Introduction to Data Science

Intro to Distance Metrics

Gordon Anderson

Distance is Similarity

- The concept of similarity is of central importance to machine learning and AI algorithms.
- It is also a keystone of our own reasoning and judgement-making abilities.
- How can we classify/cluster a set of observations?
- We have to have a way to judge their similarity to answer this question.

Similarity

- Are two colleges similar?
- Is this frozen pizza healthy?
- How long will it take to do the next homework assignment?

Are two colleges similar?

- Need to know more about the question to obtain a better, more informative answer, but, what are the attributes we could use?
    - Enrollment- an integer [0, ?)
    - Number of departments- an integer [1-?]
    - Setting- categorical, levels: rural, urban
    - Selectivity (accepted/applied)- a percent: continuous, real, in [0,100]
    - And many other possibilities…
- Then, how would we use these values to compare two colleges?
- What is the “distance” or “similarity” metric to use?

Is this frozen pizza healthy?

- I need a “model” of what is a healthy frozen pizza, or, what is “healthy” food.
- This model is a collection of attributes, perhaps with some weighting:
  - No additives
  - Organic
  - Free range
  - Low fat
  - Etc…

- Then I can compare the pizza to the healthy model and then judge it by how similar, or how *distant* it is.
- Of course, the “hungry” factor could override the similarity outcome!

Hwk assignment time?

- As with the pizza, we need a model to go by.
- There is an *apriori,* or prior model we have of how long we take to do academic tasks of a certain subject in general. Then, we apply this to a specific course.
- Attributes of such a model:
  - Difficulty of material?
  - Length of assignment?
  - Requires outside resources?
  - Etc.

Clustering

- In our world, we have a table of data, with rows- a set, or tuple, of related observations, the columns.
- We want to find clusters in these data- or groupings of tuples that are similar.
- Evaluate the within-cluster similarity vs. similarity between clusters. A good clustering has high within-cluster similarity and low between cluster similarity.
- Now we have to define how we will judge similarity.

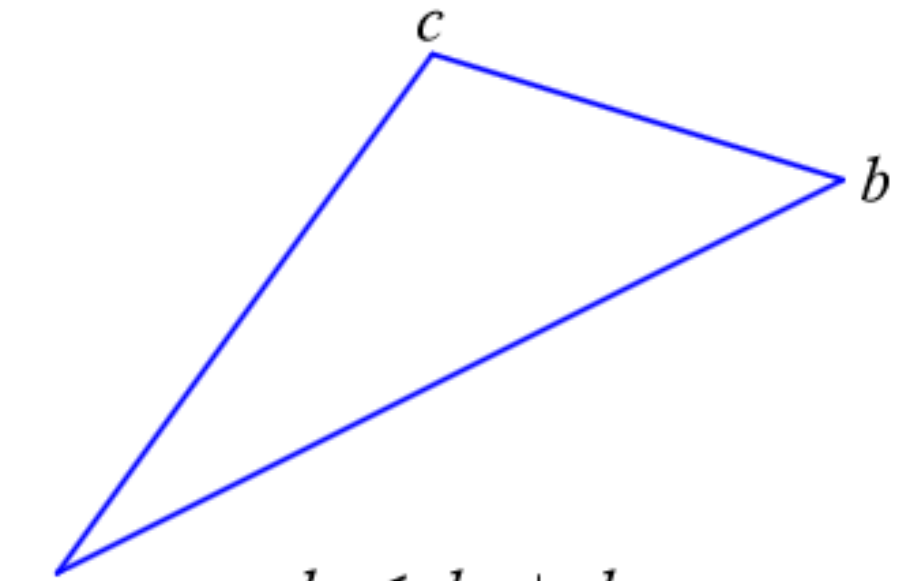
Distance metrics

- A distance metric will allow us to gauge similarity.
- Actually, this is a measure of dissimilarity!
- What is distance, though?
- A “true” distance has these properties:

