CS540 Introduction to Artificial Intelligence
Lecture 11

Young Wu
Based on lecture slides by Jerry Zhu, Yingyu Liang, and Charles Dyer

July 2, 2020

Midterm

Admin

no lecture
on Friday

Tue - Fri

- The midterms are:
- A: Too Easy
- B: Easy
- C: (B, D)
- D: Hard
- E: Too Hard

Unsupervised Learning

Motivation

- Supervised learning: $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$.
- Unsupervised learning: $x_1, x_2, \ldots, x_n$.
- There are a few common tasks without labels.

1. Clustering: separate instances into groups. group index 0,1,2 .. k.
2. Novelty (outlier) detection: find instances that are different. 0,1
3. Dimensionality reduction: represent each instance with a lower dimensional feature vector while maintaining key characteristics.

$\begin{pmatrix} 0.1 \\ 0.2 \end{pmatrix}$ $\begin{pmatrix} 0.1 \\ 0.3 \end{pmatrix}$ $\begin{pmatrix} 1.1 \\ 1.2 \end{pmatrix}$

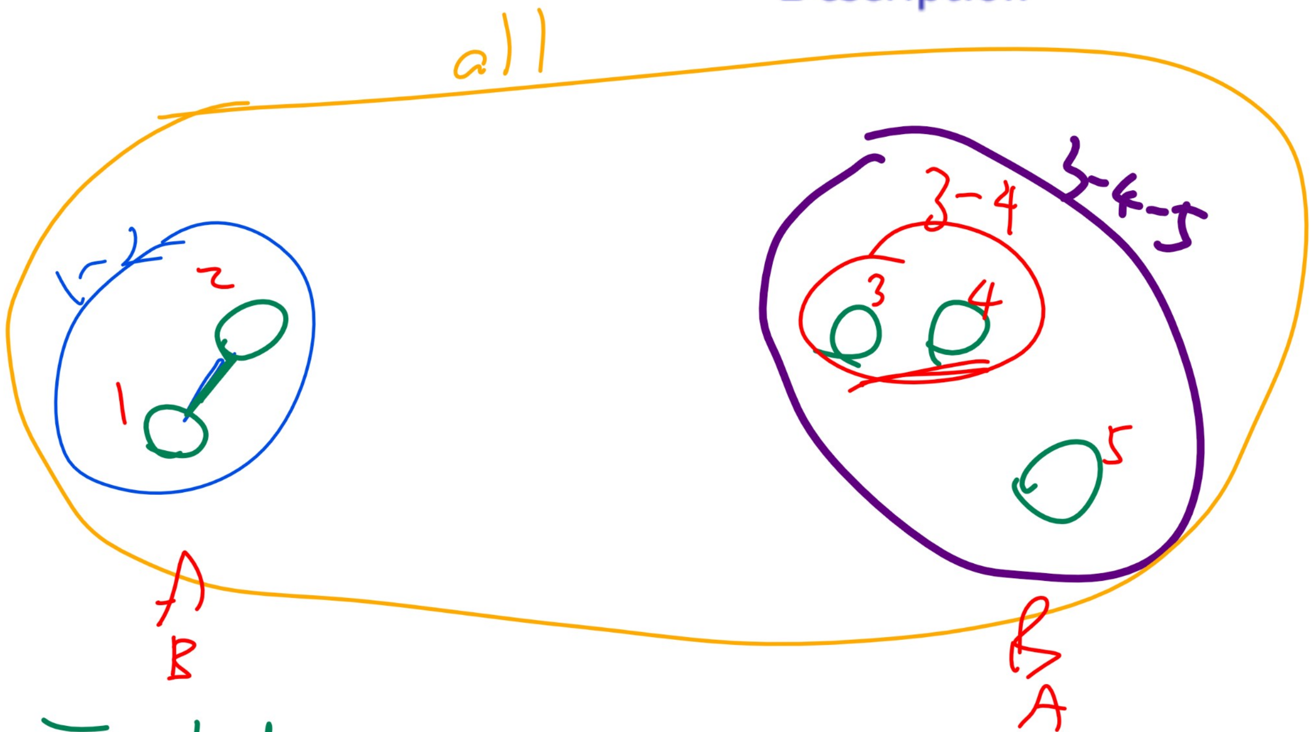
Hierarchical Clustering

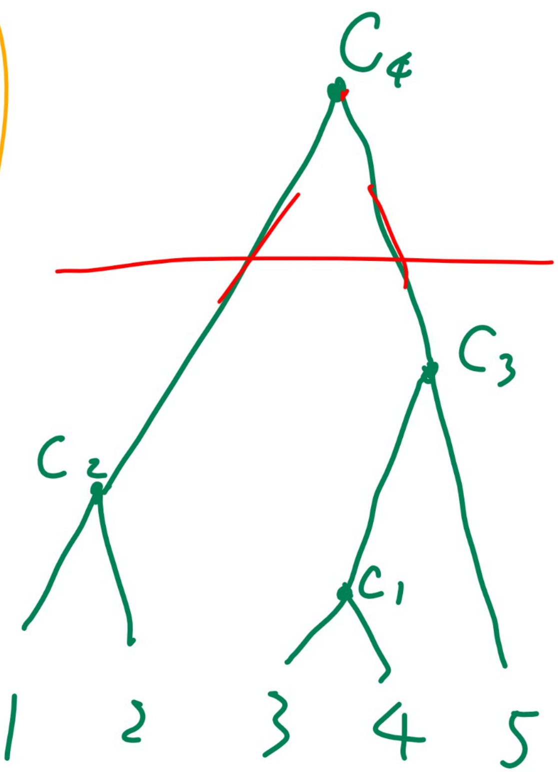
Description

- Start with each instance as a cluster.
- Merge clusters that are closest to each other.
- Result in a binary tree with close clusters as children.

