CS540 Introduction to Artificial Intelligence

Lecture 15

Young Wu

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

July 18, 2022

Midterm

Admin

Q1

- The midterm is:
- A : Too Easy
- B : Easy
- C : Just right
- D : Hard
- E : Too Hard

Socrative

Room CS540E

Midterm Discussion

Admin

- Go over some new questions at the end of the lecture.
- Post the stats later in the week.
- If you are planning to take the make-up midterm, there is no need to notify me.
- Same format, join by Zoom, *Q6* questions still on the exam (with different randomization).
- You can start the exam and not submit it, but if you submit, your current grade will be replaced.

Unsupervised Learning

Motivation

image label features

- 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.

$y_i \approx f(x_i)$

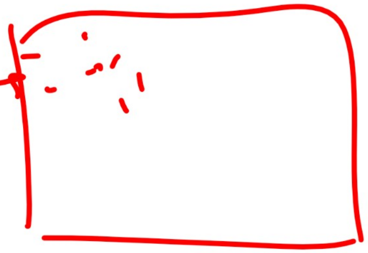
$\Pr\{y_i | x_i\} \leftarrow \Pr\{x_i | y_i\}$

Bayes Rule

0, 1, 2

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

images



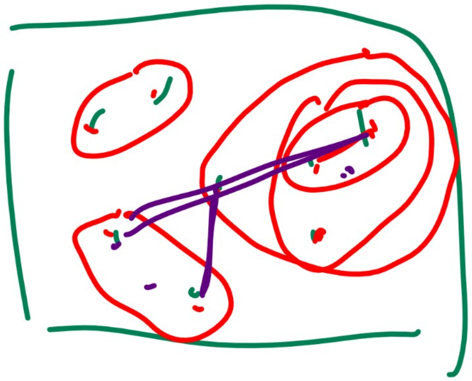
Unsupervised Learning Applications

Motivation

1. Google News
2. Google Photo
3. Image Segmentation
4. Text Processing

Hierarchical Clustering

Description



} group

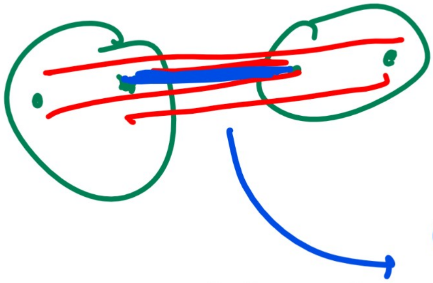
- Start with each instance as a cluster.
- Merge clusters that are closest to each other.
- Result in a binary tree with close clusters as children.

Hierarchical Clustering Diagram

Description

Single Linkage Distance

Definition



dist between clusters.

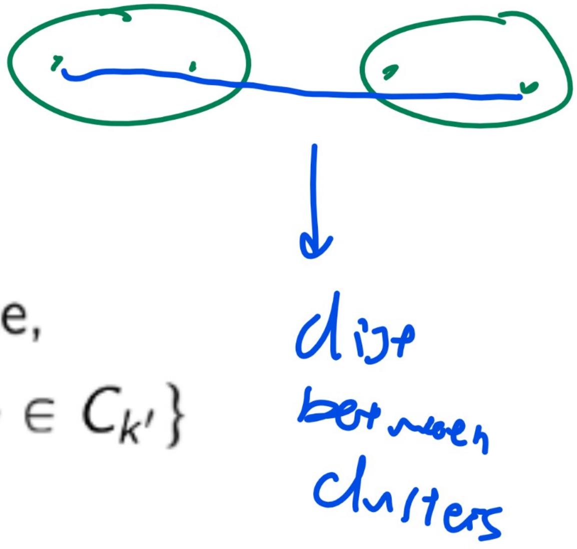
- Usually, the distance between two clusters is measured by the single-linkage distance.

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

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

Complete Linkage Distance

Definition



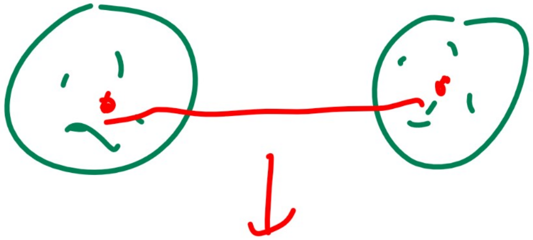
- Another measure is complete-linkage distance,

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

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

Average Linkage Distance Diagram

Definition



dist between clusters.

- Another measure is average-linkage distance.

