

The Grouping / Clustering

of Datapoints in a

Dataset

---

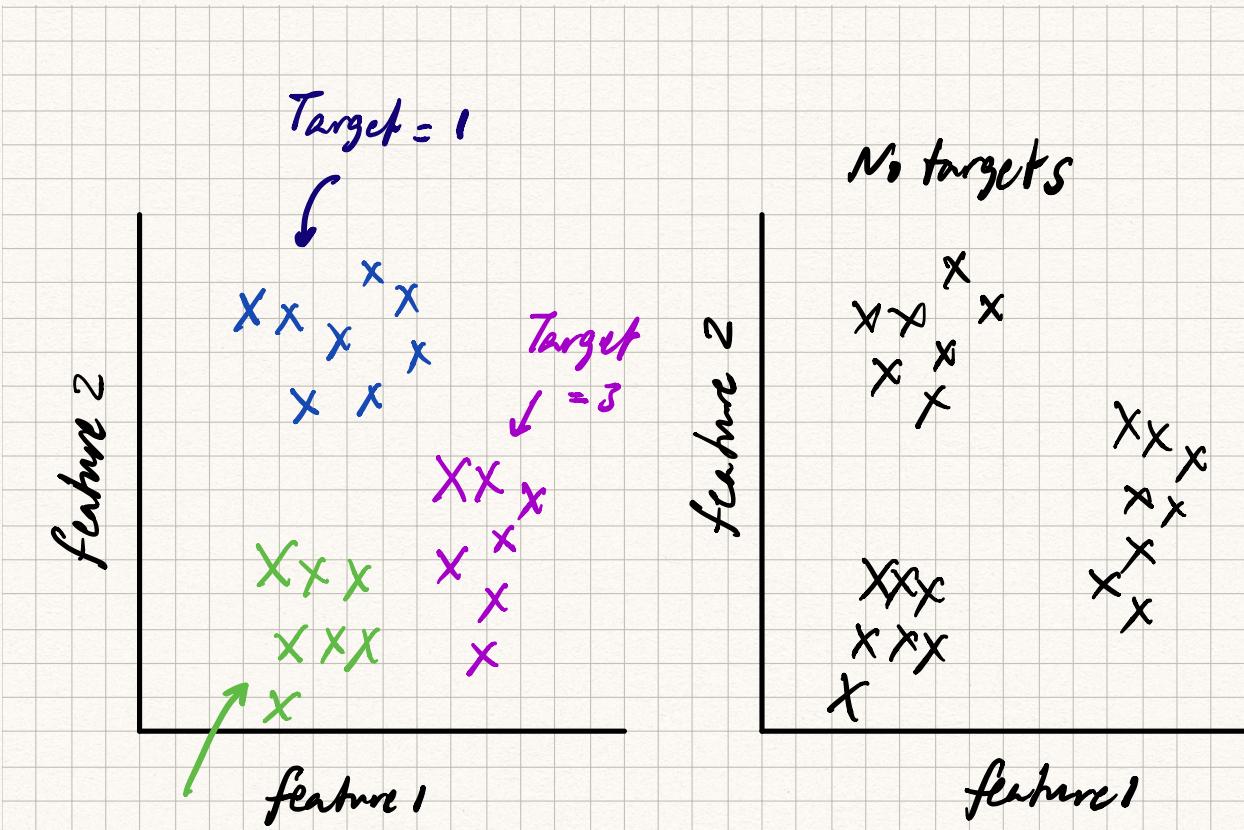
Clustering &

Unsupervised Learning

A number of business problems  
are problems that can be solved  
by finding the right clustering  
of the data.

- Which customers should receive  
this promotion?
- How many S/M/L/XL  
t-shirts should I make to  
sell at the food fair?

• • •



target = 2



f <sub>1</sub>	f <sub>2</sub>	Target
		1
		3
		2

↑  
Target Column

f <sub>1</sub>	f <sub>2</sub>

This dataset  
has no target  
column.

## Types of Machine Learning Datasets

- Supervised (targets present)
- Unsupervised (no targets, just features)
- Semi-supervised (some rows have targets and some rows don't)

What can we do with unsupervised datasets?

