CS540 Introduction to Artificial Intelligence
Lecture 15

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

July 21, 2021

Midterm

Admin

Q1

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

Midterm Discussion

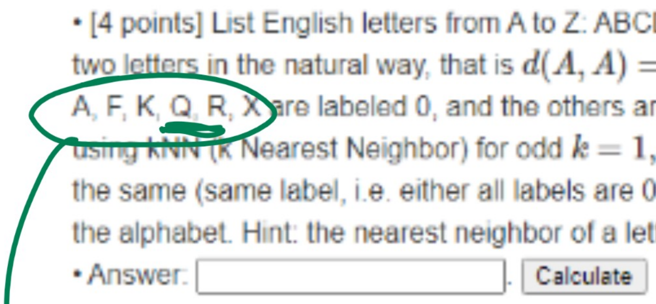
Admin

- Did not fix individual grades yet.
- Did not curve by dropping bad questions yet.
- Please report bugs on Piazza.
- Version A Part 1 average: 7.41, Part 2 average: 7.78
- Version B Part 1 average: 7.15, Part 2 average: 6.01

Midterm Questions? 1

Admin

need fix →



- [4 points] List English letters from A to Z: ABCDEFGHIJKLMNOPQRSTUVWXYZ. Define the distance between two letters in the natural way, that is $d(A, A) = 0$, $d(A, B) = 1$, $d(A, C) = 2$ and so on. Each letter has a label, A, F, K, Q, R, X are labeled 0, and the others are labeled 1. This is your training data. Now classify each letter using kNN (k Nearest Neighbor) for odd $k = 1, 3, 5, 7, \ldots$. What is the smallest k where all letters are classified the same (same label, i.e. either all labels are 0s or all labels are 1s). Break ties by preferring the earlier letters in the alphabet. Hint: the nearest neighbor of a letter is the letter itself.
- Answer: ______ . Calculate

$2n+1$

$\# = n$

3NN ⇒ 0

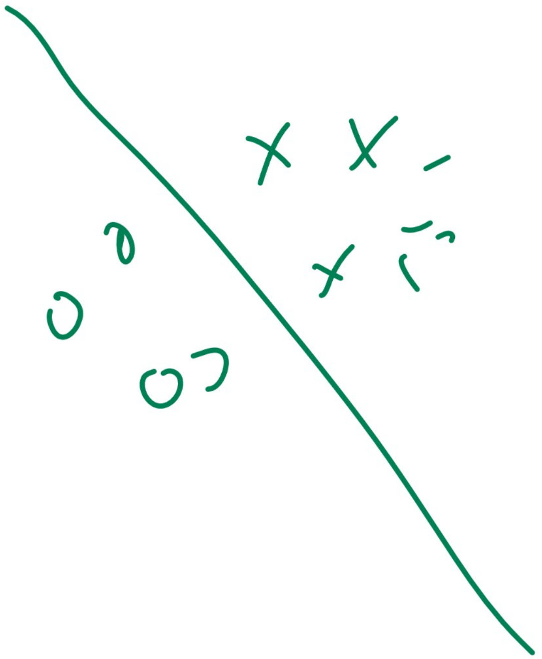
O P Q R S T

1 0 n

I am Great → training set not test.

Midterm Questions? 2

Admin



translate ✓

rotation ✓

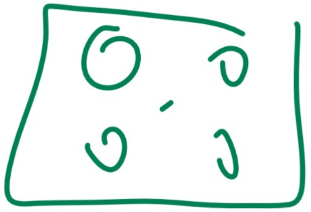
scale ?

$dae(n) = 0$

$mx + b \rightarrow b$

CNN

$\frac{10\times10}{3}$



Remind Me to Start Recording

Admin

- The messages you send in chat will be recorded: you can change your Zoom name now before I start recording.

