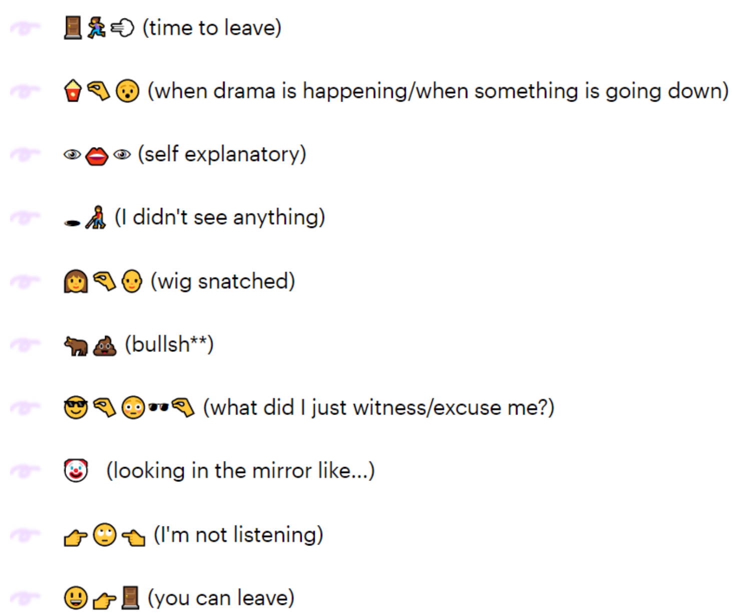
Emojis

- Emoji is a feature/attribute/dimension/variable of data object.
- What type of variable is the emoji?
  - Interval scaled variable
  - Binary variable
  - Nominal (categorical) variable
  - Ordinal variable
  - Ratio scaled variable
  - Mixed type



