

CPSC 330 Lecture 14: K-Means

Feature Engineering Recap

...

Clicker Question 13.1

Select all of the following statements which are TRUE.

Clicker Question 13.2

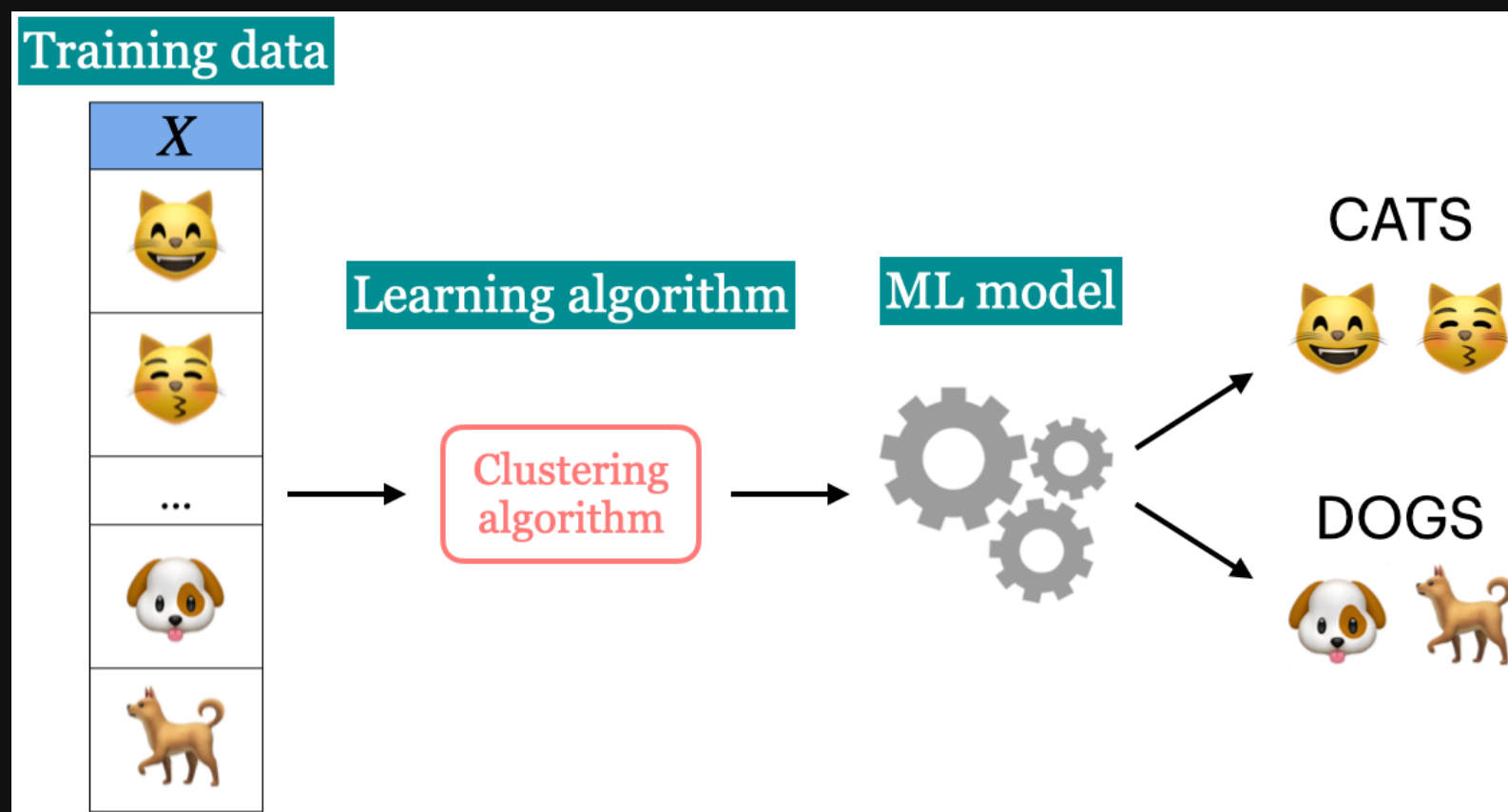
Select all of the following statements which are TRUE.

