CSE 20
DISCRETE MATH

Fall 2020

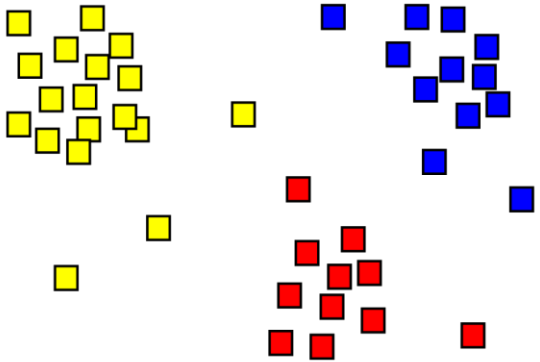
http://cseweb.ucsd.edu/classes/fa20/cse20-a/

Today's learning goals

- Define *clustering* as an application of relations and partitions
- Practice working with definitions using numerical and logical expressions

Clustering

Scenario: Good morning! You're a user experience engineer at Netflix. A product goal is to design customized home pages for groups of users who have similar interests. You task your team with designing an algorithm for producing a clustering of users based on their movie interests. Your team implements two algorithms that produce different clusterings. How do you decide which one to use? What feedback do you give the team in order to help them improve? Clearly, you will need to use math.



Clustering Criterion #1 – Partitioning

- Every user needs to be assigned to *some* group
- No user should be in *multiple* groups

Clustering Criterion #1 – Partitioning

Definition: A **partition** of a set A is a set of non-empty, disjoint subsets $A_1, A_2, \cdots, A_n$ such that $A_1 \cup A_2 \cup \cdots \cup A_n = A$.

Which of these is a partition of $\{1, 2, 3, 4, 5\}$?

A. $\{\{1, 2, 3\}, \{3, 4, 5\}\}$

B. $\{\emptyset, \{1, 2\}, \{3, 4, 5\}\}$

C. $\{\{1, 2\}, \{3, 4\}\}$

D. $\{\{1\}, \{2\}, \{3\}, \{4, 5\}\}$

E. None of the above

Clustering Criterion #1 – Partitioning

Definition: The set of movie ratings over n movies is R_n, where each element of R_n is a n-tuple with each entry in the tuple one of $\{-1, 0, 1\}$.

Which of these is a partition of R_2?

A. $\{\{(x_1, x_2) \in R_2 \mid x_1 = 0\}, \{(x_1, x_2) \in R_2 \mid x_2 = 0\}\}$

B. $\{\{(x_1, x_2) \in R_2 \mid \max(x_1, x_2) > 0\}, \{(x_1, x_2) \in R_2 \mid \max(x_1, x_2) < 0\}\}$

C. $\{\{(x_1, x_2) \in R_2 \mid x_1 = x_2\}, \{(x_1, x_2) \in R_2 \mid x_1 + 1 = x_2\}, \{(x_1, x_2) \in R_2 \mid x_1 - 1 = x_2\}\}$

D. $\{\{(x_1, x_2) \in R_2 \mid (|x_1 + x_2| < 1)\}, \{(x_1, x_2) \in R_2 \mid (|x_1 + x_2| > 1)\}, \{(x_1, x_2) \in R_2 \mid (|x_1 + x_2| = 1)\}\}$

E. None of the above

Clustering Criterion #2 – Similarity

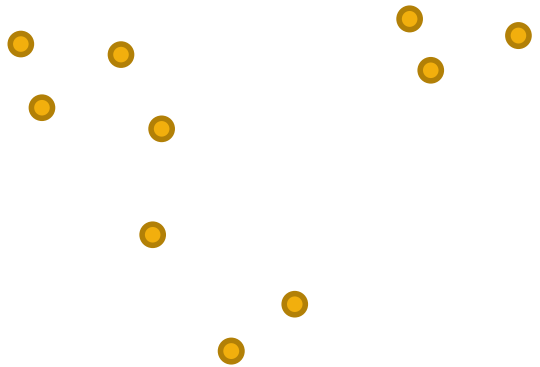
Idea: One way to measure similarity is with functions that measure **distance** between elements.

