



MEASURES

BEIYU LIN



DATA QUALITY

- Data quality problems:
 - Noise and outliers (example in yellow box)
 - Missing values (in red box)
 - Duplicate data (in green box)

<i>Tid</i>	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	10000K	Yes
6	No	NULL	60K	No
7	Yes	Divorced	220K	NULL
8	No	Single	85K	Yes
9	No	Married	90K	No
9	No	Single	90K	No

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

Proximity refers to a similarity or dissimilarity

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	$d = x - y / (n - 1)$ (values mapped to integers 0 to $n-1$, where n is the number of values)	$s = 1 - d$
Interval or Ratio	$d = x - y $	$s = -d, s = \frac{1}{1+d}, s = e^{-d},$ $s = 1 - \frac{d - \min_d}{\max_d - \min_d}$

EUCLIDEAN DISTANCE

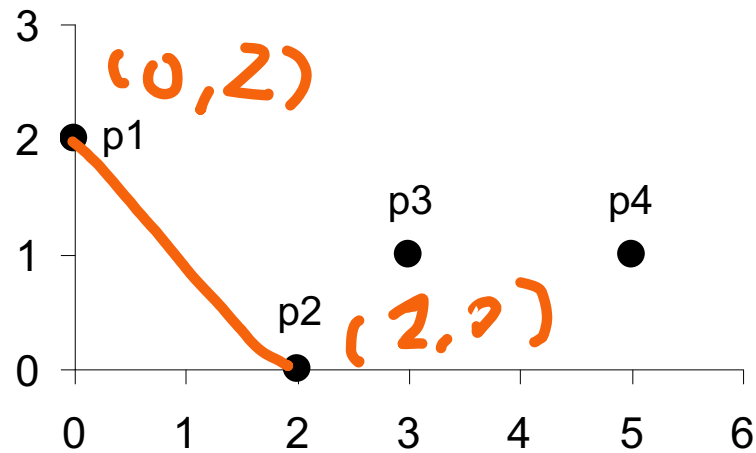
- Euclidean Distance

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

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.

EUCLIDEAN DISTANCE



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

$$\begin{aligned} &\sqrt{2^2 + 2^2} \\ &= 2\sqrt{2} = 1.414 \\ &\approx 2.828 \end{aligned}$$

	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

Euclidean Distance Matrix

EUCLIDEAN DISTANCE

- Euclidean Distance

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

MINKOWSKI DISTANCE

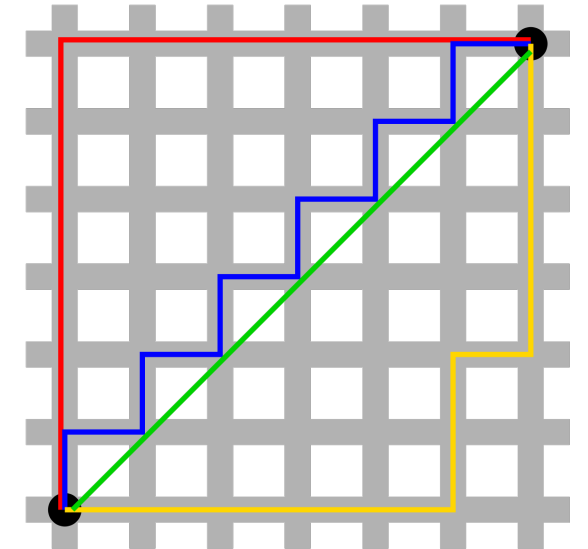
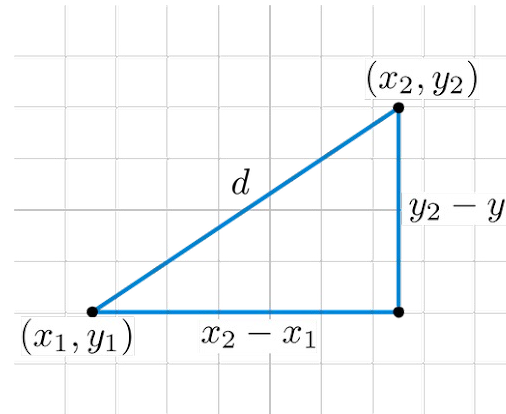
- Minkowski Distance is a **generalization** of Euclidean Distance