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

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

Hierarchical Clustering 1

Quiz

- 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 and complete linkage?

single
dist A,B = 1
A C = 5
B C = 2

complete
dist A, B = 9
A, C = 11
B, C = 8

next step merge $\boxed{A,B}$, C

next merge $\boxed{B,C}$, A

Hierarchical Clustering 2

Quiz

- Given three clusters $A = \{0, 1\}$, $B = \{4, 6\}$, $C = \{8\}$. What is the next iteration of hierarchical clustering with Euclidean distance and complete linkage?

max dist

- A : Merge A and B.
- B : Merge A and C.
- [✓ C : Merge B and C.]
- D : I don't understand.

Q2

dist A, B = 6

A, C = 8

B, C = 4

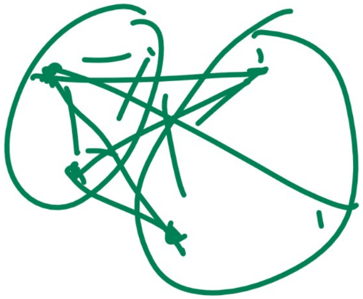
Hierarchical Clustering 3

Quiz

- Spring 2018 Midterm $Q5$
- Given three clusters $A = \{0, 1\}, B = \{4, 6\}, C = \{8\}$. 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.
- D : I don't understand.

Hierarchical Clustering 4

Quiz



- Given the distance between the clusters so far. Which pair of clusters will be merged using single 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
E	996	2037	1307	1059	0

merge CD

	AE	B	CD
AE		1075	1059
B			2687
CD			

A
B
CD
E

A	B	CD	E
	~~1075~~	~~2013~~	996
		2687	~~2037~~
			1059

merge AE

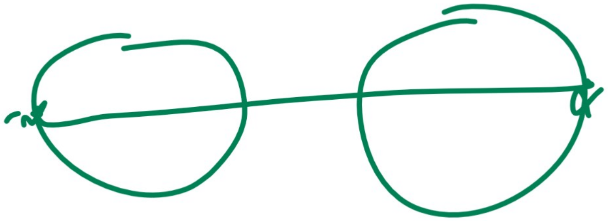
Hierarchical Clustering 4, Diagram

Quiz

–	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	1059	0

Hierarchical Clustering 5

Quiz



Q3

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

—	A	B	C	D
A	0	1075	2013	2054
B	1075	0	3272	2687
C	2013	3272	0	808
D	2054	2687	808	0

merge CD

dist(A, CD)

max d(A,C), d(A,D)

- E : I don't understand.

Q4

A

B

CD → C

merge A, B

	A	B	CD
A	0	1075	2054
B	1075	0	3272
CD	2054	3272	0

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.

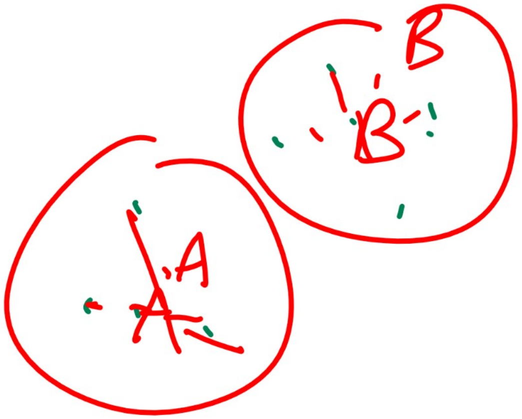
K Means Clustering

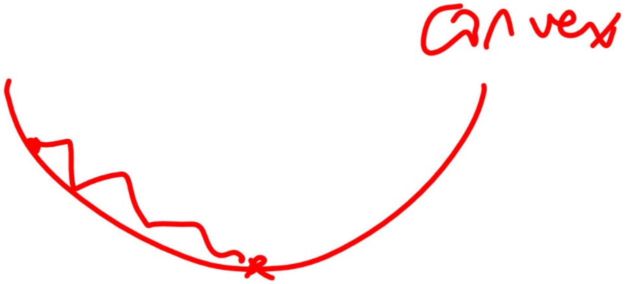
Description

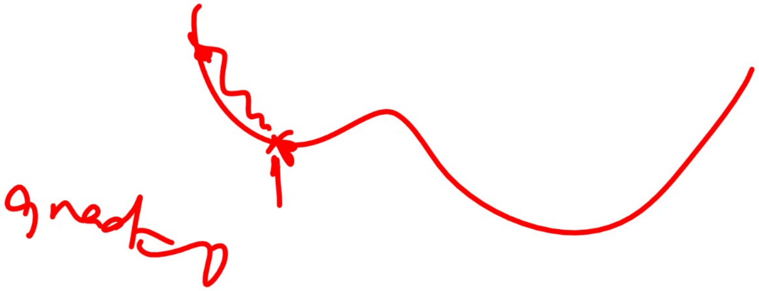
- This is not K Nearest Neighbor.
- Start with random cluster centers.
- Assign each point to its closest center.
- Update all cluster centers as the center of its points.

