CSCI 347
Homework 01

Problem 1 (2 points)

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

There are typically numerical attributes and categorical attributes in data.

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?

   (5.6 + 5.2 + 5.2 + 5.1 + 5.7 + 5.4 + 5.5) / 7 = **5.385**

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

   X1 mean = 1.36, X3 mean = 5.39.
   $\frac{1}{6}$((.3-1.36)(5.6-5.39)+(.4-1.36)(5.2-5.39)+(1.8-1.36)(5.2-5.39)+(6-1.36)(5.1-5.39)+(-.5-1.36)(5.7-5.39)+(.4-1.36)(5.4-5.39)+(1.1-1.36)(5.5-5.39) $= -\mathbf{0.347}$

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

   $\frac{1}{7}$ ((.3 23 5.6)+(.4 1 5.2)+(1.8 4 5.2)+(6 50 5.1)+(-.5 34 5.7)+(.4 19 5.4)+(1.1 11 5.5) = $(\mathbf{1.357}, \mathbf{20.286}, \mathbf{5.386})$

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

   $\frac{1}{6}$ ((23-20.29)^2+(1-20.29)^2+(4-20.29)^2+(50-20.29)^2+(34-20.29)^2+(19-20.29)^2+(11-20.29)^2 $= \mathbf{300.571}$

5. (2 points) What is the covariance matrix of D?
   $\sigma_{13} = -0.347$. (Question 2)

   $\frac{1}{6}$((.3-1.36)(23-20.29)+(.4-1.36)(1-20.29)+(1.8-1.36)(4-20.29)+(6-1.36)(50-20.29)+(-.5-1.36)(34-20.29)+(.4-1.36)(29-20.29)+(1.1-1.36)(11-20.29) = σ_{12} = 20.748.

$\frac{1}{6}$((5.6-5.39)(23-20.29)+(5.2-5.39)(1-20.29)+(5.2-5.39)(4-20.29)+(5.1-5.39)(50-20.29)+(5.7-5.39)(34-20.29)+(5.4-5.39)(29-20.29)+(5.5-5.39)(11-20.29) $= \sigma_{23} = 0.321$

$\sigma_2^2 = 300.571$. (Question 4)

$\frac{1}{6}$ ((.3-1.36)^2+(.4-1.36)^2+(1.8-1.36)^2+(6-1.36)^2+(-.5-1.36)^2+(.4-1.36)^2+(1.1-1.36)^2 $= \sigma_1^2 = 4.7$

$\frac{1}{6}$ ((5.6-5.38)^2+(5.2-5.38)^2+(5.2-5.38)^2+(5.1-5.38)^2+(5.7-5.38)^2+(5.4-5.38)^2+(5.5-5.38)^2 $=$ $\sigma_3^2 = 0.05$

$$\sum = \begin{matrix} 4.7 & 20.748 & -0.347 \\ 20.748 & 300.571 & 0.321 \\ -0.347 & 0.321 & 0.05 \end{matrix}$$

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

   $\hat{\rho_{13}} = \frac{\hat{\sigma_{13}}}{\hat{\sigma_1} * \hat{\sigma_3}}$
   $\hat{\rho_{13}} = \frac{-0.347}{2.16 * 0.224}$
   $\hat{\rho_{13}} = -\mathbf{.717}$

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

   $\sigma_1^2 = 4.7$, $\sigma_2^2 = 300.571$, $\sigma_3^2 = 0.05$
   $Var(D) = 4.7 + 300.571 + 0.05 = \mathbf{305.321}$

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 $\|a - b\|_2$?

   $\|a - b\|_2 = \sqrt{((2-15)^2 + (5-2.5)^2 + (-2.6-4)^2 + (6-4)^2} = \mathbf{14.927}$

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

   $\|a - b\|_1 = |2 - 15| + |5 - 2.5| + |-2.6 - 4| + |6 - 4| = \mathbf{24.1}$

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

   Question 2 Equation / Question 1 Equation = **0.4082**

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?

   ”cigs” stores the number of cigarettes per day smoked by the patient.

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

   There are **303** instances in the data set.

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

   This database contains **76** attributes, but all published experiments refer to using a subset of 14 of them.

Tips and Acknowledgements

Make sure to submit your answer as a PDF on Gradescope 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.