

Predictive Analytics

Chapter Four

Data Similarity Measures

“One can state, without exaggeration, that the observation of and the search for similarities and differences are the basis of all human knowledge.”

Alfred Nobel

Anasse Bari, Ph.D.

CopyRights @ Anasse Bari

Learning Outcomes

- Understanding the Notion of Similarity among Data Records.
- Learning the Mathematical Criteria For a Similarity Measure and a Similarity Distance.
- Learning Major Distances such as Euclidean Distance, Manhattan Distance, Minkowski Distance, Mahalanobis Distance, Cosine Similarity, Jacquard Similarity, Simple Matching Similarity and Pearson Correlation.

Outline

- Defining Similarity in Data
- Similarity Distance Properties
- Euclidean Distance
- Manhattan Distance
- Minkowski Distance
- Mahalanobis Distance
- Cosine Similarity
- Jacquard Similarity
- Simple Matching Similarity
- Pearson Correlation

What is “similarity” in Data?

- There exist no single definition of similarity or dissimilarity between data objects.
- One common way to derive insights from data is to identify similar data objects.
- Similarity is based on some criteria (e.g. similar documents based on topics, similar social network profile based on interests, online behavior, similar Netzines..)
- No single answer or criteria – it depends on what we want to find or emphasize in the data; this is one reason why data clustering is an “art”.
- The similarity measure is often more important than the clustering algorithm used.

Defining Similarity

- Similarity
 - Numerical measure of how alike two data objects are.
 - Is higher when objects are more alike.
 - Often normalized and falls in the range $[0,1]$
- Dissimilarity
 - Numerical measure of how different are two data objects
 - Lower when objects are more alike
 - Minimum dissimilarity is often 0 (when two data items are identical)
 - Upper limit varies
 - **Proximity refers to a similarity or dissimilarity**

(Dis)similarity measures

- Instead of talking about similarity measures, we often equivalently refer to dissimilarity measures
- Dissimilarity measure is a mathematical function $f(\mathbf{x}, \mathbf{y})$ such that $f(\mathbf{x}, \mathbf{y}) > f(\mathbf{w}, \mathbf{z})$ if and only if \mathbf{x} is less similar to \mathbf{y} than \mathbf{w} is to \mathbf{z}
- This is always a *pair-wise* measure
- Think of \mathbf{x} , \mathbf{y} , \mathbf{w} , and \mathbf{z} as gene expression profiles (rows or columns)
- Text Mining Applications (similar documents, similar tweets see slide 8)

Similarity distance Properties

Common Properties of a distance

- Distance, such as the Euclidean distance, have some well known properties.
 1. $d(p,q) \geq 0$ for all p and q and $d(p,q)=0$ only if $p=q$. (positive definiteness)
 2. $d(p,q)=d(q,p)$ for all p and q . (symmetry)
 3. $d(p,r) \leq d(p,q) + d(q,r)$ for all points p , q and r (Triangle Inequality)where $d(p,q)$ is the distance (dissimilarity) between points (data objects), p and q .
- A distance that satisfies these properties is a metric

Common Properties of Similarity

- Similarities, also have some well-known properties.
 1. $s(p,q) = 1$ (or maximum similarity) only if $p=q$
 2. $s(p,q) = s(q,p)$ for all p and q (Symmetry)Where $s(p,q)$ is the similarity between points (data objects), p and q .

Converting Raw Data into a Matrix

Before you can extract groups of similar data items from your dataset, you might need to represent your data in a tabular format known as a *data matrix*. This is a preprocessing step that comes before data clustering.

Creating a matrix of terms in documents

Suppose the dataset that you're about to analyze is contained in a set of Microsoft Word documents. The first thing you need to do is to convert the set of documents into a data matrix. Several commercial and open-source tools can handle that task, producing a matrix (often known as a *document-term matrix*), in which each row corresponds to a document in the dataset. Examples of these tools include RapidMiner, and R text-mining packages.

The next section explains how documents can be converted into a data matrix.

A *document* is, in essence, a set of words. A *term* is a set of one or multiple words.

Every term that a document contains is mentioned either once or several times in the same document. The number of times a term is mentioned in a document can be represented by *term frequency* (TF), a numerical value.

We construct the matrix of terms in the document as follows:

- * The terms that appear in all documents are listed across the top row.
- * Document titles are listed down the leftmost column
- * The numbers that appear inside the matrix cells correspond to each term's frequency.

For instance, in Table 6-1, Document A is represented as set of numbers (5,16,0,19,0,0.) where 5 corresponds to the number of times the term *predictive analytics* is repeated, 16 corresponds to the number to times *computer science* is repeated, and so on. This is the simplest way to convert a set of documents into a matrix.

