k-means clustering

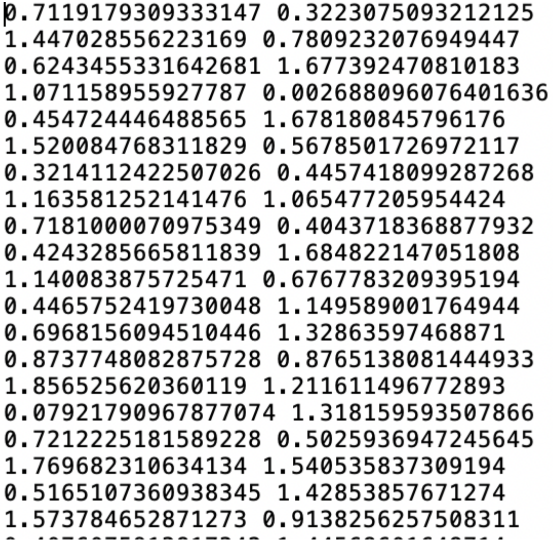
Yangming Li

University of Victoria

March 27th

motivation

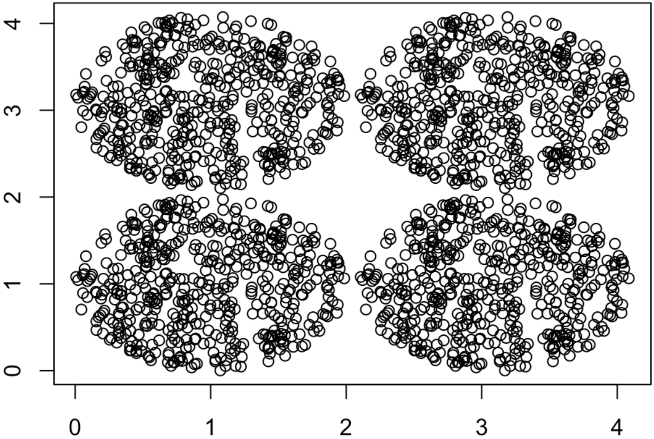
data



```
0.7119179309333147 0.3223075093212125
1.447028556223169 0.7809232076949447
0.6243455331642681 1.677392470810183
1.071158955927787 0.002688096076401636
0.454724446488565 1.678180845796176
1.520084768311829 0.5678501726972117
0.3214112422507026 0.4457418099287268
1.163581252141476 1.065477205954424
0.7181000070975349 0.4043718368877932
0.4243285665811839 1.684822147051808
1.140083875725471 0.6767783209395194
0.4465752419730048 1.149589001764944
0.6968156094510446 1.32863597468871
0.8737748082875728 0.8765138081444933
1.856525620360119 1.211611496772893
0.07921790967877074 1.318159593507866
0.7212225181589228 0.5025936947245645
1.769682310634134 1.540535837309194
0.5165107360938345 1.42853857671274
1.573784652871273 0.9138256257508311
```

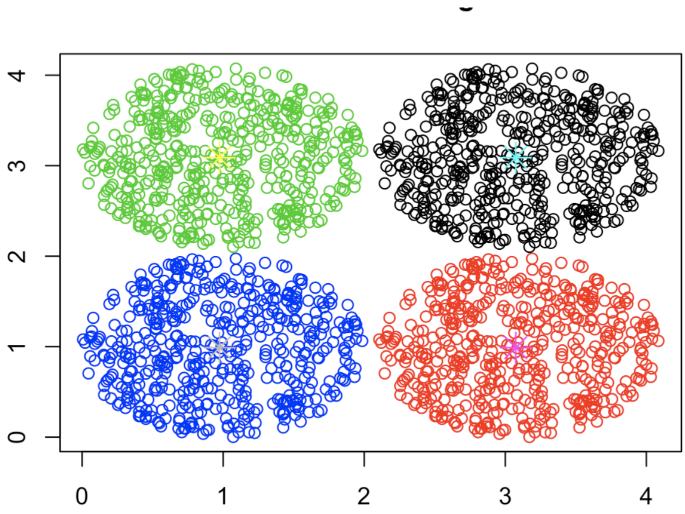
motivation

plot



motivation

plot



What is clustering

- **unsupervised learning**
  classifcation:no predefined classes and no training examples
  clustering:Grouping a set of data objects into clusters such that the data objects are similar to one another within the same cluster, dissimilar to the objects in other clusters

why clustering

- **goals** Find representatives for homogeneous groups
  Find “natural” clusters and describe their properties
  Find suitable and useful grouping ”useful” Data Classes and
  find unusual data object :Outlier Detection