Hierarchical Clustering Diagram

Description



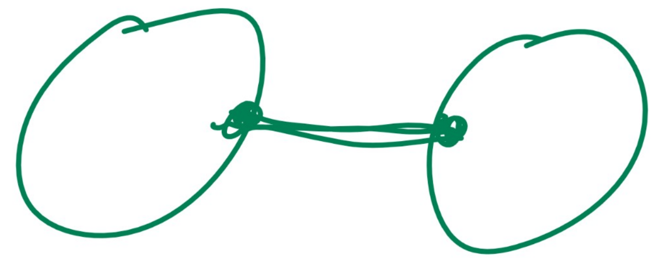


Euclidean ←

Manhattan ←

Single Linkage Distance

Definition



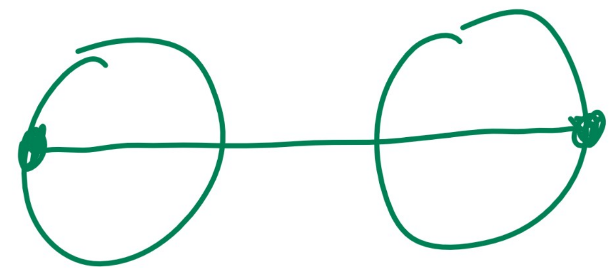
- Usually, the distance between two clusters is measured by the single-linkage distance.

$$d\left(C_k, C_{k'}\right) = \min\left\{d\left(x_i, x_{i'}\right) : x_i \in C_k, x_{i'} \in C_{k'}\right\}$$

- It is the shortest distance from any instance in one cluster to any instance in the other cluster.

Complete Linkage Distance

Definition



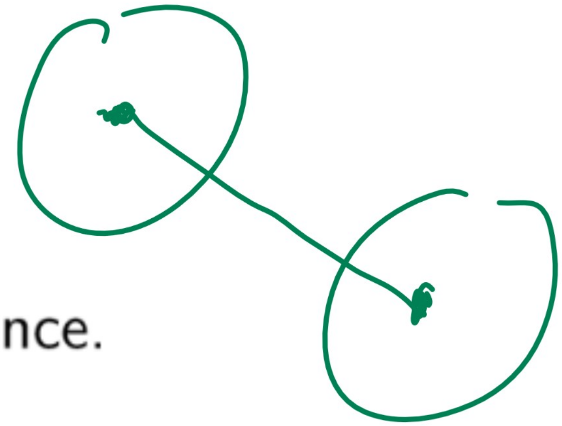
- Another measure is complete-linkage distance,

$$d\left(C_k, C_{k'}\right) = \max\left\{d\left(x_i, x_{i'}\right) : x_i \in C_k, x_{i'} \in C_{k'}\right\}$$

- It is the longest distance from any instance in one cluster to any instance in the other cluster.

Average Linkage Distance Diagram

Definition



- Another measure is average-linkage distance.

$$d\left(C_k, C_{k'}\right) = \frac{1}{|C_k|\,|C_{k'}|} \sum_{x_i \in C_k, x_{i'} \in C_{k'}} d\left(x_i, x_{i'}\right)$$

- It is the average distance from any instance in one cluster to any instance in the other cluster.