- a. You can carry out feature selection using linear models by pruning the features which have very small weights (i.e., coefficients less than a threshold).
- b. The order of features removed given by `rfe.ranking_` is the same as the order of original feature importances given by the model.

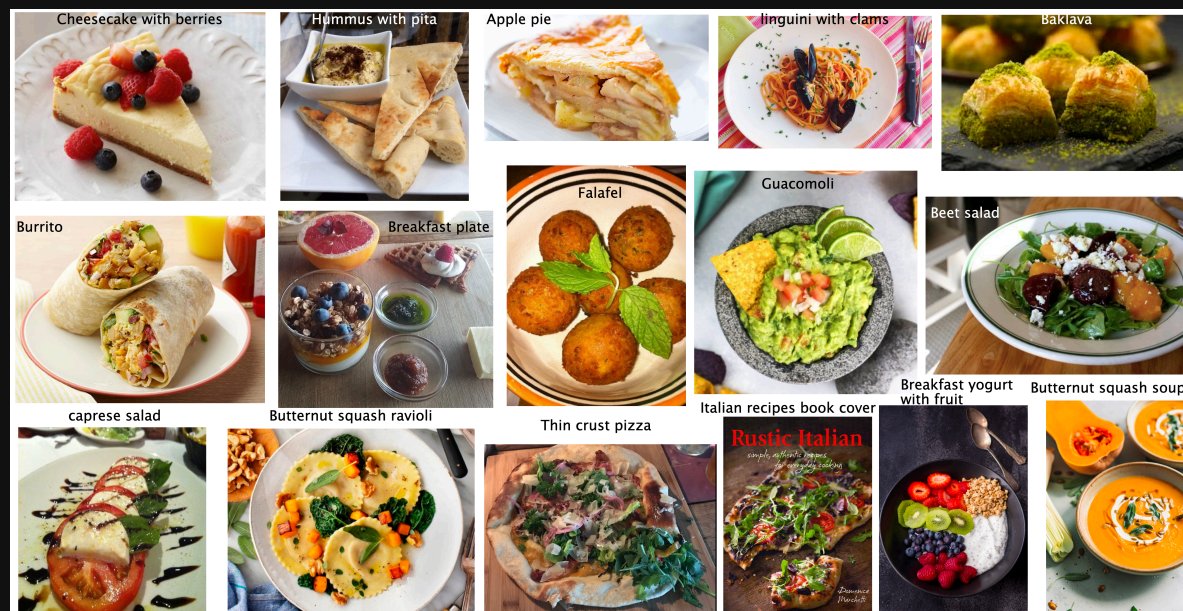
Supervised learning

Unsupervised learning

- Training data consists of observations (X) without any corresponding targets.
- Unsupervised learning could be used to group similar things together in X or to find underlying structure in the data.



Clustering activity



- Categorize the food items in the image and write your categories. Do you think there is one correct way to cluster these images? Why or why not?
- If you want to build a machine learning model to cluster such images how would you represent such images?
- Write your answers here:
https://docs.google.com/document/d/12GXA9Efi_19WiRnCi8FiOmdpxM-I_wq2nz97LjvXILg/edit?usp=sharing

The “perfect” spaghetti sauce

Suppose you are a hypothetical spaghetti sauce company and you’re asked to create the “perfect” spaghetti sauce which makes all your customers happy. The truth is humans are diverse and there is no “perfect” spaghetti sauce. There are “perfect” spaghetti sauces that cater to different tastes!

The “perfect” spaghetti sauce

Howard Moskowitz found out that Americans fall into one of the following three categories:

- people who like their spaghetti sauce **plain**
- people who like their spaghetti sauce **spicy**
- people who like their spaghetti sauce **extra chunky**



Reference: Malcolm Gladwell's Ted talk

The “perfect” spaghetti sauce

- If one “perfect” authentic sauce satisfies 60% of the people on average, creating several tailored sauce clusters could increase average happiness to between 75% to 78%.
- Can we apply this concept of clustering and tailoring solutions to specific groups in machine learning?