Data Mining

Lecture-3

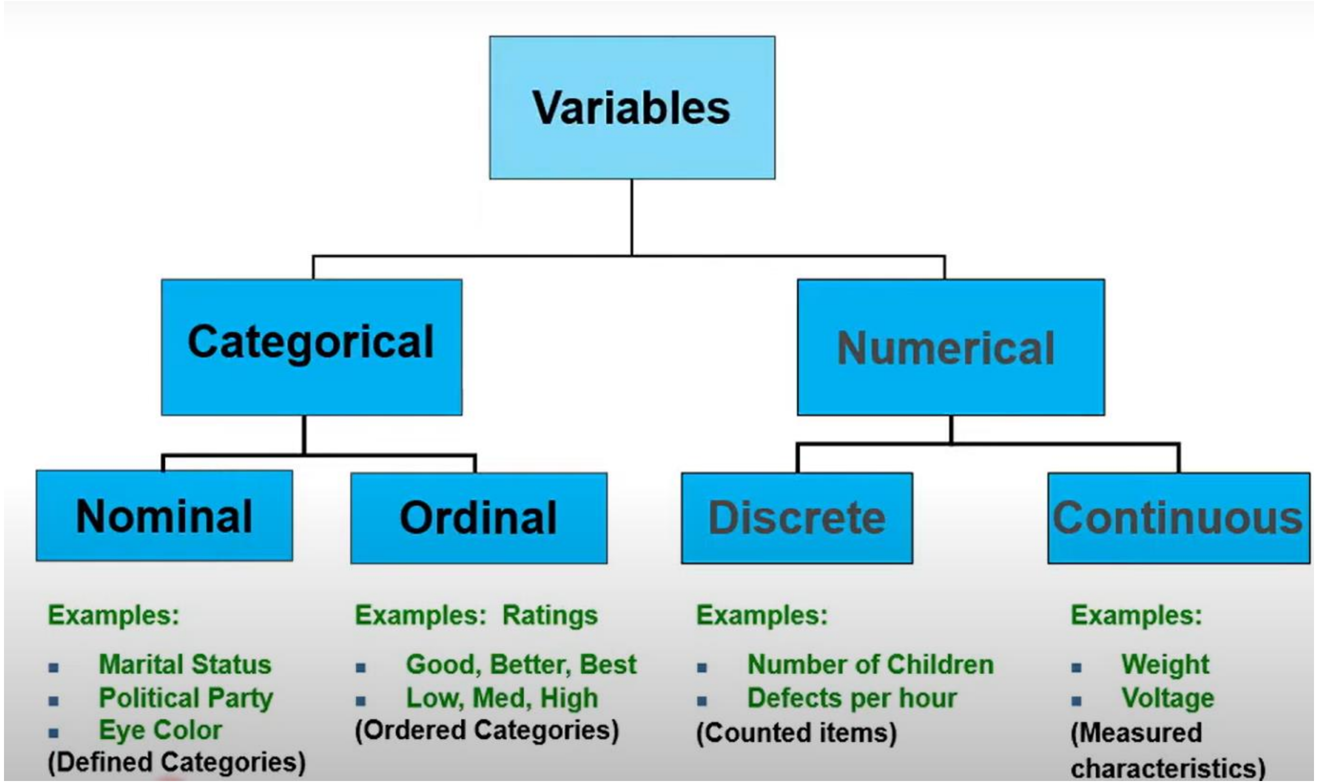
Similarity and Dissimilarity Measures

Dr. Salem Othman

Summer 2023

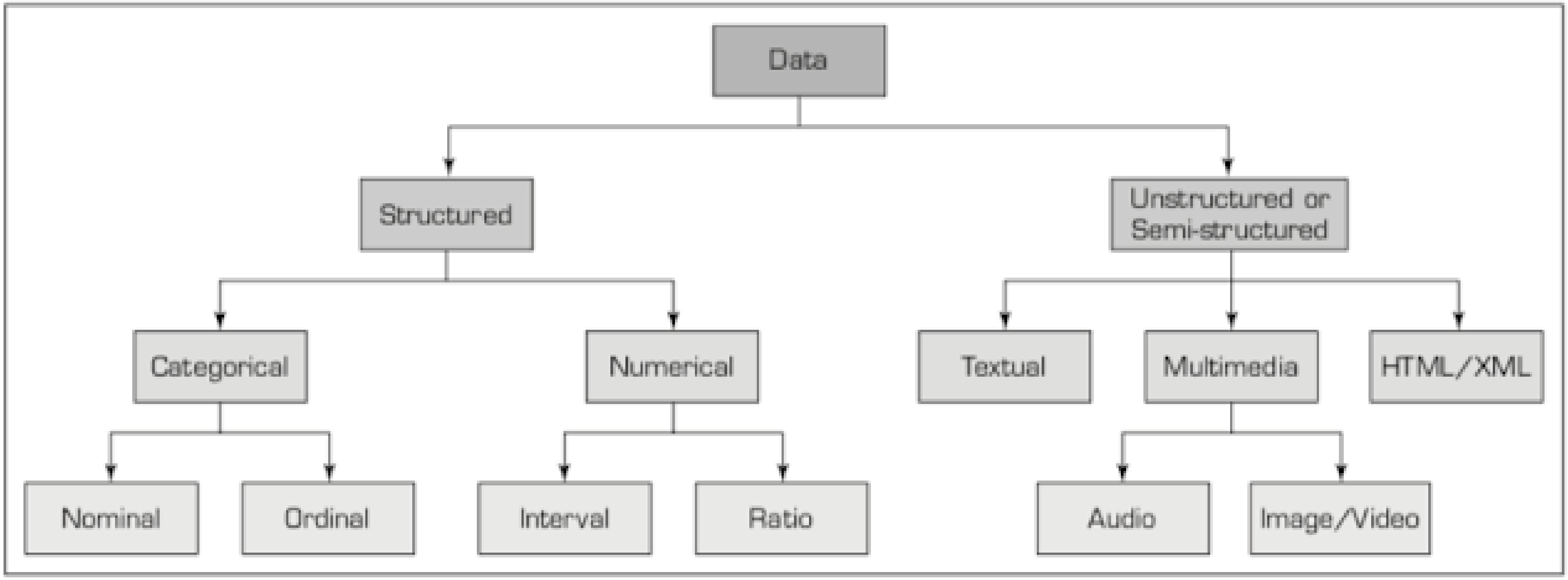
Outline

- Similarity and Distance



A Simple Taxonomy of Data

- http://www.wemiibidun.com/2018/05/a-simple-taxonomy-of-data.html



Definitions - Similarity and Dissimilarity Measures

- Similarity: A numerical measure of how alike two objects are. It is usually non-negative and often between 0 (no similarity) and 1 (complete similarity).
- Dissimilarity: A numerical measure of how different two objects are. Often synonymous with 'distance'. It can range from 0 to ∞, but commonly falls in the interval [0, 1].

Similarity and Dissimilarity Measures

- Similarity measure
  - Numerical measure of how alike two data objects are.
  - Is higher when objects are more alike.
  - Often falls in the range [0,1]
- Dissimilarity measure
  - Numerical measure of how different two data objects are
  - Lower when objects are more alike
  - Minimum dissimilarity is often 0
  - Upper limit varies based on the exact metric you are using.
- Proximity refers to a similarity or dissimilarity

Real-Life Example Use-case

Predicting COVID-19 patients on the basis of their symptoms

- With the rise of COVID-19 cases, many people are not being able to seek proper medical advice due to the shortage of both human and infrastructure resources. As a result, we as engineers can contribute our bit to solve this problem by providing a basic diagnosis to help in identifying the people suffering from COVID-19. To help us we can make use of Machine Learning algorithms to ease out this task, among which clustering algorithms come in handy to use.
- For this, we make two clusters based on the *symptoms of the patients* who are COVID-19 positive or negative and then predict whether a new incoming patient is suffering from COVID-19 or not by *measuring the similarity/dissimilarity of the observed symptoms (features)* with that of the infected person's symptoms. [1]

Similarity/Dissimilarity Transformations

- Transformations in data mining are frequently applied to:
  - Convert similarities into dissimilarities and vice versa.
  - Adjust proximity measures to fall within a specific range, such as [0,1].
- This can be particularly useful when using certain algorithms or software packages which operate within these bounds. Two common transformations include:
  - Linear Transformation: Used when original proximity measures have a finite range. This preserves relative distances between points.
  - Non-Linear Transformation: Used when original proximity measures take values from $[0, \infty)$. This can compress larger values into a range near 1.

However, transforming proximity measures can alter their meaning, and this should be considered carefully.

Transformations: Example

Movie	Person A rating	Person B rating
Movie 1	8	7
Movie 2	9	9
Movie 3	7	6
Movie 4	6	8
Movie 5	7	6

case, the transformation of similarities to the interval [0, 1] is given by the expression $s'=(s-min\_s)/(max\_s-min\_s)$, where *max_s* and *min_s* are the maximum and minimum similarity values, respectively. Likewise, dissimilarity measures with a finite range can be mapped to the interval [0,1] by using the formula $d'=(d-min\_d)/(max\_d-min\_d)$. This is an example of a linear

Dissimilarity: One simple measure of dissimilarity is the absolute difference in ratings. For Movie 1, the dissimilarity d would be |8 - 7| = 1. For Movie 2, d = |9 - 9| = 0. This gives us dissimilarities ranging from 0 to 2 in this example.

Linear transformation: We could normalize these dissimilarities to fall between 0 and 1 by subtracting the minimum dissimilarity and dividing by the range (max - min). For Movie 1, the transformed dissimilarity d' would be (1 - 0) / (2 - 0) = 0.5. For Movie 2, d' = (0 - 0) / 2 = 0.

Conversion to similarity: We can convert these dissimilarities to similarities. One simple method is s = 1 - d. For Movie 1, the similarity s would be 1 - 0.5 = 0.5. For Movie 2, s = 1 - 0 = 1.

Similarity/Dissimilarity for Simple Attributes

The following table shows the similarity and dissimilarity between two objects, *x* and *y*, with respect to a single, simple attribute.