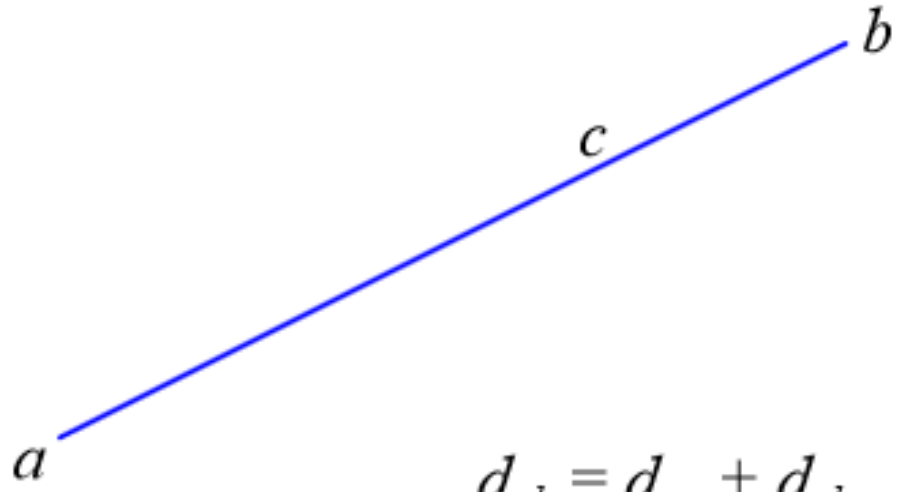
1. $d_{ab} = d_{ba}$
2. $d_{ab} \geq 0$ and $= 0$ if and only if $a = b$
3. $d_{ab} \leq d_{ac} + d_{ca}$

Distance Metrics

- The third property is called the “triangle inequality”:



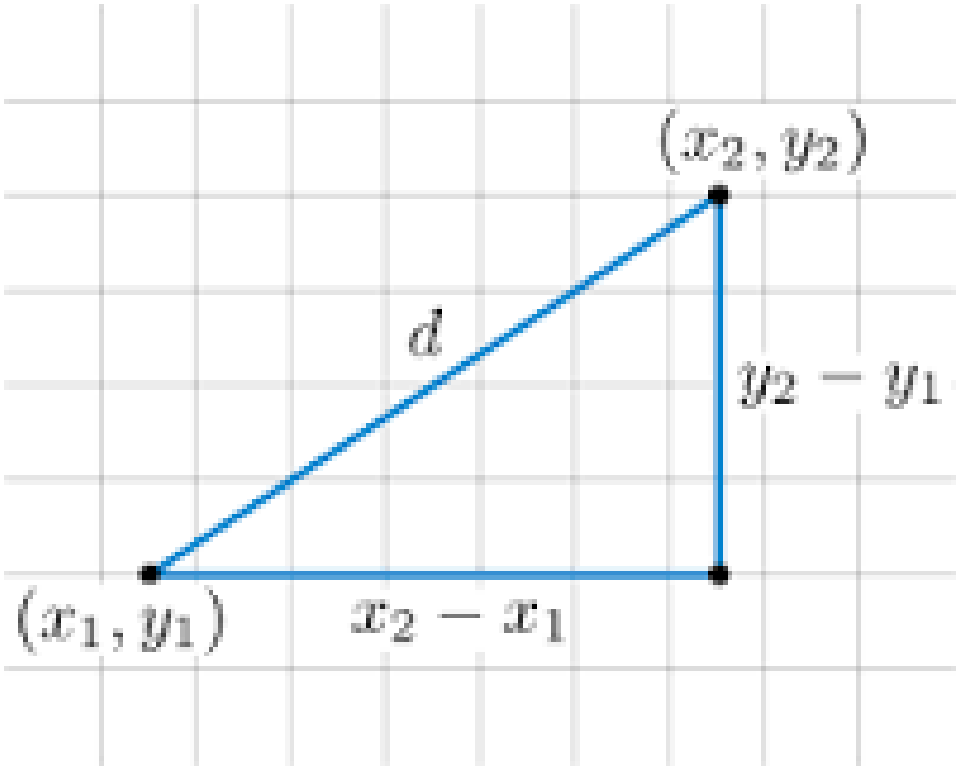
$d_{ab} \leq d_{ac} + d_{cb}$



$d_{ab} = d_{ac} + d_{cb}$

Distance Metrics

- Distance is easy to imagine for numerical values, especially in a 2D space.