K-Means

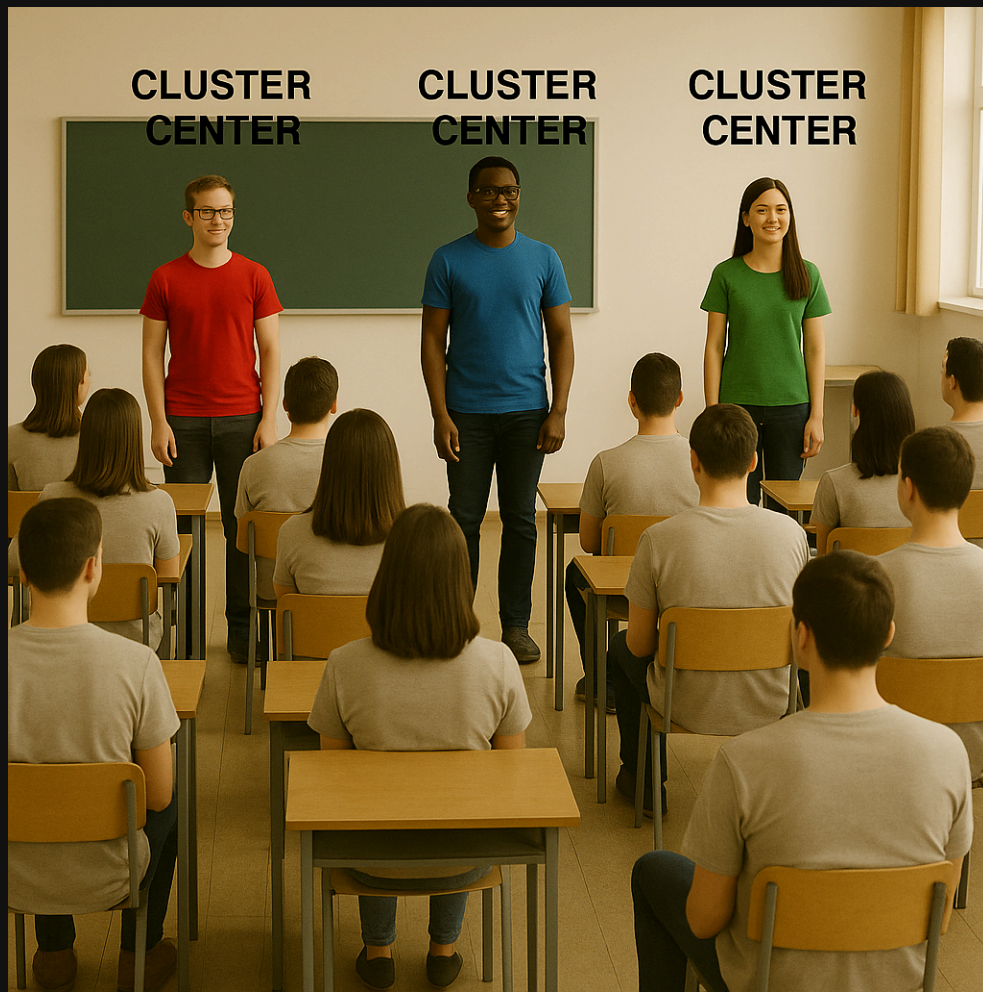
K-Means clustering algorithm

1. Select K initial centroids.
2. Assign each data point to the nearest centroid.
3. Recalculate centroids based on assigned points.
4. Repeat until centroids stabilize or reach a maximum number of iterations.

K-Means activity

Initialization

In K-Means, we start with random initial cluster centers.
Who would like to be cluster centers?

