CSCI 347: Introduction to Data Mining

Week 4b - Categorical Data

BUT FIRST…. RETRO ON HOMEWORK 1

Estimated variance of X_j: $\hat{\sigma}_j^2 = \dfrac{1}{n-1}\sum_{i=1}^{n}(x_{ij} - \hat{\mu}_j)^2$

Estimated standard deviation of X_j: $\hat{\sigma}_j = \sqrt{\hat{\sigma}_j^2}$

Covariance of X_i and X_j : $\hat{\sigma}_{ij} = \dfrac{1}{n-1}\sum_{k=1}^{n}(x_{ki} - \hat{\mu}_i)(x_{kj} - \hat{\mu}_j)$

Person's correlation coefficient of X_i and X_j : $\hat{\rho}_{ij} = \dfrac{\hat{\sigma}_{ij}}{\hat{\sigma}_i\hat{\sigma}_j}$

COMMON DATA TYPES

- Data is most often either *numerical* or *categorical*

$D =$

	temperature	length	type	weight
specimen 1	0.2	23	A	5.7
specimen 2	0.4	1	B	5.4
specimen 3	1.8	0.5	C	5.2
specimen 4	5.6	50	A	5.1
specimen 5	−0.5	34	A	5.3
specimen 6	0.4	19	B	5.4
specimen 7	1.1	11	A	5.5

RECALL: EUCLIDEAN DISTANCE

L_2 norm:

$$\|x_i - x_j\|_2 = \sqrt{\sum_{k=1}^{m} (x_{ik} - x_{jk})^2}$$ **where x_i and x_j are vectors, and there are m dimensions**

$$D = \begin{array}{cccc} & X_1 & X_2 & X_3 \\ x_1 & 0.2 & 23 & 5.7 \\ x_2 & 0.4 & 1 & 5.4 \\ x_3 & 1.8 & 0.5 & 5.2 \\ x_4 & 5.6 & 50 & 5.1 \\ x_5 & -0.5 & 34 & 5.3 \\ x_6 & 0.4 & 19 & 5.4 \\ x_7 & 1.1 & 11 & 5.5 \end{array}$$

$$\|x_1 - x_2\|_2 = \sqrt{\sum_{k=1}^{3} (x_{1k} - x_{2k})^2}$$

$$= \sqrt{(x_{11} - x_{21})^2 + (x_{12} - x_{22})^2 + (x_{13} - x_{23})^2}$$

$$= \sqrt{(0.2 - 0.4)^2 + (23 - 1)^2 + (5.7 - 5.4)^2}$$

$$= \sqrt{(-0.2)^2 + (22)^2 + (0.3)^2}$$

$$= 22.0$$

WHAT IF WE ALSO HAVE CATEGORICAL VARIABLES

L_2 **norm:**

$$\|x_i - x_j\|_2 = \sqrt{\sum_{k=1}^{m} (x_{ik} - x_{jk})^2}$$ **where x_i and x_j are vectors, and there are m dimensions**

$$D = \begin{array}{c|cccc} & X_1 & X_2 & X_3 & X_4 \\ x_1 & 0.2 & 23 & 5.7 & A \\ x_2 & 0.4 & 1. & 5.4 & B \\ x_3 & 1.8 & 0.5 & 5.2 & C \\ x_4 & 5.6 & 50 & 5.1 & A \\ x_5 & -0.5 & 34 & 5.3 & B \\ x_6 & 0.4 & 19 & 5.4 & C \\ x_7 & 1.1 & 11 & 5.5 & C \end{array}$$

$$\|x_1 - x_2\|_2 = \sqrt{\sum_{k=1}^{4} (x_{1k} - x_{2k})^2}$$

$$= \sqrt{(x_{11} - x_{21})^2 + (x_{12} - x_{22})^2 + (x_{13} - x_{23})^2 + (x_{14} - x_{24})^2}$$

$$= \sqrt{(0.2 - 0.4)^2 + (23 - 1)^2 + (5.7 - 5.4)^2 + (A - B)^2}$$

LABEL ENCODING

L_2 norm:

$$\|x_i - x_j\|_2 = \sqrt{\sum_{k=1}^{m} (x_{ik} - x_{jk})^2}$$ **where x_i and x_j are vectors, and there are m dimensions**

$A => 0$

$B => 1$

$C => 2$

$$D = \begin{array}{c|cccc} & X_1 & X_2 & X_3 & X_4 \\ x_1 & 0.2 & 23 & 5.7 & A \\ x_2 & 0.4 & 1. & 5.4 & B \\ x_3 & 1.8 & 0.5 & 5.2 & C \\ x_4 & 5.6 & 50 & 5.1 & A \\ x_5 & -0.5 & 34 & 5.3 & B \\ x_6 & 0.4 & 19 & 5.4 & C \\ x_7 & 1.1 & 11 & 5.5 & C \end{array}$$

$$\|x_1 - x_2\|_2 = \sqrt{\sum_{k=1}^{4} (x_{1k} - x_{2k})^2}$$

$$= \sqrt{(x_{11} - x_{21})^2 + (x_{12} - x_{22})^2 + (x_{13} - x_{23})^2 + (x_{14} - x_{24})^2}$$

$$= \sqrt{(0.2 - 0.4)^2 + (23 - 1)^2 + (5.7 - 5.4)^2 + (A - B)^2}$$

LABEL ENCODING

L_2 **norm:**

$$\|x_i - x_j\|_2 = \sqrt{\sum_{k=1}^{m} (x_{ik} - x_{jk})^2}$$ **where x_i and x_j are vectors, and there are m dimensions**

$A => 0$

$B => 1$

$C => 2$

$$D = \begin{array}{c|cccc} & X_1 & X_2 & X_3 & X_4 \\ x_1 & 0.2 & 23 & 5.7 & 0 \\ x_2 & 0.4 & 1. & 5.4 & 1 \\ x_3 & 1.8 & 0.5 & 5.2 & 2 \\ x_4 & 5.6 & 50 & 5.1 & 0 \\ x_5 & -0.5 & 34 & 5.3 & 1 \\ x_6 & 0.4 & 19 & 5.4 & 2 \\ x_7 & 1.1 & 11 & 5.5 & 2 \end{array}$$

$$\|x_1 - x_2\|_2 = \sqrt{\sum_{k=1}^{4} (x_{1k} - x_{2k})^2}$$

$$= \sqrt{(x_{11} - x_{21})^2 + (x_{12} - x_{22})^2 + (x_{13} - x_{23})^2 + (x_{14} - x_{24})^2}$$

$$= \sqrt{(0.2 - 0.4)^2 + (23 - 1)^2 + (5.7 - 5.4)^2 + (A - B)^2}$$

LABEL ENCODING

L_2 **norm:**

$$\|x_i - x_j\|_2 = \sqrt{\sum_{k=1}^{m} (x_{ik} - x_{jk})^2}$$ **where x_i and x_j are vectors, and there are m dimensions**