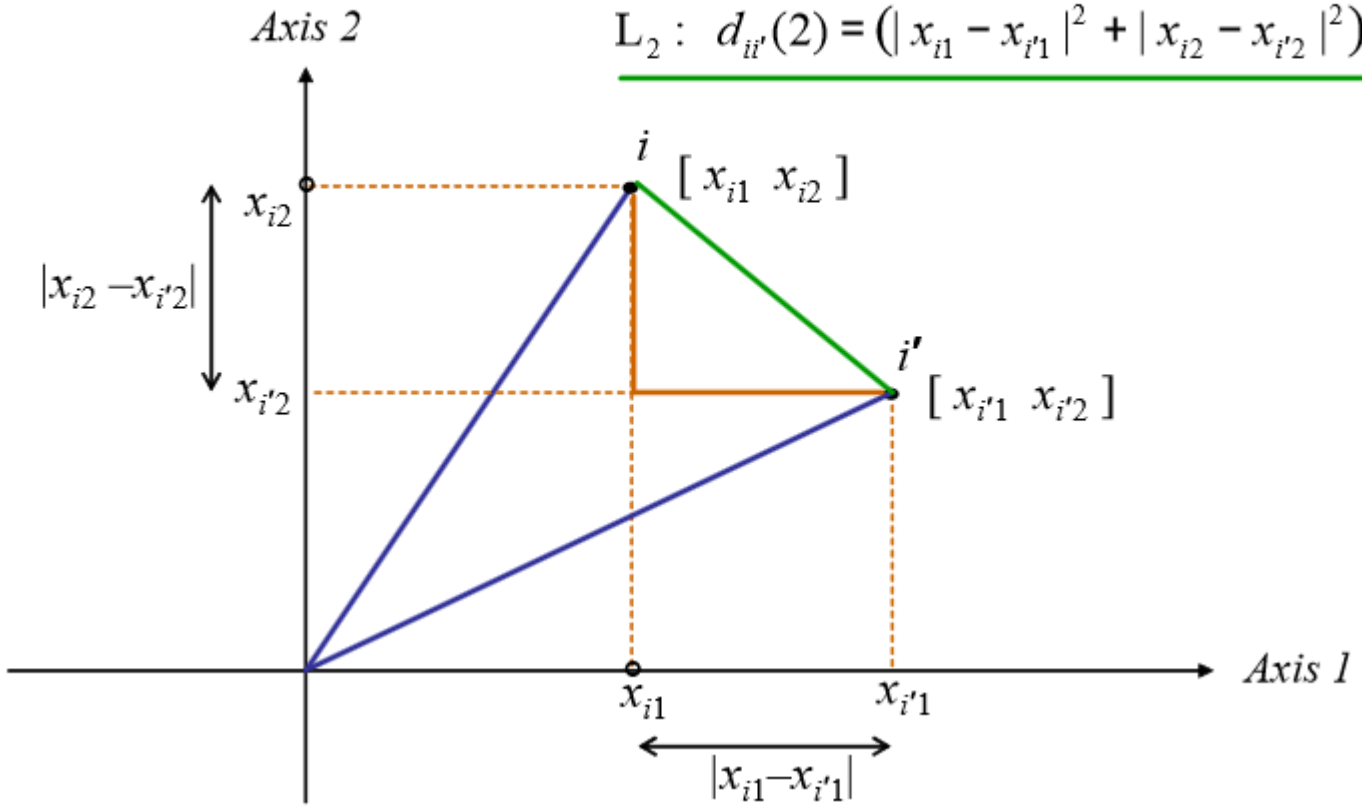


Distance Metrics

- Two common distance measures: city-block, L_1, and Euclidean, L_2.

$$L_1: \; d_{ii'}(1) = |x_{i1} - x_{i'1}| + |x_{i2} - x_{i'2}|$$

$$L_2: \; d_{ii'}(2) = \left(|x_{i1} - x_{i'1}|^2 + |x_{i2} - x_{i'2}|^2\right)^{1/2}$$



Distance matrix

- Given data and a distance metric, we calculate a distance matrix to use in clustering.