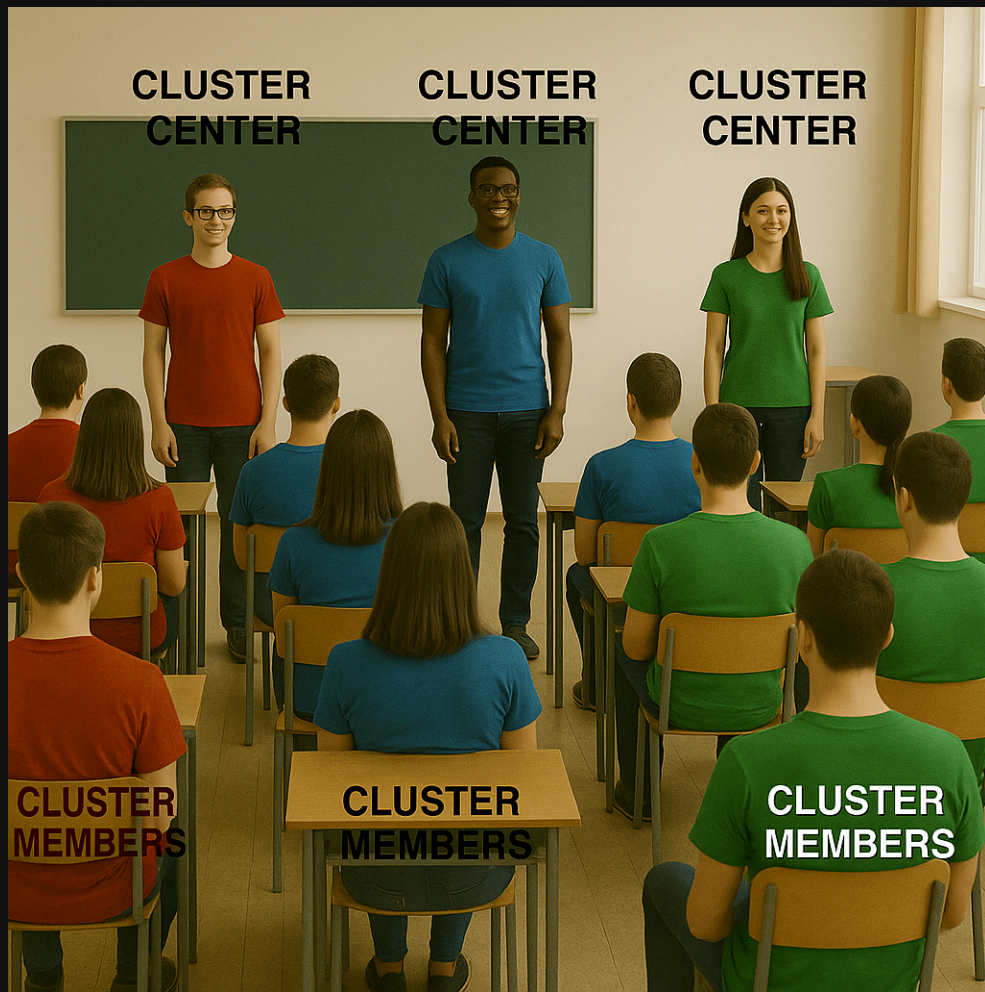


Let's pretend you are data points scattered across a 2D feature space.

Source: Image created by ChatGPT 5.0

Cluster assignment

Which cluster do you belong to?



Look around and decide which cluster center is closest to you, just by eyeballing distance in the room.

Source: Image created by ChatGPT 5.0

Recompute cluster centers

- Now each cluster needs to find its new mean location. Let's compute new centroids!
- For each cluster, identify the "average" position of your members. For example, pick someone roughly in the middle or calculate the average row/column number.
- Move your old "cluster center" to these new locations.