Hierarchical Clustering 1

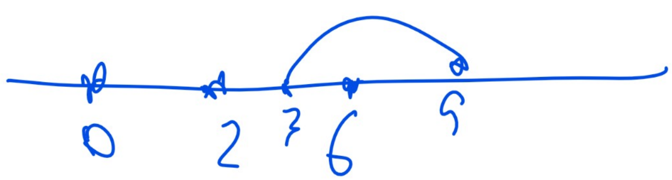
Quiz

- Spring 2018 Midterm Q5
- Given three clusters $A = \{0, 2, 6\}$, $B = \{3, 9\}$, $C = \{11\}$. What is the next iteration of hierarchical clustering with Euclidean distance and single linkage?
- A: Merge A and B.
- B: Merge A and C.
- C: Merge B and C.

	single dist	complete dist
AB	1	9
BC	2	8
AC	5	11

complete linkage

Merge B and C.



Hierarchical Clustering 2

Quiz

Q 2

- ~~Spring 2018 Midterm Q5~~

8 7

- Given three clusters $A = \{0, 1\}$, $B = \{4, 8\}$, $C = \{10, 11\}$. What is the next iteration of hierarchical clustering with Euclidean distance and complete linkage?
- A: Merge A and B.
- B: Merge A and C.
- C: Merge B and C.

Hierarchical Clustering 3

Quiz

Q3

- Spring 2018 Midterm Q5
- Given three clusters $A = \{0, 1\}$, $B = \{4, 8\}$, $C = \{10, 11\}$. What is the next iteration of hierarchical clustering with Euclidean distance and single linkage?
- A: Merge A and B.
- B: Merge A and C.
- C: Merge B and C.

Hierarchical Clustering 4

Quiz

- Spring 2018 Midterm Q5
- Given three clusters $A = \{0, 2, 6\}, B = \{3, 9\}, C = \{11\}$. What is the next iteration of hierarchical clustering with Euclidean distance and complete linkage?
- A: Merge A and B.
- B: Merge A and C.
- C: Merge B and C.

Hierarchical Clustering 3

Quiz

	A	B	CD	E
A	0	1075	2013	996
B		0	2687	2037
CD	symmetric		0	1059
E				0

P4

- Spring 2017 Midterm Q4
- Given the distance between the clusters so far. Which pair of clusters will be merged using single linkage.

cluster CD

pairwise dist

–	A	B	C	D	E
A	0	1075	2013	~~2054~~	996
B	1075	0	~~3272~~	2687	2037
C	2013	3272	0	808	1307
D	2054	2687	808	0	1059
E	996	2037	~~1307~~	$\boxed{1059}$	0

min in table

complete 2054

Hierarchical Clustering 4

Quiz

- Given the distance between the clusters so far. Which pair of clusters will be merged using complete linkage.

–	A	B	C	D	E
A	0	1075	2013	2054	996
B	1075	0	3272	2687	2037
C	2013	3272	0	808	1307
D	2054	2687	808	0	1059

Hierarchical Clustering 5

Quiz

Q 4

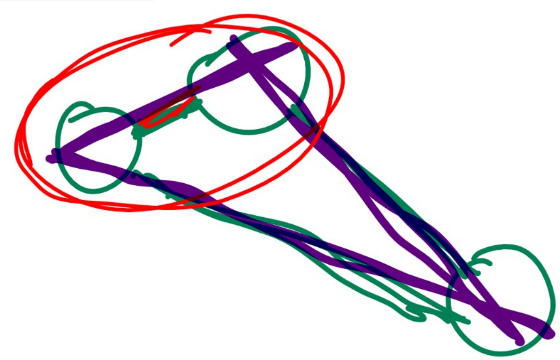
select one of them

- Given the distance between the clusters so far. Which pair of clusters will be merged using single linkage.

–	A	B	CD	E
A	0	1075	2013	996
B	1075	0	2687	2037
CD	2013	2687	0	1059
E	996	2037	1059	0

0 2 1 3 4 5

tie break by smaller index



Hierarchical Clustering

Algorithm

- Input: instances: $\{x_i\}_{i=1}^{n}$, the number of clusters K, and a distance function d.
- Output: a list of clusters $C = C_1, C_2, ..., C_K$.
- Initialize for $t = 0$.

$$C^{(0)} = C_1^{(0)}, ..., C_n^{(0)}, \text{ where } C_k^{(0)} = \{x_k\}, k = 1, 2, ..., n$$

- Loop for $t = 1, 2, ..., n - k + 1$.