Attribute Type	**Dissimilarity**	**Similarity**
Nominal	$d = \begin{cases} 0 & \text{if } x = y \\ 1 & \text{if } x \neq y \end{cases}$	$s = \begin{cases} 1 & \text{if } x = y \\ 0 & \text{if } x \neq y \end{cases}$
Ordinal	$d = \lvert x - y \rvert / (n - 1)$ (values mapped to integers 0 to $n-1$, where n is the number of values)	$s = 1 - d$
Interval or Ratio	$d = \lvert x - y \rvert$	$s = -d$, $s = \frac{1}{1+d}$, $s = e^{-d}$, $s = 1 - \frac{d - min\_d}{max\_d - min\_d}$

Dissimilarity Matrix

Similarity Matrix

	A	B	C
A	Nan	1	0
B	1	Nan	0
C	0	0	Nan

Distance Matrix

	A	B	C
A	0	0	1
B	0	0	1
C	1	1	0

- It is a matrix of pairwise dissimilarity among the data points. It is often desirable to keep only lower triangle or upper triangle of a dissimilarity matrix to reduce the space and time complexity.

1. ***It's square and symmetric(A^T*= A for a square matrix A, where A^T represents its transpose)**.
2. ***The diagonals members are zero, meaning that zero is the measure of dissimilarity between an element and itself***.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
(1)	0											
(2)	6.32	0										
(3)	4.78	6.78	0									
(4)	7.93	7.73	3.3	0								
(5)	8.82	9.79	4.07	2.24	0							
(6)	4.42	2.07	4.94	6.54	8.38	0						
(7)	5.03	7.4	0.62	3.39	3.8	5.54	0					
(8)	6.3	4.38	3.48	3.34	5.47	3.44	4.04	0				
(9)	5.3	1.13	6.42	7.84	9.78	1.47	7.01	4.59	0			
(10)	6.41	2.87	4.77	4.93	7.09	2.54	5.38	1.62	3.31	0		
(11)	0.66	6.95	4.78	8.02	8.75	5.01	4.94	6.65	5.95	6.89	0	
(12)	1.3	6.11	3.48	6.65	7.52	4.07	3.73	5.27	5.24	5.64	1.41	0

Euclidean Distance

- Euclidean Distance

$$d(\mathbf{x}, \mathbf{y}) = \sqrt{\sum_{k=1}^{n} (x_k - y_k)^2}$$

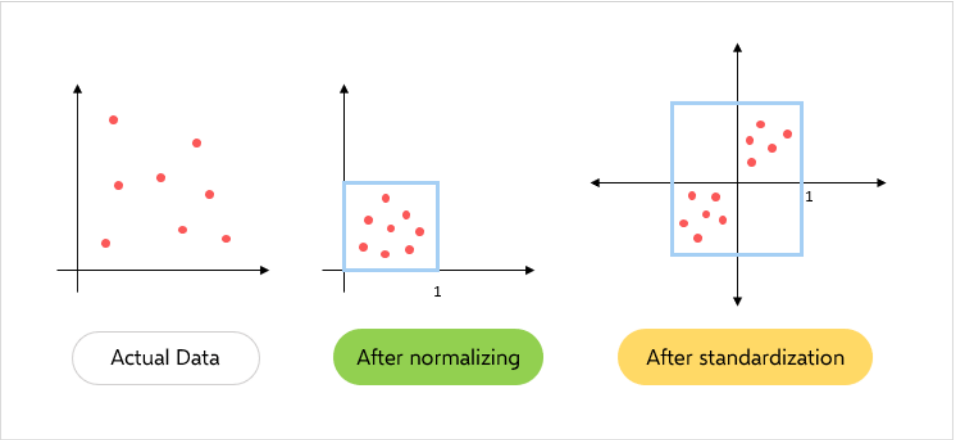
where *n* is the number of dimensions (attributes) and x_k and y_k are, respectively, the k^{th} attributes (components) or data objects $\mathbf{x}$ and $\mathbf{y}$.

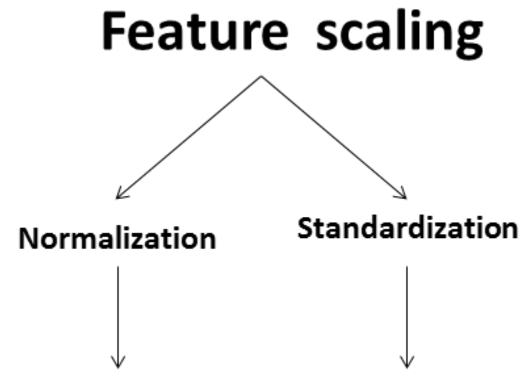
- **Standardization** is necessary, if scales differ.

Standardization vs. Normalization: What's the Difference?

- **Standardization** rescales a dataset to have a mean of 0 and a standard deviation of 1.
- https://www.statology.org/standardization-vs-normalization/

- **Normalization** rescales a dataset so that each value falls between 0 and 1.

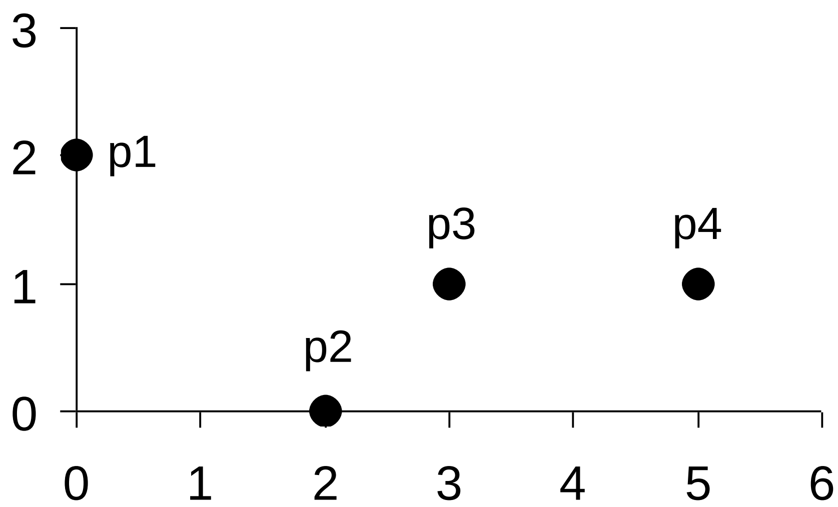




$$X_{new} = \frac{X - X_{min}}{X_{max} - X_{min}} \qquad X' = \frac{\text{X} - \text{Mean}}{\text{Standard deviation}}$$

Euclidean Distance

$$d(\mathbf{x}, \mathbf{y}) = \sqrt{\sum_{k=1}^{n} (x_k - y_k)^2}$$



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

	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

Distance Matrix

Minkowski Distance

- Minkowski Distance is a generalization of Euclidean Distance

$$d(\mathbf{x}, \mathbf{y}) = \left(\sum_{k=1}^{n} |x_k - y_k|^r \right)^{1/r}$$

Where r is a parameter, n is the number of dimensions (attributes) and x_k and y_k are, respectively, the k^{th} attributes (components) or data objects $\boldsymbol{x}$ and $\boldsymbol{y}$.

