CSCI 347: Introduction to Data Mining

Lecture 2a - Stats Review

COMMON DATA FORMATS

- Data can often be represented by a *data matrix D*

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

COMMON DATA FORMATS

- Data can often be represented by a *data matrix D*

The columns commonly represent attributes/properties of the data

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

COMMON DATA FORMATS

- Data can often be represented by a *data matrix D*

The columns commonly represent attributes/properties of the data

The rows commonly represent entities and their observed values for each attribute

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

COMMON DATA FORMATS

➤ Example:

$D =$

	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

COMMON DATA FORMATS

- Real Example from UCI Machine Learning Repository
    - link: https://archive.ics.uci.edu/ml/datasets/Online+Retail+II
    - Data set information: "This Online Retail II data set contains all the transactions occurring for a UK-based and registered, non-store online retail between 01/12/2009 and 09/12/2011.The company mainly sells unique all-occasion gift-ware. Many customers of the company are wholesalers."
    - 1067371 rows (entities), 8 columns (attributes)

$D =$

InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	*CustomerID*	*Country*
536365	85123*A*	WHITE HANGING HEART T-LIGHT HOLDER	6	12/1/108 : 26	2.55	17850	*UnitedKingdom*
536365	71053	WHITE METAL LANTERN	6	12/1/108 : 26	3.39	17850	*UnitedKingdom*
536365	84406*B*	CREAM CUPID HEARTS COAT HANGER	8	12/1/108 : 26	2.75	17850	*UnitedKingdom*
536365	84029*G*	KNITTED UNION FLAG HOT WATER BOTTLE	6	12/1/108 : 26	3.39	17850	*UnitedKingdom*
536365	84029*E*	RED WOOLLY HOTTIE WHITE HEART.	6	12/1/108 : 26	3.39	17850	*UnitedKingdom*
536365	22752	SET 7 BABUSHKA NESTING BOXES	2	12/1/108 : 26	7.65	17850	*UnitedKingdom*
536365	21730	GLASS STAR FROSTED T-LIGHT HOLDER	6	12/1/108 : 26	4.25	17850	*UnitedKingdom*
536366	22633	HAND WARMER UNION JACK	6	12/1/108 : 28	1.85	17850	*UnitedKingdom*
⋮	⋮	⋮	⋮	⋮	⋮		

COMMON DATA TYPES

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

$$D = \begin{array}{rcccc} & \text{temperature} & \text{length} & \text{type} & \text{weight} \\ \text{specimen 1} & 0.2 & 23 & A & 5.7 \\ \text{specimen 2} & 0.4 & 1 & B & 5.4 \\ \text{specimen 3} & 1.8 & 0.5 & C & 5.2 \\ \text{specimen 4} & 5.6 & 50 & A & 5.1 \\ \text{specimen 5} & -0.5 & 34 & A & 5.3 \\ \text{specimen 6} & 0.4 & 19 & B & 5.4 \\ \text{specimen 7} & 1.1 & 11 & A & 5.5 \end{array}$$

WHAT CAN WE LEARN FROM NUMERICAL DATA?

Statistics: **Mean**

Estimated mean (sample mean) of attribute j: $\hat{\mu}_j = \frac{1}{n} \sum_{i=1}^{n} x_{ij}$

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

WHAT CAN WE LEARN FROM NUMERICAL DATA?

Statistics: **Mean**

Estimated mean (sample mean) of attribute j: $\hat{\mu}_j = \frac{1}{n} \sum_{i=1}^{n} x_{ij}$

x_{32}

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

WHAT CAN WE LEARN FROM NUMERICAL DATA?

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

So:

$$\hat{\mu}_2 = \frac{1}{7}\sum_{i=1}^{7} x_{i2} = x_{12} + x_{22} + x_{32} + x_{42} + x_{52} + x_{62} + x_{72}$$

$$= \frac{1}{7}(23 + 1 + 0.5 + 50 + 34 + 19 + 11) = 19.79$$

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

WHAT CAN WE LEARN FROM NUMERICAL DATA?

Statistics: **Variance**

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

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

WHAT CAN WE LEARN FROM NUMERICAL DATA?

Statistics: **Variance**

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

Thus the estimated variance of X_2 in this example is:

$$\hat{\sigma}_2^2 = \frac{1}{6}((23 - 19.79)^2 + (1 - 19.79)^2 + (0.5 - 19.79)^2 + (50 - 19.79)^2 + (34 - 19.79)^2 + (19 - 19.79)^2 + (11 - 19.79)^2)$$

$$= 321.32$$

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

WHAT CAN WE LEARN FROM NUMERICAL DATA?

Statistics: **Standard deviation**

Estimated standard deviation of X_j: $\hat{\sigma}_j = \sqrt{\hat{\sigma}_j^2}$ (square root of the estimated variance)

Thus the estimated standard deviation of X_2 in this example is:

$$\hat{\sigma}_2 = \sqrt{\hat{\sigma}_2^2} = \sqrt{321.32} = 17.93$$

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

WHAT CAN WE LEARN FROM NUMERICAL DATA?