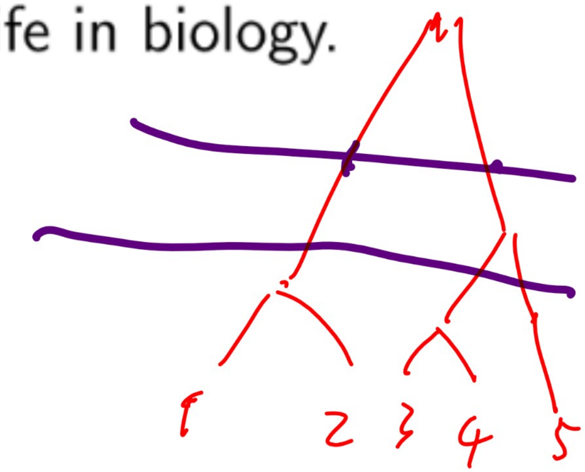
$$(k_1^\star, k_2^\star) = \arg\min_{k_1, k_2} d\left(C_{k_1}^{(t-1)}, C_{k_2}^{(t-1)}\right)$$

$$C^{(t)} = \left(C_{k_1^\star}^{(t-1)} \cup C_{k_2^\star}^{(t-1)}\right), C_1^{(t-1)}, \ldots \text{ no } k_1^\star, k_2^\star ..., C_n^{(t-1)}$$

Number of Clusters

Discussion

- K can be chosen using prior knowledge about X.
- The algorithm can stop merging as soon as all the between-cluster distances are larger than some fixed R.
- The binary tree generated in the process is often called dendrogram, or taxonomy, or a hierarchy of data points.
- An example of a dendrogram is the tree of life in biology.



K Means Clustering

Description

- This is not K ~~Nearest~~ Neighbor.
- Start with random cluster centers.
- Assign each point to its closest center.
- Update all cluster centers as the center of its points.

K Means Clustering Diagram

Description

Distortion

Distortion

- Distortion for a point is the distance from the point to its cluster center.
- Total distortion is the sum of distortion for all points.

min

$$\boxed{D_K = \sum_{i=1}^{n} d\left(x_i, c_{k^\star(x_i)}(x_i)\right)}$$

$$k^\star(x) = \arg \min_{k=1,2,\ldots K} d\left(x, c_k\right)$$

Objective Function

Definition

- When using Euclidean distance, sometimes total distortion is defined as sum of squared distances.

$$\min \quad D_K = \sum_{i=1}^{n} d_2\left(x_i, c_{k^*(x_i)} \cancel{(x_i)}\right)^2 \quad \text{by GD}$$

index

- This algorithm stop in finite steps.
- This algorithm is trying to minimize the total distortion but fails.

local ~~min~~

P4 start with multiple random initial centers

Objective Function Counterexample

Definition

K Means Clustering 1

Quiz

2 8

- Spring 2018 Midterm Q5
- Given data $\{5, 7, 10, 12\}$ and initial cluster centers $c_1 = 3, c_2 = 13$, what is the initial clusters?
- A: $\{5, 7\}$ and $\{10, 12\}$
- B: $\{5\}$ and $\{7, 10, 12\}$
- C: $\{5, 7, 10\}$ and $\{12\}$

c_1 $\qquad$ c_2

$(5, 7)$, $\quad$ $(10, 12)$

$\Downarrow$ $\qquad$ $\Downarrow$

6 $\qquad$ 11

K Means Clustering 2

Quiz

- Spring 2018 Midterm Q5
- Given data $\{5, 7, 10, 12\}$ and initial cluster centers $c_1 = 3, c_2 = 13$, what are the clusters in the next iteration?
- A: $\{5, 7\}$ and $\{10, 12\}$
- B: $\{5\}$ and $\{7, 10, 12\}$
- C: $\{5, 7, 10\}$ and $\{12\}$

$c_1 = 6$, $c_2 = 11$

K Means Clustering 3

Quiz

Q5

- Given data -2, 0, 10 and initial cluster centers $c_1 = -4, c_2 = 1$, what is the initial clusters?
- A: $\{\varnothing\}$ and -2, 0, 10
- B: -2 and $\{0, 10\}$
- C: -2, 0 and $\{10\}$
- D: -2, 0, 10 and $\{\varnothing\}$

Total Distortion 1

Quiz

Q6

- Given data -2, 0, 10 and initial cluster centers $c_1 = -4, c_2 = 1$, what is the ~~initial~~ total distortion?
- A: 0
- B: 2
- C: 12
- D: 13
- E: 15

$\{-2\}$, $\{0, 10\}$.