- Note that setting r = 1 is equivalent to calculating the Manhattan distance and setting r = 2 is equivalent to calculating the Euclidean distance.

Minkowski Distance: Examples

- $r = 1$. City block (Manhattan, taxicab, L_1 norm) distance.
  - A common example of this for binary vectors is the Hamming distance, which is just the number of bits that are different between two binary vectors

- $r = 2$. Euclidean distance

- $r \to \infty$. "supremum" (L_{max} norm, L_{∞} norm) distance.
  - This is the maximum difference between any component of the vectors

$$d(x, y) = \lim_{r \to \infty} \left(\sum_{k=1}^{n} |x_k - y_k|^r \right)^{\frac{1}{r}}$$

- Do not confuse *r* with *n*, i.e., all these distances are defined for all numbers of dimensions.

Minkowski Distance

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

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

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

Distance Matrix

Manhattan Distance

- p1=(0,2)= (x1, y1)   p2=(2,0)=(x2, y2)
- Distance= |x1-x2|+|y1-y2|

  = |0-2|+|2-0|

  = 2+2

  = 4
- Distance from p1 to p2 is 4
- And Distance from p2 to p1 is 4
- Similarly for the other points

Manhattan Distance 1D

A	B	C	D
1	5	7	9

$$d(\mathbf{x}, \mathbf{y}) = \left(\sum_{k=1}^{n} |x_k - y_k|^r \right)^{1/r}$$

Manhattan Distance d(A,B)= |1-5|=4

Dissimilarity Matrix

A(1)	0			
B(5)	4	0		
C(7)	6	2	0	
D(9)	8	4	2	0
	A(1)	B(5)	C(7)	D(9)

Min=0 Max=8

Normalized form of Dissimilarity Matrix

A(1)	0			
B(5)	0.5	0		
C(7)	0.75	0.25	0	
D(9)	1	0.5	0.25	0
	A(1)	B(5)	C(7)	D(9)

Similarity Matrix

A(1)	1			
B(5)	0.5	1		
C(7)	0.25	0.75	1	
D(9)	0	0.5	0.75	1
	A(1)	B(5)	C(7)	D(9)

Similarity = 1- Dissimilarity

Manhattan Distance 2D

A	B	C	D
(1,5)	(3,4)	(7,4)	(9,6)

$$d(\mathbf{x},\mathbf{y}) = \left(\sum_{k=1}^{n} |x_k - y_k|^r \right)^{1/r}$$

$$d(A,B) = (\sqrt{(1-3)^2 + (5-4)^2} = 2.24$$

Dissimilarity Matrix

A(1,5)	0			
B(3,4)	2.24	0		
C(7,4)	6.08	4.00	0	
D(9,6)	8.06	6.32	2.83	0
	A(1,5)	**B(3,4)**	**C(7,4)**	**D(9,6)**

Min=0, Max=8.06

Normalized form of Dissimilarity Matrix

A(1,5)	0			
B(3,4)	0.28	0		
C(7,4)	0.76	0.5	0	
D(9,6)	1.00	0.79	0.35	0
	A(1,5)	**B(3,4)**	**C(7,4)**	**D(9,6)**

Similarity Matrix

A(1)	1			
B(5)	0.72	1		
C(7)	0.24	0.50	1	
D(9)	0.00	0.21	0.65	1
	A(1)	**B(5)**	**C(7)**	**D(9)**

Manhattan Distance 3D

A(1,4,6)

B(5,7,8)

$$d(A,B) = (\sqrt{(1-5)^2 + (4-7)^2 + (6-8)^2} = 5.39$$

Binary Variables

- It has only two states
- **Symmetric**: if both of its states are equally valuable and carry same weight. For example, gender: Male and Female.
- **Asymmetric**: if both of its states are not equally important For example, positive and Negative outcome of disease test, examination.

Contingency Table for Binary Variable

		Object J		
		1	0	sum
Object I	1	q	r	q+r
	0	s	t	s+t
	sum	q+s	r+t	p=(q+r+s+t)

Binary Variable Example

Dissimilarity Measure

Object I=[1 0 0 1 0 1]

Object J=[0 1 1 0 0 1]

Symmetric Binary Variable

d(I, J)=(r+s)/(q+r+s+t)

Asymmetric Binary Variable

d(I, J)=(r+s)/(q+r+s)

		Object J		
Object I		1	0	sum
	1	1	2	3
	0	2	1	3
	sum	3	3	p=6

		Object J		
Object I		1	0	sum
	1	q	r	q+r
	0	s	t	s+t
	sum	q+s	r+t	p=(q+r+s+t)

sim(I,J)=1-d(I,J)

I, J: Symmetric	**I, J: Asymmetric**
d(I,J)=4/6=0.67	d(I,J)=4/5=0.8
Sim(I,J)=0.33	Sim(I,J)=0.2

Calculating Dissimilarity Between Asymmetric Binary Variables

Name	(M/F)	Fever	Cough	Test-I	Test-2	Test-3	Test-4
Jack	M	1	0	1	0	0	0
Mary	F	1	0	1	0	1	0
Jim	M	1	1	0	0	0	0

		Object J		
		1	0	sum
Object I	1	q	r	q+r
	0	s	t	s+t
	sum	q+s	r+t	p=(q+r+s+t)