	1	2	3	4	5
1	0.0				
2	2.0	0.0			
3	6.0	5.0	0.0		
4	10.0	9.0	4.0	0.0	
5	9.0	8.0	5.0	3.0	0.0

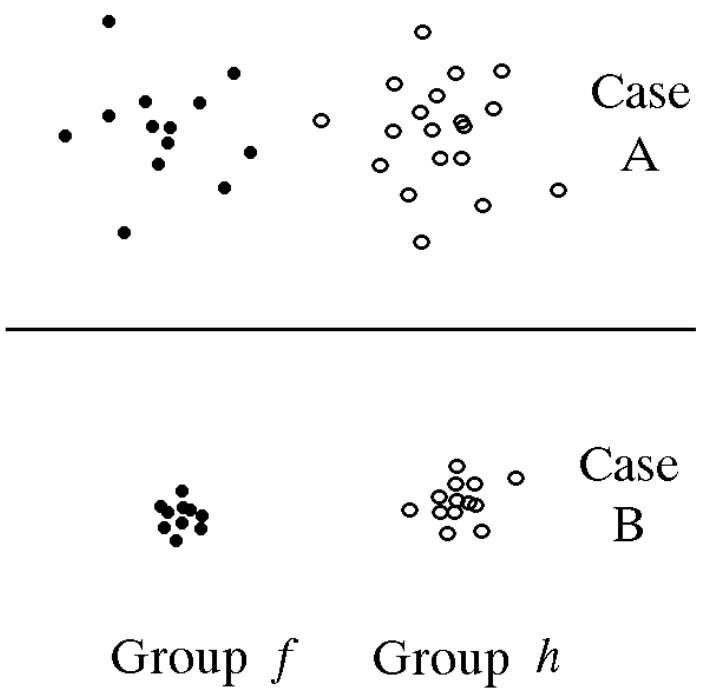
Generalized Euclidean and City Block (p dimensions)

$$ED_{i,h} = \sqrt{\sum_{j=1}^{p} (a_{i,j} - a_{h,j})^2}$$

$$CB_{i,h} = \sum_{j=1}^{p} |a_{i,j} - a_{h,j}|$$

When to use one or the other? Depends on what makes sense for your data and the questions you are addressing.

Another metric: Mahalanobis Distance



Mahalanobis distance is euclidian distance which take into account the covariance of data.

In MD, the cluster centroids are more distant in case B. MD inversely weights distance between centroids by their variance. It takes into account correlation between variables.

More on this under the topic Principal Component Analysis.

What about categorical data?

- Hard to imagine a geometric interpretation for a categorical space.

- Technically, if the measure of distance is not a “true” distance, it is called a dissimilarity measure, and we calculate a dissimilarity matrix (although Euclidean is also a measure of dissimilarity).

  Usually: similarity = 1-dissimilarity

- There are many ways to do this, and we are not constrained by the 3 distance axioms.

- The typical approach is to use a matching index or transform the data into a numerical form and calculate a distance metric like Euclidean.

Categorical example:

Data: 5 categorical variables with levels ‘a’, ‘b’, ‘c’.

	C1	C2	C3	C4	C5
1	a	c	c	b	a
2	b	c	b	a	a

How to measure distance between sample 1 and 2?
One way is matching:
Number of matches = 2, therefore, number of mismatches = 3.
Simple Matching Coefficient = mismatches/number of variables
smc = 3/5 = 0.6, note the smc is in [0,1] and is a measure of dissimilarity. For similarity, 1 - 0.6 = 0.4

Adapted from Greenacre, 2009

Categorical example:

Consider using Euclidean distances. Need numbers, so create “dummy” or indicator variables for each level of each variable:

	C1a	C1b	C1c	C2a	C2b	C2c	C3a	C3b	C3c	C4a	C4b	C4c	C5a	C5b	C5c
1	1	0	0	0	0	1	0	0	1	0	1	0	1	0	0
2	0	1	0	0	0	1	0	1	0	1	0	0	1	0	0

Compute Euclidean distance: the only terms that matter will be disagreements, when neither has the same level. (This is equivalent to an AND operation).
The Euclidean distance is $\sqrt{6}$. Let’s square that (no problem using that operation)
And we have a dissimilarity of 6. Divide by the total possible dissimilarities and we get the same value as the matching:

6/10 = 0.6

The chi-square distance is a weighted version of the smc and Euclidean weighting.