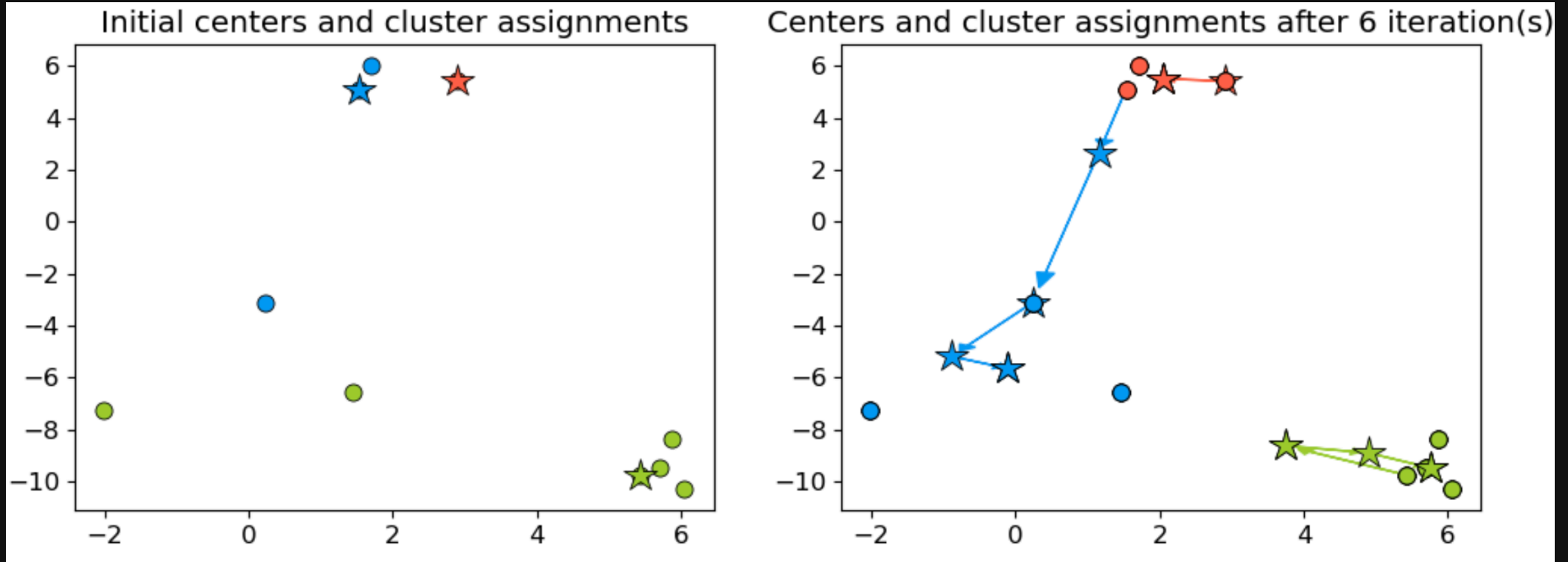
Repeat cluster assignment

Now that cluster centers have moved, you may need to switch clusters! Check again: which center are you closest to now?

Discussion questions

- Did anyone switch clusters in the second round?
- When do you think K-Means stops?
- Could we end up in a different final clustering if we started with different centers?

K-Means example



K-Means pros and cons

- **Advantages:**
 - Simple and efficient for large datasets.
 - Works well with spherical clusters.
- **Limitations:**
 - Needs pre-defined k .
 - Sensitive to outliers and initial centroid placement.

Clicker Question 14.1

Select all of the following statements which are **True**

- a. K-Means algorithm always converges to the same solution.
- b. K in K-Means should always be \leq # of features.
- c. In K-Means, it makes sense to have $K \leq$ # of examples.
- d. In K-Means, in some iterations some points may be left unassigned.

Clicker Question 14.2

Select all of the following statements which are **True**

- a. K-Means is sensitive to initialization and the solution may change depending upon the initialization.
- b. K-means terminates when the number of clusters does not increase between iterations.
- c. K-means terminates when the centroid locations do not change between iterations.
- d. K-Means is guaranteed to find the optimal solution.

Choosing k (the number of clusters)

How many clusters?

- Clustering is not about perfect scores, but **insight**.
- Typically we do not want too few or way too many clusters.
- There's no single "**correct**" value of k – only *useful* ones that make sense for your problem.
- Some **preferred** values may exist (e.g., based on domain knowledge).
- The goal is to find **interpretable, meaningful** clusters. So pick k that helps tell a **useful story** about your data.
- We typically go with **quantitative** and **qualitative** approaches.

