River Kelly
CSCI 347
Homework 01
Feb 3, 2022

Problem 1 (2 points)

What are the two main types of attributes typically found in data?

Problem 1

The two main types of attributes typically found in data are: **Qualitative and Quantitative**.

Problem 2 (14 points)

Consider the following data matrix

$$D = \begin{array}{cccc} & X_1 & X_2 & X_3 \\ x_1 & 0.3 & 23 & 5.6 \\ x_2 & 0.4 & 1 & 5.2 \\ x_3 & 1.8 & 4 & 5.2 \\ x_4 & 6.0 & 50 & 5.1 \\ x_5 & -0.5 & 34 & 5.7 \\ x_6 & 0.4 & 19 & 5.4 \\ x_7 & 1.1 & 11 & 5.5 \end{array}$$

1. (2 points) What is the estimated mean of X_3?

$$\begin{aligned} \hat{\mu_3} &= \frac{1}{7}\sum_{i=7}^{7} x_{i3} \\ &= \frac{1}{7}(5.6 + 5.2 + 5.2 + 5.1 + 5.7 + 5.4 + 5.5) \\ &= 5.3857... \end{aligned}$$

Problem 2.1

$\hat{\mu_3} = \mathbf{5.4}$

2. (2 points) What is the estimated covariance between X_1 and X_3?

First, we find $\hat{\mu_1}$ and $\hat{\mu_3}$.

$$\begin{aligned}\hat{\mu_1} &= \frac{1}{7}\sum_{i=7}^{7} x_{i1} \\ &= \frac{1}{7}(0.3 + 0.4 + 1.8 + 6.0 + -0.5 + 0.4 + 1.1) \\ &= 1.3571\end{aligned}$$

$\hat{\mu_1} = 1.4$ and $\hat{\mu_3} = 5.4$ *($\hat{\mu_3}$ from above)*

Now, covariance between X_1 and X_3.

$$\hat{\sigma}_{13} = \frac{1}{n-1}\sum_{i=1}^{n}(x_{i1} - \hat{\mu_1})(x_{i3} - \hat{\mu_3})$$

$$\begin{aligned}\hat{\sigma}_{13} = \frac{1}{6}((0.3 - 1.4)(5.6 - 5.4) + (0.4 - 1.4)(5.2 - 5.4) + (1.8 - 1.4)(5.2 - 5.4) \\ + (6.0 - 1.4)(5.1 - 5.4) + (-0.5 - 1.4)(5.7 - 5.4) + (0.4 - 1.4)(5.4 - 5.4) \\ + (1.1 - 1.4)(5.5 - 5.4))\end{aligned}$$

Problem 2.2

$\hat{\sigma}_{13} =$ **-0.35**

3. (2 points) What is the estimated multi-dimensional mean of D?

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

$$\begin{aligned}\hat{\mu} = \frac{1}{7}((0.3 \quad 23 \quad 5.6) + (0.4 \quad 1 \quad 5.2) + (1.8 \quad 4 \quad 5.2) + (6.0 \quad 50 \quad 5.1) + (-0.5 \quad 34 \quad 5.7) \\ + (0.4 \quad 19 \quad 5.4) + (1.1 \quad 11 \quad 5.5))\end{aligned}$$

Problem 2.3

$\hat{\mu} =$ **(1.4 20.3 5.4)**

4. (2 points) What is the estimated variance of X_2?

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

First, we must find $\hat{\mu}$ *(i.e.* $\hat{\mu_2}$*)*.

$$\begin{aligned}\hat{\mu} &= \frac{1}{7}\sum_{i=7}^{7} x_{i2} \\ &= \frac{1}{7}(23 + 1 + 4 + 50 + 34 + 19 + 11) \\ &= 20.3\end{aligned}$$

Now, to find the estimated variance of X_2.

$$\begin{aligned}\hat{\sigma}_2^2 &= \frac{1}{6}((23-20.3)^2 + (1-20.3)^2 + (4-20.3)^2 + (50-20.3)^2 + (34-20.3)^2 + (19-20.3)^2 + (11-20.3)^2) \\ &= 300.571667\end{aligned}$$

Problem 2.4

$\hat{\sigma}_2^2 =$ **300.6**

5. (2 points) What is the covariance matrix of D?

