CSCI 550: Advanced Data Mining

02- Data Mining and Analysis (Part 2)

MONTANA STATE UNIVERSITY

GEOMETRIC VIEW: DISTANCE BETWEEN VECTORS

- We are often interested in some measure of distance between vectors representing separate entities.
- First, some notation:

L_2 norm of a vector x_i with m dimensions (columns/attributes). The length of a vector is a nonnegative number that describes the extent of the vector in space. Also known as *Euclidian norm* or *length* of a vector

$$||x_i||_2 = \sqrt{\sum_{k=1}^{m} x_{ik}^2}$$

MONTANA STATE UNIVERSITY

DISTANCE BETWEEN VECTORS

- We are often interested in some measure of distance between vectors representing separate entities.
- First, some notation:

L_2 norm of a vector x_i with m dimensions (columns/attributes).

The length of a vector is a nonnegative number that describes the extent of the vector in space

$$||x_i||_2 = \sqrt{\sum_{k=1}^{m} x_{ik}^2}$$

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

$$||x_2||_2 = \sqrt{\sum_{k=1}^{3} x_{2k}^2} = \sqrt{(x_{21}^2 + x_{22}^2 + x_{23}^2)} = \sqrt{(0.4^2 + 1^2 + 5.4^2)} = 5.5$$

DISTANCE BETWEEN VECTORS

- We are often interested in some measure of distance between vectors representing separate entities.
- L_2 norm or Euclidian distance between two vectors:

$$||x_i - x_j||_2 = \sqrt{\sum_{k=1}^{m} (x_{ik} - x_{jk})^2}$$

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

DISTANCE BETWEEN VECTORS

- We are often interested in some measure of distance between vectors representing separate entities.
- L_2 norm or Euclidian distance:

$$||x_i - x_j||_2 = \sqrt{\sum_k^m (x_{ik} - x_{jk})^2}$$

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

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

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

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

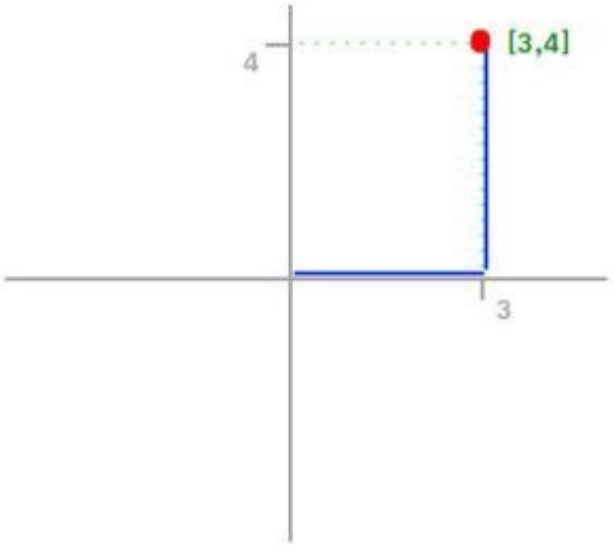
$$= \sqrt{(-0.2)^2 + (22)^2 + (0.3)^2}$$

$$= 22.0$$

DISTANCE BETWEEN VECTORS

- We are often interested in some measure of distance between vectors representing separate entities.
- L_1 norm Also known as Manhattan Distance or Taxicab norm:

$$||x_i - x_j||_1 = \sum_{k=1}^{m} |x_{ik} - x_{jk}|$$



DISTANCE BETWEEN VECTORS

- We are often interested in some measure of distance between vectors representing separate entities.
- *L_1* norm:

$$||x_i - x_j||_1 = \sum_{k}^{m} |x_{ik} - x_{jk}|$$

where x_i and x_j are vectors, and there are m dimensions

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

$$||x_1 - x_2||_1 = \sum_{k=1}^{3} |x_{1k} - x_{2k}|$$

$$= |x_{11} - x_{21}| + |x_{12} - x_{22}| + |x_{13} - x_{23}|$$

$$= |0.2 - 0.4| + |23 - 1| + |5.7 - 5.4|$$

$$= |-0.2| + |22| + |0.3|$$

$$= 22.5$$

DISTANCE BETWEEN VECTORS

- We are often interested in some measure of distance between vectors representing separate entities.
- L_1 norm: L_2 norm:

$$||x_i - x_j||_1 = \sum_{k=1}^{m} |x_{ik} - x_{jk}|$$

$$||x_i - x_j||_2 = \sqrt{\sum_{k=1}^{m} (x_{ik} - x_{jk})^2}$$

iClicker

1- Which one is always true:

A: $L_1 =< L_2$

B: $L_2 =< L_1$

C: $L_2 < L_1$

D: It depends

DISTANCE BETWEEN VECTORS

- We are often interested in some measure of distance between vectors representing separate entities.
- Dot Product is a measure of how closely two vectors align, in terms of the directions they point. The measure is a scalar number (single value) that can be used to compare the two vectors and to understand the impact of repositioning one or both of them.

$$\mathbf{a} = \begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_m \end{pmatrix} \qquad \mathbf{b} = \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_m \end{pmatrix}$$

$$a.b = a^T b = \sum_1^m a_k b_k$$

where a and b are vectors,

and there are m dimensions

DISTANCE BETWEEN VECTORS

- We are often interested in some measure of distance between vectors representing separate entities.
- Dot Product:

$$a.b = a^T b = \sum_1^m a_k b_k$$

where a and b are vectors, and there are m dimensions

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

$$x_3^T x_4 = \sum_{k=1}^{3} x_{3k} x_{4k}$$

$$= x_{31}x_{41} + x_{32}x_{42} + x_{33}x_{43}$$

$$= (1.8)(5.6) + (0.5)(50) + (5.2)(5.1)$$

$$= 61.6$$

DISTANCE BETWEEN VECTORS

- We are often interested in some measure of distance between vectors representing separate entities.
- Dot Product:

$$a.b = a^T b = \sum_1^m a_k b_k$$

where a and b are vectors,

and there are m dimensions

iClicker

True or False:

$|a.b| \leq ||a||_2 . \ ||b||_2$

A: True

B: False

DISTANCE BETWEEN VECTORS

- We are often interested in some measure of distance between vectors representing separate entities.
- Cosine Similarity: of the smallest angle between two vectors xi and x_j:

$$cos(\theta) = \frac{x_i^T x_j}{||x_i||_2 ||x_j||_2}$$ where x_i and x_j are vectors and $x_i^T x_j$ is their dot product

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

DISTANCE BETWEEN VECTORS

- We are often interested in some measure of distance between vectors representing separate entities.
- Cosine of the angle between two vectors x_i and x_j:

$$cos(\theta) = \frac{x_i^T x_j}{||x_i||_2 ||x_j||_2}$$ **where x_i and x_j are vectors and $x_i^T x_j$ is their dot product**

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

cosine of the angle between x_2 and x_3 is:

$$\frac{x_2^T x_3}{||x_2||_2 ||x_3||_2}$$

DISTANCE BETWEEN VECTORS

- We are often interested in some measure of distance between vectors representing separate entities.
- Cosine of the angle between two vectors x_i and x_j:

$$cos(\theta) = \frac{x_i^T x_j}{||x_i||_2 ||x_j||_2}$$ **where x_i and x_j are vectors and $x_i^T x_j$ is their dot product**

