THINKFUL

Intro to Unsupervised Learning

DATA SCIENCE

Warm Up

- Up until now, we have primarily considered datasets with **responses** (outcomes of interest) and **features** that could be used to predict the response. For example, with the MNIST dataset, we have the pixel representation of handwritten digits and the corresponding “true” labels.

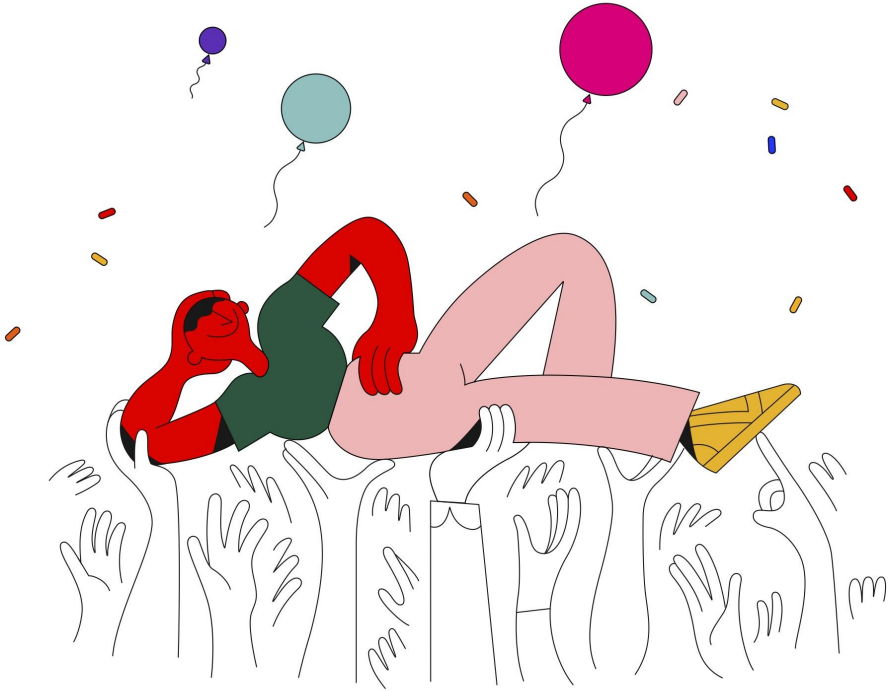
- However, labeled data is often unavailable for a variety of context-dependent reasons - labeling can be expensive, time-consuming, inaccurate, etc.. In these scenarios, we only have the features. **What questions could we ask with an unlabeled dataset? What insights could we gain from asking these questions?**

Agenda

- Unsupervised learning: overview and broad classes of algorithms
- Clustering: overview
- Similarity measures: overview
- Quantitative similarity measures

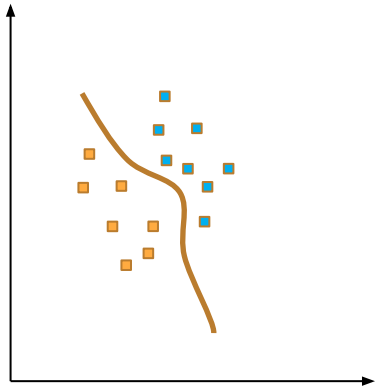
UP NEXT

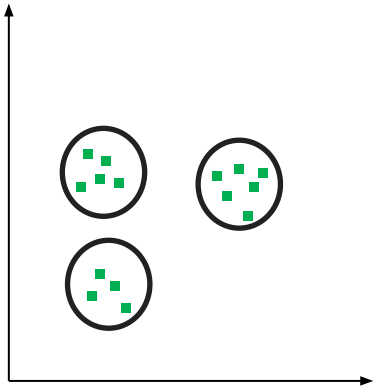
Unsupervised Learning



Unsupervised Learning - Overview

Unsupervised Learning: attempting to find patterns and/or structure in **unlabeled** data

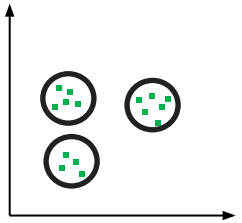




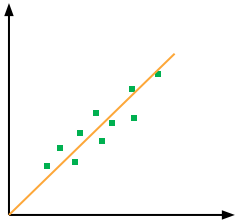
Unsupervised Learning Algorithms

Unsupervised learning can come in a variety of forms:

- **Clustering**: identifying groups of similar data points
  - Use cases: document classification, customer segmentation, fraud detection
- **Association rules:** identifying pairs (or triples, quadruples, etc.) of feature values that tend to co-occur frequently
  - Use cases: market basket analysis