what is k-means do

- `partition based algorithm`
- `history`: K-means(MacQueen,1967)
- **cluster**:The organization of unlabeled data into similarity groups called clusters.
- **centroid**:Each cluster has a cluster center

kmeans algorithm

- `key idea`: minimize the squated error
- `input`: The number of k and a database containing n objects.
- `output`: A set of k-clusters that minimize the squared-error criterion.
- `idea` initialize k centroids
  iterate two steps to convergence
  a) Assign step
  b) recompute centroid steps

algorithm

Select K points as the intial centroids

Repeat:
Assignment steps: Form K clusters by assigning all points to the closet centroids.
Recompute centroid steps:move each centroid to the center of the data-points assigned it.

Until: the centroid dones't change

when k-means convergence

1.no re-assignments of data points to different clusters
2.no change of centroids
3.minimum decease in SSE

insight algorithm

data points set D:$\left\{x_1, x_2, ..., x_n\right\}$ xi$\in R^d$

k clusters,k centers,$\mu_1, ...\mu_k \in R^d$

$$object = \sum_{j=1}^{k} \sum_{x_i \to \mu_j} \left\|x_i - \mu_j\right\|^2$$

how to express $the \sum_{x_i \to \mu_j} mathmatrically$?
hint: STAT450 therom 10.2.3

insight algorithm

adding indicator function(variable)

$$a_{ij} = \begin{cases} 1 & \mathit{if}\ x_i \rightarrow j \\ \\ 0 & \mathit{else} \end{cases}$$

simplify to

$$\mathit{object} = \sum_{j=1}^{k} \sum_{i=1}^{n} a_{ij} \left\| x_i - \mu_j \right\|^2$$

insight algorithm

minimize

$$object = \sum_{j=1}^{k} \sum_{i=1}^{n} a_{ij} \left\| x_i - \mu_j \right\|^2$$

idea:

$$1. fix\ \mu,$$

then

$$2. fix\ a$$

insight algorithm

step 1: *fix* μ

minimize

$$object = \sum_{j=1}^{k} \sum_{i=1}^{n} a_{ij} \left\| x_i - \mu_j \right\|^2$$

with respect to a , say μ_l is the nearest center to the point a

$$a_{ij} = \begin{cases} 1 & j = argmin \left\| x_i - \mu_l \right\| \\ \\ 0 & else \end{cases}$$

insight algorithm

step 2: *fix a*

minimize

$$object = \sum_{j=1}^{k} \sum_{i=1}^{n} a_{ij} \left\| x_i - \mu_j \right\|^2$$

with respect to μ
for$\mu_m \neq \mu_j$,$m \neq j$

$$object = \sum_{i=1}^{n} a_{ij} \bigtriangledown_{\mu_j} (x_i - \mu_j)^T (x_i - \mu_j) + \sum_{i=1}^{n} a_{ij} \bigtriangledown \mu_j (x_i - \mu_m)^T (x_i - \mu_m)$$

second term=0 $object = \sum_{i=1}^{n} a_{ij} \bigtriangledown_{\mu_j} (x_i - \mu_j)^T (x_i - \mu_j)$

insight algorithm

$$object = \sum_{i=1}^{n} a_{ij} \bigtriangledown_{\mu_j} (x_i - \mu_j)^T (x_i - \mu_j)$$

$$\bigtriangledown(x_i^T x_i - 2\mu_j^T xi + \mu_j^T \mu_j) = -2x_i + 2\mu_j$$

insight algorithm

$$object = \sum_{i=1}^{n} a_{ij} \bigtriangledown_{\mu_j} (x_i - \mu_j)^T (x_i - \mu_j)$$

$$\bigtriangledown(x_i^T x_i - 2\mu_j^T xi + \mu_j^T \mu_j) = -2x_i + 2\mu_j$$

$$object = (x_i^T x_i - 2\mu_j^T xi + \mu_j^T \mu_j) = -2x_i + 2\mu_j$$

$$-2\sum_{i=1}^{n} a_{ij} x_i + 2\mu_j \sum_{i=1}^{n} a_{ij} = 0$$

$$\mu_j = \frac{\sum_{i=1}^{n} a_{ij} x_i}{\sum_{i=1}^{n} a_{ij}}$$

insight algorithm

$$\frac{\partial}{\partial \mu_j} \triangledown_{\mu_j} \, object = 2(\sum_{i=1}^{n} a_{ij}) e_k$$

