The Grouping / Clustering of Datapoints in a Dataset

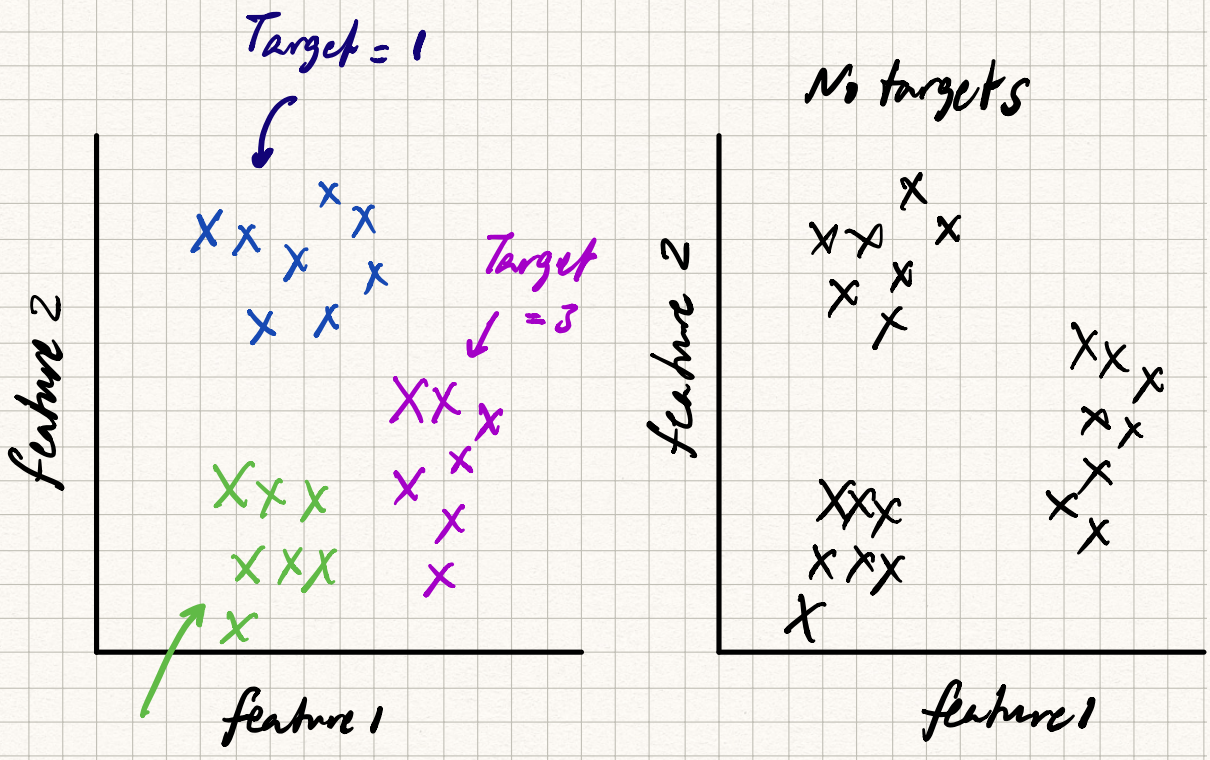
Clustering & Unsupervised Learning

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 tea-shirts should I make to sell at the food fair?

...



f1	f2	Target
		1
		3
		2

Target Column

f1	f2

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?

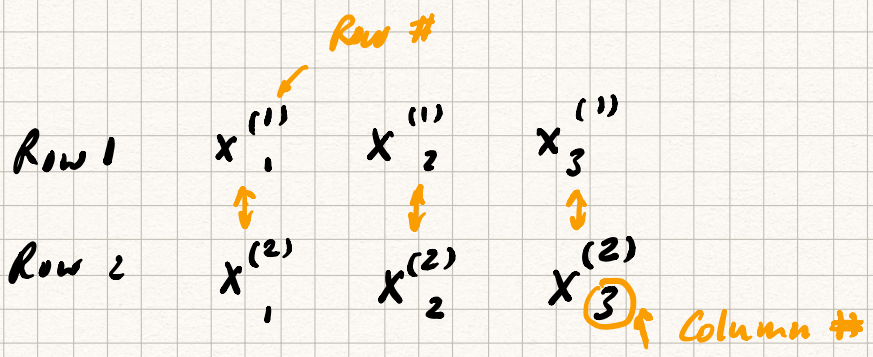
Similarity ~ How close together are the datapoints?