$A => 0$

$B => 1$

$C => 2$

$$D = \begin{array}{c|cccc} & X_1 & X_2 & X_3 & X_4 \\ x_1 & 0.2 & 23 & 5.7 & 0 \\ x_2 & 0.4 & 1. & 5.4 & 1 \\ x_3 & 1.8 & 0.5 & 5.2 & 2 \\ x_4 & 5.6 & 50 & 5.1 & 0 \\ x_5 & -0.5 & 34 & 5.3 & 1 \\ x_6 & 0.4 & 19 & 5.4 & 2 \\ x_7 & 1.1 & 11 & 5.5 & 2 \end{array}$$

$$\|x_1 - x_2\|_2 = \sqrt{\sum_{k=1}^{4} (x_{1k} - x_{2k})^2}$$

$$= \sqrt{(x_{11} - x_{21})^2 + (x_{12} - x_{22})^2 + (x_{13} - x_{23})^2 + (x_{14} - x_{24})^2}$$

$$= \sqrt{(0.2 - 0.4)^2 + (23 - 1)^2 + (5.7 - 5.4)^2 + (0 - 2)^2}$$

$$= \sqrt{(0.2)^2 + (22)^2 + (0.3)^2 + (-2)^2}$$

$$= 22.09$$

PROBLEM WITH LABEL ENCODING AND DIST

L_2 **norm:**

$$\|x_i - x_j\|_2 = \sqrt{\sum_{k=1}^{m} (x_{ik} - x_{jk})^2}$$ **where x_i and x_j are vectors, and there are m dimensions**

A => 0

B => 1

C => 2

$$D = \begin{array}{c|cccc} & X_1 & X_2 & X_3 & X_4 \\ \hline x_1 & 0.2 & 23 & 5.7 & 0 \\ x_2 & 0.4 & 1. & 5.4 & 1 \\ x_3 & 1.8 & 0.5 & 5.2 & 2 \\ x_4 & 5.6 & 50 & 5.1 & 0 \\ x_5 & -0.5 & 34 & 5.3 & 1 \\ x_6 & 0.4 & 19 & 5.4 & 2 \\ x_7 & 1.1 & 11 & 5.5 & 2 \end{array}$$

$$\|x_1 - x_2\|_2 = \sqrt{\sum_{k=1}^{4} (x_{1k} - x_{2k})^2}$$

$$= \sqrt{(x_{11} - x_{21})^2 + (x_{12} - x_{22})^2 + (x_{13} - x_{23})^2 + (x_{14} - x_{24})^2}$$

$$= \sqrt{(0.2 - 0.4)^2 + (23 - 1)^2 + (5.7 - 5.4)^2 + (0 - 2)^2}$$

$$= \sqrt{(0.2)^2 + (22)^2 + (0.3)^2 + (-2)^2}$$

$$= 22.09$$

ONE-HOT ENCODING

$$\|x_i - x_j\|_2 = \sqrt{\sum_{k=1}^{m} (x_{ik} - x_{jk})^2}$$ **where x_i and x_j are vectors, and there are m dimensions**

$$D = \begin{array}{c|cccc} & X_1 & X_2 & X_3 & X_4 \\ x_1 & 0.2 & 23 & 5.7 & A \\ x_2 & 0.4 & 1. & 5.4 & B \\ x_3 & 1.8 & 0.5 & 5.2 & C \\ x_4 & 5.6 & 50 & 5.1 & A \\ x_5 & -0.5 & 34 & 5.3 & B \\ x_6 & 0.4 & 19 & 5.4 & C \\ x_7 & 1.1 & 11 & 5.5 & C \end{array} \longrightarrow D = \begin{array}{c|cccccc} & X_1 & X_2 & X_3 & X_{4A} & X_{4B} & X_{4C} \\ x_1 & 0.2 & 23 & 5.7 & 1 & 0 & 0 \\ x_2 & 0.4 & 1. & 5.4 & 0 & 1 & 0 \\ x_3 & 1.8 & 0.5 & 5.2 & 0 & 0 & 1 \\ x_4 & 5.6 & 50 & 5.1 & 1 & 0 & 0 \\ x_5 & -0.5 & 34 & 5.3 & 0 & 1 & 0 \\ x_6 & 0.4 & 19 & 5.4 & 0 & 0 & 1 \\ x_7 & 1.1 & 11 & 5.5 & 0 & 0 & 1 \end{array}$$

ONE-HOT ENCODING

$$\|x_i - x_j\|_2 = \sqrt{\sum_{k=1}^{m} (x_{ik} - x_{jk})^2}$$ **where x_i and x_j are vectors, and there are m dimensions**

$$D = \begin{array}{c|cccc} & X_1 & X_2 & X_3 & X_4 \\ x_1 & 0.2 & 23 & 5.7 & A \\ x_2 & 0.4 & 1. & 5.4 & B \\ x_3 & 1.8 & 0.5 & 5.2 & C \\ x_4 & 5.6 & 50 & 5.1 & A \\ x_5 & -0.5 & 34 & 5.3 & B \\ x_6 & 0.4 & 19 & 5.4 & C \\ x_7 & 1.1 & 11 & 5.5 & C \end{array} \longrightarrow D = \begin{array}{c|cccccc} & X_1 & X_2 & X_3 & X_{4A} & X_{4B} & X_{4C} \\ x_1 & 0.2 & 23 & 5.7 & 1 & 0 & 0 \\ x_2 & 0.4 & 1. & 5.4 & 0 & 1 & 0 \\ x_3 & 1.8 & 0.5 & 5.2 & 0 & 0 & 1 \\ x_4 & 5.6 & 50 & 5.1 & 1 & 0 & 0 \\ x_5 & -0.5 & 34 & 5.3 & 0 & 1 & 0 \\ x_6 & 0.4 & 19 & 5.4 & 0 & 0 & 1 \\ x_7 & 1.1 & 11 & 5.5 & 0 & 0 & 1 \end{array}$$

