CS506 : Clustering

- want to group similar objects together
- dissimilar objects in different clusters
- notion of cluster can be ambiguous

Types of Clustering:

- partitional: each object belongs in exactly 1 cluster
- hierarchical : nested clusters organized in a tree
- density based: based on local density of points

Partitional Algorithms

- n objects into K clusters
- need to know K in advance

K-Means Problem

- set $X = \{x_1 \dots x_n\}$ of n points in R^d
- K is given
- cost function:

center

L_2 norm / size of difference vector $x - c_j$

$$\sum_{i=1}^{n} \min_j \{L_2^2(x_i, c_i)\} = \sum_{i=1}^{n} \min \|x_i - c_j\|_2^2$$

L_2 distance

Goal

$$\text{minimize} \sum_{j=1}^{K} \sum_{x_i \in c_j} d_2(x, c_j)^2$$

all points x in cluster c_j

*want to minimize sum of distances between x and center of cluster

- randomly pick K cluster centers (pick K points in data set) $\{c_1, \dots c_k\}$
- for each j, set cluster X_j to be the set of points in X that are the closest to center c_j
  (for each point in data, see which cluster its closest to, assign it to that cluster)
- adjust center of cluster based on center of mass
- repeat (starting 2) until it converges (cost function is at its minimum)

Possible scenarios where it does not converge:
- finite points = finite partitions and costs
- we can be in a loop (go back to same clustering you had before)
  → won't happen because you cannot move unless it improves your cost

* we will be be coding this algorithm

Statement:

$$\mu^* = \operatorname{argmin} \sum_{x \in S} \|x - \mu\|_2^2 = \bar{x} = \frac{1}{|S|} \sum_{x \in S} x$$

point μ^* is $\qquad \|x-\mu\| = \sqrt{\langle x-\mu, x-\mu \rangle}$

the point that minimizes intracluster distance is the mean

How to prove this: expand, identity of inner products / differences

$$\mu^* = \underset{\mu}{\operatorname{argmin}} \sum_{x \in S} \|x - \mu\|_2^2$$

$$= \underset{\mu}{\operatorname{arg\,min}} \sum_{x \in S} \left(\sqrt{\langle x-\mu, x-\mu \rangle}\right)^2$$

Scanned with CamScanner

want to show $\langle x, x \rangle$ = constant independent of μ

$\|x\|_2^2$ → distance/norm between x and 0

does not have x's

$$= \arg\min_{\mu} \sum_{x \in S} [\langle x, x \rangle - 2\langle x, \mu \rangle + \langle \mu, \mu \rangle]$$

$|S| = n$

$$= \arg\min_{\mu} n\langle \mu, \mu \rangle - 2 \sum_{x \in S} \langle x, \mu \rangle$$

why did we get rid of $\langle x, x \rangle$?

??

$$= \arg\min_{\mu} n\langle \mu, \mu \rangle - 2n\langle \tfrac{1}{n}\sum_{x \in S} x, \mu \rangle$$

n can be factored out, also a constant so can be taken away

$$= \arg\min_{\mu} \langle \mu, \mu \rangle - 2\langle \bar{x}, \mu \rangle$$

$$= \arg\min_{\mu} \langle \mu, \mu \rangle - 2\langle \bar{x}, \mu \rangle + \langle \bar{x}, \bar{x} \rangle$$

→ why did we add this?

$$= \arg\min_{\mu} \langle \bar{x} - \mu, \bar{x} - \mu \rangle$$

*added $\bar{x}$, don't need to subtract it because it's constant w/ respect to μ

$$= \arg\min_{\mu} \|\bar{x} - \mu\|_2^2$$

$$= \bar{x}$$

*works in dimensions higher than 2, but inefficient

Properties of K-means algorithm

- finds a local optimum
- often converges quickly (but not always)
- choice of initial points can have large influence in result

Limitations of K-Means

- non-spherical shapes won't work

*only way to determine best # of clusters is to try it out