Quantitative methods

- **The Elbow plot**
 - Plot inertia vs. K
 - Look for the “elbow” — diminishing returns
- **Silhouette score**
 - Measures cluster cohesion and separation
 - Higher score: better-defined clusters

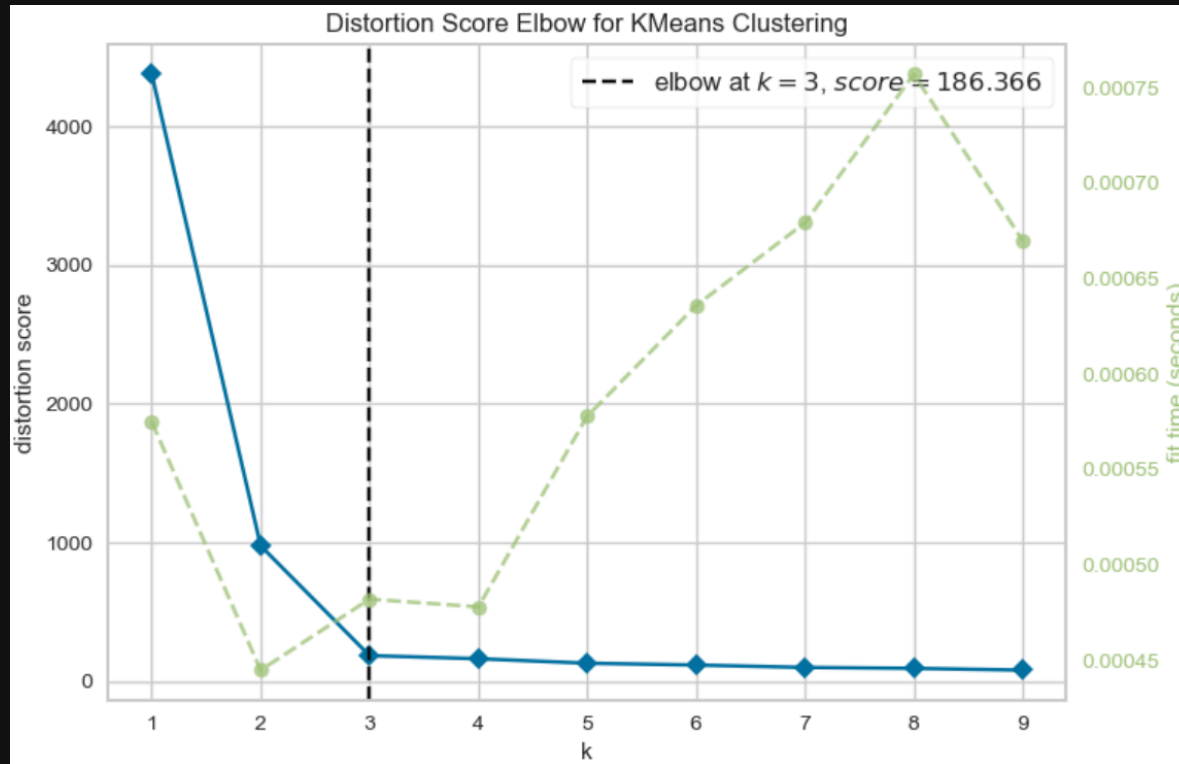
Qualitative methods

- **Manually inspect** clusters:
 - Do they make sense?
 - Are they interpretable and actionable?
- The ultimate goal: **Human-understandable themes** for real-world decisions.
- There's *no escape* from manual interpretation!

The Elbow method

- **Purpose:** Identify the optimal number of clusters k .
- **How it Works:**
 - Plot intra-cluster distances for different values of k .
 - Look for the “elbow” point where the intra-cluster reduction slows.
- **Interpretation:**
 - The point of diminishing returns suggests a good k .