ONE-HOT ENCODING

$$\|x_i - x_j\|_2 = \sqrt{\sum_{k=1}^{m} (x_{ik} - x_{jk})^2}$$ **where x_i and x_j are vectors, and there are m dimensions**

$$D = \begin{array}{c|cccc} & X_1 & X_2 & X_3 & X_4 \\ x_1 & 0.2 & 23 & 5.7 & A \\ x_2 & 0.4 & 1. & 5.4 & B \\ x_3 & 1.8 & 0.5 & 5.2 & C \\ x_4 & 5.6 & 50 & 5.1 & A \\ x_5 & -0.5 & 34 & 5.3 & B \\ x_6 & 0.4 & 19 & 5.4 & C \\ x_7 & 1.1 & 11 & 5.5 & C \end{array} \longrightarrow D = \begin{array}{c|cccccc} & X_1 & X_2 & X_3 & X_{4A} & X_{4B} & X_{4C} \\ x_1 & 0.2 & 23 & 5.7 & 1 & 0 & 0 \\ x_2 & 0.4 & 1. & 5.4 & 0 & 1 & 0 \\ x_3 & 1.8 & 0.5 & 5.2 & 0 & 0 & 1 \\ x_4 & 5.6 & 50 & 5.1 & 1 & 0 & 0 \\ x_5 & -0.5 & 34 & 5.3 & 0 & 1 & 0 \\ x_6 & 0.4 & 19 & 5.4 & 0 & 0 & 1 \\ x_7 & 1.1 & 11 & 5.5 & 0 & 0 & 1 \end{array}$$

ONE-HOT ENCODING

$$\|x_i - x_j\|_2 = \sqrt{\sum_{k=1}^{m} (x_{ik} - x_{jk})^2}$$ where x_i and x_j are vectors, and there are m dimensions

$$D = \begin{array}{c|cccc} & X_1 & X_2 & X_3 & X_4 \\ x_1 & 0.2 & 23 & 5.7 & A \\ x_2 & 0.4 & 1. & 5.4 & B \\ x_3 & 1.8 & 0.5 & 5.2 & C \\ x_4 & 5.6 & 50 & 5.1 & A \\ x_5 & -0.5 & 34 & 5.3 & B \\ x_6 & 0.4 & 19 & 5.4 & C \\ x_7 & 1.1 & 11 & 5.5 & C \end{array} \longrightarrow D = \begin{array}{c|cccccc} & X_1 & X_2 & X_3 & X_{4A} & X_{4B} & X_{4C} \\ x_1 & 0.2 & 23 & 5.7 & 1 & 0 & 0 \\ x_2 & 0.4 & 1. & 5.4 & 0 & 1 & 0 \\ x_3 & 1.8 & 0.5 & 5.2 & 0 & 0 & 1 \\ x_4 & 5.6 & 50 & 5.1 & 1 & 0 & 0 \\ x_5 & -0.5 & 34 & 5.3 & 0 & 1 & 0 \\ x_6 & 0.4 & 19 & 5.4 & 0 & 0 & 1 \\ x_7 & 1.1 & 11 & 5.5 & 0 & 0 & 1 \end{array}$$

ONE-HOT ENCODING

$$\|x_i - x_j\|_2 = \sqrt{\sum_{k=1}^{m} (x_{ik} - x_{jk})^2}$$ **where x_i and x_j are vectors, and there are m dimensions**

$$D = \begin{array}{c} \\ x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \\ x_6 \\ x_7 \end{array} \begin{array}{cccc} X_1 & X_2 & X_3 & X_4 \\ 0.2 & 23 & 5.7 & A \\ 0.4 & 1. & 5.4 & B \\ 1.8 & 0.5 & 5.2 & C \\ 5.6 & 50 & 5.1 & A \\ -0.5 & 34 & 5.3 & B \\ 0.4 & 19 & 5.4 & C \\ 1.1 & 11 & 5.5 & C \end{array} \longrightarrow D = \begin{array}{c} \\ x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \\ x_6 \\ x_7 \end{array} \begin{array}{cccccc} X_1 & X_2 & X_3 & X_{4A} & X_{4B} & X_{4C} \\ 0.2 & 23 & 5.7 & 1 & 0 & 0 \\ 0.4 & 1. & 5.4 & 0 & 1 & 0 \\ 1.8 & 0.5 & 5.2 & 0 & 0 & 1 \\ 5.6 & 50 & 5.1 & 1 & 0 & 0 \\ -0.5 & 34 & 5.3 & 0 & 1 & 0 \\ 0.4 & 19 & 5.4 & 0 & 0 & 1 \\ 1.1 & 11 & 5.5 & 0 & 0 & 1 \end{array}$$

$$\|x_1 - x_2\|_2 = \sqrt{\sum_{k=1}^{6} (x_{1k} - x_{2k})^2} = \sqrt{(x_{11} - x_{21})^2 + (x_{12} - x_{22})^2 + (x_{13} - x_{23})^2 + (x_{14} - x_{24})^2 + (x_{15} - x_{25})^2 + (x_{16} - x_{26})^2}$$

$$= \sqrt{(0.2 - 0.4)^2 + (23 - 1)^2 + (5.7 - 5.4)^2 + (1 - 0)^2 + (0 - 1)^2 + (0 - 0)^2}$$

$$= \sqrt{(0.2)^2 + (22)^2 + (0.3)^2 + (1)^2 + (-1)^2 + (0)^2} = 22.05$$

ONE-HOT ENCODING: DOT PRODUCT

For one-hot encoded data,
the number of matching categorical values is the dot product of their vectors

$$D = \begin{array}{cc} & X_4 \\ x_1 & A \\ x_2 & B \\ x_3 & C \\ x_4 & A \\ x_5 & B \\ x_6 & C \\ x_7 & C \end{array} \quad \longrightarrow \quad D = \begin{array}{cccc} & X_{4A} & X_{4B} & X_{4C} \\ x_1 & 1 & 0 & 0 \\ x_2 & 0 & 1 & 0 \\ x_3 & 0 & 0 & 1 \\ x_4 & 1 & 0 & 0 \\ x_5 & 0 & 1 & 0 \\ x_6 & 0 & 0 & 1 \\ x_7 & 0 & 0 & 1 \end{array}$$

$$x_1 \cdot x_2 = 1 * 0 + 0 * 1 + 0 * 0 = 0$$

ONE-HOT ENCODING: DOT PRODUCT

For one-hot encoded data,
the number of matching categorical values is the dot product of their vectors

$$D = \begin{array}{cc} & X_4 \\ x_1 & A \\ x_2 & B \\ x_3 & C \\ x_4 & A \\ x_5 & B \\ x_6 & C \\ x_7 & C \end{array} \longrightarrow D = \begin{array}{cccc} & X_{4A} & X_{4B} & X_{4C} \\ x_1 & 1 & 0 & 0 \\ x_2 & 0 & 1 & 0 \\ x_3 & 0 & 0 & 1 \\ x_4 & 1 & 0 & 0 \\ x_5 & 0 & 1 & 0 \\ x_6 & 0 & 0 & 1 \\ x_7 & 0 & 0 & 1 \end{array}$$

$$x_1 \cdot x_2 = 1 * 0 + 0 * 1 + 0 * 0 = 0$$
$$x_6 \cdot x_7 = 0 * 0 + 0 * 0 + 1 * 1 = 1$$