		Mary		
		1	0	Sum
Jack	1	2	0	2
	0	1	3	4
	sum	3	3	p=6

		Jim		
		1	0	Sum
Jack	1	1	1	2
	0	1	3	4
	sum	2	4	p=6

- d(Jack, Mary)=(0+1)/(2+0+1)=1/3=0.33
- d(Jack, Jim)=(1+1)/(1+1+1)=2/3=0.67
- d(Mary, Jim)=(2+1)/(1+1+2)=3/4=0.75

Symmetric Binary Variable

d(I, J)=(r+s)/(q+r+s+t)

Asymmetric Binary Variable

d(I, J)=(r+s)/(q+r+s)

Nominal (Categorical) Variables

- It is generalization of binary variables. It takes more than two states.
- Dissimilarity measure is **d(i,j)=(p-m)/p**
  - **m** is number of matches
  - **p** is the number of variables

p =1
m is either 0 or 1

Name	Blood Pressure
Jack	High
Mary	Low
Jim	Medium

d(High, High)=(1-1)/1=0

d(Low, Low)=(1-1)/1=0

d(Medium, Medium)=(1-1)/1=0

d(High, Low)=(1-0)/1=1

d(High, Medium)=(1-0)/1=1

d(Low, Medium)=(1-0)/1=1

Jack	0		
Mary	1	0	
Jim	1	1	0
	Jack	Mary	Jim

Nominal Variable Example

Name	(Color1, Color2)
Jack	Red, Green
Mary	Red, Yellow
Jim	Blue, Blue

p=2

m is either 0 , 1 or 2

Jack	0		
Mary	0.5	0	
Jim	1	1	0
	Jack	**Mary**	**Jim**

d(Jack, Jack)=(2-2)/2=0

d(Jack, Mary)=(2-1)/2=0.5

d(Jack, Jim)=(2-0)/2=1

d(Mary, Jim)=(2-0)/2=1

Ordinal Variables

- It similar to categorical variables, but numbers have meaningful order. For example, Sports (Gold, Silver, Bronze).

$M_f=4$

$Z_{if}=r_{if}-1/(M_f-1)=r_{if}-1/3$

Name	Rank	r_{if}	Z_{if}
Jack	Excellent	4	(4-1)/3=1
Mary	Better	3	(3-1)/3=0.67
Jim	Good	2	(2-1)/3=0.33
Pat	Average	1	(1-1)/3=0

To get dissimilarity matrix, apply Euclidean or any other distance.

Jack(1)	0			
Mary(0.67)	0.33	0		
Jim(0.33)	0.67	0.34	0	
Pat(0)	1	0.67	0.33	0
	Jack(1)	**Mary(0.67)**	**Jim(0.33)**	**Pat(0)**

Z_{if} means normalized (0-1) rank value

Common Properties of a Distance

- Distances, such as the Euclidean distance, have some well known properties.

  1. $d(\mathbf{x}, \mathbf{y}) \geq 0$ for all *x* and *y* and $d(\mathbf{x}, \mathbf{y}) = 0$ if and only if $\mathbf{x} = \mathbf{y}$. (positive definiteness)
  2. $d(\mathbf{x}, \mathbf{y}) = d(\mathbf{y}, \mathbf{x})$ for all $\mathbf{x}$ and $\mathbf{y}$. (Symmetry)
  3. $d(\mathbf{x}, \mathbf{z}) \leq d(\mathbf{x}, \mathbf{y}) + d(\mathbf{y}, \mathbf{z})$ for all points $\mathbf{x}$, $\mathbf{y}$, and $\mathbf{z}$. (Triangle Inequality)

  where $d(\mathbf{x}, \mathbf{y})$ is the distance (dissimilarity) between points (data objects), $\mathbf{x}$ and $\mathbf{y}$.

- A distance that satisfies these properties is a metric. So, it can be used as a measure for dissimilarity

Common Properties of a Similarity

- Similarities, also have some well known properties.

  1. $s(\mathbf{x}, \mathbf{y}) = 1$ (or maximum similarity) only if $\mathbf{x} = \mathbf{y}$.
     (does not always hold, e.g., cosine)
  2. $s(\mathbf{x}, \mathbf{y}) = s(\mathbf{y}, \mathbf{x})$ for all $\mathbf{x}$ and $\mathbf{y}$. (Symmetry)

  where $s(\mathbf{x}, \mathbf{y})$ is the similarity between points (data objects), $\mathbf{x}$ and $\mathbf{y}$.

Similarity Between Binary Vectors

- Common situation is that objects, $\mathbf{x}$ and $\mathbf{y}$, have only binary attributes

- Compute similarities using the following quantities

  f_{01} = the number of attributes where $\mathbf{x}$ was 0 and $\mathbf{y}$ was 1

  f_{10} = the number of attributes where $\mathbf{x}$ was 1 and $\mathbf{y}$ was 0

  f_{00} = the number of attributes where $\mathbf{x}$ was 0 and $\mathbf{y}$ was 0

  f_{11} = the number of attributes where $\mathbf{x}$ was 1 and $\mathbf{y}$ was 1

- Simple Matching and Jaccard Coefficients

  SMC = number of matches / number of attributes

  $= (f_{11} + f_{00}) / (f_{01} + f_{10} + f_{11} + f_{00})$

  J = number of 11 matches / number of non-zero attributes

  $= (f_{11}) / (f_{01} + f_{10} + f_{11})$

SMC versus Jaccard: Example

$\mathbf{x} =$ 1 0 0 0 0 0 0 0 0 0

$\mathbf{y} =$ 0 0 0 0 0 0 1 0 0 1

$f_{01} = 2$ (the number of attributes where **x** was 0 and **y** was 1)

$f_{10} = 1$ (the number of attributes where **x** was 1 and **y** was 0)

$f_{00} = 7$ (the number of attributes where **x** was 0 and **y** was 0)

$f_{11} = 0$ (the number of attributes where **x** was 1 and **y** was 1)