$2 + 1 + 9$

$c_1 = \boxed{-2}$ $\quad$ $c_2 = \boxed{5}$

$c_1 = -1 \leftarrow \{-2, 0\}$ $\qquad$ $\{10\} \rightarrow c_2 = 10$

final total distortion $\boxed{2}$

K Means Clustering 4

∞

$$\begin{bmatrix} \phi(x_1)^T\phi(x_1) & \phi(x_1)^T\phi(x_2) \\ & \phi(x_2)^T\phi(x_1) \end{bmatrix} \Rightarrow \text{Kernel}$$

$n \times n$

for 2 instances.

Quiz

- Given data -2, 0, 10 and initial cluster centers $c_1 = -4, c_2 = 1$, what are the clusters in the next iteration?
- A: $\{\emptyset\}$ and -2, 0, 10
- B: -2 and $\{0, 10\}$
- C: -2, 0 and $\{10\}$
- D: -2, 0, 10 and $\{\emptyset\}$

5 classes

generative, $Y = 0, 1, 2, 3, 4$

estimate $\Pr\{X, Y\} \xrightarrow{\text{compute}} \Pr\{Y \mid X\}$

$\Pr\{Y = 4\} = 1 - \Pr\{0\} - \Pr\{3\}$

$4 + 20 = 4$

$X_1 \mid Y=0,\ X_1 \mid Y=1 \ \ldots$

$2 \times 5 \times 2$

Naive Bayes

Total Distortion 2

$X_1 = 0, 1, 2$

Quiz

$Y = 0, 1, 2, 3, 4$

X_1, X_2

X_1, X_2

$2 \times 5 \times 2$

- Given data -2, 0, 10 and initial cluster centers $c_1 = -4, c_2 = 1$, what is the final total distortion?
- A: 0
- B: 2
- C: 12
- D: 13
- E: 15

K Means Clustering

Algorithm

- Input: instances: $\{x_i\}_{i=1}^n$, the number of clusters K, and a distance function d.
- Output: a list of clusters $C = C_1, C_2, ..., C_K$.
- Initialize $t = 0$.

$$c_k^{(0)} = K \text{ random points}$$

- Loop until $c^{(t)} = c^{(t-1)}$.

$$C_k^{(t-1)} = \left\{x : k = \arg \min_{k' \in 1,2,\ldots,K} d\left(x, c_k^{(t-1)}\right)\right\}$$

$$c_k^{(t)} = \frac{1}{\left|C_k^{(t-1)}\right|} \sum_{x \in C_k^{(t-1)}} x$$

Number of Clusters

Discussion

- There are a few ways to pick the number of clusters K.

1. K can be chosen using prior knowledge about X.
2. K can be the one that minimizes distortion? No, when $K = n$, distortion $= 0$.
3. K can be the one that minimizes distortion + regularizer.

$$K^{\star} = \arg\min_{k} \left(D_k + \lambda \cdot m \cdot k \cdot \log n \right)$$

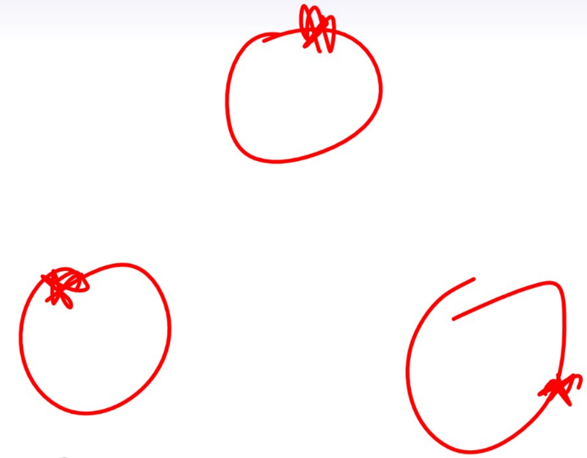
trade off # chosen

D

- λ is a fixed constant chosen arbitrarily.

Initial Clusters

Discussion



- There are a few ways to initialize the clusters.

1. K uniform random points in $\{x_i\}_{i=1}^{n}$. repeat many
2. 1 uniform random point in $\{x_i\}_{i=1}^{n}$ as $c_1^{(0)}$, then find the farthest point in $\{x_i\}_{i=1}^{n}$ from $c_1^{(0)}$ as $c_2^{(0)}$, and find the farthest point in $\{x_i\}_{i=1}^{n}$ from the closer of $c_1^{(0)}$ and $c_2^{(0)}$ as $c_3^{(0)}$, and repeat this K times.

try P4