CSCI 550: Advanced Data Mining

02- Data Mining and Analysis

MONTANA STATE UNIVERSITY

WHAT CAN WE LEARN FROM NUMERICAL DATA?

- Data can often be represented by an n*d *data matrix* D

$$\mathbf{D} = \left(\begin{array}{c|cccc} & X_1 & X_2 & \cdots & X_d \\ \hline \mathbf{x}_1 & x_{11} & x_{12} & \cdots & x_{1d} \\ \mathbf{x}_2 & x_{21} & x_{22} & \cdots & x_{2d} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \mathbf{x}_n & x_{n1} & x_{n2} & \cdots & x_{nd} \end{array}\right)$$

- where x_i denotes the i^{th} row:

$$\mathbf{x}_i = (x_{i1}, x_{i2}, \ldots, x_{id})$$

- where X_j denotes the j^{th} column:

$$X_j = (x_{1j}, x_{2j}, \ldots, x_{nj})$$

WHAT CAN WE LEARN FROM NUMERICAL DATA?

- Review: Statistics, central tendency
- The estimator of expected value (mean) of attribute j :

$$\hat{\mu}_j = \frac{1}{n} \sum_{i=1}^{n} x_{ij}$$

- one-number summary of the location or central tendency for the distribution of X
- What are other measures for central tendency?
- Which one is preferred?

WHAT CAN WE LEARN FROM NUMERICAL DATA?

- Review: multi-dimensional mean

What is the sample mean of the entire (numerical) data set?

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

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

$$\hat{\mu} = \frac{1}{7}((0.2 \quad 23 \quad 5.7) + (0.4 \quad 1 \quad 5.4) + (1.8 \quad 0.5 \quad 5.2) + (5.6 \quad 50 \quad 5.1) + (-0.5 \quad 34 \quad 5.3) + (0.4 \quad 19 \quad 5.4) + (1.1 \quad 11 \quad 5.5))$$

$$= (1.3 \quad 19.8 \quad 5.4)$$

MEAN CENTERING

- Mean-centering shifts the data matrix mean to 0.
- Mean-centering:
- $z_i = x_i - \hat{\mu}$ (for each attribute, subtract the mean from the instance value)

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

$$z_{11} = x_{11} - \hat{\mu}_1 = 0.2 - 1.3 = -1.1$$

for the first attribute

MEAN CENTERING

- Mean-centering shifts the data matrix mean to 0.
- Mean-centering:
- $z_i = x_i - \hat{\mu}$ (for each attribute, subtract the mean from the instance value)

$$D = \begin{array}{c} \\ x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \\ x_6 \\ x_7 \end{array} \begin{array}{ccc} X_1 & X_2 & X_3 \\ 0.2 & 23 & 5.7 \\ 0.4 & 1 & 5.4 \\ 1.8 & 0.5 & 5.2 \\ 5.6 & 50 & 5.1 \\ -0.5 & 34 & 5.3 \\ 0.4 & 19 & 5.4 \\ 1.1 & 11 & 5.5 \end{array} \longrightarrow Z = \begin{array}{c} \\ z_1 \\ z_2 \\ z_3 \\ z_4 \\ z_5 \\ z_6 \\ z_7 \end{array} \begin{array}{ccc} X_1 & X_2 & X_3 \\ -1.1 & 3.2 & 0.3 \\ -0.9 & -18.8 & 0.0 \\ 0.5 & -19.3 & -0.2 \\ 4.3 & 30.2 & -0.3 \\ -1.8 & 14.2 & -0.1 \\ -0.9 & -0.8 & 0.0 \\ -0.2 & -8.8 & 0.1 \end{array}$$

WHAT CAN WE LEARN FROM NUMERICAL DATA?

- Review: Statistics, measures of dispersion
- Sample variance of of attribute j :

$$\hat{\sigma}_j^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_{ij} - \hat{\mu})^2$$

Why does the sample variance have n-1 in the denominator?

- What are other measures for dispersion?
-

MONTANA STATE UNIVERSITY

WHAT CAN WE LEARN FROM NUMERICAL DATA?

- Review: Total variance
- What is the total variance in a numerical data set?

$$\text{Var}(D) = \hat{\sigma}_1^2 + \hat{\sigma}_2^2 + \ldots + \hat{\sigma}_n^2$$

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

$$Var(D) = \hat{\sigma}_1^2 + \hat{\sigma}_2^2 + \hat{\sigma}_3^2 = 4.1 + 321.3 + 0.0 = 325.4$$

WHAT CAN WE LEARN FROM NUMERICAL DATA?

- Review: Measures of Association
- covariance
- What is the covariance between two attributes in a numerical data set?

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

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

WHAT CAN WE LEARN FROM NUMERICAL DATA?

