David Hamlin January 21, 2021

CSCI 347 Problem 1

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

The two main types of attributes typically found in data are numerical and categorical. Numerical data is expressed in numbers, whereas categorical data is by groups

David Hamlin January 21, 2021

CSCI 347 Problem 2

[14 points] Consider the following data matrix D:

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

A) [2 points] What is the sample mean of X_3?
Sample Mean= 5.386

B) [2 points] What is the sample covariance between X_1 and X_3?
Sample Covariance= -0.347

C) [2 points] What is the (multivariate) sample mean $\hat{\mu}$ of the data set (your answer should be a vector)?
Sample Mean of data set= (1.357 20.285 5.385714285714286)

D) [2 points] What is the sample variance $\hat{\delta_2^2}$ of X_2?
Sample Variance= 300.571

E) [2 points] What is the covariance matrix for this data?
Covariance Matrix= $\begin{bmatrix} 4.703 & 20.748 & -0.347 \\ 20.748 & 300.571 & .321 \\ -0.347 & .321 & 0.0514 \end{bmatrix}$

F) [2 points] What is the correlation between X_1 and X_3
Correlation= 0.0818

G) [2 points] What is the total variance of D?
Total Variance= 0.051

Code **is** on **next** page

```
import math
import numpy as np
np.set_printoptions(suppress=True)

#PROBLEM A)
def sampleMean(df, column):
    column = column-1
    #I do this at the start so i can enter
    # column 1 because index of array starts at 0
    sum = 0
    for row in df:
        sum += row[column]
    #Adds up values in the column number and divides it by total number of rows
    mean = sum/len(df)
    return mean

#PROBLEM B)
def sampleCovariance(df, col1, col2):
    sum = 0
    mean1 = sampleMean(df, col1)
    mean2 = sampleMean(df, col2)
    col1 = col1 - 1
    col2 = col2 - 1

    for row in df: #formula for sample covariance
        sum+=((row[col1] - mean1) * (row[col2] - mean2))

    return sum/(len(df)-1)

#PROBLEM C)
def totSampleMean(df):
    return str(sampleMean(df, 1)) +"   " \
           +str(sampleMean(df, 2)) + "   "+str(sampleMean(df, 3))

#PROBLEM D)
def sampleVariance(df, col):
    return sampleCovariance(df, col, col)

#PROBLEM E)
def covarianceMatrix(df):
    s = np.zeros(shape=(len(df[0]),len(df[0])))
    #s is our covariance matrix, that populates it with 0s

    # and this populates it with the sample covariance
    for row in range (len(s[0])):
        for col in range (len(s[0])):
            s[row][col] = sampleCovariance(df, row+1, col+1)
            #its expecting human vals so i have to add one
    return s
```

```
#PROBLEM F)
def correlation(df, col1, col2):
    col1 = col1-1
    col2 = col2-1 #formula for correlation
    return sampleCovariance(df, col1, col2)/(math.sqrt(sampleVariance(
        df, col1))*math.sqrt(sampleVariance(df, col2)))

#PROBLEM G)
def totVariance(df):
    sum = 0
    for row in range(len(df)):
        sum+=sampleVariance(df, row)
        return sum

def main():

    #Create data matrix
    d = [[.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]]

    print("Data Matrix D =")
    for row in d:
        print()
        for element in row:
            print(str(element) + " ", end='')

    #Prints out answers by calling methods with appropriate params
    print("A) Sample Mean= " + str(sampleMean(d, 3)))
    print("B) Sample Covariance= " + str(sampleCovariance(d, 1, 3)))
    print("C) Sample Mean of data set= " + "("+ totSampleMean(d) + ")")
    print("D) Sample Variance= " + str(sampleVariance(d, 2)))
    print("E) Covariance Matrix= ")
    print((covarianceMatrix(d)))
    print("F) Correlation= " + str(correlation(d, 1, 3)))
    print("G) Total Variance= " + str(totVariance(d)))

main()
```

David Hamlin January 21, 2021

CSCI 347 Problem 3

[6 points] 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 $||a - b||_2$?

$$||x_i - x_j||_2 = \sqrt{\sum_{k=1}^{m} (x_{ik} - x_{jk})^2}$$

$$||x_i - x_j||_2 = \sqrt{\sum_{k=1}^{4} (x_{1k} - x_{2k})^2}$$

$$= \sqrt{(x_{11} - x_{21})^2 + (x_{12} - x_{22})^2 + (x_{13} - x_{23})^2 + (x_{14} - x_{24})^2}$$

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

$$= \sqrt{(169 + 6.25 + 43.56 + 4}$$

$$= \mathbf{14.927}$$

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

$$||x_i - x_j||_1 = \sum_{k=1}^{m} |(x_{ik} - x_{jk})|$$

$$||x_i - x_j||_1 = \sum_{k=1}^{4} |(x_{1k} - x_{2k})|$$

$$= |x_{11} - x_{21}| + |x_{12} - x_{22}| + |x_{13} - x_{23}| + |x_{14} - x_{24}|$$

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

$$= | - 13| + |2.5| + | - 6.6| + |2|$$

$$= \mathbf{24.1}$$

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

$$\frac{x_i^T x_j}{||x_i||_2 ||x_j||_2}$$

$$\frac{(2 \quad 5 \quad -2.6 \quad 6)^T (15 \quad 2.5 \quad 4 \quad 4)}{\sqrt{(2^2 + 5^2 + -2.6^2 + 6^2} \sqrt{(15^2 + 2.5^2 + 4^2 + 4^2}}$$

$$= \frac{56.1}{8.471 * 16.225}$$

$$= \mathbf{.408}$$

David Hamlin January 21, 2021

CSCI 347 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`

Answer the following questions about the data set:

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

The number of cigarettes per day

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

There are 303 rows in the dataset

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

There are 76 attributes, but only 14 attributes are used.