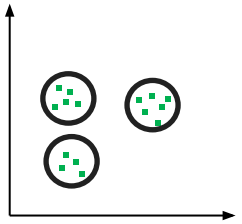


$$\{milk, bread\} \rightarrow \{cereal\}$$

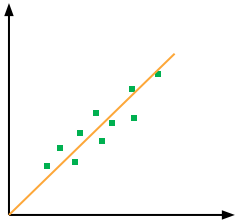


Unsupervised Learning Algorithms

- **Dimensionality reduction:** identifying innate structure in the features that allows for lower-dimension representation with minimal loss of information
  - Use cases: tackling the curse of dimensionality for high-dimensional data, visualization



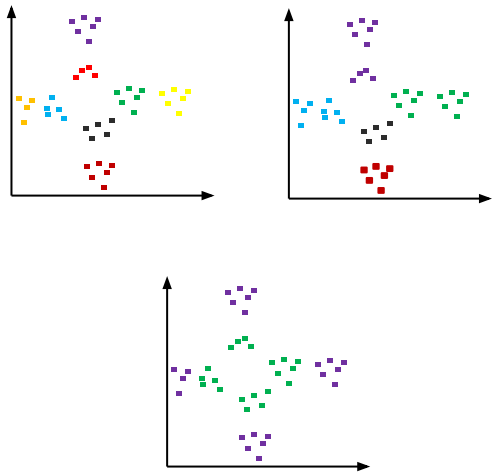
$\{milk, bread\} \rightarrow \{cereal\}$



Clustering

Clustering: identifying groups of similar observations

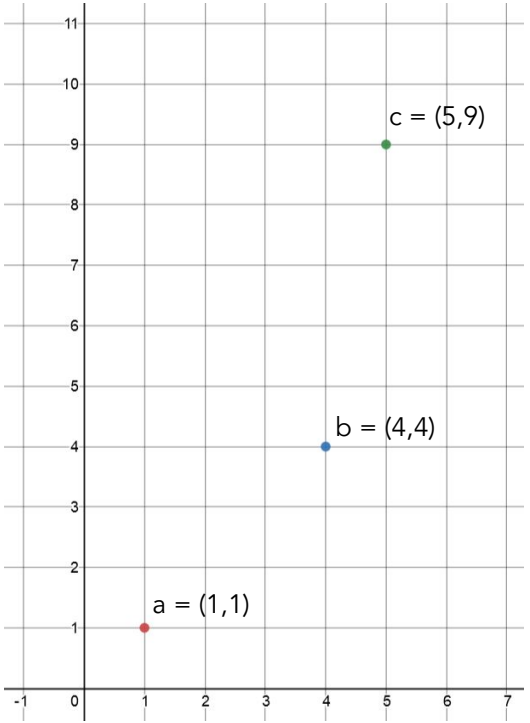
- **Groups**: What is a “group”? No precise, universal definition; classes of algorithms assume different definitions, which are suitable for different contexts
  - For example, centroid-based models define groups in terms of individual points called “centroids”; observations near a given centroid form a cluster
- **Similarity:** How do we quantify similarity between two observations?
  - Special considerations for categorical features