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

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

Problem 2.5

$$\Sigma = \begin{pmatrix} 4.7 & 20.75 & -0.35 \\ 20.75 & 300.6 & 0.32 \\ -0.35 & 0.32 & 0.052 \end{pmatrix}$$

6. (2 points) What is the estimated correlation between X_1 and X_3?

$$\begin{aligned}\hat{\rho}_{13} &= \frac{\hat{\sigma}_{13}}{\hat{\sigma}_1 \hat{\sigma}_3} \\ &= \frac{-0.35}{\sqrt{4.7} \times \sqrt{0.052}} \\ &= \frac{-0.35}{0.49} \\ &= -0.70797\end{aligned}$$

Problem 2.6

$\hat{\rho}_{13} = -0.71$

7. (2 points) What is the total variance D?

$$\begin{aligned}\mathrm{Var}(D) &= \hat{\sigma}_1^2 + \hat{\sigma}_2^2 + \cdots + \hat{\sigma}_n^2 \\ &= 4.7 + 300.6 + 0.052 \\ &= 305.35\end{aligned}$$

Problem 2.7

$\mathrm{Var}(D) = 305.35$

Problem 3 (6 points)

Given $a, b \in \mathbb{R}^4$ (that is a fancy way of saying that a and b are 4-dimensional vectors with real values) where

$$\begin{aligned} a &= \begin{bmatrix} 2.0 & 5.0 & -2.6 & 6.0 \end{bmatrix} \\ b &= \begin{bmatrix} 15.0 & 2.5 & 4.0 & 4.0 \end{bmatrix} \end{aligned}$$

1. (2 points) What is $\|a - b\|_2$?

$$\begin{aligned} \|a - b\|_2 &= \sqrt{\sum_{k=1}^{4} (a_k - b_k)^2} \\ &= \sqrt{(2.0 - 15.0)^2 + (5.0 - 2.5)^2 + (-2.6 - 4.0)^2 + (6.0 - 4.0)^2} \\ &= \sqrt{222.81} \\ &= 14.9268 \end{aligned}$$

Problem 3.1

$\|a - b\|_2 = 14.93$

2. (2 points) What is $\|a - b\|_1$?

$$\begin{aligned} \|a - b\|_1 &= \sum_{k=1}^{m} \mid x_{ik} - x_{jk} \mid \\ &=\mid 2.0 - 15.0 \mid + \mid 5.0 - 2.5 \mid + \mid -2.6 - 4.0 \mid + \mid 6.0 - 4.0 \mid \\ &=\mid -13 \mid + \mid 2.5 \mid + \mid -6.6 \mid + \mid 2.0 \mid \\ &= 13 + 2.5 + 6.6 + 2.0 \\ &= 24.1 \end{aligned}$$

Problem 3.2

$\|a - b\|_1 = 24.1$

3. (2 points) What is the cosine of the angle between a and b?

$$\begin{aligned}
cos(\theta) &= \frac{a \cdot b}{\|a\|_2 \|b\|_2} \\
&= \frac{\begin{bmatrix}2.0 & 5.0 & -2.6 & 6.0\end{bmatrix} \cdot \begin{bmatrix}15.0 & 2.5 & 4.0 & 4.0\end{bmatrix}}{\sqrt{((2.0)^2 + (5.0)^2 + (-2.6)^2 + (6.0)^2)}\sqrt{((15.0)^2 + (2.5)^2 + (4.0)^2 + (4.0)^2)}} \\
&= \frac{(2.0)(15.0) + (5.0)(2.5) + (-2.6)(4.0) + (6.0)(4.0)}{\sqrt{71.76}\sqrt{263.25}} \\
&= 0.408167
\end{aligned}$$

Problem 3.2

$cos(\theta) = 0.41$

Problem 4 (3 points)

The following questions reference the Heart Disease data set from the UCI Machine Learning Repository: `https://archive.ics.uci.edu/ml/datasets/Heart+Disease`

1. (1 point) One attribute is named “cigs”. What information is stored in the “cigs” attribute?

Problem 4.1

The “cigs” attribute stores information about the number of **cigarettes per day**.

2. (1 point) How man rows (i.e., observations, entities, instances) are there in the data set?

Problem 4.2

Number of Instances: **303**

3. (1 point) How man attributes are there in the data set?

Problem 4.3

Number of Attributes: **75**

Tips and Acknowledgements

Make sure to submit your answer as a PDF on Gradscope and Brightspace. Make sure to show your work. Include any code snippets you used to generate an answer, using comments in the code to clearly indicate which problem corresponds to which code.

Acknowledgements: Homework problems adapted from assignments of Veronika Strnadova-Neeley.