ONE-HOT ENCODING: DOT PRODUCT

For one-hot encoded data,
the number of matching categorical values is the dot product of their vectors

$$D = \begin{array}{c|cc} & X_1 & X_2 \\ \hline x_1 & A & H \\ x_2 & B & L \\ x_3 & C & L \\ x_4 & A & L \\ x_5 & B & H \\ x_6 & C & L \\ x_7 & C & H \end{array} \longrightarrow D = \begin{array}{c|ccccc} & X_{1A} & X_{1B} & X_{1C} & X_{2A} & X_{2B} \\ \hline x_1 & 1 & 0 & 0 & 1 & 0 \\ x_2 & 0 & 1 & 0 & 0 & 1 \\ x_3 & 0 & 0 & 1 & 0 & 1 \\ x_4 & 1 & 0 & 0 & 0 & 1 \\ x_5 & 0 & 1 & 0 & 1 & 0 \\ x_6 & 0 & 0 & 1 & 0 & 1 \\ x_7 & 0 & 0 & 1 & 1 & 0 \end{array}$$

$$x_1 \cdot x_2 = 1 * 0 + 0 * 1 + 0 * 0 + 1 * 0 + 0 * 1 = 0$$

ONE-HOT ENCODING: DOT PRODUCT

For one-hot encoded data,
the number of matching categorical values is the dot product of their vectors

$$D = \begin{array}{ccc} & X_1 & X_2 \\ x_1 & A & H \\ x_2 & B & L \\ x_3 & C & L \\ x_4 & A & L \\ x_5 & B & H \\ x_6 & C & L \\ x_7 & C & H \end{array} \longrightarrow D = \begin{array}{cccccc} & X_{1A} & X_{1B} & X_{1C} & X_{2A} & X_{2B} \\ x_1 & 1 & 0 & 0 & 1 & 0 \\ x_2 & 0 & 1 & 0 & 0 & 1 \\ x_3 & 0 & 0 & 1 & 0 & 1 \\ x_4 & 1 & 0 & 0 & 0 & 1 \\ x_5 & 0 & 1 & 0 & 1 & 0 \\ x_6 & 0 & 0 & 1 & 0 & 1 \\ x_7 & 0 & 0 & 1 & 1 & 0 \end{array}$$

$$x_1 \cdot x_2 = 1 * 0 + 0 * 1 + 0 * 0 + 1 * 0 + 0 * 1 = 0$$
$$x_2 \cdot x_3 = 0 * 0 + 1 * 0 + 0 * 1 + 0 * 0 + 1 * 1 = 1$$

ONE-HOT ENCODING: DOT PRODUCT

For one-hot encoded data,
the number of matching categorical values is the dot product of their vectors

$$D = \begin{array}{ccc} & X_1 & X_2 \\ x_1 & A & H \\ x_2 & B & L \\ x_3 & C & L \\ x_4 & A & L \\ x_5 & B & H \\ x_6 & C & L \\ x_7 & C & H \end{array} \longrightarrow D = \begin{array}{cccccc} & X_{1A} & X_{1B} & X_{1C} & X_{2A} & X_{2B} \\ x_1 & 1 & 0 & 0 & 1 & 0 \\ x_2 & 0 & 1 & 0 & 0 & 1 \\ x_3 & 0 & 0 & 1 & 0 & 1 \\ x_4 & 1 & 0 & 0 & 0 & 1 \\ x_5 & 0 & 1 & 0 & 1 & 0 \\ x_6 & 0 & 0 & 1 & 0 & 1 \\ x_7 & 0 & 0 & 1 & 1 & 0 \end{array}$$

$$x_1 \cdot x_2 = 1*0 + 0*1 + 0*0 + 1*0 + 0*1 = 0$$
$$x_2 \cdot x_3 = 0*0 + 1*0 + 0*1 + 0*0 + 1*1 = 1$$
$$x_3 \cdot x_6 = 0*0 + 0*0 + 1*1 + 0*0 + 1*1 = 2$$

ONE-HOT ENCODING: HAMMING DISTANCE

Hamming Distance,
the number of mismatches between two vectors

$$\delta_H(x_i, x_j) = sum(xi \oplus xj)$$

ONE-HOT ENCODING: HAMMING DISTANCE

Hamming Distance,
the number of mismatches between two vectors

$$\delta_H(x_i, x_j) = sum(xi \oplus xj)$$

Recall XOR $\oplus$

a	b	$a \oplus b$
0	0	0
0	1	1
1	0	1
1	1	0

ONE-HOT ENCODING: HAMMING DISTANCE

Hamming Distance,
the number of mismatches between two vectors

Recall XOR $\oplus$

a	b	$a \oplus b$
0	0	0
0	1	1
1	0	1
1	1	0

$$\delta_H(x_i, x_j) = sum(xi \oplus xj)$$

$$D = \begin{array}{c|cc} & X_1 & X_2 \\ \hline x_1 & A & H \\ x_2 & B & L \\ x_3 & C & L \\ x_4 & A & L \\ x_5 & B & H \\ x_6 & C & L \\ x_7 & C & H \end{array} \longrightarrow D = \begin{array}{c|ccccc} & X_{1A} & X_{1B} & X_{1C} & X_{2A} & X_{2B} \\ \hline x_1 & 1 & 0 & 0 & 1 & 0 \\ x_2 & 0 & 1 & 0 & 0 & 1 \\ x_3 & 0 & 0 & 1 & 0 & 1 \\ x_4 & 1 & 0 & 0 & 0 & 1 \\ x_5 & 0 & 1 & 0 & 1 & 0 \\ x_6 & 0 & 0 & 1 & 0 & 1 \\ x_7 & 0 & 0 & 1 & 1 & 0 \end{array}$$

$$\delta_H(x_1, x_2)$$

ONE-HOT ENCODING: HAMMING DISTANCE

Hamming Distance,
the number of mismatches between two vectors

Recall XOR $\oplus$

a	b	$a \oplus b$
0	0	0
0	1	1
1	0	1
1	1	0

$$\delta_H(x_i, x_j) = sum(xi \oplus xj)$$

$$D = \begin{array}{c|cc} & X_1 & X_2 \\ \hline x_1 & A & H \\ x_2 & B & L \\ x_3 & C & L \\ x_4 & A & L \\ x_5 & B & H \\ x_6 & C & L \\ x_7 & C & H \end{array} \longrightarrow D = \begin{array}{c|ccccc} & X_{1A} & X_{1B} & X_{1C} & X_{2A} & X_{2B} \\ \hline x_1 & 1 & 0 & 0 & 1 & 0 \\ x_2 & 0 & 1 & 0 & 0 & 1 \\ x_3 & 0 & 0 & 1 & 0 & 1 \\ x_4 & 1 & 0 & 0 & 0 & 1 \\ x_5 & 0 & 1 & 0 & 1 & 0 \\ x_6 & 0 & 0 & 1 & 0 & 1 \\ x_7 & 0 & 0 & 1 & 1 & 0 \end{array}$$