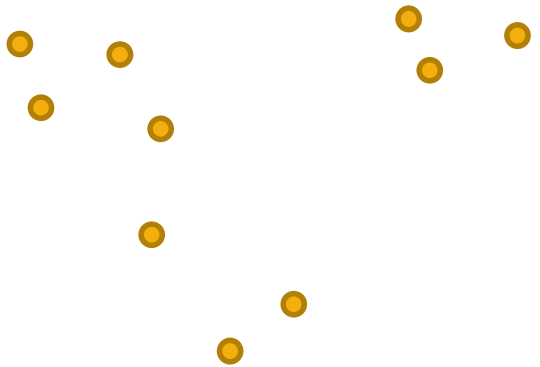
$$d(\ (x_1, \ldots, x_n), (y_1, \ldots, y_n)\ ) = \sum_{1 \leq i \leq n} |x_i - y_i|$$

Consider $x = (1, 0, 1, 0, 1)$, $y = (1, 1, 1, 0, 1)$, $z = (-1, -1, 0, 0, 0)$, $w = (0, 0, 0, 1, 0)$.

What is $d(x, y)$? $d(x, z)$? $d(z, w)$?

Clustering Criterion #2 – Similarity





Clustering Criterion #2 – Similarity

Definition: For a cluster of ratings $C = \{r_1, r_2, \cdots, r_n\} \subseteq U$, the **diameter** of the cluster is defined by:

$$diameter(C) = \max_{1 \leq i,j \leq n} (d(r_i, r_j))$$

Consider $x = (1, 0, 1, 0, 1)$, $y = (1, 1, 1, 0, 1)$, $z = (-1, -1, 0, 0, 0)$, $w = (0, 0, 0, 1, 0)$.

What is $diameter(\{x, y, z\})$? $diameter(\{x, y\})$? $diameter(\{x, z, w\})$?

Clustering Criterion #2 – Similarity

Assume we have a partition $C_1, \cdots, C_m$ of U and we can calculate diameter *on a single cluster.* We can aggregate over all the clusters by summing:

$$\sum_{1 \leq i \leq m} (diameter(C_i))$$

Clustering Criterion #2 – Similarity

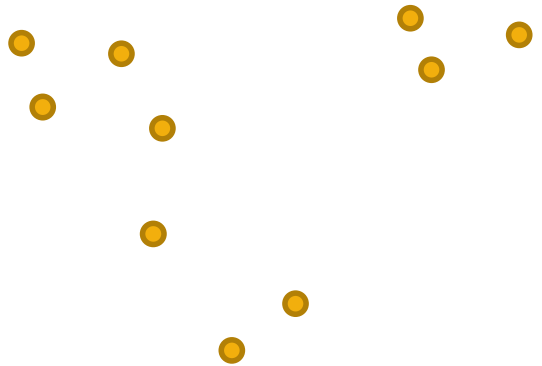
Is there a partition $C_1, \cdots, C_m$ of U that is easy to construct that makes $\sum_{1 \leq i \leq m} (diameter(C_i))$ be as small as possible?

A. Yes, choose $C_1 = U$

B. Yes, choose $C_1 = \{r_1\}, C_2 = \{r_2\}, \cdots, C_n = \{r_n\}$

C. Yes, but it isn't one of these

D. It depends on what "easy" means



Real consequences

Business

Apple Card algorithm sparks gender bias allegations against Goldman Sachs

Entrepreneur David Heinemeier Hansson says his credit limit was 20 times that of his wife, even though she has the higher credit score

Clustering Criterion #3 – Density

How can we express the idea of **many elements within a small area**? Key idea: "give credit" to small diameter clusters with many elements.

Which of these is a good definition of the **density** of a cluster $C = \{r_1, r_2, \cdots, r_n\}$?