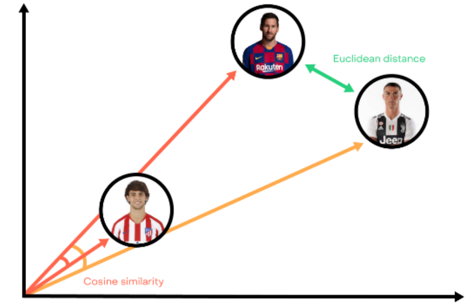
$$\text{SMC} = (f_{11} + f_{00}) / (f_{01} + f_{10} + f_{11} + f_{00})$$
$$= (0+7) / (2+1+0+7) = 0.7$$

$$J = (f_{11}) / (f_{01} + f_{10} + f_{11}) = 0 / (2 + 1 + 0) = 0$$

SMC & JC

- SMC counts both presences and absences equally. Used for objects with symmetric binary attributes.
- Can be used to find students who answered similarly in a test – True/False questions.
- JC is used to handle objects with asymmetric binary attributes.
- Ex:
- No. of products not purchased is far more than purchased.
- SMC would say all transactions are very similar.
- Use JC

Cosine Similarity



- If $\mathbf{d}_1$ and $\mathbf{d}_2$ are two document vectors, then

$$\cos(\mathbf{d}_1, \mathbf{d}_2) = \langle \mathbf{d}_1, \mathbf{d}_2 \rangle / \|\mathbf{d}_1\| \, \|\mathbf{d}_2\| ,$$

where $\langle \mathbf{d}_1, \mathbf{d}_2 \rangle$ indicates inner product or vector dot product of vectors, $\mathbf{d}_1$ and $\mathbf{d}_2$, and $\| \mathbf{d} \|$ is the length of vector $\mathbf{d}$.

- Example:

$$\cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\| \|\mathbf{B}\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2} \sqrt{\sum_{i=1}^{n} B_i^2}}$$

$\mathbf{d}_1 = \mathbf{3\ 2\ 0\ 5\ 0\ 0\ 0\ 2\ 0\ 0}$

$\mathbf{d}_2 = \mathbf{1\ 0\ 0\ 0\ 0\ 0\ 0\ 1\ 0\ 2}$

$\langle \mathbf{d}_1, \mathbf{d2} \rangle = 3*1 + 2*0 + 0*0 + 5*0 + 0*0 + 0*0 + 0*0 + 2*1 + 0*0 + 0*2 = 5$

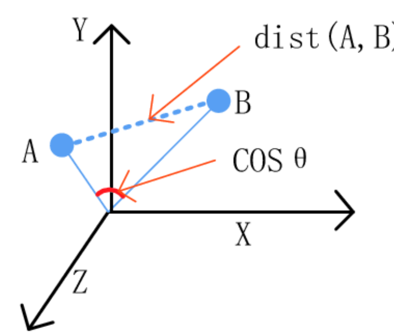
$| \mathbf{d}_1 \| = (3*3+2*2+0*0+5*5+0*0+0*0+0*0+2*2+0*0+0*0)^{0.5} = (42)^{0.5} = 6.481$

$\| \mathbf{d}_2 \| = (1*1+0*0+0*0+0*0+0*0+0*0+0*0+1*1+0*0+2*2)^{0.5} = (6)^{0.5} = 2.449$

$\cos(\mathbf{d}_1, \mathbf{d}_2) = 0.3150$

$\cos(\mathbf{d}_1, \mathbf{d}_2)=0$ indicates both are dissimilar

$\cos(\mathbf{d}_1, \mathbf{d}_2)=1$ indicates both are similar



Extended Jaccard Coefficient (Tanimoto)

- Variation of JC
- Used for document data
- Reduces to Jaccard for binary attribute

$$T(p, q) = \frac{p \bullet q}{\|p\|^2 + \|q\|^2 - p \bullet q}$$

Correlation measures the linear relationship between objects

$$\text{corr}(\mathbf{x}, \mathbf{y}) = \frac{\text{covariance}(\mathbf{x}, \mathbf{y})}{\text{standard\_deviation}(\mathbf{x}) * \text{standard\_deviation}(\mathbf{y})} = \frac{s_{xy}}{s_x \; s_y}, \quad (2.11$$

where we are using the following standard statistical notation and definitions

$$\text{covariance}(\mathbf{x}, \mathbf{y}) = s_{xy} = \frac{1}{n-1} \sum_{k=1}^{n} (x_k - \overline{x})(y_k - \overline{y}) \quad (2.12$$

$$\text{standard\_deviation}(\mathbf{x}) = s_x = \sqrt{\frac{1}{n-1} \sum_{k=1}^{n} (x_k - \overline{x})^2}$$

$$\text{standard\_deviation}(\mathbf{y}) = s_y = \sqrt{\frac{1}{n-1} \sum_{k=1}^{n} (y_k - \overline{y})^2}$$

