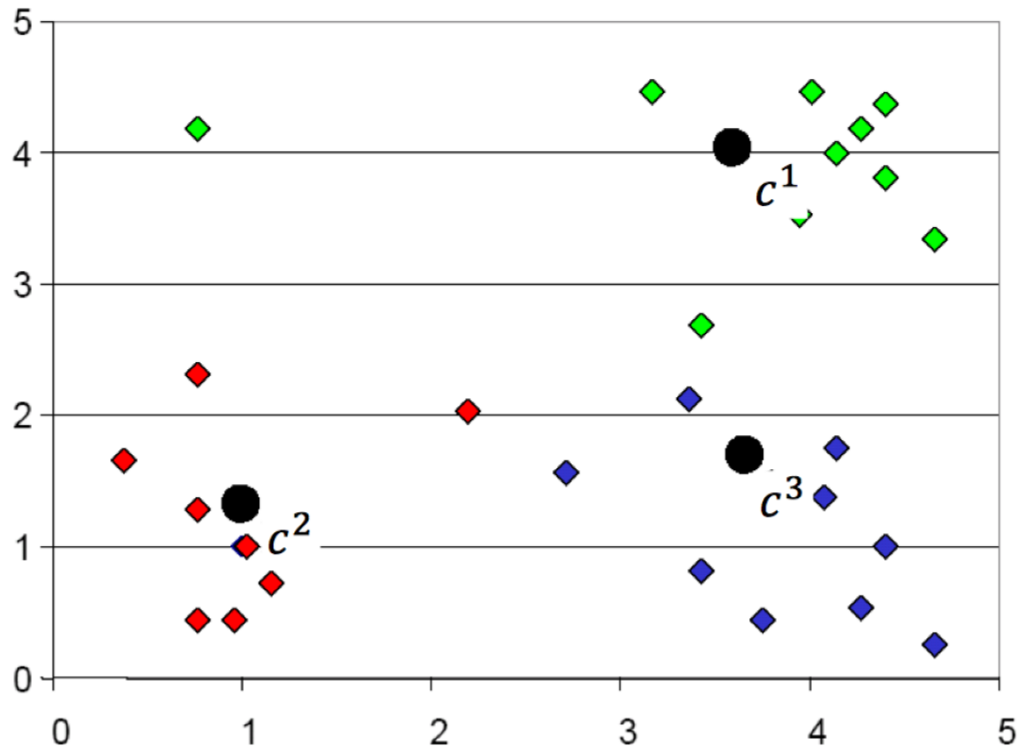
K-Means step 1



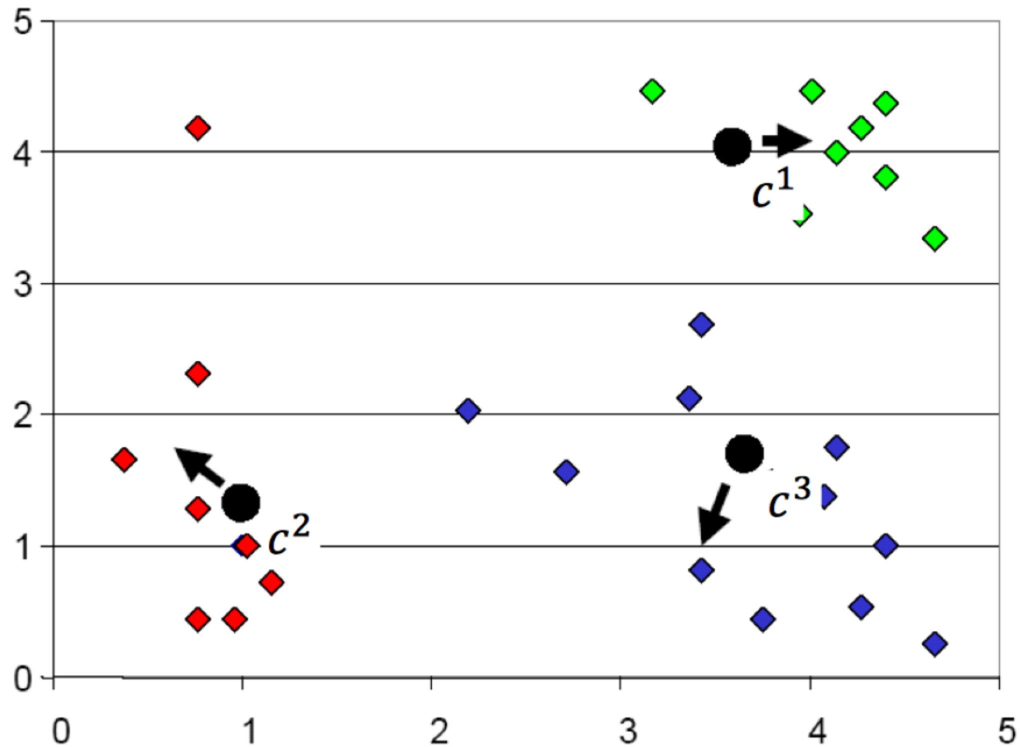
K-Means step 2



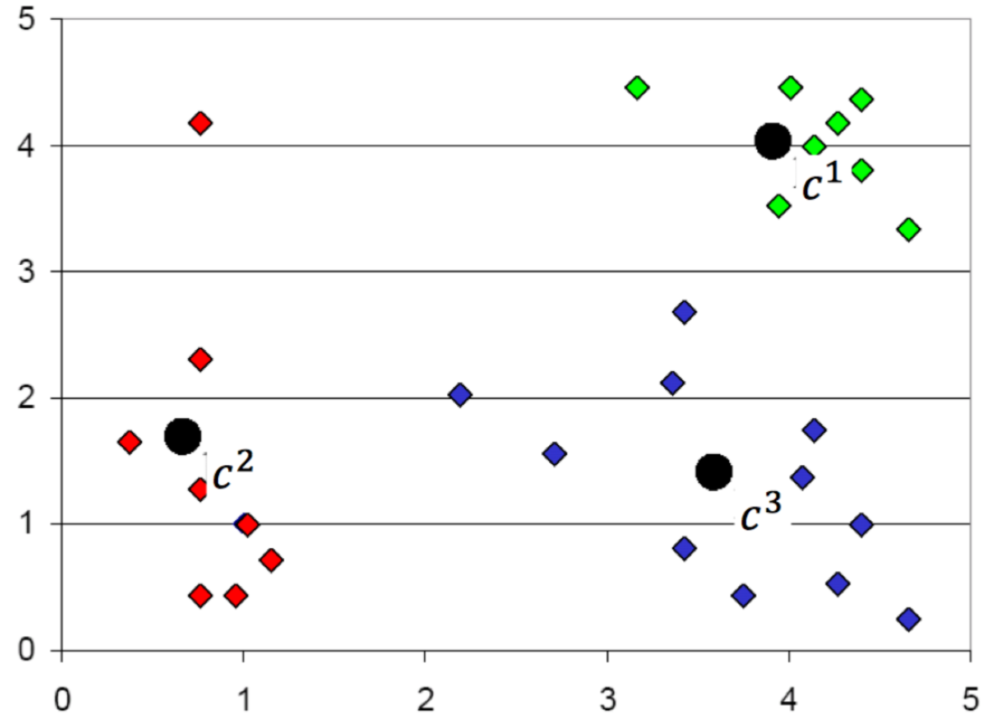
K-Means step 3



K-Means step 4



K-Means step 5



Outline

- Clustering
- Distance function
- K-means algorithm
- Analysis of K-means

Questions

- Will different initialization lead to different results?
  - Yes
  - No
  - Sometimes
- Will the algorithm always stop after some iterations?
  - Yes
  - No (We have to set a maximum number of iterations)
  - Sometime

Yes. Does it always converge to a optimum?

=> No, it is likely to converge to a local optimum.

