K-means clustering

Hung-Hsuan Chen 陳弘軒
Computer Science and Information Engineering
National Central University
hhchen@ncu.edu.tw

Slides adapted from David Sontag (NYU), Andrew W. Moore (CMU), Elise Arnaud (INRIA)

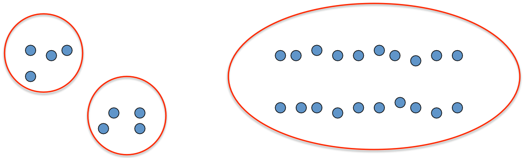
Clustering

Clustering:

- – Unsupervised learning
- – Requires data, but no labels
- – Detect patterns e.g. in
  - Group emails or search results
  - Customer shopping patterns
  - Regions of images
- – Useful when don't know what you're looking for
- – But: can get gibberish

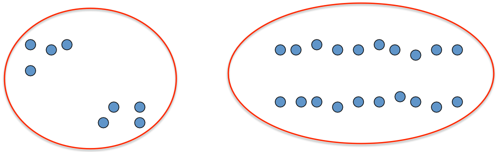
Clustering

- Basic idea: group together similar instances
- Example: 2D point patterns



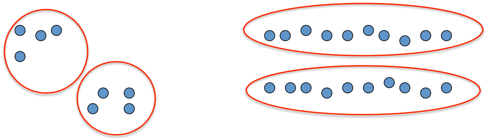
Clustering

- Basic idea: group together similar instances
- Example: 2D point patterns



Clustering

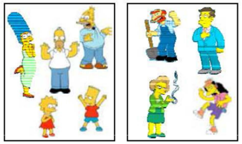
- Basic idea: group together similar instances
- Example: 2D point patterns



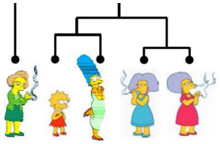
- What could "similar" mean?
  - One option: small Euclidean distance (squared)
  - Clustering results are crucially dependent on the measure of similarity (or distance) between the "points" to be clustered

Clustering algorithms

- Partition algorithm (Flat)
  - K-means
  - Mixture Gaussian
  - Spectral clustering



- Hierarchical algorithm
  - Bottom up – agglomerative
  - Top down -- divisive



Clustering examples

Image segmentation

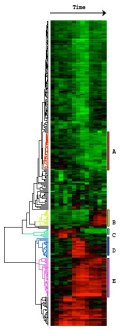
Goal: Break up the image into meaningful or perceptually similar regions

similar RGB will in same group.

[Slide from James Hayes]