Statistics: M**ulti-dimensional mean**

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

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

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

WHAT CAN WE LEARN FROM NUMERICAL DATA?

Statistics: **multi-dimensional mean**

What is the estimated multi-dimensional mean of the (numerical) data?

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

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

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

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

WHAT CAN WE LEARN FROM NUMERICAL DATA?

Statistics: **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?

Statistics: **covariance**

What is the covariance between two attributes in a numerical data set?

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

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

First, we find $\hat{\mu}_1$ and $\hat{\mu}_2$:

$\hat{\mu}_2 = 1.3$ *and* $\hat{\mu}_2 = 19.8$

WHAT CAN WE LEARN FROM NUMERICAL DATA?

Statistics: **covariance**

What is the covariance between two attributes in a numerical data set?

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

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

First, we find $\hat{\mu}_1$ *and* $\hat{\mu}_2$ *:*

$\hat{\mu}_2 = 1.3$ *and* $\hat{\mu}_2 = 19.8$

Next, we use $\hat{\mu}_1$ ***and*** $\hat{\mu}_2$ ***to find*** $\hat{\sigma}_{12}$**:**

$$\hat{\sigma}_{12} = \frac{1}{6}((0.2 - 1.3)(23 - 19.8) + (0.4 - 1.3)(1 - 19.8) + (1.8 - 1.3)(0.5 - 19.8)$$

$$+(5.6 - 1.3)(50 - 19.8) + (-0.5 - 1.3)(34 - 19.8) + (0.4 - 1.3)(19 - 19.8) + (1.1 - 1.3)(11 - 19.8))$$

$$\hat{\sigma}_{12} = 18.4$$

IN-CLASS PROBLEM:

Find the covariance between X_2 and X_3 in the data matrix below:

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

IN-CLASS PROBLEM:

Find the covariance between X_2 and X_3 in the data matrix below:

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

We $\hat{\mu}_2 = 19.8$ and $\hat{\mu}_3 = 5.4$ to find $\hat{\sigma}_{23}$:

$$\hat{\sigma}_{23} = \frac{1}{6}((23 - 19.8)(5.7 - 5.4) + (1 - 19.8)(5.4 - 5.4) + (0.5 - 19.8)(5.2 - 5.4)+$$

$$+(50 - 19.8)(5.1 - 5.4) + (34 - 19.8)(5.3 - 5.4) + (19 - 19.8)(5.4 - 5.4) + (11 - 19.8)(5.5 - 5.4))$$

$$\hat{\sigma}_{23} = -1.09$$

COVARIANCE MATRIX

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

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

COVARIANCE MATRIX

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

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

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

$$\Sigma = \begin{pmatrix} 4.1 & \hat{\sigma}_{12} & \hat{\sigma}_{13} \\ \hat{\sigma}_{21} & 321.3 & \hat{\sigma}_{23} \\ \hat{\sigma}_{31} & \hat{\sigma}_{32} & 0.0 \end{pmatrix}$$

COVARIANCE MATRIX

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

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

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

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

WHAT CAN WE LEARN FROM NUMERICAL DATA?

Statistics: **total variance**

What is the **total variance** in a numerical data set?

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

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

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

WHAT CAN WE LEARN FROM NUMERICAL DATA?

Statistics: **correlation (Pearson's Correlation Coefficient)**

What is the correlation between two attributes in a numerical data set?

$$\hat{\rho}_{12} = \frac{\hat{\sigma}_{12}}{\hat{\sigma}_1 \hat{\sigma}_2}$$

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

DATA NORMALIZATION

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]

$$x'_{ij} = \frac{x_{ij} - \min_i\{x_{ij}\}}{\max_i\{x_{ij}\} - \min_i\{x_{ij}\}}$$

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

DATA NORMALIZATION

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]

$$x'_{ij} = \frac{x_{ij} - \min_i\{x_{ij}\}}{\max_i\{x_{ij}\} - \min_i\{x_{ij}\}}$$

$$D = \begin{array}{cccc} & X_1 & X_2 & X_3 \\ x_1 & 0.2 & 23 & 5.7 \\ x_2 & 0.4 & 1 & 5.4 \\ x_3 & 1.8 & 0.5 & 5.2 \\ x_4 & 5.6 & 50 & 5.1 \\ x_5 & -0.5 & 34 & 5.3 \\ x_6 & 0.4 & 19 & 5.4 \\ x_7 & 1.1 & 11 & 5.5 \end{array} \longrightarrow \quad x'_{21} = \frac{0.4 - (-0.5)}{5.6 - (-0.5)} = 0.1$$