Similarity Measures

Clustering involves identifying groups of similar observations

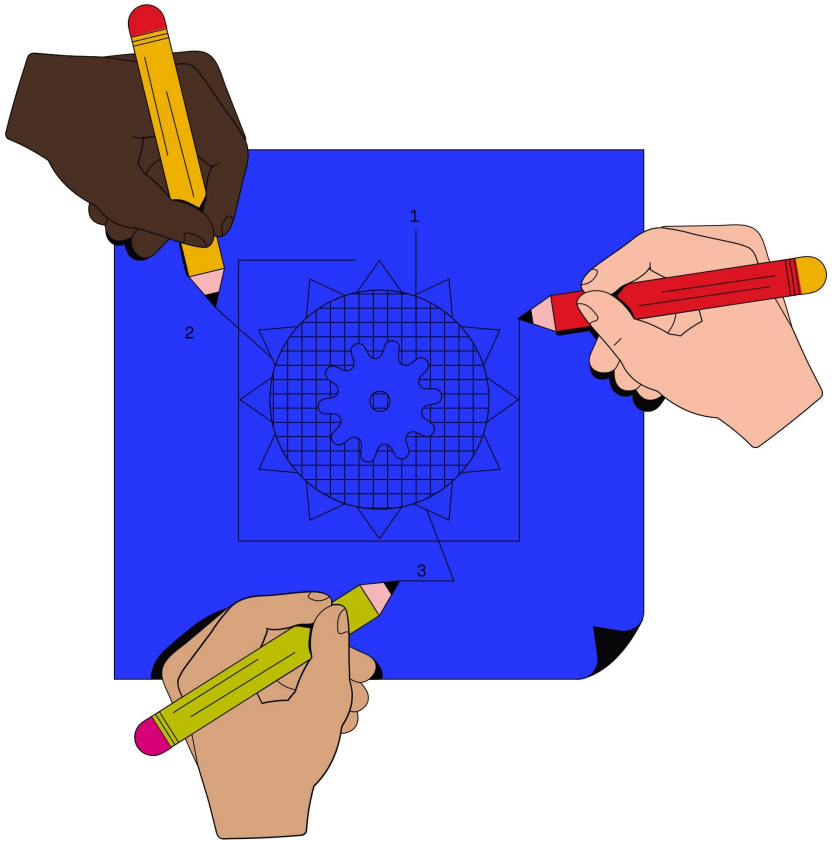
- In order to accomplish this goal, we must quantify the similarity between observations
- Choice of appropriate similarity measure is critical and strongly influences clustering results!



"Specifying an appropriate dissimilarity measure is far more important in obtaining success with clustering than choice of clustering algorithm. This aspect of the problem is emphasized less in literature than the algorithms themselves, since it depends on knowledge specifics and is less amenable to general research."
The Elements of Statistical Learning (Hastie, Tibshirani, and Friedman), p. 506

UP NEXT

Quantitative Measures



Vector Representations

We can think of observations consisting of *k* quantitative features as vectors in k-dimensional space

- We can then define various distance metrics that satisfy certain theoretical constraints, but also have key differences that make them suitable for differing contexts
- Given two points x and y, a distance metric between x and y must satisfy the following conditions:
  - $d(x,y) \geq 0$
  - If $d(x,y) = 0$, then $x = y$
  - $d(x,y) = d(y,x)$
  - $d(x,y) \leq d(x,z) + d(y,z)$ (Triangle Inequality)

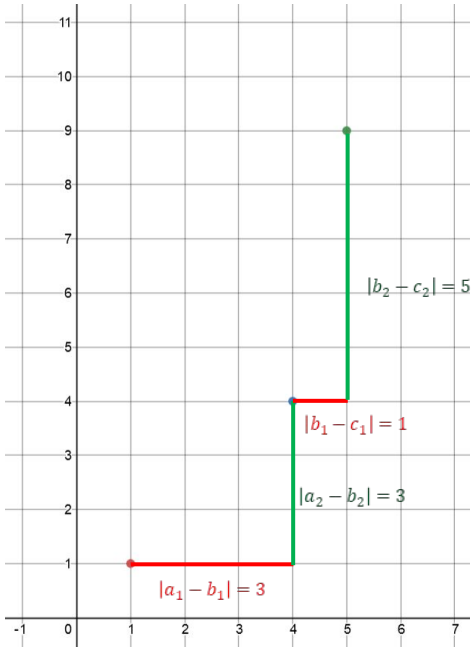
Vector Representations

- Alternatively, we can define functions that do not necessarily meet all of these constraints but still capture the similarity between two vectors in some meaningful way, called similarity measures

L-Norms: L_1 Norm

The L_1 norm, also known as the taxicab metric or Manhattan distance, can be calculated as follows (assuming x and y are n-dimensional points $(x_1,x_2,\ldots x_n,)$ and $(y_1,y_2,\ldots y_n,)$:

$$L_1(x,y) = \sum_{i=1}^{n} |x_i - y_i|$$



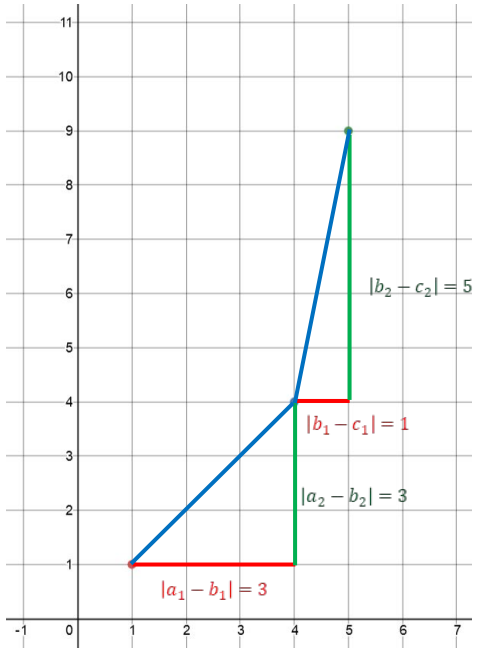
$L_1(a,b) = 3 + 3 = 6$

$L_1(b,c) = 5 + 1 = 6$

L-Norms: L_2 Norm

The L_2 norm, also known as the Euclidean Distance or straight-line distance, can be calculated as follows (assuming x and y are n-dimensional points $(x_1,x_2,...x_n,)$ and $(y_1,y_2,...y_n,)$:

$$L_2(x,y) = \sqrt{\sum_{i=1}^{n} |x_i - y_i|^2}$$



$L_2(a,b) = \sqrt{3^2 + 3^2} = \sqrt{18}$ *or* ~4.24
$L_2(b,c) = \sqrt{5^2 + 1^2} = \sqrt{26}$ *or* ~5.10

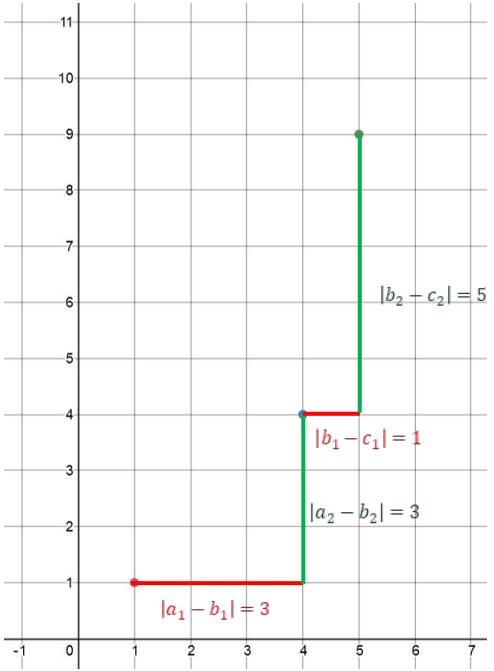
Minkowski Distance and L_∞ Norm

The L_1 and L_2 norms can be generalized to the L_p norm, otherwise known as the Minkowski distance:

$$L_p(x,y) = \sqrt[p]{\sum_{i=1}^{n} |x_i - y_i|^p}$$

A special generalization is the L_∞ norm, which corresponds to the limit of the above equation as $p \to \infty$:

$$L_\infty(x,y) = \max_i(|x_i - y_i|)$$



$L_\infty(a,b) = \max(3,3) = 3$

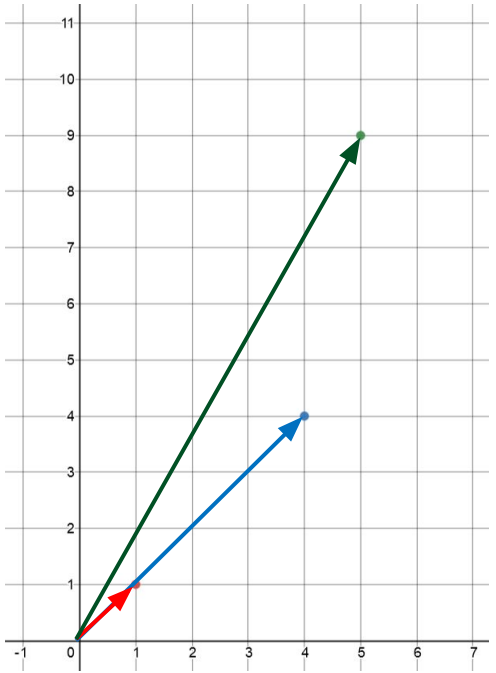
$L_\infty(b,c) = \max(5,1) = 5$

Cosine Similarity

An alternative to the previously discussed distance metrics is cosine similarity, which considers the angles between the vectors:

Cosine similarity $= cos(\theta)$

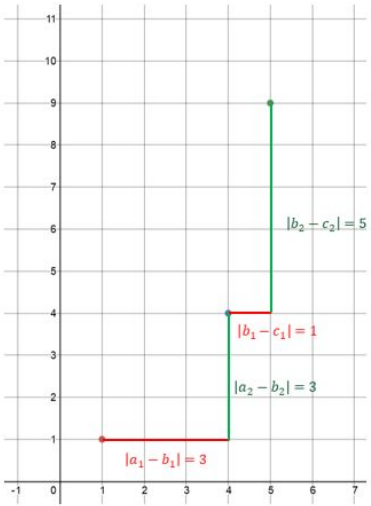
$$= \frac{X \cdot Y}{\|X\| \|Y\|}$$



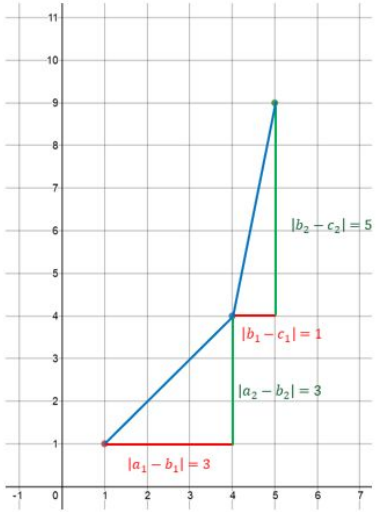
$Cosine(a, b) = 1$

$Cosine(b, c) \neq 1$

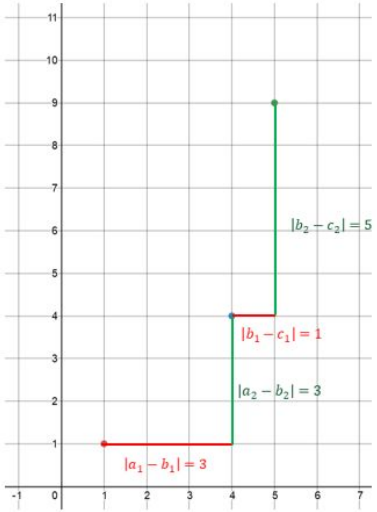
Comparing Quantitative Measures



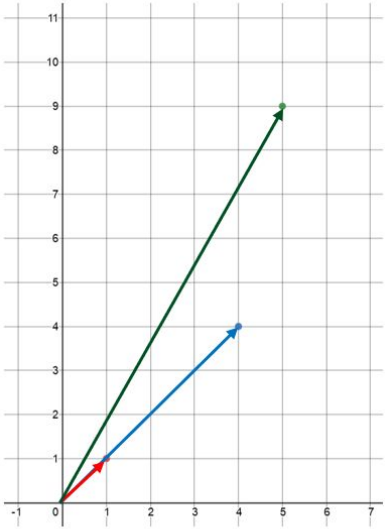
$L_1(a,b) = 3 + 3 = 6$
$L_1(b,c) = 5 + 1 = 6$



$L_2(a,b) = \sqrt{3^2 + 3^2} = \sqrt{18}$ *or* ~4.24
$L_2(b,c) = \sqrt{5^2 + 1^2} = \sqrt{26}$ *or* ~5.10



$L\infty(a,b) = \max(3,3) = 3$
$L\infty(b,c) = \max(5,1) = 5$



$Cosine(a,b) = 1$
$Cosine(b,c) \neq 1$

Importance Of Standardization

Calculating similarities on the original quantitative values can lead to misleading results, because quantities can be measured on different scales that have inherently different numeric variation

ID	Age	Number of children	Years of Post-Secondary Education	Income
1	30	0	3	$80k
2	32	0	4	$75k
3	55	3	12	$83k

Importance Of Standardization

In this example, 1 and 2 are clearly more similar to each other than 3, but income is on a much larger scale (thousands) than the other variables, so it would dominate any distance calculation, leading to the incorrect conclusion than 1 and 3 are more similar

To account for this, we usually standardize quantitative variables prior to calculating similarities. There is no universal method for such standardization; common methods include standard score normalization and min-max scaling

ID	Age	Number of children	Years of Post-Secondary Education	Income
1	30	0	3	$80k
2	32	0	4	$75k
3	55	3	12	$83k

Summary

- Primary goal of unsupervised learning is to find patterns in unlabeled data
- Most common type of unsupervised learning is clustering, which involves identifying groups of “similar” observations
- Main topic of this week is clustering, with focus on similarity measures, algorithms, and anomaly detection as a use case
- Commonly used quantitative similarity measures include L-norms and cosine distance

EXERCISE

1. Jupyter notebook
2. Data set: Starbucks locations in the U.S.
3. Choose a region of reasonable size and calculate the distance matrix between all observations in that region
4. Identify the nearest neighbor of each Starbucks
5. Identify the Starbucks locations whose nearest neighbors are the further away (“on an island”, so to speak)

Thank You