$$\delta_H(x_1, x_2) = sum(x1 \oplus x2)$$

ONE-HOT ENCODING: HAMMING DISTANCE

Hamming Distance,
the number of mismatches between two vectors

Recall XOR $\oplus$

a	b	$a \oplus b$
0	0	0
0	1	1
1	0	1
1	1	0

$$\delta_H(x_i, x_j) = sum(xi \oplus xj)$$

$$D = \begin{array}{ccc} & X_1 & X_2 \\ x_1 & A & H \\ x_2 & B & L \\ x_3 & C & L \\ x_4 & A & L \\ x_5 & B & H \\ x_6 & C & L \\ x_7 & C & H \end{array} \longrightarrow D = \begin{array}{cccccc} & X_{1A} & X_{1B} & X_{1C} & X_{2A} & X_{2B} \\ x_1 & 1 & 0 & 0 & 1 & 0 \\ x_2 & 0 & 1 & 0 & 0 & 1 \\ x_3 & 0 & 0 & 1 & 0 & 1 \\ x_4 & 1 & 0 & 0 & 0 & 1 \\ x_5 & 0 & 1 & 0 & 1 & 0 \\ x_6 & 0 & 0 & 1 & 0 & 1 \\ x_7 & 0 & 0 & 1 & 1 & 0 \end{array}$$

$$\delta_H(x_1, x_2) = sum(x1 \oplus x2) = (1 \oplus 0) + (0 \oplus 1) + (0 \oplus 0) + (1 \oplus 0) + (0 \oplus 1)$$

ONE-HOT ENCODING: HAMMING DISTANCE

Hamming Distance,
the number of mismatches between two vectors

Recall XOR $\oplus$

a	b	$a \oplus b$
0	0	0
0	1	1
1	0	1
1	1	0

$\delta_H(x_i, x_j) = sum(xi \oplus xj)$

$$D = \begin{array}{c|cc} & X_1 & X_2 \\ \hline x_1 & A & H \\ x_2 & B & L \\ x_3 & C & L \\ x_4 & A & L \\ x_5 & B & H \\ x_6 & C & L \\ x_7 & C & H \end{array} \longrightarrow D = \begin{array}{c|ccccc} & X_{1A} & X_{1B} & X_{1C} & X_{2A} & X_{2B} \\ \hline x_1 & 1 & 0 & 0 & 1 & 0 \\ x_2 & 0 & 1 & 0 & 0 & 1 \\ x_3 & 0 & 0 & 1 & 0 & 1 \\ x_4 & 1 & 0 & 0 & 0 & 1 \\ x_5 & 0 & 1 & 0 & 1 & 0 \\ x_6 & 0 & 0 & 1 & 0 & 1 \\ x_7 & 0 & 0 & 1 & 1 & 0 \end{array}$$

$$\delta_H(x_1, x_2) = sum(x1 \oplus x2) = (1 \oplus 0) + (0 \oplus 1) + (0 \oplus 0) + (1 \oplus 0) + (0 \oplus 1) = 1 + 1 + 0 + 1 + 1 = 4$$

ONE-HOT ENCODING: HAMMING DISTANCE

Hamming Distance,
the number of mismatches between two vectors

Recall XOR $\oplus$

a	b	$a \oplus b$
0	0	0
0	1	1
1	0	1
1	1	0

$$\delta_H(x_i, x_j) = sum(xi \oplus xj)$$

$$D = \begin{array}{c|cc} & X_1 & X_2 \\ \hline x_1 & A & H \\ x_2 & B & L \\ x_3 & C & L \\ x_4 & A & L \\ x_5 & B & H \\ x_6 & C & L \\ x_7 & C & H \end{array} \longrightarrow D = \begin{array}{c|ccccc} & X_{1A} & X_{1B} & X_{1C} & X_{2A} & X_{2B} \\ \hline x_1 & 1 & 0 & 0 & 1 & 0 \\ x_2 & 0 & 1 & 0 & 0 & 1 \\ x_3 & 0 & 0 & 1 & 0 & 1 \\ x_4 & 1 & 0 & 0 & 0 & 1 \\ x_5 & 0 & 1 & 0 & 1 & 0 \\ x_6 & 0 & 0 & 1 & 0 & 1 \\ x_7 & 0 & 0 & 1 & 1 & 0 \end{array}$$

$$\delta_H(x_1, x_2) = sum(x1 \oplus x2) = (1 \oplus 0) + (0 \oplus 1) + (0 \oplus 0) + (1 \oplus 0) + (0 \oplus 1) = 1 + 1 + 0 + 1 + 1 = 4$$

$$\delta_H(x_2, x_3) = ??$$

$$\delta_H(x_3, x_6) = ??$$

ONE-HOT ENCODING: HAMMING DISTANCE

Hamming Distance,
the number of mismatches between two vectors

Recall XOR $\oplus$

a	b	$a \oplus b$
0	0	0
0	1	1
1	0	1
1	1	0

$$\delta_H(x_i, x_j) = sum(xi \oplus xj)$$

$$D = \begin{array}{c|cc} & X_1 & X_2 \\ \hline x_1 & A & H \\ x_2 & B & L \\ x_3 & C & L \\ x_4 & A & L \\ x_5 & B & H \\ x_6 & C & L \\ x_7 & C & H \end{array} \longrightarrow D = \begin{array}{c|ccccc} & X_{1A} & X_{1B} & X_{1C} & X_{2A} & X_{2B} \\ \hline x_1 & 1 & 0 & 0 & 1 & 0 \\ x_2 & 0 & 1 & 0 & 0 & 1 \\ x_3 & 0 & 0 & 1 & 0 & 1 \\ x_4 & 1 & 0 & 0 & 0 & 1 \\ x_5 & 0 & 1 & 0 & 1 & 0 \\ x_6 & 0 & 0 & 1 & 0 & 1 \\ x_7 & 0 & 0 & 1 & 1 & 0 \end{array}$$

$$\delta_H(x_1, x_2) = sum(x1 \oplus x2) = (1 \oplus 0) + (0 \oplus 1) + (0 \oplus 0) + (1 \oplus 0) + (0 \oplus 1) = 1 + 1 + 0 + 1 + 1 = 4$$

$$\delta_H(x_2, x_3) = 0 \oplus 0 + 1 \oplus 0 + 0 \oplus 1 + 0 \oplus 0 + 1 \oplus 1 = 2$$