DATA NORMALIZATION

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]

$$x'_{ij} = \frac{x_{ij} - \min_i\{x_{ij}\}}{\max_i\{x_{ij}\} - \min_i\{x_{ij}\}}$$

$$D = \begin{array}{cccc} & X_1 & X_2 & X_3 \\ x_1 & 0.2 & 23 & 5.7 \\ x_2 & 0.4 & 1 & 5.4 \\ x_3 & 1.8 & 0.5 & 5.2 \\ x_4 & 5.6 & 50 & 5.1 \\ x_5 & -0.5 & 34 & 5.3 \\ x_6 & 0.4 & 19 & 5.4 \\ x_7 & 1.1 & 11 & 5.5 \end{array} \longrightarrow D' = \begin{array}{cccc} & X_1 & X_2 & X_3 \\ x'_1 & 0.1 & 0.5 & 1.0 \\ x'_2 & 0.1 & 0.0 & 0.5 \\ x'_3 & 0.4 & 0.0 & 0.2 \\ x'_4 & 1.0 & 1.0 & 0.0 \\ x'_5 & 0.0 & 0.7 & 0.3 \\ x'_6 & 0.1 & 0.4 & 0.5 \\ x'_7 & 0.3 & 0.2 & 0.7 \end{array}$$

MEAN-CENTERING

Mean-centering shifts the data matrix mean to 0.

Mean-centering:

$$x'_{ij} = x_{ij} - \hat{\mu}_j$$

$$D = \begin{array}{cccc} & X_1 & X_2 & X_3 \\ x_1 & 0.2 & 23 & 5.7 \\ x_2 & 0.4 & 1 & 5.4 \\ x_3 & 1.8 & 0.5 & 5.2 \\ x_4 & 5.6 & 50 & 5.1 \\ x_5 & -0.5 & 34 & 5.3 \\ x_6 & 0.4 & 19 & 5.4 \\ x_7 & 1.1 & 11 & 5.5 \end{array} \longrightarrow x'_{21} = x_{21} - \hat{\mu}_1 = 0.4 - 1.3 = -0.9$$

MEAN-CENTERING

Mean-centering shifts the data matrix mean to 0.

Mean-centering:

$$x'_{ij} = x_{ij} - \hat{\mu}_j$$

$$D = \begin{array}{c|ccc} & X_1 & X_2 & X_3 \\ \hline x_1 & 0.2 & 23 & 5.7 \\ x_2 & 0.4 & 1 & 5.4 \\ x_3 & 1.8 & 0.5 & 5.2 \\ x_4 & 5.6 & 50 & 5.1 \\ x_5 & -0.5 & 34 & 5.3 \\ x_6 & 0.4 & 19 & 5.4 \\ x_7 & 1.1 & 11 & 5.5 \end{array} \longrightarrow D' = \begin{array}{c|ccc} & X_1 & X_2 & X_3 \\ \hline x'_1 & -1.1 & 3.2 & 0.3 \\ x'_2 & -0.9 & -18.8 & 0.0 \\ x'_3 & 0.5 & -19.3 & -0.2 \\ x'_4 & 4.3 & 30.2 & -0.3 \\ x'_5 & -1.8 & 14.2 & -0.1 \\ x'_6 & -0.9 & -0.8 & 0.0 \\ x'_7 & -0.2 & -8.8 & 0.1 \end{array}$$

DATA NORMALIZATION

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

$$x'_{ij} = \frac{x_{ij} - \hat{\mu}_j}{\hat{\sigma}_j}$$

$$D = \begin{array}{c|ccc} & X_1 & X_2 & X_3 \\ x_1 & 0.2 & 23 & 5.7 \\ x_2 & 0.4 & 1 & 5.4 \\ x_3 & 1.8 & 0.5 & 5.2 \\ x_4 & 5.6 & 50 & 5.1 \\ x_5 & -0.5 & 34 & 5.3 \\ x_6 & 0.4 & 19 & 5.4 \\ x_7 & 1.1 & 11 & 5.5 \end{array} \longrightarrow D' = \begin{array}{c|ccc} & X_1 & X_2 & X_3 \\ x'_1 & -0.5 & 0.2 & 1.7 \\ x'_2 & -0.4 & -1.0 & 0.1 \\ x'_3 & 0.3 & -1.1 & -0.9 \\ x'_4 & 2.1 & 1.7 & -1.4 \\ x'_5 & -0.9 & 0.8 & -0.4 \\ x'_6 & -0.4 & -0.0 & 0.1 \\ x'_7 & -0.1 & -0.5 & 0.7 \end{array}$$

AND NOW FOR SOMETHING DIFFERENT

- Show that the mean of the centered data matrix is **0.**

AND NOW FOR SOMETHING DIFFERENT

- Show that the mean of the centered data matrix is **0.**

- **Answer: Let z_i be the ith row of the centered data matrix. Then:**

$$\frac{1}{n}\sum_{i=1}^{n} z_i = \frac{1}{n}\sum_{i=1}^{n}(x_i - \hat{\mu}) = \frac{1}{n}\sum_{i=1}^{n}(x_i) - \frac{1}{n}\sum_{i=1}^{n}(\hat{\mu})$$

$$= \frac{1}{n}\sum_{i=0}^{n} x_i - \left(\frac{1}{n}\right)(n)(\hat{\mu}) = \hat{\mu} - \hat{\mu} = 0$$

Next Time

- Review of some Linear Algebra