Is there some structure in the  
dataset?

Do datapoints group/cluster together?

OR

Are there datapoints that have  
similar values for each of the  
features?



For any 2 rows in the dataset, how  
similar are the rows?

## Features

	$x_1$	$x_2$	$x_3$	$x_4$	
Row 1	2	6	5	3	?
Row 2	1	0	6	2	How similar are these rows?

Similarity ~ How close together  
are the datapoints?

What's the distance

between them?

Distance between row 1 and row 2:

$$\sqrt{(2-1)^2 + (6-0)^2 + (5-6)^2 + (3-2)^2}$$

Euclidean Distance

	Row #		
Row 1	$x_1^{(1)}$	$x_2^{(1)}$	$x_3^{(1)}$
Row 2	$x_1^{(2)}$	$x_2^{(2)}$	$x_3^{(2)}$
			Column #

General form of distance (one family of it)

$$\left[ |x_1^{(1)} - x_1^{(2)}|^k + |x_2^{(1)} - x_2^{(2)}|^k + |x_3^{(1)} - x_3^{(2)}|^k \right]^{1/k}$$

↑  
Absolute Value

$k=1 \rightarrow$  Manhattan distance

$k=2 \rightarrow$  Euclidean distance

$k > 2 \rightarrow$  Minkowski distance of  
Order  $k$

We'll  
use  
this

$k = \infty \rightarrow$  Chebyshev distance

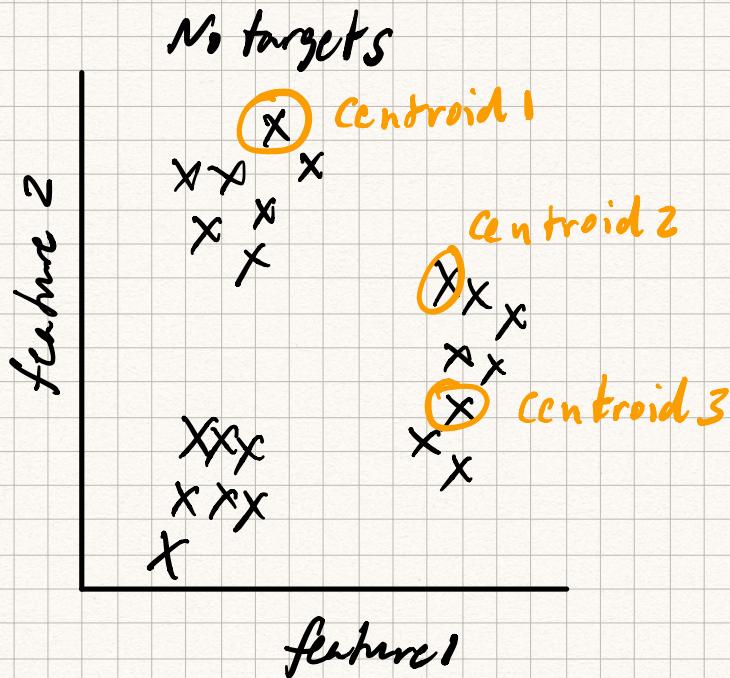
Once we can calculate distance, we can calculate similarity. The greater the distance, the less similar two rows are.

$$\text{Similarity} = \frac{1}{\text{Distance}}$$

Nice trick!

To measure something, measure its opposite and take the reciprocal.