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



General form of distance (one family of it)

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

Absolute Value

$k = 1 \rightarrow$ Manhattan distance

$k = 2 \rightarrow$ Euclidean distance

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

We'll use this

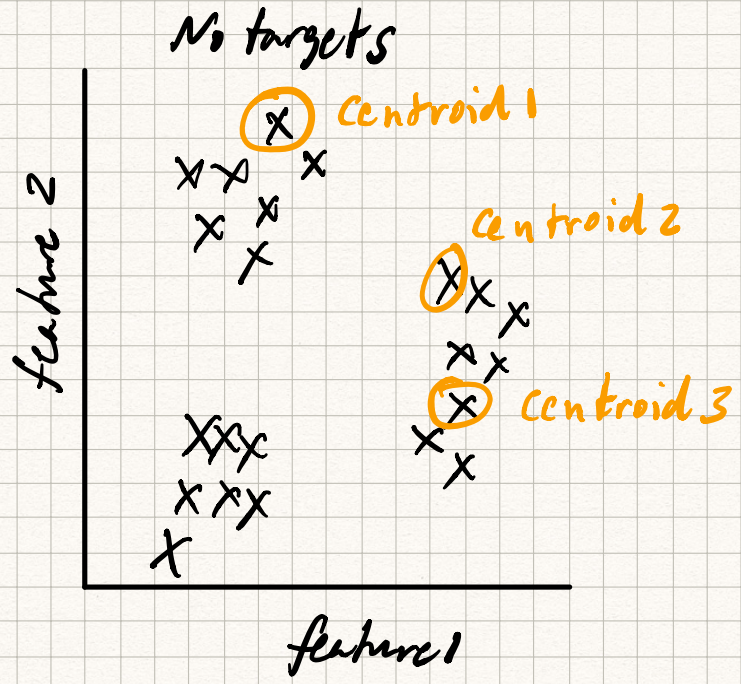
$k = \infty \rightarrow$ Chebyshev distance

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



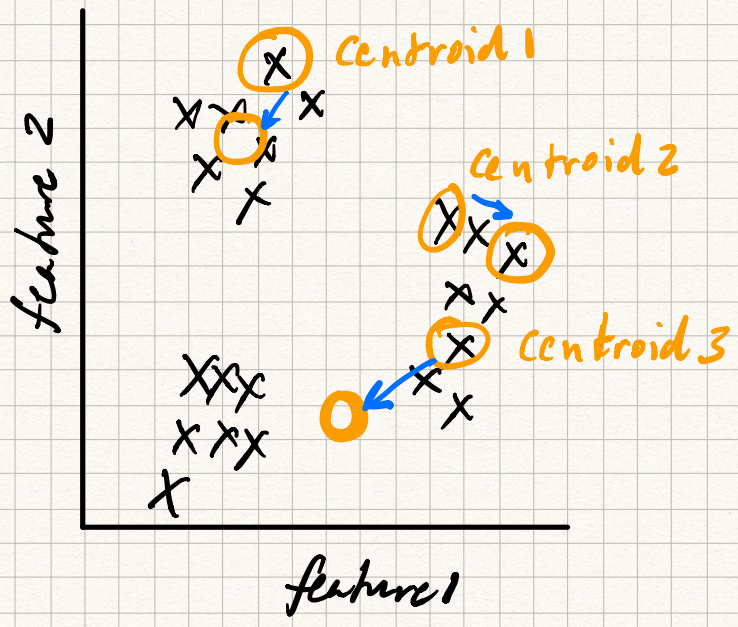
Can be any reasonable #

1) Randomly choose 3 points in the dataset. (X)

2) For each point in the dataset, find its distance to each (X)

3) Assign the point to the <u>closest</u> 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