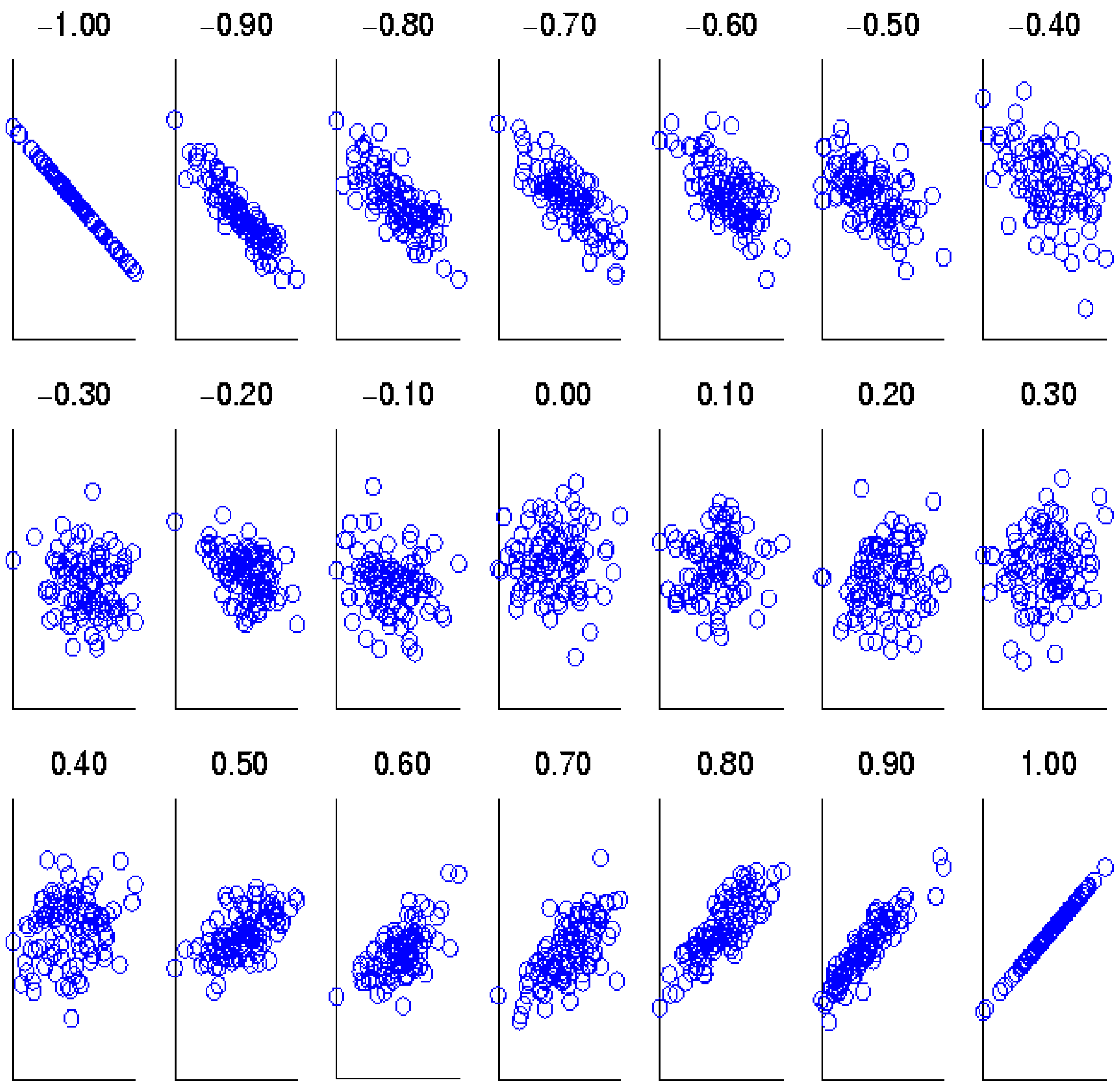
$$\overline{x} = \frac{1}{n} \sum_{k=1}^{n} x_k \text{ is the mean of } \mathbf{x}$$

$$\overline{y} = \frac{1}{n} \sum_{k=1}^{n} y_k \text{ is the mean of } \mathbf{y}$$

Pearson's correlation

- If correlation between two variables x and y is -1, they are negatively correlated.
  - If one increases, the other decreases and vice versa.
- If correlation between two variables x and y is +1, they are positively correlated.
  - Either both increase or both decrease.

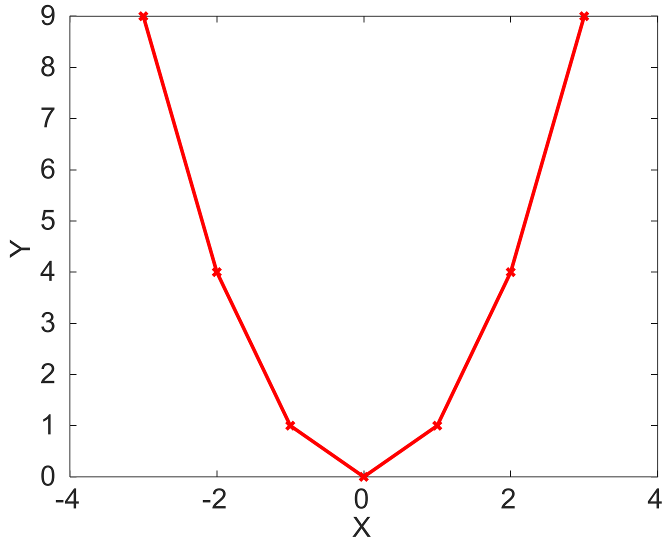
Visually Evaluating Correlation



Scatter plots showing the similarity from –1 to 1.

Drawback of Correlation

- $\mathbf{x} = (-3, -2, -1, 0, 1, 2, 3)$
- $\mathbf{y} = (9, 4, 1, 0, 1, 4, 9)$



$y_i = x_i^2$

- $\text{mean}(\mathbf{x}) = 0$, $\text{mean}(\mathbf{y}) = 4$
- $\text{std}(\mathbf{x}) = 2.16$, $\text{std}(\mathbf{y}) = 3.74$

- corr = (-3)(5)+(-2)(0)+(-1)(-3)+(0)(-4)+(1)(-3)+(2)(0)+3(5) / (6 * 2.16 * 3.74)

  = 0

Correlation vs Cosine vs Euclidean Distance

- Compare the three proximity measures according to their behavior under variable transformation
  - scaling: multiplication by a value
  - translation: adding a constant

Property	Cosine	Correlation	Euclidean Distance
Invariant to scaling (multiplication)	Yes	Yes	No
Invariant to translation (addition)	No	Yes	No

- Consider the example
  - $\mathbf{x} = (1, 2, 4, 3, 0, 0, 0)$, $\mathbf{y} = (1, 2, 3, 4, 0, 0, 0)$
  - $\mathbf{y_s = y * 2}$ (scaled version of y), $\mathbf{y_t = y + 5}$ (translated version)

Measure	(x, y)	(x, y_s)	(x, y_t)
Cosine	0.9667	0.9667	0.7940
Correlation	0.9429	0.9429	0.9429
Euclidean Distance	1.4142	5.8310	14.2127

Correlation vs cosine vs Euclidean distance

- Choice of the right proximity measure depends on the domain
- What is the correct choice of proximity measure for the following situations?
  - Comparing documents using the frequencies of words
    - Documents are considered similar if the word frequencies are similar
  - Comparing the temperature in Celsius of two locations
    - Two locations are considered similar if the temperatures are similar in magnitude
  - Comparing two time series of temperature measured in Celsius
    - Two time series are considered similar if their “shape” is similar, i.e., they vary in the same way over time, achieving minimums and maximums at similar times, etc.