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

cosine of the angle between x_2 and x_3 is:

$$\frac{x_2^T x_3}{||x_2||_2 ||x_3} = \frac{(0.4 \quad 1 \quad 5.4)^T (1.8 \quad 0.5 \quad 5.2)}{\sqrt{(0.4^2 + 1^2 + 5.4^2)}\sqrt{(1.8^2 + 0.5^2 + 5.2^2)}}$$

$$= \frac{(0.4)(1.8) + (1)(0.5) + (5.4)(5.2))}{\sqrt{(0.4^2 + 1^2 + 5.4^2)}\sqrt{(1.8^2 + 0.5^2 + 5.2^2)}}$$

$$= 0.96$$

MONTANA STATE UNIVERSITY

GEOMETRIC INTERPRETATION OF SAMPLE COVARIANCE

- Consider the the mean-centered data matrix:

$$\overline{X}_1 = X_1 - \hat{\mu}_1 \cdot \mathbf{1} = \begin{pmatrix} x_{11} - \hat{\mu}_1 \\ x_{21} - \hat{\mu}_1 \\ \vdots \\ x_{n1} - \hat{\mu}_1 \end{pmatrix} \qquad \overline{X}_2 = X_2 - \hat{\mu}_2 \cdot \mathbf{1} = \begin{pmatrix} x_{12} - \hat{\mu}_2 \\ x_{22} - \hat{\mu}_2 \\ \vdots \\ x_{n2} - \hat{\mu}_2 \end{pmatrix}$$

- And remember sample covariance between X_1 and X_2 is given as:

$$\boxed{\hat{\sigma}_{12} = \frac{1}{n} \sum_{i=1}^{n} (x_{i1} - \hat{\mu}_1)(x_{i2} - \hat{\mu}_2)}$$

- We can show:

$$\boxed{\hat{\sigma}_{12} = \frac{\overline{X}_1^T \overline{X}_2}{n}}$$

MONTANA STATE UNIVERSITY

GEOMETRIC INTERPRETATION OF SAMPLE CORRELATION

- Remember: $cos(\theta) = \dfrac{x_i^T x_j}{||x_i||_2 ||x_j||_2}$ and $\rho_{12} = \dfrac{\sigma_{12}}{\sigma_1 \sigma_2} = \dfrac{\sigma_{12}}{\sqrt{\sigma_1^2 \sigma_2^2}}$

- Sample correlation can be written as:

$$\boxed{\hat{\rho}_{12} = \frac{\overline{X}_1^T \overline{X}_2}{\sqrt{\overline{X}_1^T \overline{X}_1}\sqrt{\overline{X}_2^T \overline{X}_2}} = \frac{\overline{X}_1^T \overline{X}_2}{\|\overline{X}_1\| \, \|\overline{X}_2\|} = \left(\frac{\overline{X}_1}{\|\overline{X}_1\|}\right)^T \left(\frac{\overline{X}_2}{\|\overline{X}_2\|}\right) = \cos\theta}$$

RECALL: EUCLIDEAN DISTANCE

- Euclidean distance:

$$||x_i - x_j||_2 = \sqrt{\sum_{k=1}^{m} (x_{ik} - x_{jk})^2}$$

where x_i and x_j are vectors, and there are m dimensions

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

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

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

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

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

$$= 22.0$$

WHAT IF WE ALSO HAVE CATEGORICAL VARIABLES?

- Euclidean distance:

$$||x_i - x_j||_2 = \sqrt{\sum_{k=1}^{m} (x_{ik} - x_{jk})^2}$$

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

- Euclidean distance:

$$||x_i - x_j||_2 = \sqrt{\sum_{k=1}^{m} (x_{ik} - x_{jk})^2}$$

where x_i and x_j are vectors, and there are m dimensions

$$D = \begin{array}{c|cccc} & X_1 & X_2 & X_3 & X_4 \\ x_1 & 0.2 & 23 & 5.7 & 1 \\ x_2 & 0.4 & 1 & 5.4 & 2 \\ x_3 & 1.8 & 0.5 & 5.2 & 3 \\ x_4 & 5.6 & 50 & 5.1 & 1 \\ x_5 & -0.5 & 34 & 5.3 & 2 \\ x_6 & 0.4 & 19 & 5.4 & 3 \\ x_7 & 1.1 & 11 & 5.5 & 3 \end{array}$$

