Clustering with the k-means algorithm II

Sanjoy Dasgupta

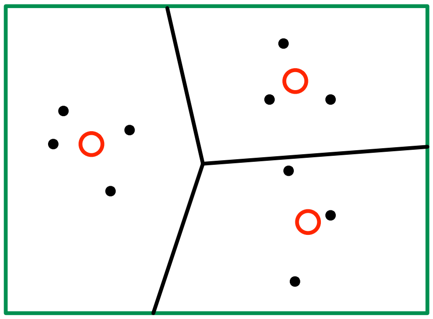
University of California, San Diego

Topics we'll cover

1. Two uses of k-means clustering
2. Clustering in a streaming or online setting
3. The good and bad of k-means

Lloyd's k-means algorithm

- Initialize centers $\mu_1, \ldots, \mu_k$ in some manner.
- Repeat until convergence:
    - Assign each point to its closest center.
    - Update each μ_j to the mean of the points assigned to it.



Each iteration reduces the cost ⇒ convergence to a local optimum.

Two common uses of clustering

- Vector quantization
  Find a finite set of representatives that provides good coverage of a complex, possibly infinite, high-dimensional space.

- Finding meaningful structure in data
  Finding salient grouping in data.

Representing images using k-means codewords

How to represent a collection of images as fixed-length vectors?

- Take all $\ell \times \ell$ patches in all images. Extract features for each.
- Run k-means on this entire collection to get k centers.
- Now associate any image patch with its nearest center.
- Represent an image by a histogram over $\{1, 2, \ldots, k\}$.

Looking for natural groups in data

"Animals with attributes" data set

- 50 animals: antelope, grizzly bear, beaver, dalmatian, tiger, ...
- 85 attributes: longneck, tail, walks, swims, nocturnal, forager, desert, bush, plains, ...
- Each animal gets a score $(0 - 100)$ along each attribute
- 50 data points in $\mathbb{R}^{85}$

Apply k-means with $k = 10$ and look at grouping obtained.

1. zebra
2. spider monkey, gorilla, chimpanzee
3. tiger, leopard, wolf, bobcat, lion
4. hippopotamus, elephant, rhinoceros
5. killer whale, blue whale, humpback whale, seal, walrus, dolphin
6. giant panda
7. skunk, mole, hamster, squirrel, rabbit, bat, rat, weasel, mouse, raccoon
8. antelope, horse, moose, ox, sheep, giraffe, buffalo, deer, pig, cow
9. beaver, otter
10. grizzly bear, dalmatian, persian cat, german shepherd, siamese cat, fox, chihuahua, polar bear, collie

1. zebra
2. spider monkey, gorilla, chimpanzee
3. tiger, leopard, fox, wolf, bobcat, lion
4. hippopotamus, elephant, rhinoceros, buffalo, pig
5. killer whale, blue whale, humpback whale, seal, otter, walrus, dolphin
6. dalmatian, persian cat, german shepherd, siamese cat, chihuahua, giant panda, collie
7. beaver, skunk, mole, squirrel, bat, rat, weasel, mouse, raccoon
8. antelope, horse, moose, ox, sheep, giraffe, deer, cow
9. hamster, rabbit
10. grizzly bear, polar bear

Streaming and online computation

Streaming computation: for data too large to fit in memory.

- Make one pass (or maybe a few passes) through the data.
- On each pass:
  - See data points one at a time, in order.
  - Update models/parameters along the way.
- Only enough space to store a tiny fraction of data, or perhaps a short summary.

Online computation: even more lightweight, for data continuously being collected.

- Initialize a model.
- Repeat forever:
  - See a new data point.
  - Update model if need be.

Example: sequential k-means

1. Set the centers $\mu_1, \ldots, \mu_k$ to the first k data points
2. Set their counts to $n_1 = n_2 = \cdots = n_k = 1$
3. Repeat, possibly forever:
    - Get next data point x
    - Let μ_j be the center closest to x
    - Update μ_j and n_j:

$$\mu_j = \frac{n_j \mu_j + x}{n_j + 1} \quad \text{and} \quad n_j = n_j + 1$$

K-means: the good and the bad

The good:

- Fast and easy.
- Effective in quantization.

The bad:

- Geared towards spherical clusters of roughly the same radius.

How to accommodate clusters of more general shape?