CSC 5800: Intelligent Systems

Homework- 1

Name: Soumyadeep Chatterjee Access ID: HQ8682

Problem 1:

(a) Sorting a student database based on student identification numbers. **No; this task mostly involves database operation.**

(b) By looking at a CT scan, a doctor wants to identify if a patient has cancer or not. There are a lot of labeled CT scans that the doctor will use for making the decision. **Yes, it belongs to classification category, because here the doctor is classifying a CT scan doctor wants to identify if a patient has cancer or not.**

(c) An image analyst obtains some new images and wants to automatically detect the number of distinct objects in the image. He doesn't have any prior information about these objects: **Yes, it belongs to the clustering group of data mining because the analyst counting/grouping similar objects.**

(d) Predicting the outcomes of tossing a (fair) pair of dice. **No, it does not involve a large dataset or predicting any kind of patterns.**

(e) Predicting the future stock price of a company using historical records. **Yes, it belongs to the regression group of data mining because using numerical values(historical records) we are trying to predict the future stock price.**

(f) In an Internet search engine company, there is a need to find potential users who will click a particular advertisement on the webpage. **Yes, it belongs to the classification group of data mining as it is classifying two kinds of users; one who will click and the other not.**

(g) Monitoring the heart rate of a patient for abnormalities: **Yes, it is anomaly detection; detecting abnormal patterns.**

(h) Extracting the frequencies of a sound wave: **No, it is a kind of signal processing**

Problem 2:

(a) Brightness as measured by a light meter: **Continuous and Ratio; brightness can be measured on a continuous scale and it has a true zero point.**

(b) Angles as measured in degrees between 0 and 360: **Continuous and interval. Interval between degree have no true zero point**

(c) Bronze, Silver, and Gold medals as awarded at the Olympics: **Discrete data; because medals represent distinct categories. Qualitative(nominal) has a meaningful order or ranking.**

(d) Time in terms of AM or PM: **this does not fit because for binary; AM/PM does not necessarily represent two options. Coming to discrete; if considered AM/PM it can be but if we consider 24 hour format with their continuous values. Continuous does n0t apply at any angle. It is also neither qualitative nor quantitative; it has a kind of hybrid order; because AM/PM simply are labels without any order nor AM/PM represent numerical values and no true zero point.**

(e) Military rank. **Discrete because military ranks have distinct categories. Ordinal, because military ranks have a well defined order.**

Problem 3:

a)

Basic description of the “iris” data matrix:

-> The dataset contains 150 rows(data points); representing 150 individual iris flowers
-> There are four features(columns) for each row:
- Sepal Length
- Sepal Width
- Petal Length
- Petal width
(all in centimeters)

-> The dataset includes a *categorical* variable species, which signifies the types of iris flowers
-> Setosa
-> Versicolor
-> Virginica
-> few rows of the dataset

	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
1	5.1	3.5	1.4	0.2	setosa
2	4.9	3.0	1.4	0.2	setosa
3	4.7	3.2	1.3	0.2	setosa
4	4.6	3.1	1.5	0.2	setosa
5	5.0	3.6	1.4	0.2	setosa
6	5.4	3.9	1.7	0.4	setosa
7	4.6	3.4	1.4	0.3	setosa
8	5.0	3.4	1.5	0.2	setosa
9	4.4	2.9	1.4	0.2	setosa
10	4.9	3.1	1.5	0.1	setosa

```r
# (a) Load iris.dat file (available in R) – Give the basic description of the
# data matrix; no.of data points, no. of features, no. of classes

view(iris)

# number of data points(rows)
nrow(iris)

# number of features(columns)
ncol(iris)

# number of classes/unique variables
length(unique(iris$Sepal.Length))
length(unique(iris$Sepal.Width))
length(unique(iris$Petal.Length))
length(unique(iris$Petal.Width))
```

```
> # number of data points(rows)
> nrow(iris)
[1] 150
>
> # number of features(columns)
> ncol(iris)
[1] 5
>
> # number of classes/unique variables
> length(unique(iris$Sepal.Length))
[1] 35
> length(unique(iris$Sepal.Width))
[1] 23
> length(unique(iris$Petal.Length))
[1] 43
> length(unique(iris$Petal.Width))
[1] 22
```

b)

-> Sepal Length
- Minimum Value: 4.30
- Maximum Value: 7.90
- Mean: 5.843
- Median: 5.80
- Standard Deviation: 0.8280
  a) A smaller standard deviation suggests that the data points are closer to the mean, which show less variable.
  b) *Sepal lengths are relatively consistent or tightly clustered around the mean.*

-> Sepal Width
- Minimum Value: 2.0
- Maximum Value: 4.40
- Mean: 3.057
- Median: 3.0
- Standard Deviation: 0.4358
  Sepal widths are relatively consistent or tightly clustered around the mean.