Clustering examples

Clustering gene expression data

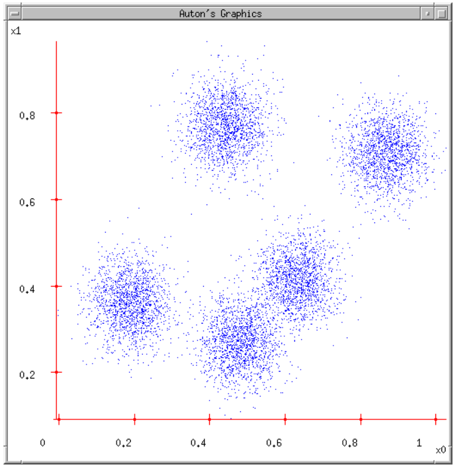


Eisen et al, PNAS 1998

groups or clusters

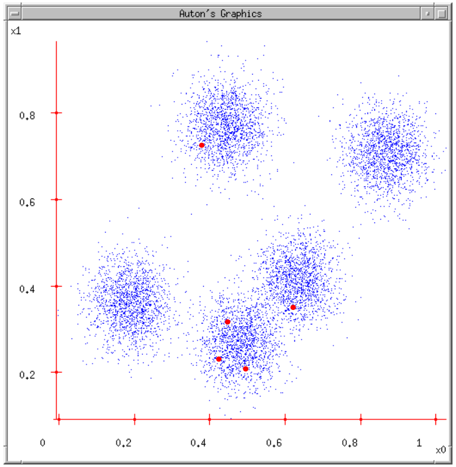
K-means

1. Ask user how many clusters they'd like. (e.g. k=5)



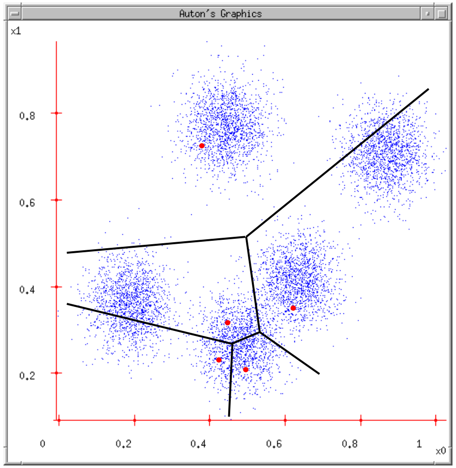
K-means

1. Ask user how many clusters they'd like. (e.g. k=5)
2. Randomly guess k cluster Center locations



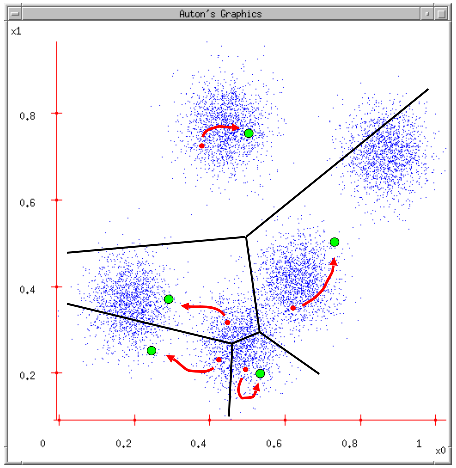
K-means

1. Ask user how many clusters they'd like. (e.g. k=5)
2. Randomly guess k cluster Center locations
3. Each datapoint finds out which Center it's closest to. (Thus each Center "owns" a set of datapoints)



K-means

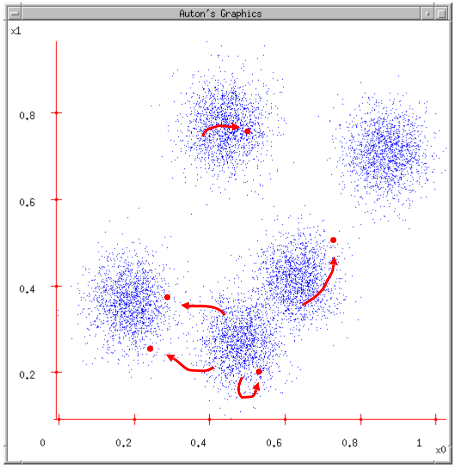
1. Ask user how many clusters they'd like. (e.g. k=5)
2. Randomly guess k cluster Center locations
3. Each datapoint finds out which Center it's closest to.
4. Each Center finds the centroid of the points it owns



K-means

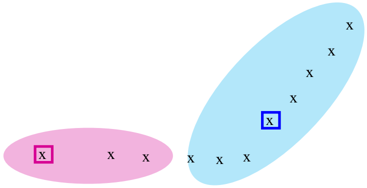
1. Ask user how many clusters they'd like. (e.g. k=5)
2. Randomly guess k cluster Center locations
3. Each datapoint finds out which Center it's closest to.
4. Each Center finds the centroid of the points it owns…
5. …and jumps there
6. …Repeat until terminated!

converge



Example: Assigning Clusters

k=2

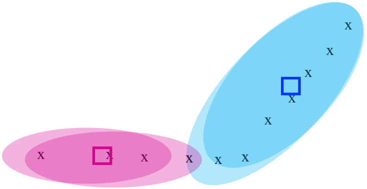


x … data point
□ … centroid

Clusters after round 1

Adopted from "Mining massive datasets" by J. Leskovec et al.`

Example: Assigning Clusters

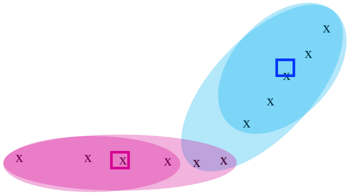


x … data point
□ … centroid

Clusters after round 2

Adopted from "Mining massive datasets" by J. Leskovec et al.`

Example: Assigning Clusters



x … data point
□ … centroid

Clusters at the end

Adopted from "Mining massive datasets" by J. Leskovec et al.`

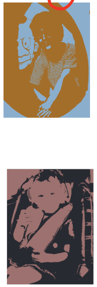
K-means is guaranteed to converge

- We ignore the proof here
- However, given different initializations, the converged results could be different

Example: K-Means for Segmentation

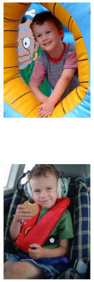
only have two different color

K=2



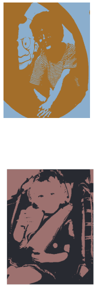
Goal of Segmentation is to partition an image into regions each of which has reasonably homogenous visual appearance.