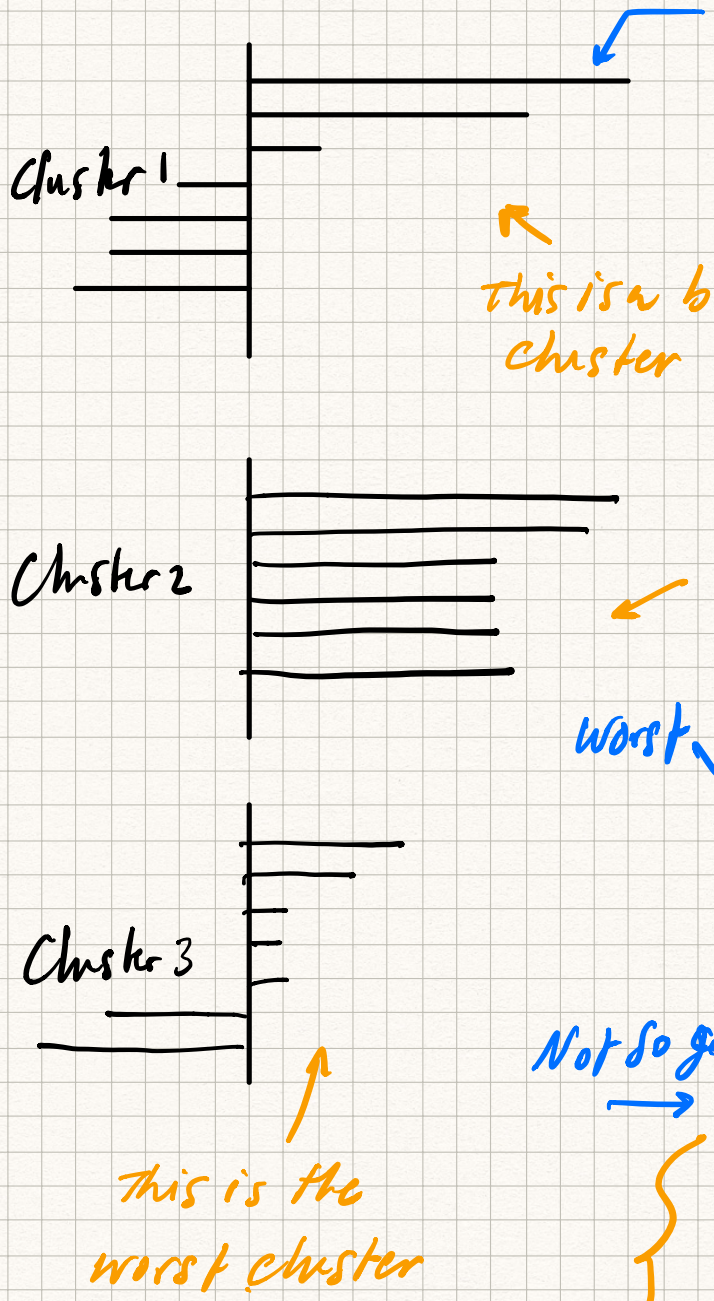
- a certain # of repetitions are completed.

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.