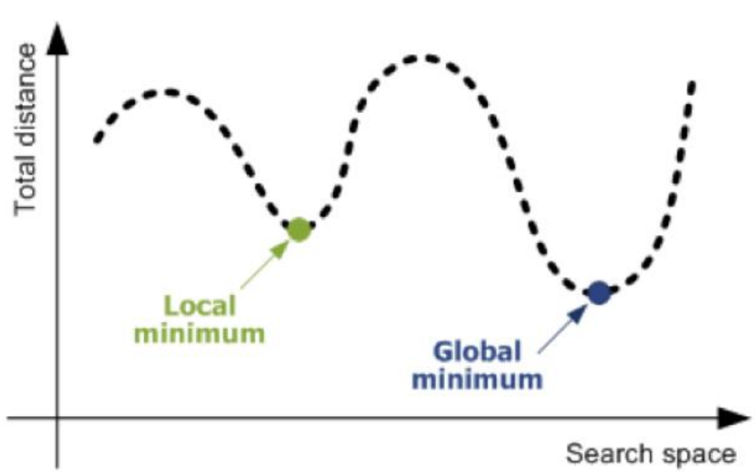
Formal statement of the clustering problem

- Given n data points, $\{x^1, x^2, \dots, x^n\} \in R^d$
- Find k cluster centers, $\{c^1, c^2, \dots, c^k\} \in R^d$
- And assign each data point i to one cluster, $\pi(i) \in \{1, \dots, k\}$
- Such that the averaged square distances from each data point to its respective cluster center is small

$$min \frac{1}{n} \sum_{i=1}^{n} ||x^i - c^{\pi(i)}||^2$$

Clustering is NP-Hard

- Find k cluster centers, $\{c^1, c^2, \dots, c^k\} \in R^d$ and assign each data point to one cluster, $\pi(i) \in \{1, \dots, k\}$, minimize
$$min \frac{1}{n} \sum_{i=1}^{n} ||x^i - c^{\pi(i)}||^2$$
- A search problem over the space of discrete assignments
  - For all n data points together, there are k^n possibility
  - The cluster assignment determines cluster centers.



An example

- For all n data points together, there are k^n possibilities, where k is the number of clusters.
- An example:
  - X={A, B, C}, n=3 (data points) k = 2 clusters

Convergence of K-Means

- Will K-Means objective oscillate?

$$min \frac{1}{n} \sum_{i=1}^{n} ||x^i - c^{\pi(i)}||^2$$

- The minimum value of the objective is finite.

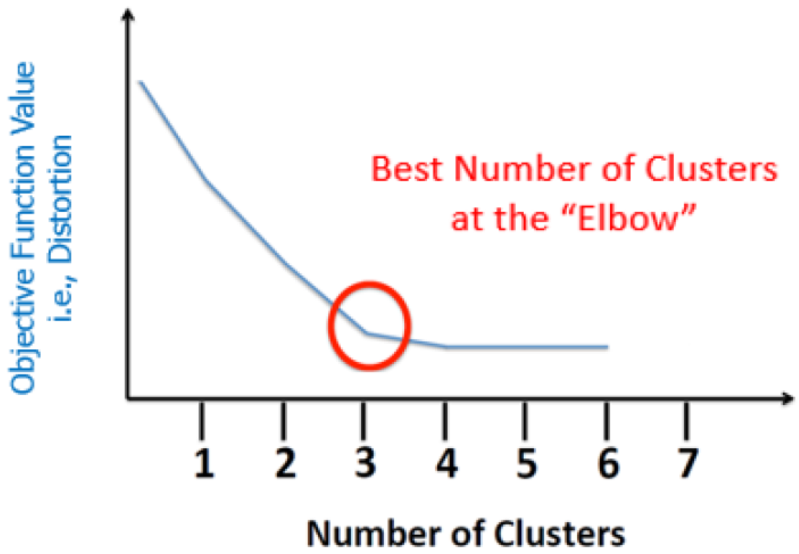
- Each iteration of K-means algorithm decrease the objective.
  - Both cluster assignment step and center adjustment step decrease objective $argmin_{j=1,\ldots,k} ||x^i - c^j||^2$ for each data point i

Time Complexity

- Assume computing distance between two instances is O(d) where d is the dimensionality of the vectors.
- Reassigning clusters for all datapoints:
  - O(kn) distance computations (when there is one feature)
  - O(knd) (when there is d features)
- Computing centroids: Each instance vector gets added once to some centroid (Finding centroid for each feature): O(nd).
- Assume these two steps are each done once for I iterations: O(Iknd).

Slide credit: Ray Mooney.

How to choose K?



Distortion score: computing the sum of squared distances from each point to its assigned center.