$$\delta_H(x_3, x_6) = ??$$

ONE-HOT ENCODING: HAMMING DISTANCE

Hamming Distance,
the number of mismatches between two vectors

Recall XOR $\oplus$

a	b	$a \oplus b$
0	0	0
0	1	1
1	0	1
1	1	0

$$\delta_H(x_i, x_j) = sum(xi \oplus xj)$$

$$D = \begin{array}{c|cc} & X_1 & X_2 \\ \hline x_1 & A & H \\ x_2 & B & L \\ x_3 & C & L \\ x_4 & A & L \\ x_5 & B & H \\ x_6 & C & L \\ x_7 & C & H \end{array} \longrightarrow D = \begin{array}{c|ccccc} & X_{1A} & X_{1B} & X_{1C} & X_{2A} & X_{2B} \\ \hline x_1 & 1 & 0 & 0 & 1 & 0 \\ x_2 & 0 & 1 & 0 & 0 & 1 \\ x_3 & 0 & 0 & 1 & 0 & 1 \\ x_4 & 1 & 0 & 0 & 0 & 1 \\ x_5 & 0 & 1 & 0 & 1 & 0 \\ x_6 & 0 & 0 & 1 & 0 & 1 \\ x_7 & 0 & 0 & 1 & 1 & 0 \end{array}$$

$$\delta_H(x_1, x_2) = sum(x1 \oplus x2) = (1 \oplus 0) + (0 \oplus 1) + (0 \oplus 0) + (1 \oplus 0) + (0 \oplus 1) = 1 + 1 + 0 + 1 + 1 = 4$$

$$\delta_H(x_2, x_3) = 0 \oplus 0 + 1 \oplus 0 + 0 \oplus 1 + 0 \oplus 0 + 1 \oplus 1 = 2$$

$$\delta_H(x_3, x_6) = ??$$

ONE-HOT ENCODING: HAMMING DISTANCE

Hamming Distance,
the number of mismatches between two vectors

Recall XOR $\oplus$

a	b	$a \oplus b$
0	0	0
0	1	1
1	0	1
1	1	0

$$\delta_H(x_i, x_j) = sum(xi \oplus xj)$$

$$D = \begin{array}{c|cc} & X_1 & X_2 \\ \hline x_1 & A & H \\ x_2 & B & L \\ x_3 & C & L \\ x_4 & A & L \\ x_5 & B & H \\ x_6 & C & L \\ x_7 & C & H \end{array} \longrightarrow D = \begin{array}{c|ccccc} & X_{1A} & X_{1B} & X_{1C} & X_{2A} & X_{2B} \\ \hline x_1 & 1 & 0 & 0 & 1 & 0 \\ x_2 & 0 & 1 & 0 & 0 & 1 \\ x_3 & 0 & 0 & 1 & 0 & 1 \\ x_4 & 1 & 0 & 0 & 0 & 1 \\ x_5 & 0 & 1 & 0 & 1 & 0 \\ x_6 & 0 & 0 & 1 & 0 & 1 \\ x_7 & 0 & 0 & 1 & 1 & 0 \end{array}$$

$$\delta_H(x_1, x_2) = sum(x1 \oplus x2) = (1 \oplus 0) + (0 \oplus 1) + (0 \oplus 0) + (1 \oplus 0) + (0 \oplus 1) = 1 + 1 + 0 + 1 + 1 = 4$$

$$\delta_H(x_2, x_3) = 0 \oplus 0 + 1 \oplus 0 + 0 \oplus 1 + 0 \oplus 0 + 1 \oplus 1 = 2$$

$$\delta_H(x_3, x_6) = 0 \oplus 0 + 0 \oplus 0 + 1 \oplus 1 + 0 \oplus 0 + 1 \oplus 1 = 0$$

SET COMPARISON: JACCARD SIMILARITY

Jaccard Similarity,
the size of the intersection over the size of the union

$$J(x_i, x_j) = \frac{|x_i \cap x_j|}{|x_i \cup x_j|} = \frac{sum(x_i \wedge x_j)}{sum(x_i \vee x_j)}$$

SET COMPARISON: JACCARD SIMILARITY

Jaccard Similarity,
the size of the intersection over the size of the union

$$J(x_i, x_j) = \frac{|x_i \cap x_j|}{|x_i \cup x_j|} = \frac{sum(x_i \wedge x_j)}{sum(x_i \vee x_j)}$$

$$D = \begin{array}{c|ccccc} & X_1 & X_2 & X_3 & X_4 & X_5 \\ x_1 & 1 & 0 & 0 & 1 & 0 \\ x_2 & 0 & 1 & 0 & 0 & 1 \\ x_3 & 0 & 0 & 1 & 0 & 1 \\ x_4 & 1 & 0 & 0 & 0 & 1 \\ x_5 & 0 & 1 & 0 & 1 & 0 \\ x_6 & 0 & 0 & 1 & 0 & 1 \\ x_7 & 0 & 0 & 1 & 1 & 0 \end{array}$$

$$J(x_1, x_2) = \frac{sum(x1 \wedge x2)}{sum(x1 \vee x2)}$$

SET COMPARISON: JACCARD SIMILARITY

Jaccard Similarity,
the size of the intersection over the size of the union

$$J(x_i, x_j) = \frac{|x_i \cap x_j|}{|x_i \cup x_j|} = \frac{sum(x_i \wedge x_j)}{sum(x_i \vee x_j)}$$

$$D = \begin{array}{c|ccccc} & X_1 & X_2 & X_3 & X_4 & X_5 \\ x_1 & 1 & 0 & 0 & 1 & 0 \\ x_2 & 0 & 1 & 0 & 0 & 1 \\ x_3 & 0 & 0 & 1 & 0 & 1 \\ x_4 & 1 & 0 & 0 & 0 & 1 \\ x_5 & 0 & 1 & 0 & 1 & 0 \\ x_6 & 0 & 0 & 1 & 0 & 1 \\ x_7 & 0 & 0 & 1 & 1 & 0 \end{array}$$

$$J(x_1, x_2) = \frac{sum(x1 \wedge x2)}{sum(x1 \vee x2)} = \frac{(1 \wedge 0) + (0 \wedge 1) + (0 \wedge 0) + (1 \wedge 0) + (0 \wedge 1)}{(1 \vee 0) + (0 \vee 1) + (0 \vee 0) + (1 \vee 0) + (0 \vee 1)}$$