Scale goes -1 to 1

This is a bad cluster

This is a good cluster

This is the worst cluster

Worst

-1 : close to neighboring cluster

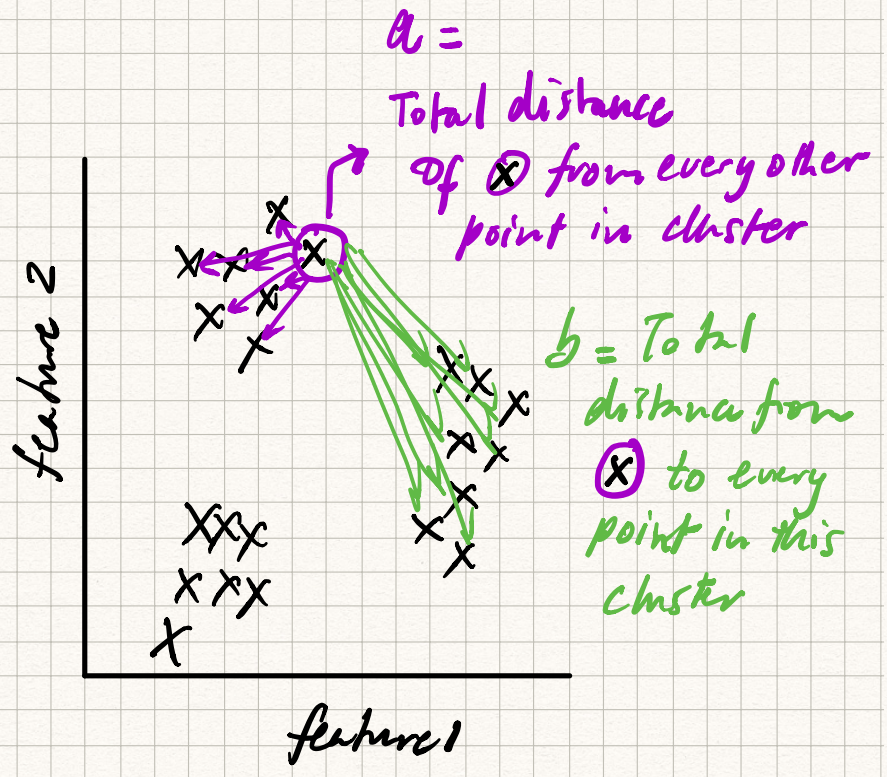
Not so good

0 = In the middle of the clusters

1 = In the middle of the right cluster

Best

Calculation of Silhouette Scores



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

Convince yourself that this is always between -1 and 1

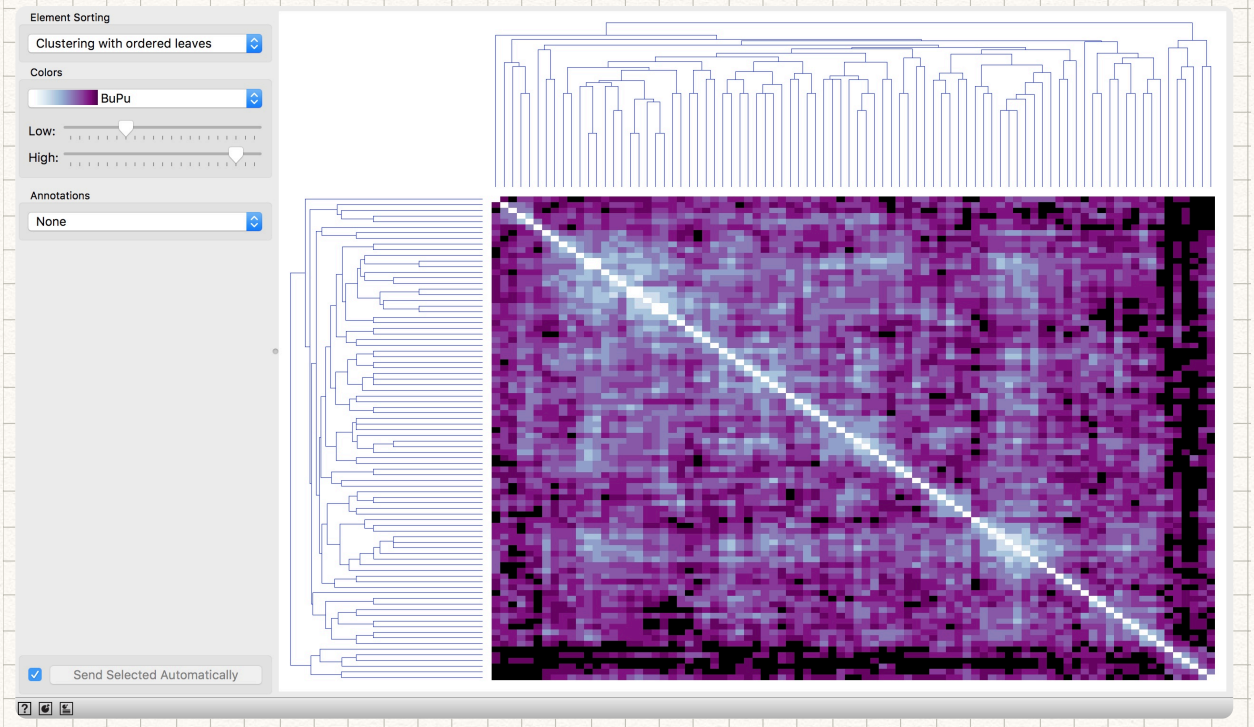
How is clustering a machine learning problem?

- 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.