K Means Clustering Demo

Description







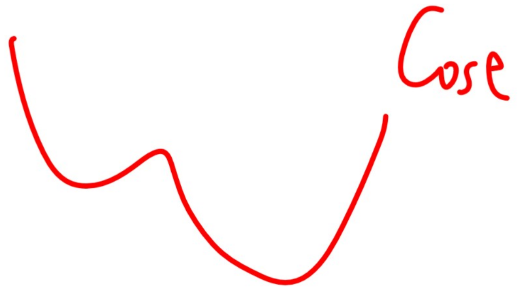
Distortion

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.

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

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



Objective Function Counterexample

Definition

Gradient Descent

Definition

- When d is the Euclidean distance. K Means algorithm is the gradient descent when distortion is the objective (cost) function.

$$\frac{\partial}{\partial c_k} \sum_{k=1}^{K} \sum_{x \in C_k} \|x - c_k\|_2^2 = 0$$

$$\Rightarrow -2 \sum_{x \in C_k} (x - c_k) = 0$$

$$\Rightarrow \boxed{c_k = \frac{1}{|C_k|} \sum_{x \in C_k} x}$$

$$w = w - \frac{\partial c}{\partial w}$$

K Means Clustering 1

Quiz

- Given data $x = \{-1, 0, 2\}$ and initial cluster centers $c_1 = 0, c_2 = 1$, what is the initial clusters and what is the initial total distortion (sum of squares without square root)? $\sqrt{3}$

	dist to C_1	dist to C_2	cluster	distortion
-1	1	2	C_1	1^2
0	0	1	C_1	1^2
2	2	1	C_2	1^2
				3

distortion

$(\frac{1}{2})^2 + (\frac{1}{2})^2 + 0^2 = \frac{1}{2}$

$\boxed{\{-1, 0\}}$ $\boxed{\{2\}}$

$c_1' = -\frac{1}{2}$

$c_2' = 2$

K Means Clustering 2

Quiz

$$\begin{pmatrix} c_{11} \\ c_{12} \\ c_{13} \end{pmatrix} \qquad \begin{pmatrix} x_{11} \\ x_{12} \\ x_{13} \end{pmatrix}$$

Q5

- Given data $x = \{-1, 0, 2\}$ and initial cluster centers $c_1 = 0, c_2 = 5$, what is the initial clusters?
- A : $\{\varnothing\}$ and $\{-1, 0, 2\}$
- B : $\{-1\}$ and $\{0, 2\}$
- C : $\{-1, 0\}$ and $\{2\}$
- [✓ D : $\{-1, 0, 2\}$ and $\{\varnothing\}$]
- E : I don't understand.

	$dist_1$	$dist_2$	?
-1			c_1
0			c_1
2	$\boxed{2}$	3	c_1

P4 → k-means → try diff initial cluster

Total Distortion 2

Quiz

$\frac{1}{7}$

Q6

- Given data $x = \{-1, 0, 2\}$ and initial cluster centers $c_1 = 0, c_2 = 5$, what is the initial total distortion (sum of squares without square root)?
- *A* : 2
- [✓ *B* : 5]
- *C* : 10
- *D* : 50
- *E* : *I* don't understand.

$1^2 + 0^2 + 2^2 = 5$

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.

W

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

~~$\lambda k \sqrt{m}$~~ n

- λ is a fixed constant chosen arbitrarily.

min distortion
vs min K.

Initial Clusters

Discussion

- There are a few ways to initialize the clusters.

1. K uniform random points in $\{x_i\}_{i=1}^{n}$.
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.

Gaussian Mixture Model

Discussion

- In K means, each instance belong to one cluster with certainty.
- One continuous version is called the Gaussian mixture model: each instance belongs to one of the clusters with a positive probability.
- The model can be trained using Expectation Maximization Algorithm (EM Algorithm).

back 7:15

Gaussian Mixture Model Demo

Discussion