SET COMPARISON: JACCARD SIMILARITY

Jaccard Similarity,
the size of the intersection over the size of the union

$$J(x_i, x_j) = \frac{|x_i \cap x_j|}{|x_i \cup x_j|} = \frac{sum(x_i \wedge x_j)}{sum(x_i \vee x_j)}$$

$$D = \begin{array}{cccccc} & X_1 & X_2 & X_3 & X_4 & X_5 \\ x_1 & 1 & 0 & 0 & 1 & 0 \\ x_2 & 0 & 1 & 0 & 0 & 1 \\ x_3 & 0 & 0 & 1 & 0 & 1 \\ x_4 & 1 & 0 & 0 & 0 & 1 \\ x_5 & 0 & 1 & 0 & 1 & 0 \\ x_6 & 0 & 0 & 1 & 0 & 1 \\ x_7 & 0 & 0 & 1 & 1 & 0 \end{array}$$

$$J(x_1, x_2) = \frac{sum(x1 \wedge x2)}{sum(x1 \vee x2)} = \frac{(1 \wedge 0) + (0 \wedge 1) + (0 \wedge 0) + (1 \wedge 0) + (0 \wedge 1)}{(1 \vee 0) + (0 \vee 1) + (0 \vee 0) + (1 \vee 0) + (0 \vee 1)} = \frac{0+0+0+0+0}{1+1+0+1+1} = 0$$

SET COMPARISON: JACCARD SIMILARITY

Jaccard Similarity,
the size of the intersection over the size of the union

$$J(x_i, x_j) = \frac{|x_i \cap x_j|}{|x_i \cup x_j|} = \frac{sum(x_i \wedge x_j)}{sum(x_i \vee x_j)}$$

$$D = \begin{array}{cccccc} & X_1 & X_2 & X_3 & X_4 & X_5 \\ x_1 & 1 & 0 & 0 & 1 & 0 \\ x_2 & 0 & 1 & 0 & 0 & 1 \\ x_3 & 0 & 0 & 1 & 0 & 1 \\ x_4 & 1 & 0 & 0 & 0 & 1 \\ x_5 & 0 & 1 & 0 & 1 & 0 \\ x_6 & 0 & 0 & 1 & 0 & 1 \\ x_7 & 0 & 0 & 1 & 1 & 0 \end{array}$$

$$J(x_1, x_2) = \frac{sum(x1 \wedge x2)}{sum(x1 \vee x2)} = \frac{(1 \wedge 0) + (0 \wedge 1) + (0 \wedge 0) + (1 \wedge 0) + (0 \wedge 1)}{(1 \vee 0) + (0 \vee 1) + (0 \vee 0) + (1 \vee 0) + (0 \vee 1)} = \frac{0 + 0 + 0 + 0 + 0}{1 + 1 + 0 + 1 + 1} = 0$$

$$J(x_2, x_3) = ??$$

$$J(x_3, x_6) = ??$$

SET COMPARISON: JACCARD SIMILARITY

Jaccard Similarity,

the size of the intersection over the size of the union

$$J(x_i, x_j) = \frac{|x_i \cap x_j|}{|x_i \cup x_j|} = \frac{sum(x_i \wedge x_j)}{sum(x_i \vee x_j)}$$

$$D = \begin{array}{c|ccccc} & X_1 & X_2 & X_3 & X_4 & X_5 \\ x_1 & 1 & 0 & 0 & 1 & 0 \\ x_2 & 0 & 1 & 0 & 0 & 1 \\ x_3 & 0 & 0 & 1 & 0 & 1 \\ x_4 & 1 & 0 & 0 & 0 & 1 \\ x_5 & 0 & 1 & 0 & 1 & 0 \\ x_6 & 0 & 0 & 1 & 0 & 1 \\ x_7 & 0 & 0 & 1 & 1 & 0 \end{array}$$

$$J(x_1, x_2) = \frac{sum(x1 \wedge x2)}{sum(x1 \vee x2)} = \frac{(1 \wedge 0) + (0 \wedge 1) + (0 \wedge 0) + (1 \wedge 0) + (0 \wedge 1)}{(1 \vee 0) + (0 \vee 1) + (0 \vee 0) + (1 \vee 0) + (0 \vee 1)} = \frac{0+0+0+0+0}{1+1+0+1+1} = 0$$

$$J(x_2, x_3) = ??$$

$$J(x_3, x_6) = ??$$

SET COMPARISON: JACCARD SIMILARITY

Jaccard Similarity,
the size of the intersection over the size of the union

$$J(x_i, x_j) = \frac{|x_i \cap x_j|}{|x_i \cup x_j|} = \frac{sum(x_i \wedge x_j)}{sum(x_i \vee x_j)}$$

$$D = \begin{array}{cccccc} & X_1 & X_2 & X_3 & X_4 & X_5 \\ x_1 & 1 & 0 & 0 & 1 & 0 \\ x_2 & 0 & 1 & 0 & 0 & 1 \\ x_3 & 0 & 0 & 1 & 0 & 1 \\ x_4 & 1 & 0 & 0 & 0 & 1 \\ x_5 & 0 & 1 & 0 & 1 & 0 \\ x_6 & 0 & 0 & 1 & 0 & 1 \\ x_7 & 0 & 0 & 1 & 1 & 0 \end{array}$$

$$J(x_1, x_2) = \frac{sum(x1 \wedge x2)}{sum(x1 \vee x2)} = \frac{(1 \wedge 0) + (0 \wedge 1) + (0 \wedge 0) + (1 \wedge 0) + (0 \wedge 1)}{(1 \vee 0) + (0 \vee 1) + (0 \vee 0) + (1 \vee 0) + (0 \vee 1)} = \frac{0+0+0+0+0}{1+1+0+1+1} = 0$$

$$J(x_2, x_3) = \frac{0 \wedge 0 + 1 \wedge 0 + 0 \wedge 1 + 0 \wedge 0 + 1 \wedge 1}{0 \vee 0 + 1 \vee 0 + 0 \vee 1 + 0 \vee 0 + 1 \vee 1}$$

$$J(x_3, x_6) = ??$$

SET COMPARISON: JACCARD SIMILARITY

Jaccard Similarity,
the size of the intersection over the size of the union

$$J(x_i, x_j) = \frac{|x_i \cap x_j|}{|x_i \cup x_j|} = \frac{sum(x_i \wedge x_j)}{sum(x_i \vee x_j)}$$

$$D = \begin{array}{c|ccccc} & X_1 & X_2 & X_3 & X_4 & X_5 \\ x_1 & 1 & 0 & 0 & 1 & 0 \\ x_2 & 0 & 1 & 0 & 0 & 1 \\ x_3 & 0 & 0 & 1 & 0 & 1 \\ x_4 & 1 & 0 & 0 & 0 & 1 \\ x_5 & 0 & 1 & 0 & 1 & 0 \\ x_6 & 0 & 0 & 1 & 0 & 1 \\ x_7 & 0 & 0 & 1 & 1 & 0 \end{array}$$

$$J(x_1, x_2) = \frac{sum(x1 \wedge x2)}{sum(x1 \vee x2)} = \frac{(1 \wedge 0) + (0 \wedge 1) + (0 \wedge 0) + (1 \wedge 0) + (0 \wedge 1)}{(1 \vee 0) + (0 \vee 1) + (0 \vee 0) + (1 \vee 0) + (0 \vee 1)} = \frac{0+0+0+0+0}{1+1+0+1+1} = 0$$