APPENDIX DISTANCE MAPS/ DENDOGRAMS



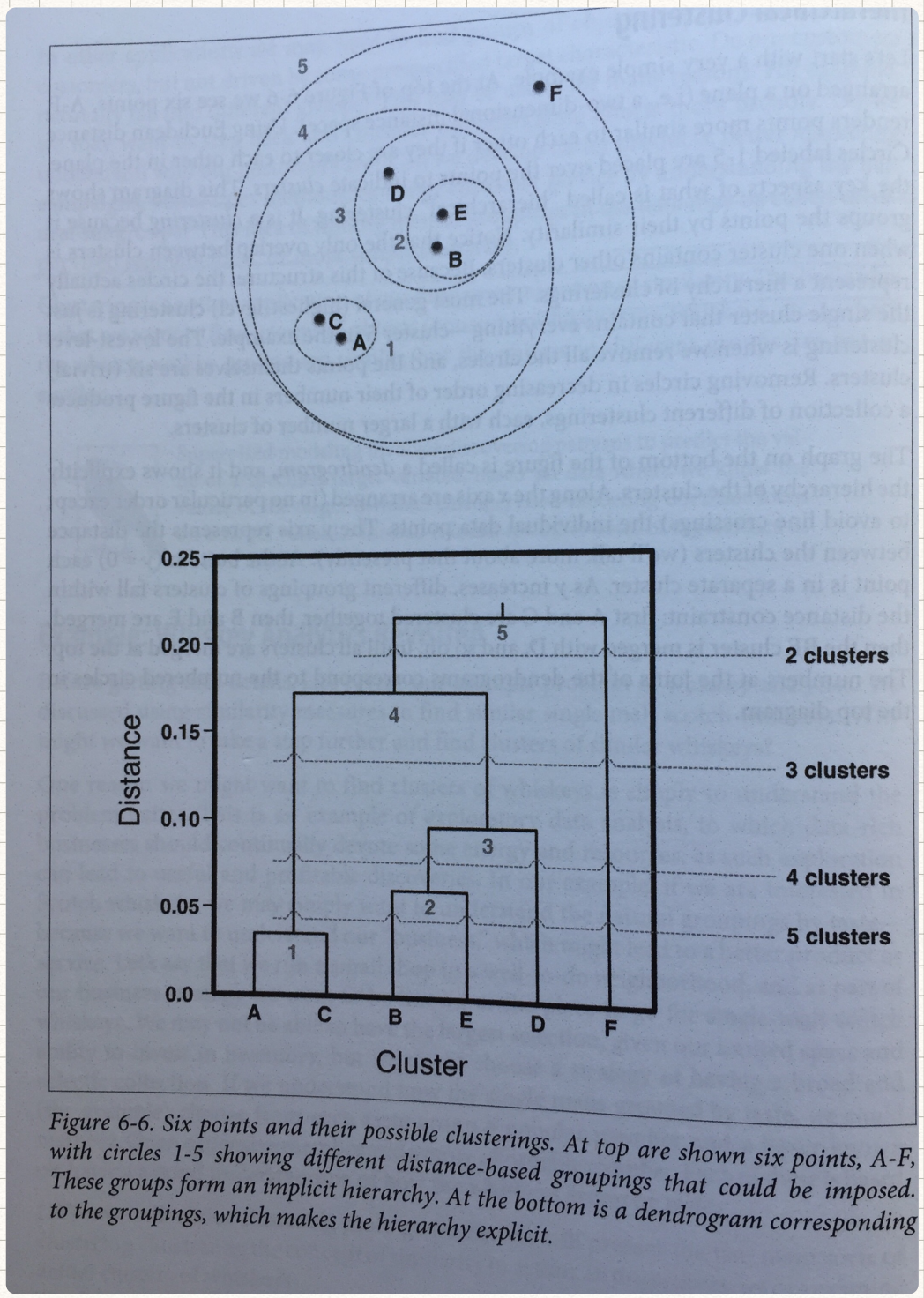


Figure 6-6. Six points and their possible clusterings. At top are shown six points, A-F, with circles 1-5 showing different distance-based groupings that could be imposed. These groups form an implicit hierarchy. At the bottom is a dendrogram corresponding to the groupings, which makes the hierarchy explicit.

Provost and Fawcett, Data Science for Business, p. 166.