Unsupervised Learning

Motivation

$\Rightarrow y \approx \hat{f}(x)$

$\hat{f}(x^{new}) \rightarrow$ guess y

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

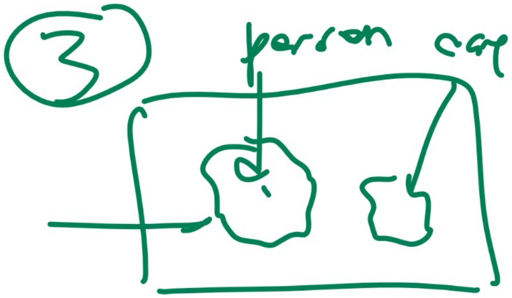
today

1. Clustering: separate instances into groups. labels
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.

PCA

eigen face

$1000 \longrightarrow$ ③



similar
color
Location

Hierarchical Clustering

Description

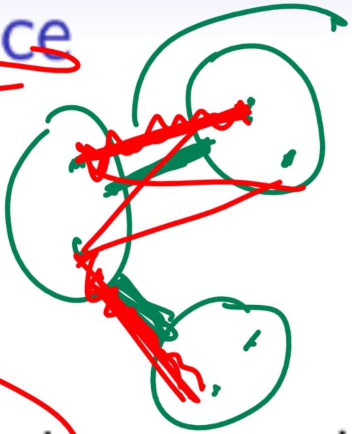
- Start with each instance as a cluster.
- Merge clusters that are closest to each other.
- Result in a binary tree with close clusters as children.

Hierarchical Clustering Diagram

Description

Single Linkage Distance

Definition



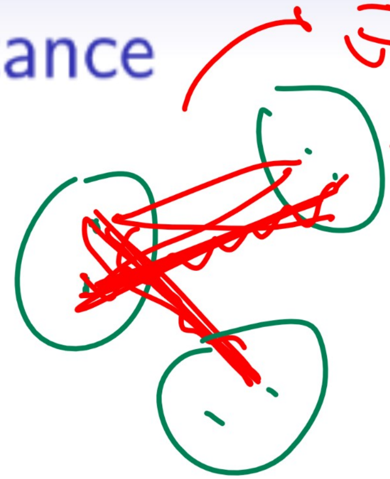
- Usually, the distance between two clusters is measured by the single-linkage distance.

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

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

Complete Linkage Distance

Definition



Euclidean

Manhattan

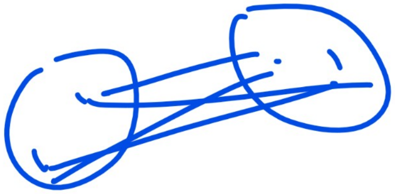
- Another measure is complete-linkage distance,

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

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

Average Linkage Distance Diagram

Definition



$\frac{\text{sum}}{4}$

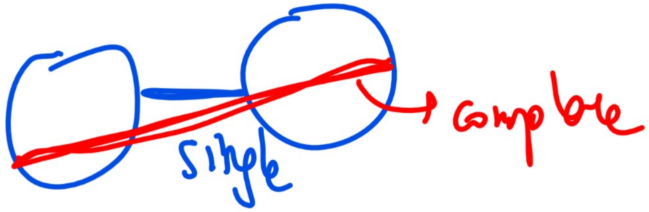
- Another measure is average-linkage distance.

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

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

Hierarchical Clustering 1

Quiz



- 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 and complete linkage?

A,B → 1

A,C → 5

B,C → 2

closest

{A,B}   C

A,B → 9

A,C → 11

B,C → 8

closest. A , {B,C}

Hierarchical Clustering 2

Quiz

Q2

- 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 complete linkage?
- A: Merge A and B.
- B: Merge A and C.
- [✓ C:] Merge B and C. ✓

AB → 6

AC → 8

BC → 4

Hierarchical Clustering 3

Quiz

Q3

3 2

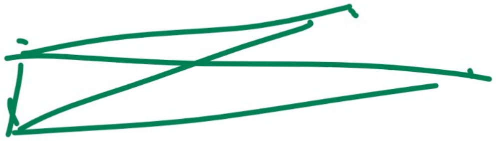
- Spring 2018 Midterm Q5
- Given three clusters $A = \{0, 1\}$, $B = \{4, 6\}$, $C = \{8\}$. What is the next iteration of hierarchical clustering with Euclidean distance and single linkage?

7

- A: Merge *A* and *B*.
- B: Merge *A* and *C*.
- [✓ C: Merge *B* and *C*.]