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

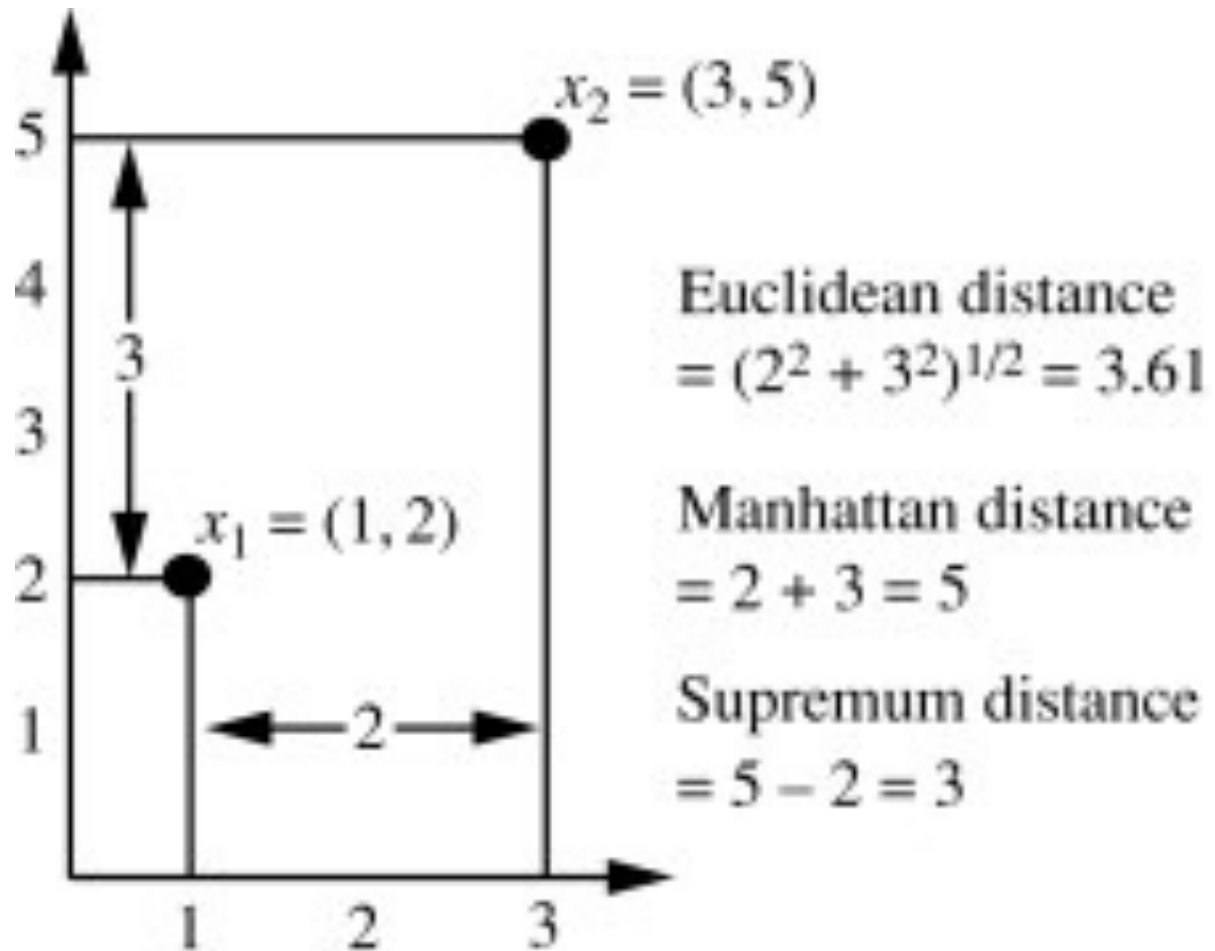
Where r is a parameter, n is the number of dimensions (attributes) and x_k and y_k are, respectively, the k^{th} attributes (components) or data objects \mathbf{x} and \mathbf{y} .