The Elbow method example

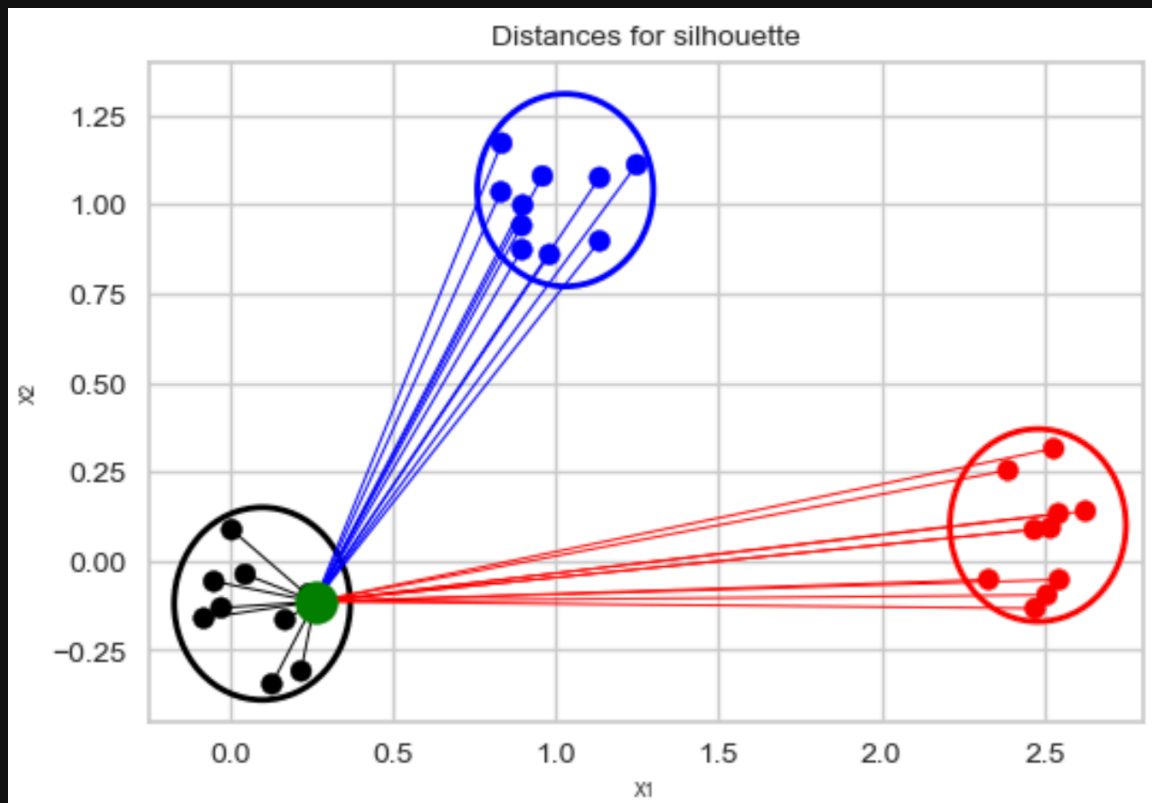


What would be the intracluster distance when

- $k = 1$
- $k = \#$ number of data points?