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

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

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

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

$$= 22.02$$

PROBLEM

$$||x_i - x_j||_2 = \sqrt{\sum_{k=1}^{m} (x_{ik} - x_{jk})^2}$$

where x_i and x_j are vectors, and there are m dimensions

iClicker

$$D = \begin{array}{c} \\ x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \\ x_6 \\ x_7 \end{array} \begin{array}{cccc} X_1 & X_2 & X_3 & X_4 \\ 0.2 & 23 & 5.7 & 1 \\ 0.4 & 1 & 5.4 & 3 \\ 1.8 & 0.5 & 5.2 & 3 \\ 5.6 & 50 & 5.1 & 1 \\ -0.5 & 34 & 5.3 & 2 \\ 0.4 & 19 & 5.4 & 3 \\ 1.1 & 11 & 5.5 & 3 \end{array}$$

1- Do you expect to get same distance if we change x_{24} to 3?

- A: Yes
- B: No

2- Find the Euclidean distance between x_1 and x_2

PROBLEM

- Find the Euclidean distance between x_1 and x_2 :

$$||x_i - x_j||_2 = \sqrt{\sum_{k=1}^{m} (x_{ik} - x_{jk})^2}$$

where x_i and x_j are vectors, and there are m dimensions

$$D = \begin{array}{c|cccc} & X_1 & X_2 & X_3 & X_4 \\ x_1 & 0.2 & 23 & 5.7 & 1 \\ x_2 & 0.4 & 1 & 5.4 & 3 \\ x_3 & 1.8 & 0.5 & 5.2 & 3 \\ x_4 & 5.6 & 50 & 5.1 & 1 \\ x_5 & -0.5 & 34 & 5.3 & 2 \\ x_6 & 0.4 & 19 & 5.4 & 3 \\ x_7 & 1.1 & 11 & 5.5 & 3 \end{array}$$

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

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

$$= \sqrt{(0.2 - 0.4)^2 + (23 - 1)^2 + (5.7 - 5.4)^2 + (1 - 3)^2}$$

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

$$= 22.09$$

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

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

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

$$D = \begin{array}{c|cccc} & X_1 & X_2 & X_3 & X_4 \\ \hline x_1 & 0.2 & 23 & 5.7 & A \\ x_2 & 0.4 & 1 & 5.4 & B \\ x_3 & 1.8 & 0.5 & 5.2 & C \\ x_4 & 5.6 & 50 & 5.1 & A \\ x_5 & -0.5 & 34 & 5.3 & B \\ x_6 & 0.4 & 19 & 5.4 & C \\ x_7 & 1.1 & 11 & 5.5 & C \end{array} \longrightarrow \begin{array}{c|cccccc} & X_1 & X_2 & X_3 & X_4 & X_5 & X_6 \\ \hline x_1 & 0.2 & 23 & 5.7 & 1 & 0 & 0 \\ x_2 & 0.4 & 1 & 5.4 & 0 & 1 & 0 \\ x_3 & 1.8 & 0.5 & 5.2 & 0 & 0 & 1 \\ x_4 & 5.6 & 50 & 5.1 & 1 & 0 & 0 \\ x_5 & -0.5 & 34 & 5.3 & 0 & 1 & 0 \\ x_6 & 0.4 & 19 & 5.4 & 0 & 0 & 1 \\ x_7 & 1.1 & 11 & 5.5 & 0 & 0 & 1 \end{array}$$

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

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

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

ONE-HOT ENCODING

$$||x_i - x_j||_2 = \sqrt{\sum_{k=1}^{m} (x_{ik} - x_{jk})^2}$$

where x_i and x_j are vectors, and there are m dimensions

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

$$\begin{array}{c|cccccc} & X_1 & X_2 & X_3 & X_4 & X_5 & X_6 \\ x_1 & 0.2 & 23 & 5.7 & 1 & 0 & 0 \\ x_2 & 0.4 & 1 & 5.4 & 0 & 0 & 1 \\ x_3 & 1.8 & 0.5 & 5.2 & 0 & 0 & 1 \\ x_4 & 5.6 & 50 & 5.1 & 1 & 0 & 0 \\ x_5 & -0.5 & 34 & 5.3 & 0 & 1 & 0 \\ x_6 & 0.4 & 19 & 5.4 & 0 & 0 & 1 \\ x_7 & 1.1 & 11 & 5.5 & 0 & 0 & 1 \end{array}$$