Mixed data example:

Example of continuous and discrete (categorical) data. The discrete levels have represented by indicator variables. Also calculate mean, sd:

	Continuous variables			*Sampled region*					*Substrate character*				
Station	*Depth*	*Temperature*	*Salinity*	*Tarehola*	*Skognes*	*Njosken*	*Storura*	*Clay*	*Silt*	*Sand*	*Gravel*	*Stone*
s3	30	3.15	33.52	1	0	0	0	0	1	0	0	1
s8	29	3.15	33.52	1	0	0	0	1	0	0	1	0
s25	30	3.00	33.45	0	1	0	0	1	0	1	0	0
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
s84	66	3.22	33.48	0	0	0	1	1	0	0	0	0
mean	58.15	3.086	33.50	0.242	0.273	0.242	0.242	0.606	0.152	0.364	0.182	0.061
s.d.	32.45	0.100	0.076	0.435	0.452	0.435	0.435	0.496	0.364	0.489	0.392	0.242

Use Gower's generalized coefficient of dissimilarity.

Mixed data example:

Example of continuous and discrete (categorical) data. The discrete levels have represented by indicator variables. Also calculate mean, sd:

	Continuous variables			*Sampled region*				*Substrate character*				
Station	Depth	Temperature	Salinity	Tarehola	Skognes	Njosken	Storura	Clay	Silt	Sand	Gravel	Stone
s3	30	3.15	33.52	1	0	0	0	0	1	0	0	1
s8	29	3.15	33.52	1	0	0	0	1	0	0	1	0
s25	30	3.00	33.45	0	1	0	0	1	0	1	0	0
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
s84	66	3.22	33.48	0	0	0	1	1	0	0	0	0
mean	58.15	3.086	33.50	0.242	0.273	0.242	0.242	0.606	0.152	0.364	0.182	0.061
s.d.	32.45	0.100	0.076	0.435	0.452	0.435	0.435	0.496	0.364	0.489	0.392	0.242

Use Gower's generalized coefficient of dissimilarity.

1. Standardize each variable
2. Scale the categorical variables by $\sqrt{2}$ (compensates for the 1,0)
3. Compute a distance metric.

Aside: scaling data-normalize vs. standardize

Normalize: scales data in the range [0,1].

$$x' = \frac{x - \min(x)}{\max(x) - \min(x)}$$

Standardize: new distribution with mean = 0 and unit variance.

$$x' = \frac{x - mean(x)}{sd(x)}$$

The effect of outliers is diminished by normalization.
Standardizing may result in extreme values as it does not bound the data as in normalization.

There are many other scaling techniques.

Mixed data example:

Applied Gower's technique- transform the data so that a distance metric can be calculated.

	Continuous variables			*Sampled region*				*Substrate character*				
Station	Depth	Temperature	Salinity	Tarehola	Skognes	Njosken	Storura	Clay	Silt	Sand	Gravel	Stone
s3	-0.868	0.615	0.260	1.231	-0.426	-0.394	-0.394	-0.864	1.648	-0.526	-0.328	2.741
s8	-0.898	0.615	0.260	1.231	-0.426	-0.394	-0.394	0.561	-0.294	-0.526	1.477	-0.177
s25	-0.868	-0.854	-0.676	-0.394	1.137	-0.394	-0.394	0.561	-0.294	0.921	-0.328	-0.177
:	:	:	:	:	:	:	:	:	:	:	:	:
:	:	:	:	:	:	:	:	:	:	:	:	:
s84	0.242	1.294	-0.294	-0.394	-0.426	-0.394	1.231	0.561	-0.294	-0.526	-0.328	-0.177

One caveat: the more "slicing and dicing" the more error and bias can creep in. It's good to be able to document and justify each step you take.

Another way to calculate a distance matrix with mixed data: random forests. More on that later on…

Distance and Similarity

- We have seen some examples for calculating distances between data points for numeric, categorical and mixed data.
- There are many, many more.
- Which I the best to use?
- Depends on the data, the tools you want to use, the questions you are asking.
- Important: this is an empirical investigation. That means, if you wonder which technique would be better- try it and see!!!!