Hierarchical Clustering 4

Quiz

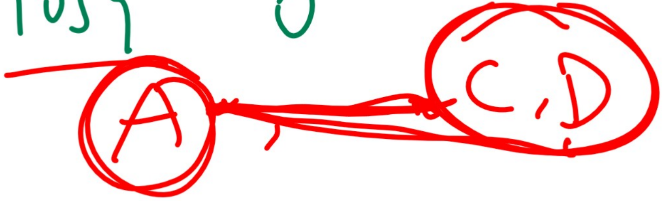


- Spring 2017 Midterm Q4
- Given the distance between the clusters so far. Which pair of clusters will be merged using single linkage.

points

–	A	B	C	D	E
A	0	1075	~~2013~~	~~2054~~	996
B	1075	0	~~3272~~	2687	2037
C	2013	~~3272~~	0	~~308~~	~~1307~~
D	~~2054~~	2687	~~308~~	0	1059
E	996	2037	~~1307~~	1059	0

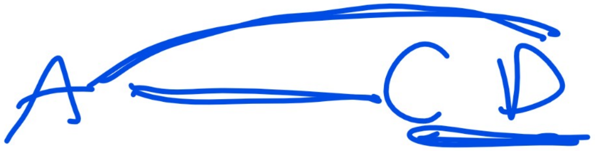
smallest



merge C and D.

Hierarchical Clustering 5

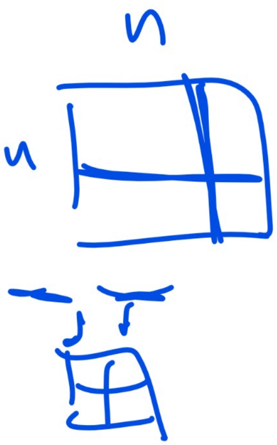
Quiz



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

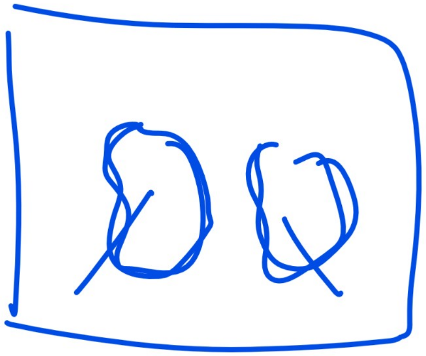
–	A	B	C	D	E
A	0	1075	~~2013~~	2054	996
B	1075	0	3272	~~2687~~	2037
C	~~2013~~	3272	0	~~808~~	1307
D	2054	~~2687~~	~~808~~	0	~~1059~~

1307 ~~1059~~



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.

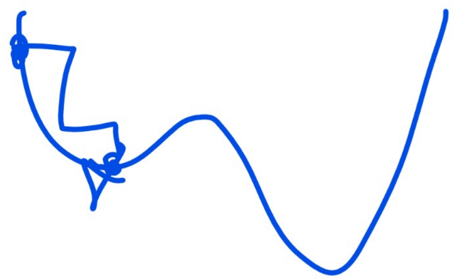
gradient descent

K means

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 Counterexample

Definition

K Means Clustering 1

Quiz

c_1 c_1 c_2

- Given data -1, 0, 2 and initial cluster centers $c_1 = 0, c_2 = 1$, 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 $\{\varnothing\}$

$$\begin{tikzcd}
-1 \arrow[r] \arrow[dr] & c_1 & 1 \\
 & c_2 & 2
\end{tikzcd}$$

Total Distortion 1

Quiz

C_1 C_2

- Given data $-1, 0, 2$ and initial cluster centers $c_1 = 0, c_2 = 1$, what is the initial total distortion (sum of squares without square root)?
- A: 0
- [✓ B: 2]
- C: 5
- D: 10
- E: 50

$1 + 0 + 1 = 2$

K Means Clustering 2

Quiz

Q4

- Given data -1, 0, 2 and initial cluster centers $c_1 = 0$, $c_2 = 5$, what is the initial clusters?
- A: $\{\emptyset\}$ and -1, 0, 2
- B: -1 and $\{0, 2\}$
- C: -1, 0 and $\{2\}$
- [✓ D: -1, 0, 2 and $\{\emptyset\}$]

→ total distortion.

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

Total Distortion 2

Quiz

Q5

- Given data -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: 0
- B: 2
- [✓ C: 5]
- D: 10
- E: 50

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.

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

  - 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 $\{x_i\}_{i=1}^n$.
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).

Gaussian Mixture Model Demo

Discussion