Name(s): ______________________________

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 and numerical

2) Consider the following data matrix D:

$$D = \begin{array}{cccc} & X_1 & X_2 & X_3 \\ x_1 & 0.3 & 23 & 5.6 \\ x_2 & 0.4 & 1 & 5.2 \\ x_3 & 1.8 & 4 & 5.2 \\ x_4 & 6 & 50 & 5.1 \\ x_5 & -0.5 & 34 & 5.7 \\ x_6 & 0.4 & 19 & 5.4 \\ x_7 & 1.1 & 11 & 5.5 \end{array}$$

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

$$\hat{\mu}_3 = \frac{1}{7}\sum_{i=1}^{7} x_{i3} = \frac{1}{7}(5.6 + 5.2 + 5.2 + 5.1 + 5.7 + 5.4 + 5.5) = 5.39$$

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

$$
\begin{aligned}
\hat{\sigma}_{13} = \frac{1}{6}&((5.6 - 5.39)(0.3 - 1.36) \\
&+(5.2 - 5.39)(0.4 - 1.36) \\
&+(5.2 - 5.39)(1.8 - 1.36) \\
&+(5.1 - 5.39)(6 - 1.36) \\
&+(5.7 - 5.39)(-0.5 - 1.36) \\
&+(5.4 - 5.39)(0.4 - 1.36) \\
&+(5.5 - 5.39)(1.1 - 1.36)) \\
= -\, 0.35
\end{aligned}
$$

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

$$\hat{\mu} = (1.36 \quad 20.29 \quad 5.39)$$

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

$$
\begin{aligned}
\hat{\mu}_2 &= \frac{1}{7}\sum_{i=1}^{7} x_{i2} = \frac{1}{7}(23 + 1 + 4 + 50 + 34 + 19 + 11) = 20.29 \\
\hat{\sigma}_2^2 &= \frac{1}{6}((23 - 20.29)^2 + (1 - 20.29)^2 \\
&+(4 - 20.29)^2 + (50 - 20.29)^2 \\
&+(34 - 20.29)^2 + (19 - 20.29)^2 \\
&+(11 - 20.29)^2) \\
&= 300.57
\end{aligned}
$$

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

$$
\Sigma = \begin{pmatrix} \hat{\sigma}_1^2 & \hat{\sigma}_{12} & \hat{\sigma}_{13} \\ \hat{\sigma}_{21} & \hat{\sigma}_2^2 & \hat{\sigma}_{23} \\ \hat{\sigma}_{31} & \hat{\sigma}_{32} & \hat{\sigma}_3^2 \end{pmatrix} = \begin{pmatrix} 4.7 & 20.75 & -0.35 \\ 20.75 & 300.57 & 0.32 \\ -0.35 & 0.32 & 0.05 \end{pmatrix}
$$

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

$$\hat{\rho}_{13} = \frac{\hat{\sigma}_{13}}{\hat{\sigma}_1 \hat{\sigma}_3} = \frac{-0.35}{(2.17)(0.22)} = -0.73$$

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

$$var(D) = \hat{\sigma}_1^2 + \hat{\sigma}_2^2 + \hat{\sigma}_3^2 = 4.70 + 300.57 + 0.05 = 305.32$$

3) Let **a** and **b** be two 4-dimensional vectors:

$a = (2,5,-2.6,6)$ and $b = (15,2.5,4,4)$

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

$$||a-b||_2 = \sqrt{\sum_{k=1}^{4} (a_k - b_k)^2} = \sqrt{(2-15)^2 + (5-2.5)^2 + (-2.6-4)^2 + (6-4)^2}$$

$$= \sqrt{222.81} = 14.93$$

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

$$||a-b||_1 = \sum_{k=1}^{4} |a_k - b_k| = |2-15| + |5-2.5| + |-2.6-4| + |6-4|$$

$$= 24.1$$

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

$$\frac{a^T b}{||a||_2 ||b||_2} = \frac{(2)(15) + (5)(2.5) + (-2.6)(4) + (6)(4))}{\sqrt{(2^2 + 5^2 + -2.6^2 + 6^2)}\sqrt{(15^2 + 2.5^2 + 4^2 + 4^2)}} = 0.45$$

4) 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?

How many cigarettes per day a person smokes.

(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?

75