- Review: Measures of Association
- covariance
- What is the covariance between two attributes in a numerical data set?

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

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

iClicker

What are the possible values of covariance?

A: Only positive values
B: between -1 to +1
C: Between -∞ to +∞

WHAT CAN WE LEARN FROM NUMERICAL DATA?

- Review: Measures of Association
- Correlation coefficient
  - $$\frac{cov\ (x,y)}{std\ (x) \times std\ (y)} = \frac{\sigma_{12}}{\sigma 1 \times \sigma 2}$$

iClicker

1- What are the possible values of Correlation coefficient?

A: Only positive values
B: between -1 to +1
C: Between -∞ to +∞

2- What does correlation coef of 1 mean?

Correlation and Casuality

iClicker

1- cor (x,y) =0.7 which one is true:

A: An increase in x will cause an increase in y

B: An increase in y will cause an increase in x

C: x and y move together

D: All above

2- Is cor(x,y) = cor (y,x) true:

A: Yes

B: No

Correlation and Casuality

- Correlation doesn't have direction, but causality has direction
- Correlation DOES NOT imply causality!
- Having doubts, check [spurious-correlation]

COVARIANCE MATRIX

- Review: Measures of Association
- covariance matrix
- The covariance matrix Σ stores the covariance between each pair of attributes, as well as the variance for each attribute:

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

$$\Sigma = \begin{pmatrix} \hat{\sigma}_1^2 & \hat{\sigma}_{12} & \hat{\sigma}_{13} \\ \hat{\sigma}_{21} & \hat{\sigma}_2^2 & \hat{\sigma}_{23} \\ \hat{\sigma}_{31} & \hat{\sigma}_{32} & \hat{\sigma}_3^2 \end{pmatrix}$$

COVARIANCE MATRIX

- The covariance matrix $\sum$ stores the covariance between each pair of attributes, as well as the variance for each attribute:

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

$$
\Sigma = \begin{pmatrix} \hat{\sigma}_1^2 & \hat{\sigma}_{12} & \hat{\sigma}_{13} \\ \hat{\sigma}_{21} & \hat{\sigma}_2^2 & \hat{\sigma}_{23} \\ \hat{\sigma}_{31} & \hat{\sigma}_{32} & \hat{\sigma}_3^2 \end{pmatrix}
$$

$$
\Sigma = \begin{pmatrix} 4.1 & 18.4 & -0.26 \\ 18.4 & 321.3 & -1.09 \\ -0.26 & -1.09 & 0.0 \end{pmatrix}
$$

DATA NORMALIZATION (LINEAR SCALING)

- Some attributes may dominate our data analysis if we're not careful (for example, those with significantly larger values). Therefore we may want to normalize the data.
- Range normalization shifts attribute values to the range [0,1]

$$x_i' = \frac{x_i - \min_i\{x_i\}}{\max_i\{x_i\} - \min_i\{x_i\}}$$

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

$$x_1' = \frac{0.2 - (-0.5)}{5.6 - (-0.5)} = 0.1 \text{ for the first attribute}$$

DATA NORMALIZATION (LINEAR SCALING)

- Some attributes may dominate our data analysis if we're not careful (for example, those with significantly larger values). Therefore we may want to normalize the data.
- Range normalization shifts attribute values to the range [0,1]

$$x_i' = \frac{x_i - \min_i\{x_i\}}{\max_i\{x_i\} - \min_i\{x_i\}}$$

$$D = \begin{array}{c|ccc} & X_1 & X_2 & X_3 \\ x_1 & 0.2 & 23 & 5.7 \\ x_2 & 0.4 & 1 & 5.4 \\ x_3 & 1.8 & 0.5 & 5.2 \\ x_4 & 5.6 & 50 & 5.1 \\ x_5 & -0.5 & 34 & 5.3 \\ x_6 & 0.4 & 19 & 5.4 \\ x_7 & 1.1 & 11 & 5.5 \end{array} \longrightarrow D' = \begin{array}{cccc} & X_1 & X_2 & X_3 \\ x_1' & 0.1 & 0.5 & 1.0 \\ x_2' & 0.1 & 0.0 & 0.5 \\ x_3' & 0.4 & 0.0 & 0.2 \\ x_4' & 1.0 & 1.0 & 0.0 \\ x_5' & 0.0 & 0.7 & 0.3 \\ x_6' & 0.1 & 0.4 & 0.5 \\ x_7' & 0.3 & 0.2 & 0.7 \end{array}$$

DATA NORMALIZATION (Z-SCORE)