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



EXAMPLE FOR DIFFERENT DISTANCES



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

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 \mathbf{x} and \mathbf{y} and $d(\mathbf{x}, \mathbf{y}) = 0$ if and only if $\mathbf{x} = \mathbf{y}$.
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**

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, x and y , have only binary attributes

- Compute similarities using the following quantities

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

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

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

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

- Simple Matching and Jaccard Coefficients

SMC = number of matches / number of attributes

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

mutual presences

+ mutual absence

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

mutual presences

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

SMC VERSUS JACCARD: EXAMPLE

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

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

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

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

$$f_{00} = 7 \quad (\text{the number of attributes where } \mathbf{x} \text{ was } 0 \text{ and } \mathbf{y} \text{ was } 0)$$

$$f_{11} = 0 \quad (\text{the number of attributes where } \mathbf{x} \text{ was } 1 \text{ and } \mathbf{y} \text{ was } 1)$$

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

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

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:

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

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

$$\langle \mathbf{d}_1, \mathbf{d}_2 \rangle = 3*1 + 2*0 + 0*0 + 5*0 + 0*0 + 0*0 + 0*0 + 2*1 + 0*0 + 0*2 = 5$$

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

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

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

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

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

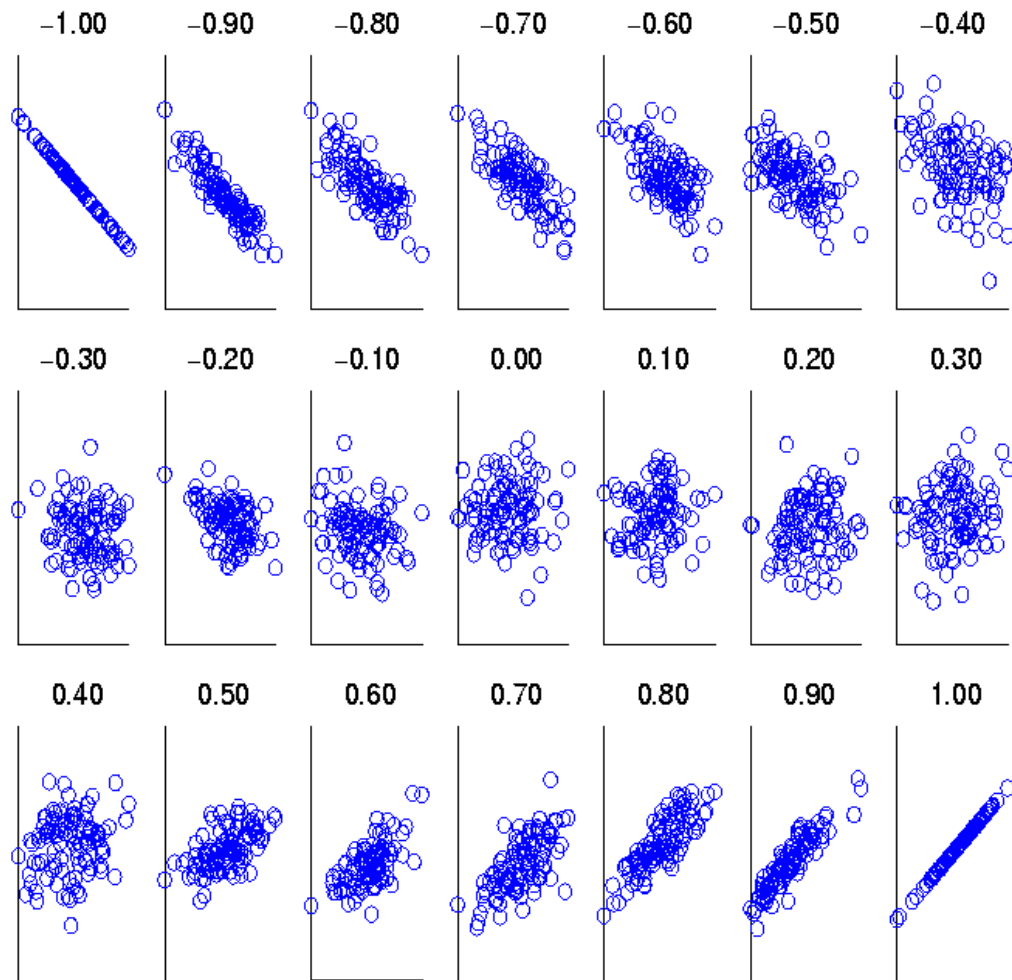
$$\text{standard_deviation}(\mathbf{x}) = s_x = \sqrt{\frac{1}{n-1} \sum_{k=1}^n (x_k - \bar{x})^2}$$