$$J(x_2, x_3) = \frac{0 \wedge 0 + 1 \wedge 0 + 0 \wedge 1 + 0 \wedge 0 + 1 \wedge 1}{0 \vee 0 + 1 \vee 0 + 0 \vee 1 + 0 \vee 0 + 1 \vee 1} = \frac{0+0+0+0+1}{0+1+1+0+1} = \frac{1}{3}$$

$$J(x_3, x_6) = ??$$

SET COMPARISON: JACCARD SIMILARITY

Jaccard Similarity,
the size of the intersection over the size of the union

$$J(x_i, x_j) = \frac{|x_i \cap x_j|}{|x_i \cup x_j|} = \frac{sum(x_i \wedge x_j)}{sum(x_i \vee x_j)}$$

$$D = \begin{array}{c|ccccc} & X_1 & X_2 & X_3 & X_4 & X_5 \\ x_1 & 1 & 0 & 0 & 1 & 0 \\ x_2 & 0 & 1 & 0 & 0 & 1 \\ x_3 & 0 & 0 & 1 & 0 & 1 \\ x_4 & 1 & 0 & 0 & 0 & 1 \\ x_5 & 0 & 1 & 0 & 1 & 0 \\ x_6 & 0 & 0 & 1 & 0 & 1 \\ x_7 & 0 & 0 & 1 & 1 & 0 \end{array}$$

$$J(x_1, x_2) = \frac{sum(x1 \wedge x2)}{sum(x1 \vee x2)} = \frac{(1 \wedge 0) + (0 \wedge 1) + (0 \wedge 0) + (1 \wedge 0) + (0 \wedge 1)}{(1 \vee 0) + (0 \vee 1) + (0 \vee 0) + (1 \vee 0) + (0 \vee 1)} = \frac{0+0+0+0+0}{1+1+0+1+1} = 0$$

$$J(x_2, x_3) = \frac{0 \wedge 0 + 1 \wedge 0 + 0 \wedge 1 + 0 \wedge 0 + 1 \wedge 1}{0 \vee 0 + 1 \vee 0 + 0 \vee 1 + 0 \vee 0 + 1 \vee 1} = \frac{0+0+0+0+1}{0+1+1+0+1} = \frac{1}{3}$$

$$J(x_3, x_6) = ??$$

SET COMPARISON: JACCARD SIMILARITY

Jaccard Similarity,
the size of the intersection over the size of the union

$$J(x_i, x_j) = \frac{|x_i \cap x_j|}{|x_i \cup x_j|} = \frac{sum(x_i \wedge x_j)}{sum(x_i \vee x_j)}$$

$$D = \begin{array}{cccccc} & X_1 & X_2 & X_3 & X_4 & X_5 \\ x_1 & 1 & 0 & 0 & 1 & 0 \\ x_2 & 0 & 1 & 0 & 0 & 1 \\ x_3 & 0 & 0 & 1 & 0 & 1 \\ x_4 & 1 & 0 & 0 & 0 & 1 \\ x_5 & 0 & 1 & 0 & 1 & 0 \\ x_6 & 0 & 0 & 1 & 0 & 1 \\ x_7 & 0 & 0 & 1 & 1 & 0 \end{array}$$

$$J(x_1, x_2) = \frac{sum(x1 \wedge x2)}{sum(x1 \vee x2)} = \frac{(1 \wedge 0) + (0 \wedge 1) + (0 \wedge 0) + (1 \wedge 0) + (0 \wedge 1)}{(1 \vee 0) + (0 \vee 1) + (0 \vee 0) + (1 \vee 0) + (0 \vee 1)} = \frac{0+0+0+0+0}{1+1+0+1+1} = 0$$

$$J(x_2, x_3) = \frac{0 \wedge 0 + 1 \wedge 0 + 0 \wedge 1 + 0 \wedge 0 + 1 \wedge 1}{0 \vee 0 + 1 \vee 0 + 0 \vee 1 + 0 \vee 0 + 1 \vee 1} = \frac{0+0+0+0+1}{0+1+1+0+1} = \frac{1}{3}$$

$$J(x_3, x_6) = \frac{0 \wedge 0 + 0 \wedge 0 + 1 \wedge 1 + 0 \wedge 0 + 1 \wedge 1}{0 \vee 0 + 0 \vee 0 + 1 \vee 1 + 0 \vee 0 + 1 \vee 1}$$

SET COMPARISON: JACCARD SIMILARITY

Jaccard Similarity,
the size of the intersection over the size of the union

$$J(x_i, x_j) = \frac{|x_i \cap x_j|}{|x_i \cup x_j|} = \frac{sum(x_i \wedge x_j)}{sum(x_i \vee x_j)}$$

$$D = \begin{array}{c|ccccc} & X_1 & X_2 & X_3 & X_4 & X_5 \\ \hline x_1 & 1 & 0 & 0 & 1 & 0 \\ x_2 & 0 & 1 & 0 & 0 & 1 \\ x_3 & 0 & 0 & 1 & 0 & 1 \\ x_4 & 1 & 0 & 0 & 0 & 1 \\ x_5 & 0 & 1 & 0 & 1 & 0 \\ x_6 & 0 & 0 & 1 & 0 & 1 \\ x_7 & 0 & 0 & 1 & 1 & 0 \end{array}$$

$$J(x_1, x_2) = \frac{sum(x1 \wedge x2)}{sum(x1 \vee x2)} = \frac{(1 \wedge 0) + (0 \wedge 1) + (0 \wedge 0) + (1 \wedge 0) + (0 \wedge 1)}{(1 \vee 0) + (0 \vee 1) + (0 \vee 0) + (1 \vee 0) + (0 \vee 1)} = \frac{0+0+0+0+0}{1+1+0+1+1} = 0$$

$$J(x_2, x_3) = \frac{0 \wedge 0 + 1 \wedge 0 + 0 \wedge 1 + 0 \wedge 0 + 1 \wedge 1}{0 \vee 0 + 1 \vee 0 + 0 \vee 1 + 0 \vee 0 + 1 \vee 1} = \frac{0+0+0+0+1}{0+1+1+0+1} = \frac{1}{3}$$

$$J(x_3, x_6) = \frac{0 \wedge 0 + 0 \wedge 0 + 1 \wedge 1 + 0 \wedge 0 + 1 \wedge 1}{0 \vee 0 + 0 \vee 0 + 1 \vee 1 + 0 \vee 0 + 1 \vee 1} = \frac{0+0+1+0+1}{0+0+1+0+1} = \frac{2}{2} = 1$$