DISCOVERY

#37: Distance Metrics and Clustering

April 24, 2019

Karle Flanagan and Wade Fagen-Ulmschneider

Distance Metrics

In many areas of Data Science, we need to define how different two rows of data are from each other. The most common way to find this difference is to define a **distance metric** that can be used to provide a numeric difference or "distance" two rows of data are from each other.

Distance Metric Example: Towards Intuitive Distance Metrics

Consider two data rows in a dataset that has only two data points (sometimes called "features"):

	Orange	Blue
0	0	0
1	3	3

Q1: How can we represent these data rows as geometric points?

Q2: How can we graph these two points?

Q3: What are the different ways of finding the distance between these two points?

Q4: How do we expand this idea out to data with 3 or more features?

	Orange	Blue	Purple
0	0	0	0
1	3	3	3

Q5: Thinking about how a 2D shape is traversed in a 3D world, is there a possible way to traverse a 3D shape in a 4D world?

Distance Metric Example: Common Distance Metrics

A few common distance metrics used in Data Science to compare **two** points together.

Distance Metric #1:	
Distance Metric #2:	
Distance Metric #3:	
Distance Metric #4:	

Distance Metric Example

Consider a new dataset where features have widely different values:

	Orange	Blue	Purple
0	0.003	8	10,000,000
1	0.003	8	20,000,000
2	0.023	88	20,000,000

Q6: Is Row 1 or Row 2 "closer" to Row 0? How do we know?

...what column is dominating the distance in every comparison?

Normalizing Data ("Feature Scaling")

It is important that no single column dominates the distance metric. The fact the underlying value is large should not give it an over-sized effect in determining the distance between two points! Many different methods:

1. Standard Score:

2. Feature Scaling:

3. ... many others: https://en.wikipedia.org/wiki/Normalization_(statistics)

Normalized Feature Scaling: Example #1

Raw Input Data:

	Orange	Blue	Purple
0	0.003	8	10,000,000
1	0.003	8	20,000,000
2	0.023	88	20,000,000
3	0.008	18	11,000,000

Normalized Feature Scaled Data:

	Orange	Blue	Purple
0			
1			
2			
3			

Normalized Feature Scaling: Example #2

Raw Input Data:

	A	B	C
0	14	0.02	7
1	-12	0.04	-3
2	39	0.01	1
3	11	0.06	-2

Normalized Feature Scaled Data:

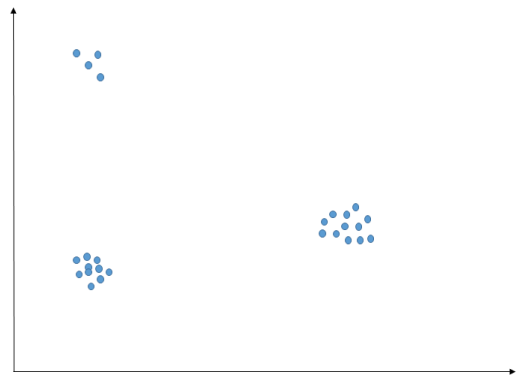
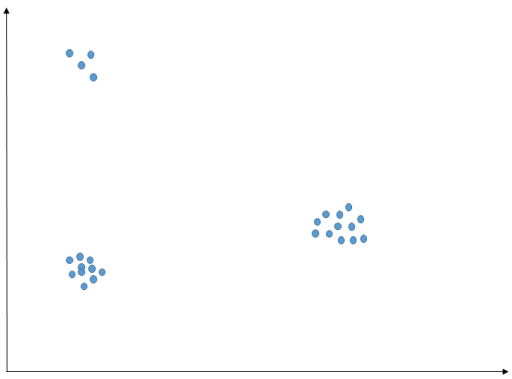
	A	B	C
0			
1			
2			
3			

Clustering

With several robust way of finding a "normalized distance" between two data points, we can begin asking more general questions about data as a whole. For example: can we **cluster similar data** together? Can we have a computer do this automatically?

Consider a simple set of points with two features. Is there a way to cluster these points? To cluster points requires us to draw **k** regions where every point is within exactly one region. The regions **cannot overlap** and every point must be part of exactly one of the **k** regions.

	
Create **two** separate clusters (regions).	Create **three** separate clusters (regions).

	
Create **four** separate clusters (regions).	Create **five** separate clusters (regions).

Given this dataset, which cluster size was most natural?