Table 6-1 Converting a Collection of Documents into a Matrix

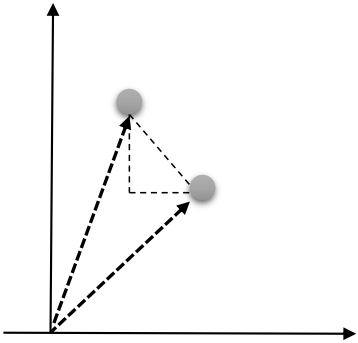
	<i>Predictive Analytics</i>	<i>Computer Science</i>	<i>Learning</i>	<i>Clustering</i>	<i>2013</i>	<i>Anthropology</i>
Document A	5	16	0	19	0	0
Document B	8	6	2	3	0	0
Document C	0	5	2	3	3	9
Document D	1	9	13	4	6	7
Document E	2	16	16	0	2	13
Document F	13	0	19	16	4	2

Pre-requisite: In most cases you will need to Convert your raw data into a matrix (tabular format)

Example: word document matrix

Reading Handout: Cb6 from *Predictive Analytics for Dummies Book*, Anasse Bari et.al. 2016 Wiley.

Euclidean distance

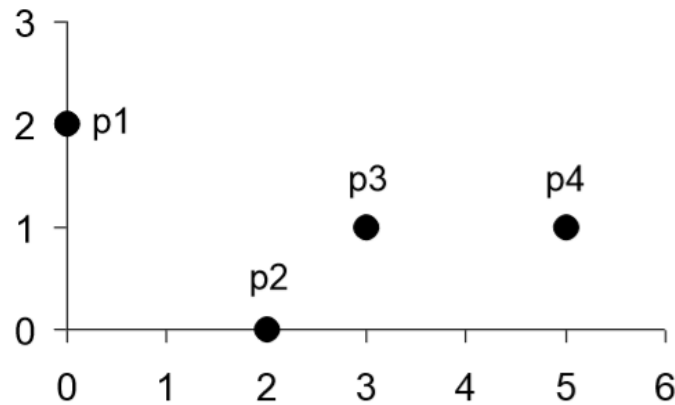


Formula:

$$dist = \sqrt{\sum_{k=1}^n (p_k - q_k)^2}$$

- Here n is the number of dimensions in the data vector. For instance:
 - Number of time-points/conditions (when clustering genes)
 - Number of genes (when clustering samples)
 - Where n is the number of dimensions (attributes) and p_k and q_k are, respectively, the k th attributes (components) or data objects p and q .

Euclidean distance (cont'd)



Dataset
(Input)

point	x	y
p1	0	2
p2	2	0
p3	3	1
p4	5	1

Distance value
between P1 and P4.
It is also called the
similarity between
P1 and P4.

Distance Matrix

	p1	p2	p3	p4
p1	0	2.828	3.162	5.099
p2	2.828	0	1.414	3.162
p3	3.162	1.414	0	2
p4	5.099	3.162	2	0

Minkowski Distance

Formula:

$$dist = \left(\sum_{k=1}^n |p_k - q_k|^r \right)^{\frac{1}{r}}$$

- Where r is a parameter, n is the number of dimensions (attributes) and p_k and q_k are, respectively, the k th attributes (components) or data objects p and q .
- $r = 1$. City block (Manhattan Distance, taxicab, L1 norm) distance.
A common example of this is the Hamming distance, which is just the number of bits that are different between two binary vectors
Read more at: <http://xlinux.nist.gov/dads/HTML/manhattanDistance.html>
- $r = 2$. Euclidean distance

Minkowski Distance (cont'd)

point	x	y
p1	0	2
p2	2	0
p3	3	1
p4	5	1

Distance Matrix

L1	p1	p2	p3	p4
p1	0	4	4	6
p2	4	0	2	4
p3	4	2	0	2
p4	6	4	2	0

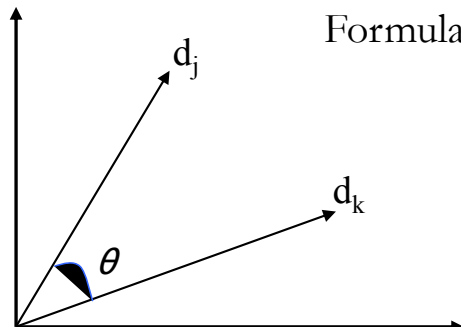
Distance value
between P1 and P4.
It is also called the
similarity between
P1 and P4.