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

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

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

VISUALLY EVALUATING CORRELATION



Scatter plots showing the similarity from -1 to 1 .



- Compare the three proximity measures according to their behavior under variable transformation

CORRELATION VS COSINE VS EUCLIDEAN DISTANCE

- 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
 - $\mathbf{x} = (1, 2, 4, 3, 0, 0, 0)$, $\mathbf{y} = (1, 2, 3, 4, 0, 0, 0)$
 - $\mathbf{y}_s = \mathbf{y} * 2$ (scaled version of \mathbf{y}), $\mathbf{y}_t = \mathbf{y} + 5$ (translated version)

Measure	(\mathbf{x}, \mathbf{y})	$(\mathbf{x}, \mathbf{y}_s)$	$(\mathbf{x}, \mathbf{y}_t)$
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 discipline 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

- The more certain an outcome, the less information that it carries
 - For example, if a coin has two heads, then an outcome of head
 - 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

■ For a variable (event), X ,

- with n possible values (outcomes), x_1, x_2, \dots, x_n
- each outcome having probability, p_1, p_2, \dots, 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

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)

MUTU

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

MP

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 = $0.9928 + 1.4406 - 2.2710 = 0.1624$

MAXIMAL INFORMATION COEFFICIENT

- Reshef, David N., Yakir A. Reshef, Hilary K. Finucane, Sharon R. Grossman, Gilean McVean, Peter J. Turnbaugh, Eric S. Lander, Michael Mitzenmacher, and Pardis C. Sabeti. "Detecting novel associations in large data sets." *science* 334, no. 6062 (2011): 1518-1524.
- Applies mutual information to two continuous variables
- Consider the possible binnings of the variables into discrete categories
 - $n_X \times n_Y \leq N^{0.6}$ where
 - n_X is the number of values of X
 - n_Y is the number of values of Y
 - N is the number of samples (observations, data objects)
- Compute the mutual information
 - Normalized by $\log_2(\min(n_X, n_Y))$
- Take the highest value

- Sometimes attributes are of many different types, but an overall similarity is needed.

1: For the k^{th} attribute, compute a similarity, $s_k(\mathbf{x}, \mathbf{y})$, in the range $[0, 1]$.

2: Define an indicator variable, δ_k , for the k^{th} attribute as follows:

$\delta_k = 0$ if the k^{th} attribute is an asymmetric attribute and

both objects have a value of 0, or if one of the objects has a missing value for the k^{th} attribute

$\delta_k = 1$ otherwise

3. Compute $\text{similarity}(\mathbf{x}, \mathbf{y}) = \frac{\sum_{k=1}^n \delta_k s_k(\mathbf{x}, \mathbf{y})}{\sum_{k=1}^n \delta_k}$

USING WEIGHTS TO COMBINE SIMILARITIES

- May not want to treat all attributes the same.

- Use non-negative weights ω_k

- $similarity(\mathbf{x}, \mathbf{y}) = \frac{\sum_{k=1}^n \omega_k \delta_k s_k(\mathbf{x}, \mathbf{y})}{\sum_{k=1}^n \omega_k \delta_k}$

- Can also define a weighted Minkowski distance

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