

Name: Michael Belmeur

Homework 1: CSCI 347: Data Mining

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.

- 1) [2 points] What are the two main types of attributes typically found in data?

Categorical & Numerical

- 2) [14 points] Consider the following data matrix D:

	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	50	5.1
x_5	-0.5	34	5.7
x_6	0.4	19	5.4
x_7	1.1	11	5.5

- (A) [2 points] What is the sample mean of X_3 ?

$$\frac{(5.6 + 5.2 + 5.2 + 5.1 + 5.7 + 5.4 + 5.5)}{7} = 5.4$$

(B) [2 points] What is the sample covariance between X_1 and X_3 ?

$$\frac{(0.3 + 0.4 + 1.8 + 6 - 0.5 + 0.4 + 1.1)}{7} = 1.36$$

$$\frac{1}{6} \left((0.3 - 1.4)(5.6 - 5.4) + (0.4 - 1.4)(5.2 - 5.4) + (1.8 - 1.4)(5.2 - 5.4) \right. \\ \left. + (6 - 1.4)(5.1 - 5.4) + (-0.5 - 1.4)(5.7 - 5.4) + (0.4 - 1.4)(5.4 - 5.4) \right. \\ \left. + (1.1 - 1.4)(5.5 - 5.4) \right) = -0.35$$

(C) [2 points] What is the (multivariate) sample mean $\hat{\mu}$ of the data set (your answer should be a vector)?

$$\frac{1}{7} \left(\begin{pmatrix} 0.3 & 23 & 5.6 \end{pmatrix} + \begin{pmatrix} 0.4 & 1 & 5.2 \end{pmatrix} + \begin{pmatrix} 1.8 & 4 & 5.2 \end{pmatrix} + \begin{pmatrix} 6 & 50 & 5.1 \end{pmatrix} \right. \\ \left. + \begin{pmatrix} -0.5 & 34 & 5.7 \end{pmatrix} + \begin{pmatrix} 0.4 & 19 & 5.4 \end{pmatrix} + \begin{pmatrix} 1.1 & 11 & 5.5 \end{pmatrix} \right) \\ = \begin{pmatrix} 1.4 & 20.3 & 5.4 \end{pmatrix}$$

(D) [2 points] What is the sample variance $\hat{\sigma}_2^2$ of X_2 ?

$$\frac{1}{7} (23 + 1 + 4 + 50 + 34 + 19 + 11) \\ = 20.3$$

$$\frac{1}{6} \left((23 - 20.3)^2 + (1 - 20.3)^2 + (4 - 20.3)^2 + (50 - 20.3)^2 + (34 - 20.3)^2 \right. \\ \left. + (19 - 20.3)^2 + (11 - 20.3)^2 \right) \\ = 300.6$$

(E) [2 points] What is the covariance matrix for this data?

$$\Sigma = \begin{pmatrix} 4.7 & 20.75 & -0.35 \\ 20.75 & 310.6 & 0.32 \\ -0.35 & 0.32 & 0.052 \end{pmatrix}$$

(F) [2 points] What is the correlation between X_1 and X_3 ?

$$\frac{-0,35}{\sqrt{4,7} \times \sqrt{0,052}} \\ = -0,71$$

(G) [2 points] What is the total variance of D ?

$$= 4,7 + 300,6 + 0,052 \\ = 305,35$$

3) [6 points] Let a and b be two 4-dimensional vectors:

$$a = (2,5, -2,6,6) \text{ and } b = (15,2,5,4,4)$$

(A) [2 points] What is $||a - b||_2$?

$$\sqrt{(2-15)^2 + (5-2,5)^2 + (-2,6-4)^2 + (6-4)^2} \\ = \sqrt{222,8} = 14,93$$

(B) [2 points] What is $\|a - b\|_1$?

$$|2-15| + |5-2.5| + |-2.6-4| + |6-4| \\ = 24.1$$

(C) [2 points] What is the cosine of the angle between a and b ?

$$\frac{2(15) + 5(2.5) + (-2.6)4 + 6(4)}{\sqrt{12^2 + 15^2 - 2.6^2 + 6^2} \sqrt{15^2 + 2.5^2 + 4^2 + 4^2}} \\ = 0.45$$

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>

Answer the following questions about the data set:

(A) [1 point] One attribute is named "cigs" What information is stored in the "cigs" attribute?

of cigarettes someone smokes each day

(B) [1 point] How many rows (entities/instances) are there in this data set?

303

(C) [1 point] How many attributes are there in this data set?

76 total 14 used