Minkowski Distance (cont'd)

L2	p1	p2	p3	p4
p1	0	2.828	3.162	5.099
p2	2.828	0	1.414	3.162
p3	3.162	1.414	0	2
p4	5.099	3.162	2	0

L_∞	p1	p2	p3	p4
p1	0	2	3	5
p2	2	0	1	3
p3	3	1	0	2
p4	5	3	2	0

Cosine Similarity



Formula:

$$\text{sim}(d_j, d_k) = \frac{\vec{d}_j \cdot \vec{d}_k}{\|\vec{d}_j\| * \|\vec{d}_k\|} = \frac{\sum_{i=1}^n w_{i,j} w_{i,k}}{\sqrt{\sum_{i=1}^n w_{i,j}^2} \sqrt{\sum_{i=1}^n w_{i,k}^2}}$$

Given that d_j and d_k are normalized: $\|\vec{d}_j\| = \sqrt{\sum_{i=1}^n w_{i,j}^2} = 1$

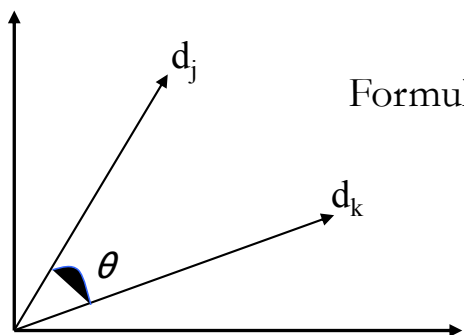
- Distance between vectors P_1 and P_2 is the value of the cosine of the angle θ between the two vectors.
- Note – this is *similarity*, not distance
 - No triangle inequality for similarity.

Note: A vector can be *normalized* (given a length of 1) by dividing each of its components by its length – here we use the L_2 norm

Copyrights @ Anasse Bari

$$\|\mathbf{x}\|_2 = \sqrt{\sum_i x_i^2}$$

Cosine Similarity (for length-normalized vectors)



Formula (dot product): $(sim) \cos(\vec{d}_j, \vec{d}_k) = \vec{d}_j \cdot \vec{d}_k$

Note: A vector can be *normalized* (given a length of 1) by dividing each of its components by its length – here we use the L_2 norm

Copyrights @ Anasse Bari

$$\|\mathbf{x}\|_2 = \sqrt{\sum_i x_i^2}$$

Cosine Similarly

- If $d1$ and $d2$ are two document vectors, then

$$\cos(d1, d2) = (d1 \bullet d2) / \|d1\| \|d2\|,$$

Where \bullet indicates vector dot product and $\|d1\|$ is the length of vector d.

- Example:

$$d1 = 3 \ 2 \ 0 \ 5 \ 0 \ 0 \ 0 \ 2 \ 0 \ 0$$

$$d2 = 1 \ 0 \ 0 \ 0 \ 0 \ 0 \ 0 \ 1 \ 0 \ 2$$

$$d1 \bullet d2 = 3*1 + 2*0 + 0*0 + 5*0 + 0*0 + 0*0 + 0*0 + 2*1 + 0*0 + 0*2 = 5$$

$$\|d1\| = \sqrt{(3*3 + 2*2 + 0*0 + 5*5 + 0*0 + 0*0 + 0*0 + 2*2 + 0*0 + 0*0)} = 6.481$$

$$\|d2\| = \sqrt{(1*1 + 0*0 + 0*0 + 0*0 + 0*0 + 0*0 + 0*0 + 1*1 + 0*0 + 2*2)} = 2.245$$

$$\cos(d1, d2) = 0.3150$$

Mahalanobis Distance

$$D^2 = (x - m)^T C^{-1} (x - m)$$

Where:

D^2 is Mahalanobis distance

x is Vector of data

m is vector of mean values of independent variables

C^{-1} is inverse covariance matrix of independent variables

T indicates vector should be transposed

Note: If the covariance matrix is the identity matrix, the Mahalanobis distance reduces to the Euclidean distance. If the covariance matrix is diagonal, then the resulting distance measure is called the normalized Euclidean distance:

Mahalanobis Distance (cont'd)

Example:

Suppose we took a single observation from a bivariate population with Variable X and Variable Y, and that our two variables had the following characteristics:

Variable X: mean = 500, SD = 79.32

Variable Y: mean = 500, SD = 79.25

Variance/Covariance Matrix		
	X	Y
X	6291.55737	3754.32851
Y	3754.32851	6280.77066

If, in our single observation, $X = 410$ and $Y = 400$, we would calculate the Mahalanobis distance for that single value as (next slide):

Mahalanobis Distance (cont'd)

Given that Mahalanobis Distance $D^2 = (x - m)^T C^{-1} (x - m)$

$$(x - m) = \begin{pmatrix} 410 - 500 \\ 400 - 500 \end{pmatrix} = \begin{pmatrix} -90 \\ -100 \end{pmatrix}$$

$$C^{-1} = \begin{pmatrix} 6291.55737 & 3754.32851 \\ 3754.32851 & 6280.77066 \end{pmatrix}^{-1} = \begin{pmatrix} 0.00025 & -0.00015 \\ -0.00015 & 0.00025 \end{pmatrix}$$

$$\text{Therefore } D^2 = (-90 \ -100) \times \begin{pmatrix} 0.00025 & -0.00015 \\ -0.00015 & 0.00025 \end{pmatrix} \times \begin{pmatrix} -90 \\ -100 \end{pmatrix} = 1.825$$

Jaccard coefficient

- A similarity measure that is often used to measure the overlap of two sets X and Y .