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

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

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

DOT PRODUCT AND BINARY DATA

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

$$s(x_i, x_j) = x_i^T x_j$$

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

$$x_1^T x_2 = 1(0) + 0(0) + 0(1) = 0$$

$s(x_1, x_2) = 0$ $\quad$ *because* $x_1 = A$ *and* $x_2 = B$ *for attribute* X_4

DOT PRODUCT AND BINARY DATA

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

$$s(x_i, x_j) = x_i^T x_j$$

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

$$x_1^T x_4 = 1(1) + 0(0) + 0(0) = 1$$

$s(x_1, x_2) = 0$ $\quad$ $because\ x_1 = A\ and\ x_2 = B\ for\ attribute\ X_4$

$s(x_1, x_4) = 1$ $\quad$ $because\ x_1 = A\ and\ x_4 = A\ for\ attribute\ X_4$

DOT PRODUCT AND BINARY DATA

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

$$s(x_i, x_j) = x_i^T x_j$$

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

$$x_3^T x_6 = 0(0) + 0(0) + 1(1) + 0(0) + 1(1) = 2$$

$s(x_3, x_6) = 2$ *because x_3 and x_6 match in 2 categorical attributes*

DOT PRODUCT AND BINARY DATA

For one-hot encoded data, the number categorical attributes d is the squared 2-norm each point:

$$d = ||x_i||_2^2 = x_i^T x_i$$

$$D = \begin{array}{cc} & X_4 \\ x_1 & A \\ x_2 & C \\ x_3 & C \\ x_4 & A \\ x_5 & B \\ x_6 & C \\ x_7 & C \end{array}$$

	X_{4A}	X_{4B}	X_{4C}
x_1	1	0	0
x_2	0	0	1
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

$$x_1^2 = 1^2 + 0^2 + 0^2 = 1 = d$$

DOT PRODUCT AND BINARY DATA

For one-hot encoded data, the number categorical attributes d is the squared 2-norm each point:

$$d = ||x_i||_2^2 = x_i^T x_i$$

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

$$x_2^2 = 0^2 + 0^2 + 1^2 = 1 = d$$

DOT PRODUCT AND BINARY DATA

For one-hot encoded data, the number categorical attributes d is the squared 2-norm each point:

$$d = ||x_i||_2^2 = x_i^T x_i$$

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

$$x_2^2 = 0^2 + 0^2 + 1^2 + 0^2 + 1^2 = 2 = d$$

HAMMING DISTANCE

- Hamming Distance: number of mismatches
- $\delta_H(x_i, x_j) = d - s$ = number of entries where and do not have the same value (Where is the number of categorical attributes and is the number of matches in categorical value)

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

$$||x_2||_2^2 = x_2^2 = 0^2 + 0^2 + 1^2 + 0^2 + 1^2 = 2 = d$$

$$s(x_1, x_2) = x_1^T x_2 = 0$$

$$\delta_H(x_i, x_j) = 2 - 0 = 2$$

In-class Problem

- Find the Hamming distance between x_1 and x_4 :

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

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

COSINE SIMILARITY

- Hamming distance: $\delta_H(x_i, x_j) = d - s$
- $\cos(\theta) = \dfrac{x_i^T x_j}{||x_i||_2 ||x_j||_2} = \dfrac{s}{\sqrt{d}\sqrt{d}} = \dfrac{s}{d}$
  - $s(x_i, x_j) = x_i^T x_j$
  - $d = ||x_i||_2^2 = x_i^T x_i$