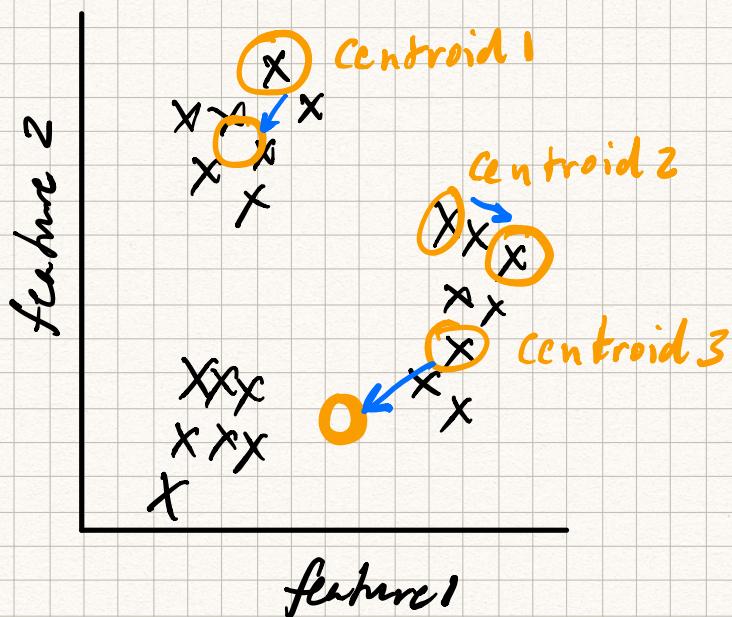
# The K-Means Algorithm



can be any  
reasonable #

- 1) Randomly choose 3 points  
in the dataset.  $\textcircled{x}$
- 2) For each point in the dataset, find  
its distance to each  $\textcircled{x}$
- 3) Assign the point to the closest  
centroid.

4) Move each centroid to the "middle" of the datapoints associated w/ the centroid.



5) Repeat 2, 3, and 4 until

- The centroids don't move any more (or move very little)  
OR
- a certain # of repetitions are completed.

Let's see what this looks  
like in Orange . . .