Formula: $\text{Jaccard}(X,Y) = (\text{Intersection } |X \cap Y|) / (\text{Union } |X \cup Y|)$

$$\text{Jaccard}(X,X) = 1$$

$$\text{Jaccard}(X,Y) = 0 \text{ if } X \cap Y = 0$$

X and Y don't have to be the same size.

The Jaccard is always between 0 and 1

Jaccard coefficient

- Example

Suppose we have two sets $A = \{7, 3, 2, 4, 1\}$ and $B = \{4, 1, 9, 7, 5\}$.

The union is $A \cup B = \{7, 3, 2, 4, 1, 9, 5\}$

The Intersection is $A \cap B = \{7, 4, 1\}$

$$S_{AB} = \frac{|A \cap B|}{|A \cup B|} = \frac{3}{7} = 0.429$$

Correlation

- Correlation measures the linear relationship between objects
- To compute correlation, we standardize data objects, p and q , and then take their dot product:

$$p'_k = \frac{p_k - \text{mean}(p)}{\text{std}(p)}$$

$$q'_k = \frac{q_k - \text{mean}(q)}{\text{std}(q)}$$

$$\text{Correlation}(p, q) = p' \bullet q'$$

Pearson Linear Correlation

Formula:

$$\rho(\mathbf{x}, \mathbf{y}) = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}}$$

$$\bar{x} = \frac{1}{n} \sum_i^n x_i$$

$$\bar{y} = \frac{1}{n} \sum_i^n y_i$$

Pearson Linear Correlation

- Pearson linear correlation (PLC) is a measure that is invariant to scaling and shifting (vertically) of the expression values
- Always between 0 and +1 (perfectly anti-correlated and perfectly correlated)
- This is a similarity measure, but we can easily make it into a dissimilarity measure:

$$d_p = \frac{1 - \rho(\mathbf{x}, \mathbf{y})}{2}$$

$$\rho_{X,Y} = \text{corr}(X, Y) = \frac{\text{cov}(X, Y)}{\sigma_X \sigma_Y} = \frac{E[(X - \mu_X)(Y - \mu_Y)]}{\sigma_X \sigma_Y},$$

Pearson Linear Correlation

- Example

Find Correlation coefficient for X and Y values are given below

X= (1,2,3,4,5)

Y= {11,22,34,43,56}

Correlation Coefficient (ρ)=0.9989

$$d_p = \frac{1 - \rho}{2} = \frac{1 - 0.9989}{2} = 0.00055$$

Simple Matching Similarity

The Simple Matching Coefficient (SMC) is a statistical technique used for comparing the similarity and diversity of sample sets.

Given two objects, A and B, each with n binary attributes, SMC is defined as:

$$SMC = \frac{\text{Number of Matching Attributes}}{\text{Number of Attributes}} = \frac{M_{00} + M_{11}}{M_{00} + M_{01} + M_{10} + M_{11}}$$

Where:

M_{00} represents the total number of attributes where A and B both have a value of 1.

M_{01} represents the total number of attributes where the attribute of A is 0 and the attribute of B is 1.

M_{10} represents the total number of attributes where the attribute of A is 1 and the attribute of B is 0.

M_{11} represents the total number of attributes where A and B both have a value of 0.²⁶

Resources and References

<https://www.coursehero.com/file/p54u3ta6/University-of-Florida-CISE-department-Gator-Engineering-Data-Mining-Sanjay/>

[Tapana Aksaranan](#), A Study on Quantifying Color Appearance of Translucent Material, ProQuest, 2008 - 342 pages

http://www.jennessent.com/arcview/mahalanobis_description.htm

Prof. Bellaachia data mining slides, <http://www.seas.gwu.edu/~bell/csci243/csci243.htm>

<http://people.revoledu.com/kardi/tutorial/Similarity/Jaccard.html>

Boston University slides for clustering: <http://www.cs.bu.edu/fac/gkollios/ada05/>

MIT slides for clustering: www.mit.edu/~georg

<http://www.statisticshowto.com/how-to-compute-pearsons-correlation-coefficients/>

https://en.wikipedia.org/wiki/Simple_matching_coefficient

http://mines.humanoriented.com/classes/2010/fall/csci568/portfolio_exports/sdaugherty/similarity.htm

Predictive Analytics

End of Chapter Four

Data Similarity Measures

“One can state, without exaggeration, that the observation of and the search for similarities and differences are the basis of all human knowledge.”

Alfred Nobel

Anasse Bari, Ph.D.

CopyRights @ Anasse Bari