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

	X_{1A}	X_{1B}	X_{1C}	X_{2H}	X_{2L}
x_1	1	0	0	1	0
x_2	0	0	1	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

$s(x_1, x_4) = x_1^T x_4 = 1$

$\delta_H(x_i, x_j) = 2 - 1 = 1$

$\cos(\theta) = \frac{1}{2}$

JACCARD SIMILARITY

The ratio of the number of matching values to the number of distinct values that appear in both data instances:

$$J(x_i, x_j) = \frac{s}{2(d-s)+s} = \frac{s}{2d-s}$$

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

$$J(x_2, x_6) = \frac{1}{3}$$

$s(x_2, x_6) = 1$ *because x_2 and x_6 match in 1 categorical attributes*

JACCARD SIMILARITY

The ratio of the number of matching values to the number of distinct values that appear in both data instances:

$$J(x_i, x_j) = \frac{s}{2(d-s)+s} = \frac{s}{2d-s}$$

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

$$J(x_2, x_5) = \frac{1}{3}$$

$s(x_2, x_5) = 1$ *because x_2 and x_5 match in 1 categorical attributes*

JACCARD SIMILARITY

The ratio of the number of matching values to the number of distinct values that appear in both data instances:

$$J(x_i, x_j) = \frac{s}{2(d-s)+s} = \frac{s}{2d-s}$$

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

$$J(x_3, x_6) = \frac{2}{2} = 1$$

$s(x_3, x_6) = 1$ *because x_2 and x_6 match in 1 categorical attributes*

HOW ELSE MIGHT WE COMBINE CATEGORICAL AND NUMERICAL DATA?

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

GOWER DISTANCE

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

$$G(x_i, x_j) = \frac{1}{d}\sum_{k=1}^{d} \mathbf{dist}_{ij}(k)$$

Categorical

$$\mathbf{dist}_{ij}(k) = \begin{cases} 0 & \textbf{if } x_{ik} = x_{jk} \\ 1 & \textbf{otherwise} \end{cases}$$

Numerical

$$\mathbf{dist}_{ij}(k) = \frac{|x_{ik} - x_{jk}|}{\mathbf{Range}(k)}$$

GOWER DISTANCE

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

$$G(x_1, x_2) = \frac{1}{4} \sum_{k=1}^{4} \textbf{dist}_{12}(k)$$

$$= \frac{1}{4}\left(\frac{0.2}{6.1} + \frac{22}{49.5} + \frac{0.3}{0.6} + 1\right)$$

$$= 0.49$$

$$G(x_i, x_j) = \frac{1}{d} \sum_{k=1}^{d} \textbf{dist}_{ij}(k)$$

Categorical

$$\textbf{dist}_{ij}(k) = \begin{cases} 0 & \textbf{if } x_{ik} = x_{jk} \\ 1 & \textbf{otherwise} \end{cases}$$

Numerical

$$\textbf{dist}_{ij}(k) = \frac{|x_{ik} - x_{jk}|}{\textbf{Range}(k)}$$

IN-CLASS PROBLEM:

Compute the Gower distance between x_1 and x_3

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

IN-CLASS PROBLEM:

Compute the Gower distance between x_1 and x_3

$D =$

	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

$$G(x_1, x_3) = \frac{1}{4}\sum_{k=1}^{4} \mathbf{dist}_{13}(k)$$

$$= \frac{1}{4}\left(\frac{1.6}{6.1} + \frac{22.5}{49.5} + \frac{0.5}{0.6} + 1\right)$$

$$= 0.64$$

$$G(x_i, x_j) = \frac{1}{d}\sum_{k=1}^{d} \mathbf{dist}_{ij}(k)$$

Categorical

$$\mathbf{dist}_{ij}(k) = \begin{cases} 0 & \textbf{if } x_{ik} = x_{jk} \\ 1 & \textbf{otherwise} \end{cases}$$

Numerical

$$\mathbf{dist}_{ij}(k) = \frac{|x_{ik} - x_{jk}|}{\mathbf{Range}(k)}$$