Comparison of Proximity Measures

- Domain of application
  - Similarity measures tend to be specific to the type of attribute and data
  - Record data, images, graphs, sequences, 3D-protein structure, etc. tend to have different measures
- However, one can talk about various properties that you would like a proximity measure to have
  - Symmetry is a common one
  - Tolerance to noise and outliers is another
  - Ability to find more types of patterns?
  - Many others possible
- The measure must be applicable to the data and produce results that agree with domain knowledge

Information Based Measures

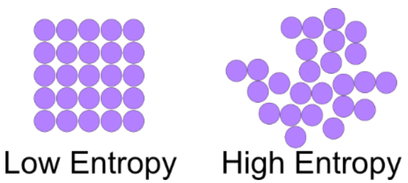
- Information theory is a well-developed and fundamental disciple with broad applications

- Some similarity measures are based on information theory
  - Mutual information in various versions
  - Maximal Information Coefficient (MIC) and related measures
  - General and can handle non-linear relationships
  - Can be complicated and time intensive to compute

Information and Probability

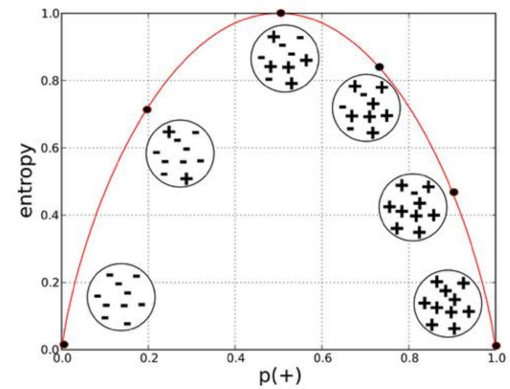
- Information relates to possible outcomes of an event
  - transmission of a message, flip of a coin, or measurement of a piece of data

- The more certain an outcome, the less information that it contains and vice-versa
  - For example, if a coin has two heads, then an outcome of heads provides no information
  - More quantitatively, the information is related the probability of an outcome
    - The smaller the probability of an outcome, the more information it provides and vice-versa
  - Entropy is the commonly used measure



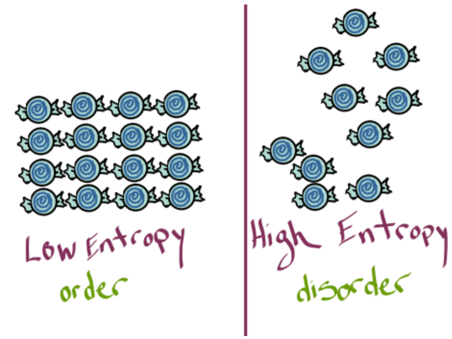
Entropy

https://planetcalc.com/2476/



- For
  - a variable (event), X,
  - with n possible values (outcomes), $x_1, x_2 \ldots, x_n$
  - each outcome having probability, $p_1, p_2 \ldots, p_n$
  - the entropy of X , $H(X)$, is given by

$$H(X) = -\sum_{i=1}^{n} p_i \log_2 p_i$$



- Entropy is between 0 and $\log_2 n$ and is measured in bits
  - Thus, entropy is a measure of how many bits it takes to represent an observation of X on average

Entropy Examples

- For a coin with probability p of heads and probability $q = 1 - p$ of tails

$$H = -p \log_2 p - q \log_2 q$$

  - For $p = 0.5$, $q = 0.5$ (fair coin) $H = 1$
  - For $p = 1$ or $q = 1$, $H = 0$

- What is the entropy of a fair four-sided die?

Entropy for Sample Data: Example

Hair Color	Count	p	$-p\log_2 p$
Black	75	0.75	0.3113
Brown	15	0.15	0.4105
Blond	5	0.05	0.2161
Red	0	0.00	0
Other	5	0.05	0.2161
Total	100	1.0	1.1540

Maximum entropy is $\log_2 5 = 2.3219$

Entropy for Sample Data

- Suppose we have
  - a number of observations (m) of some attribute, X, e.g., the hair color of students in the class,
  - where there are n different possible values
  - And the number of observation in the i^{th} category is m_i
  - Then, for this sample

$$H(X) = -\sum_{i=1}^{n} \frac{m_i}{m} \log_2 \frac{m_i}{m}$$

- For continuous data, the calculation is harder

Mutual Information

- Information one variable provides about another

Formally, $I(X,Y) = H(X) + H(Y) - H(X,Y)$, where

$H(X,Y)$ is the joint entropy of X and Y,

$$H(X,Y) = -\sum_i \sum_j p_{ij} \log_2 p_{ij}$$

Where p_{ij} is the probability that the i^{th} value of X and the j^{th} value of Y occur together

- For discrete variables, this is easy to compute

- Maximum mutual information for discrete variables is $\log_2(\min(n_X, n_Y))$, where n_X (n_Y) is the number of values of X (Y)

Mutual Information Example

Student Status	Count	p	$-p\log_2 p$
Undergrad	45	0.45	0.5184
Grad	55	0.55	0.4744
Total	100	1.00	0.9928

Grade	Count	p	$-p\log_2 p$
A	35	0.35	0.5301
B	50	0.50	0.5000
C	15	0.15	0.4105
Total	100	1.00	1.4406

Student Status	Grade	Count	p	$-p\log_2 p$
Undergrad	A	5	0.05	0.2161
Undergrad	B	30	0.30	0.5211
Undergrad	C	10	0.10	0.3322
Grad	A	30	0.30	0.5211
Grad	B	20	0.20	0.4644
Grad	C	5	0.05	0.2161
Total		100	1.00	2.2710

Mutual information of Student Status and Grade = 0.9928 + 1.4406 - 2.2710 = 0.1624

References

1. https://www.analyticsvidhya.com/blog/2021/04/proximity-measures-in-data-mining-and-machine-learning/