The Silhouette method

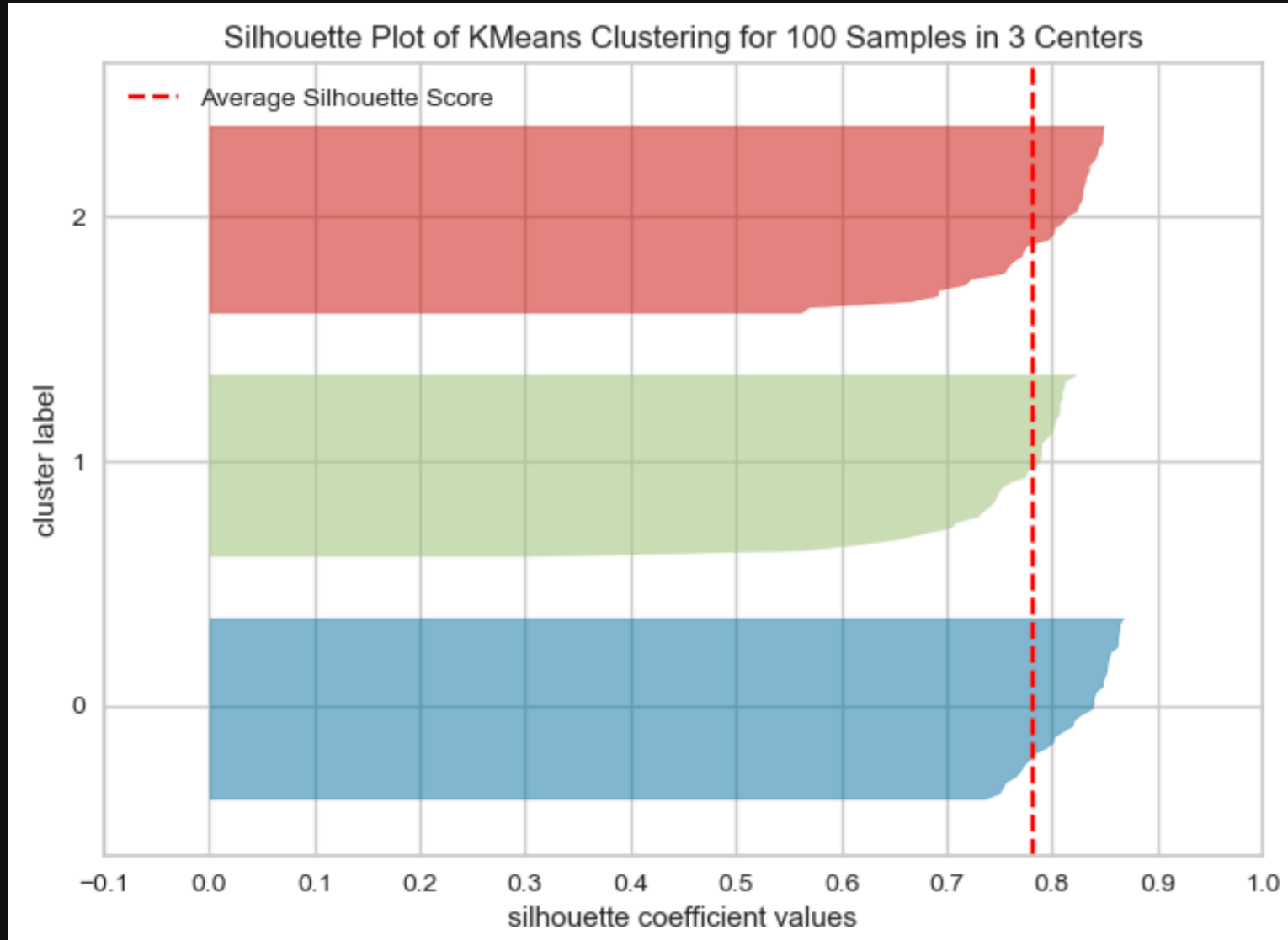
- **Silhouette Score:** Measures how well data points fit within their cluster.
 - $s(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))}$
 - $a(i)$: Mean distance to other points in the same cluster.
 - $b(i)$: Mean distance to points in the nearest neighboring cluster.



The Silhouette method

- **Range:** -1 to 1
 - **1:** Perfect clustering.
 - **0:** Overlapping clusters.
 - **Negative:** Poor clustering.
- Higher average silhouette score indicates “better” clustering.

Silhouette plot example



Clicker Question 14.3

Select all of the following statements which are **True**

- a. If you train K-Means with `n_clusters`= the number of examples, the inertia value will be 0.
- b. The elbow plot shows the tradeoff between within cluster distance and the number of clusters.
- c. Unlike the Elbow method, the Silhouette method is not dependent on the notion of cluster centers.
- d. The elbow plot is not a reliable method to obtain the optimal number of clusters in all cases.
- e. The Silhouette scores ranges between -1 and 1 where higher scores indicates better cluster assignments.

Group Work: Class Demo & Live Coding

Super cool Demo!

For this demo, each student should [click this link](#) to create a new repo in their accounts, then clone that repo locally to follow along with the demo from today.

All credit to Dr. Varada Kolhatkar for putting this together!