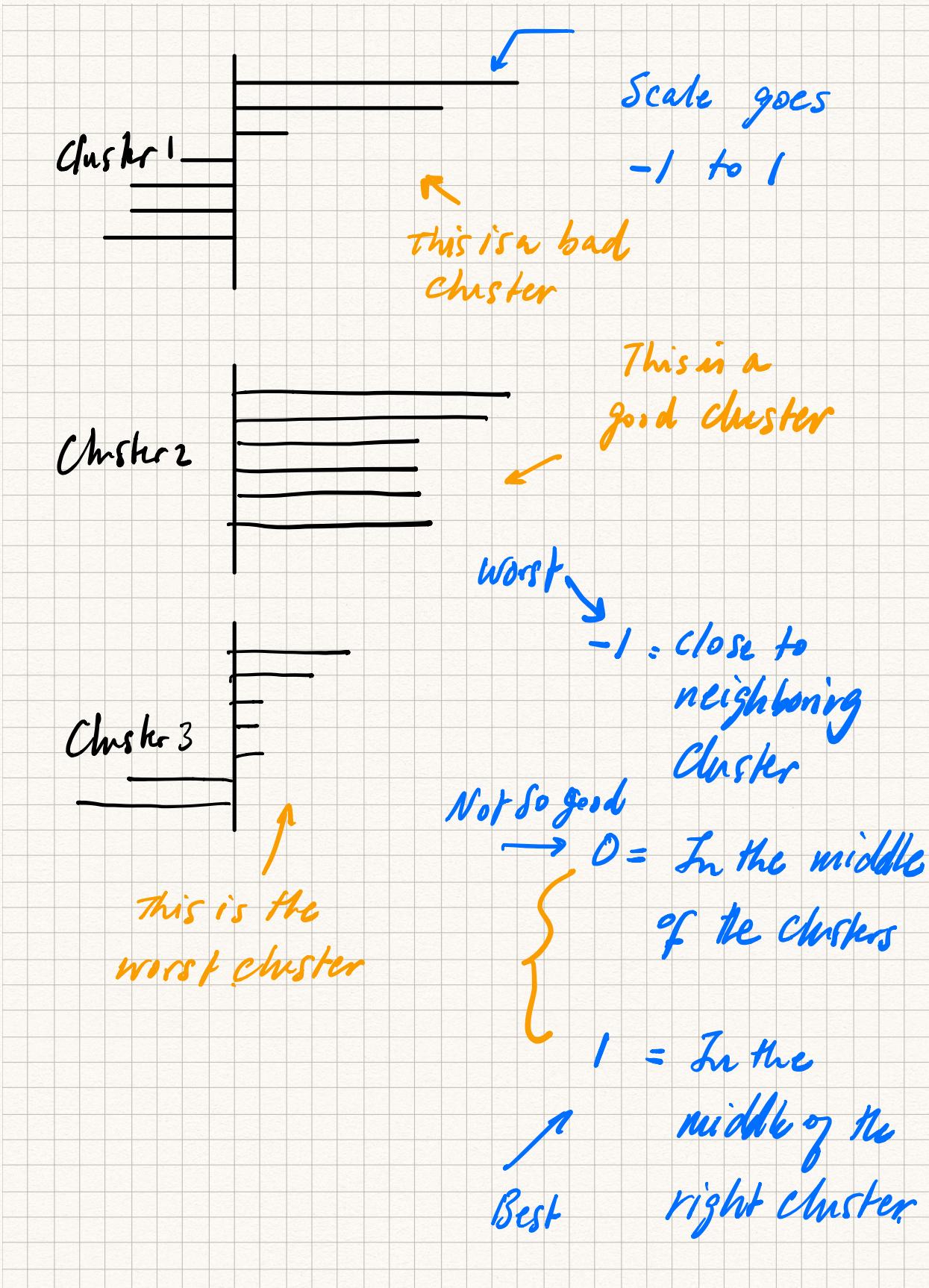
---

## Silhouette Scores

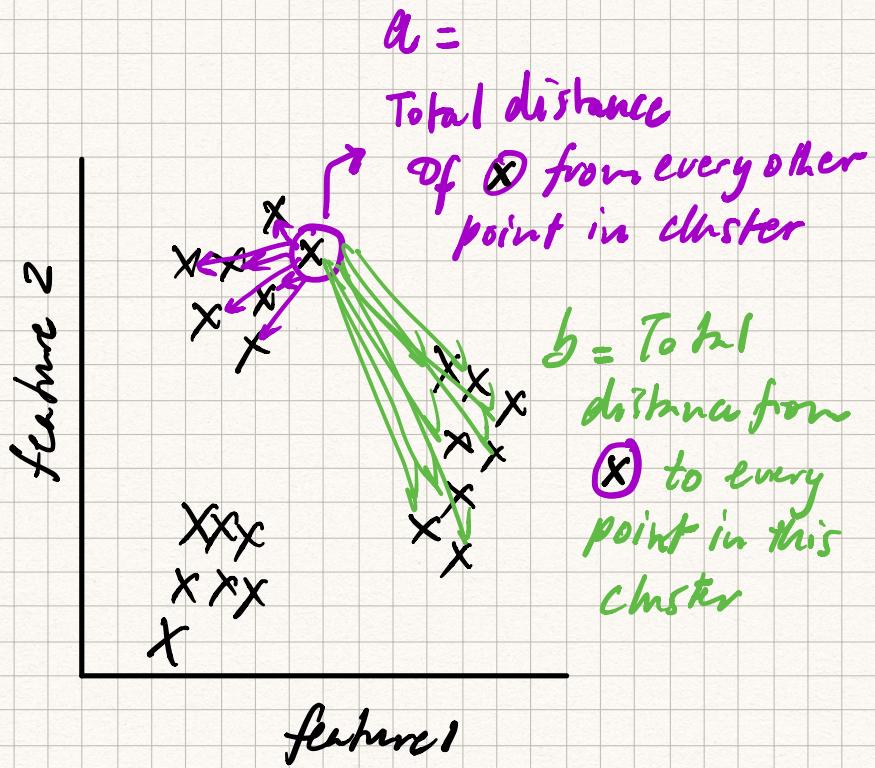
---

Measure the quality of each  
cluster.

Review Orange YouTube training  
videos - 11, 12, 13.



## Calculation of Silhouette Scores



$$\text{Silhouette Score} = \frac{b - a}{\max(a, b)}$$

Convince yourself that this  
is always between -1 and 1

How is clustering a machine learning problem?

- What's being optimized?.
- what are the parameters being learned?
- what are the hyperparameters?

Let's answer these questions in reverse order.

- Hyperparameters = # of clusters/  
centroids
- Parameters being learned =  
positions of the centroids
- What's optimized =  
Sum of distances between  
each centroid and all  
the points that belong to  
it.