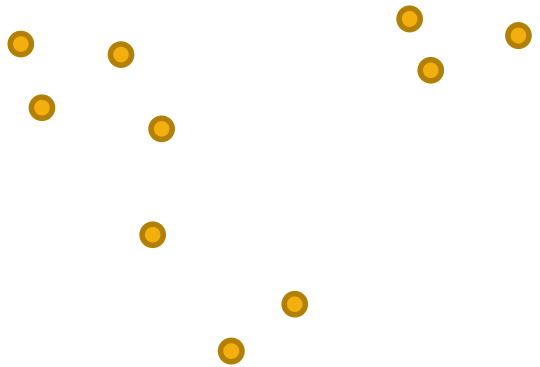
A. $n - diameter(C)$

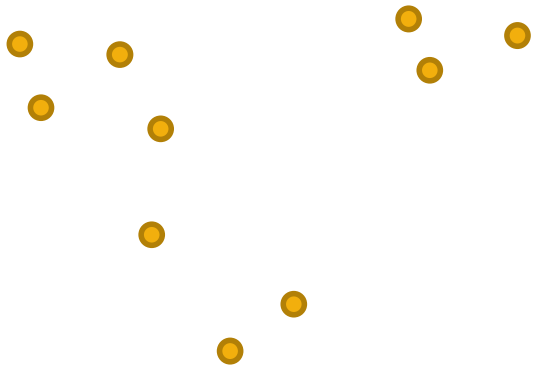
B. $\dfrac{n}{diameter(C)}$

C. $\dfrac{n}{1 + diameter(C)}$

D. $\dfrac{diameter(C)}{n}$

Clustering Criterion #3 – Density





Clustering Criterion #4 – Aggregating

Which method of aggregating across all clusters would give the team a useful metric to improve (m is the number of clusters)?

A. $\max_{1 \leq i \leq m} density(C_i)$, higher is better

B. $\sum_{1 \leq i \leq m} (density(C_i))$, higher is better

C. $\dfrac{\sum_{1 \leq i \leq m} (density(C_i))}{m}$, higher is better

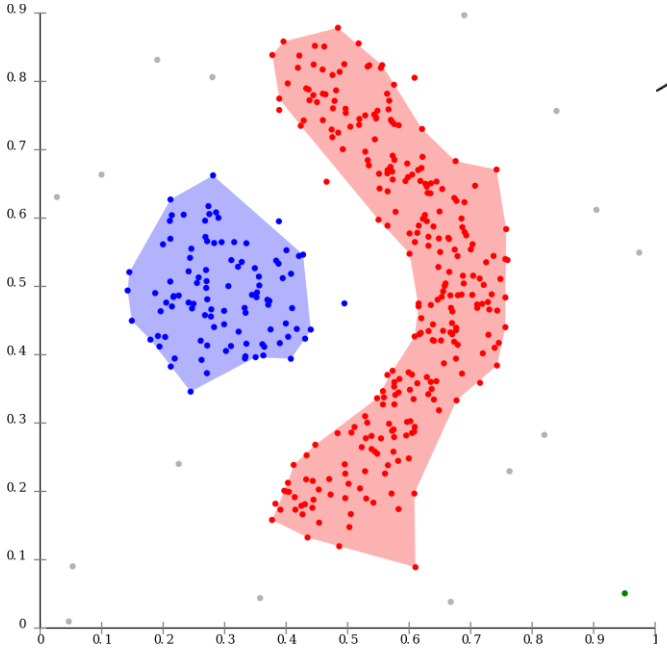
D. $\min_{1 \leq i \leq m} density(C_i)$, lower is better

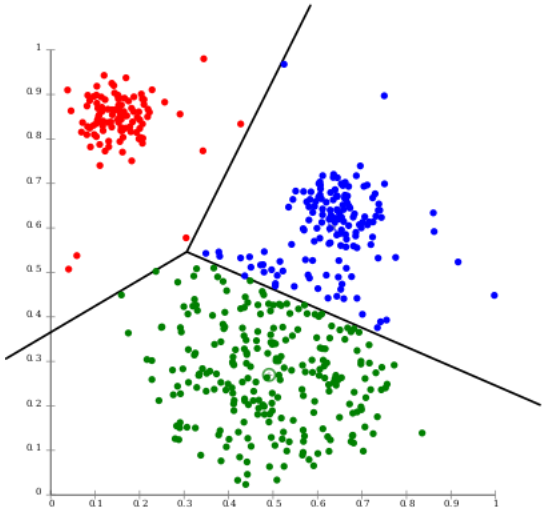
E. $\min_{1 \leq i \leq m} density(C_i)$, higher is better

Generating Clusters (Efficiently)

- CSE 150 series (AI & Machine Learning)
- https://en.wikipedia.org/wiki/K-means_clustering
- https://en.wikipedia.org/wiki/Hierarchical_clustering
- This is a big, active research area!

Clustering


0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1


1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1