$$\triangledown_{\mu_j} object = 2(\sum_{i=1}^{n} a_{ij}) \mathbb{I} > 0$$

$$a_{ij} = \begin{cases} 1 & if \; x_i \rightarrow j \\ \\ 0 & else \end{cases}$$

insight algorithm

$$\bigtriangledown_{\mu_k} object = 2(\sum_{i=1}^{n} a_{ij})\mathbb{I} > 0$$

$$a_{ij} = \begin{cases} 1 & if\ x_i\ \rightarrow j \\ \\ 0 & else \end{cases}$$

$$n_j = \sum_{i=1}^{n} a_{ij} = number\ of\ points : x_i\ assigned\ to\ j$$

$$\mu_j = \frac{\sum_{i=1}^{n} a_{ij} x_i}{\sum_{i=1}^{n} a_{ij}}$$

insight algorithm

$$n_j = \sum_{i=1}^{n} a_{ij} = number\ of\ points : x_i\ assigned\ to\ j$$

$$\mu_j = \frac{\sum_{i=1}^{n} a_{ij} x_i}{\sum_{i=1}^{n} a_{ij}}$$

$$\mu_j = \frac{1}{n_j} \sum_{i:X_i \to \mu_j} x_i$$

small example

suppose we have number set D=2,3,4,10,11,12,20,25,30, we want to cluster it into two groups,any guess for the two groups?

small example

suppose we have data set D=2,3,4,10,11,12,20,25,30, we want to cluster it into two groups
Select 2 points as the intial centroids:

$$\mu_1 = 4, \mu_2 = 12$$

Assignment steps: $\{2, 3, 4\}$,$\{10, 11, 12, 20, 25, 30\}$
recompute centroid steps

$$\mu_1 = 3, \mu_2 = 18$$

Assignment steps: $\{2, 3, 4, 10, \}$,$\{11, 12, 20, 25, 30\}$
recompute centroid steps

$$\mu_1 = 4.75, \mu_2 = 19.5$$

small example

Assignment steps: $\{2, 3, 4, 10, 11, 12\}$,$\{20, 25, 30\}$
recompute centroid steps

$$\mu_1 = 7, \mu_2 = 25$$

Assignment steps: $\{2, 3, 4, 10, 11, 12\}$,$\{20, 25, 30\}$
Converages stop!

determine value of k

1.**Elbow Method**
choose different k ,see the sum of squared error, we want a small k, get the minimize **sum of the squared errors**
plot a line chart of the SSE for each value of k.The line chart looks like an arm, then the "elbow" on the arm is the value of k that is the best.
question: what k values get the minimum the sum squared errors?

determine k,average silhouette approach

2.**Silhouette coefficient** Cohesion a(x): average distance for the point x to the points in the same cluster.
Separation b(x): average distance for point x to the points in other clusters j. Find the minimum among the clusters.
bx=min($b_{i1}, b_{i2}, ...b_{ik}$)
silhouette s(x):
$s(x) = \frac{b(x)-a(x)}{max(a(x),b(x))}$
s(x)=$\begin{cases} 1 - \frac{a(x)}{b(x)} & ax < b(x) \\ 0 & a(x) = b(x) \\ \frac{b(x)}{a(x)} & ai > b(i) \end{cases}$
s(x)=[-1, +1]:
-1=bad, 0=indifferent, 1=good
Silhouette coefficient (SC): $SC = \frac{1}{N} \sum s(x)$

big example

R example in higher dimension

instead of using the build in kmeans. write own kmeans graph,goal is to plot the original data graph and the graph clustered by color

advantage

1.easy to understand and to implement
2.efficient: and fast :k cluster n data points the running time to assign step data points to closest cluster O(KN)
The running time for change the cluster center to the average assigned points O(N)

disadvantage

**1.number of clusters, K , must be determined before hand.
2.it is sensitive to initial condition. Different initial condition may produce different result of cluster. may be trapped in the local optimum**

questions

any questions?

reference

Thank you for watching!

Hartigan, John A., and Manchek A. Wong. "Algorithm AS 136: A k-means clustering algorithm." Journal of the Royal Statistical Society. Series C (Applied Statistics) 28.1 (1979): 100-108.

Alsabti, Khaled, Sanjay Ranka, and Vineet Singh. "An efficient k-means clustering algorithm." (1997).

NG, Andrew. Lecture 13.2 — Clustering — KMeans Algorithm — [Machine Learning — Andrew Ng]. 2008, towardsdatascience.com/andrew-ngs-machine-learning-course-in-python-kmeans-clustering-pca-b7ba6fafa74.

K mean clustering algorithm with solve e. www.youtube.com/watch?v=YWgcKSa$_2 agt = 621s$.