- Z-score or standard score normalization tells us how many standard deviations each entity value is from the attribute mean:

$$x_i' = \frac{x_i - \hat{\mu}}{\hat{\sigma}}$$

$$D = \begin{array}{c|ccc} & X_1 & X_2 & X_3 \\ x_1 & 0.2 & 23 & 5.7 \\ x_2 & 0.4 & 1 & 5.4 \\ x_3 & 1.8 & 0.5 & 5.2 \\ x_4 & 5.6 & 50 & 5.1 \\ x_5 & -0.5 & 34 & 5.3 \\ x_6 & 0.4 & 19 & 5.4 \\ x_7 & 1.1 & 11 & 5.5 \end{array} \longrightarrow D' = \begin{array}{c|ccc} & X_1 & X_2 & X_3 \\ x_1' & -0.5 & 0.2 & 1.7 \\ x_2' & -0.4 & -1.0 & 0.1 \\ x_3' & 0.3 & -1.1 & -0.9 \\ x_4' & 2.1 & 1.7 & -1.4 \\ x_5' & -0.9 & 0.8 & -0.4 \\ x_6' & -0.4 & -0.0 & 0.1 \\ x_7' & -0.1 & -0.5 & 0.7 \end{array}$$

DATA NORMALIZATION (Z-SCORE)

- Z-score or standard score normalization tells us how many standard deviations each entity value is from the attribute mean:

$$x_i' = \frac{x_i - \hat{\mu}}{\hat{\sigma}}$$

iClicker

1- What is the variance of a standard normalized attribute:

A: 0

B: 1

C: between 0 and 1

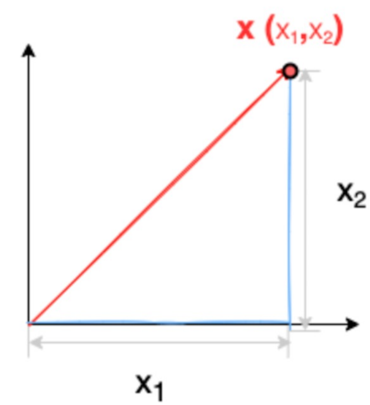
D: It depends

GEOMETRIC VIEW: DISTANCE BETWEEN VECTORS

- We are often interested in some measure of distance between vectors representing separate entities.
- First, some notation: norm of a vector with dimensions (columns/attributes). The length of a vector is a nonnegative number that describes the extent of the vector in space

$$||x||_2 = \sqrt{x_1^2 + x_2^2}$$

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



DISTANCE BETWEEN VECTORS

- We are often interested in some measure of distance between vectors representing separate entities.
- First, some notation: the L_2 norm of a vector with dimensions (columns/attributes):

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

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

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

DISTANCE BETWEEN VECTORS

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

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

where x_i and x_j are vectors, and there are dimensions

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

STATE UNIVERSITY

DISTANCE BETWEEN VECTORS

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

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

where x_i and x_j are vectors, and there are dimensions

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

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

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

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

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

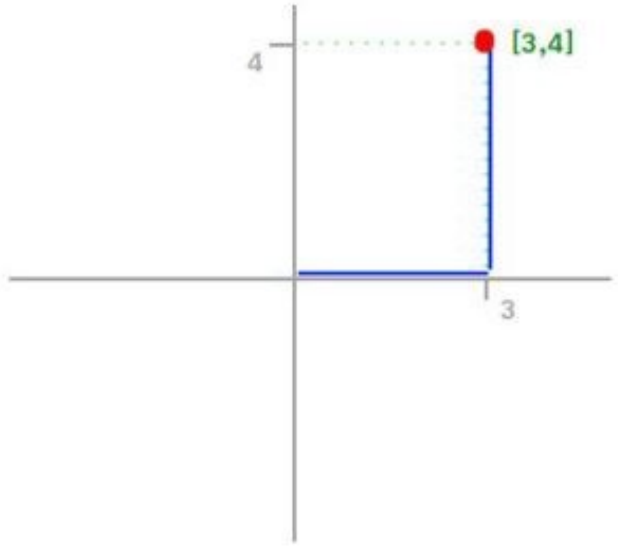
$$= 22.0$$

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:

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

where x_i and x_j are vectors, and there are dimensions



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 dimensions

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

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

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

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

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

$$= 22.5$$

MONTANA STATE UNIVERSITY

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.

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

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.
- is a measure of how closely two vectors align, in terms of the directions they point.
- Dot Product:

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

$$\vec{a} \cdot \vec{b} = \|\vec{a}\|_2 \, \|\vec{b}\|_2 \, \cos\theta$$

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:

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

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

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

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

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:

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

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

cosine of the angle between x_2 and x_3 is:

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

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

$$= 0.96$$

GEOMETRIC INTERPRETATION OF SAMPLE COVARIANCE

- Consider the the mean-centered data matrix:

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

- 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:

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

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:

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