Original

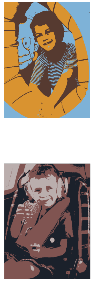


Example: K-Means for Segmentation

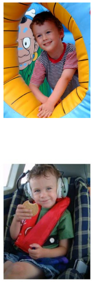
K=2



K=3

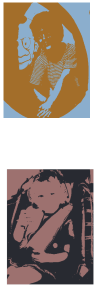


Original



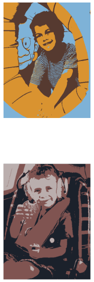
Example: K-Means for Segmentation

K=2



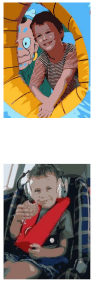
Size: 4% of original

K=3



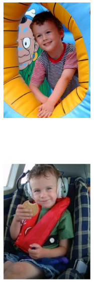
8%

K=10

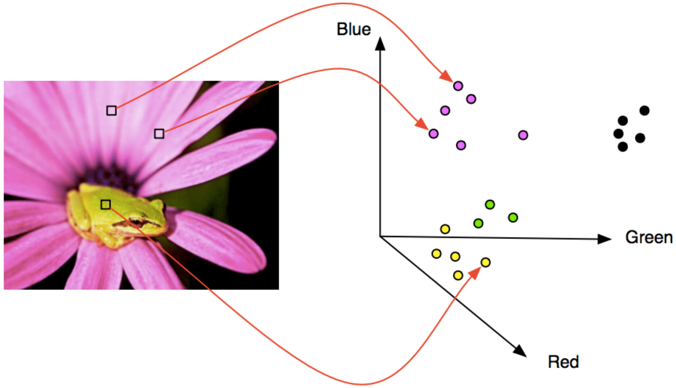


17%

Original



K-Means for Segmentation

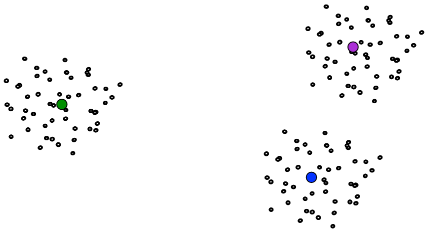


Initialization

- K-means **algorithm** is a heuristic
  - Requires initial means
  - It does matter what you pick!
  - What can go wrong?
  - Various schemes for preventing this kind of thing: variance-based split / merge, initialization heuristics

Example of K-Means Getting Stuck

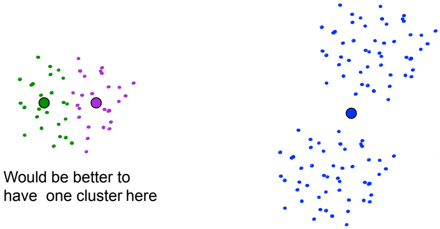
Ideally



Example of K-Means Getting Stuck

A local optimum:

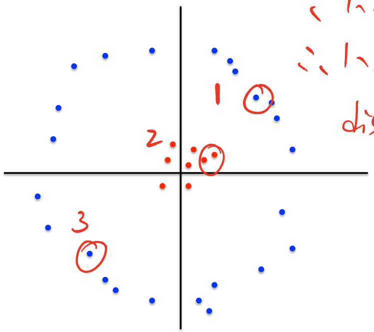
* the initial data point should choose as far as possible



Would be better to have one cluster here

and two clusters here…

K-means not able to properly cluster

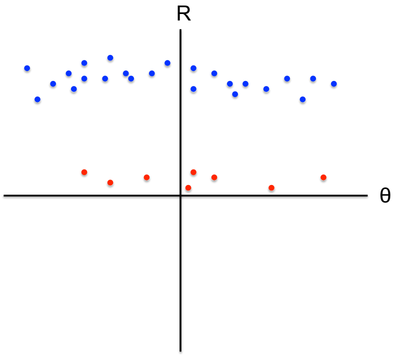


∵ 1, 2, diffrent

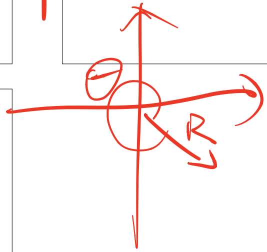
∵ 1, 3, different

dist(1,2) > dist(1,3)

Changing the features can help



transport



Quiz

- What is the difference between k-means and KNN?
- Given a set of students with their heights, weights, and genders, you are asked to build a model to predict the gender of a new student
  - Which one is more appropriate? KNN or k-means?

→ k-mean unsupervise

KNN supervise

→ KNN cause we have label