-> Petal Length
- Minimum Value: 1.0
- Maximum Value: 6.90
- Mean: 3.758
- Median: 4.35
- Standard Deviation: 1.7652
    a) Larger SD suggests data points are more spread out from the mean
    b) *Petal Lengths are not tightly clustered around mean. Some being longer or shorter*

-> Petal Width
- Minimum Value: 0.10
- Maximum Value: 2.50
- Mean: 1.19
- Median: 1.30
- Standard Deviation: 0.7622

A large SD signify that the petal widths vary significantly from the average; being wider or narrower

```
# (b) Give some basic statistics (such as mean, median, standard deviation,
# min, max) for each of these features

summary(iris)

sd(iris$Sepal.Length)
sd(iris$Sepal.Width)
sd(iris$Petal.Length)
sd(iris$Petal.Width)
```

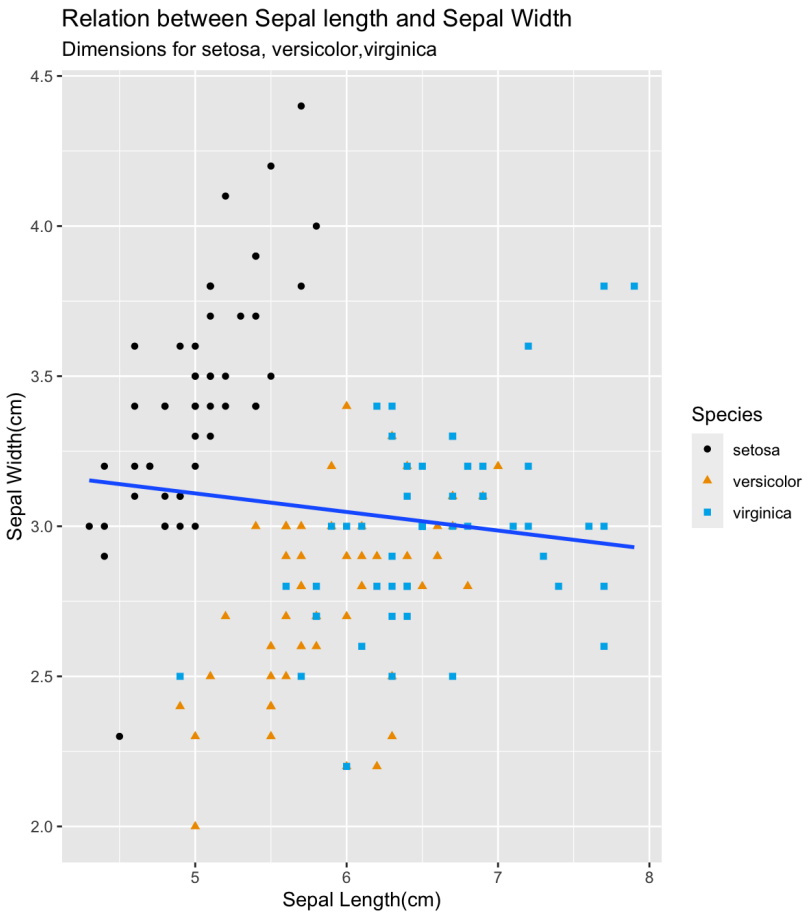
```
> summary(iris)
  Sepal.Length    Sepal.Width     Petal.Length    Petal.Width          Species
 Min.   :4.300   Min.   :2.000   Min.   :1.000   Min.   :0.100   setosa    :50
 1st Qu.:5.100   1st Qu.:2.800   1st Qu.:1.600   1st Qu.:0.300   versicolor:50
 Median :5.800   Median :3.000   Median :4.350   Median :1.300   virginica :50
 Mean   :5.843   Mean   :3.057   Mean   :3.758   Mean   :1.199
 3rd Qu.:6.400   3rd Qu.:3.300   3rd Qu.:5.100   3rd Qu.:1.800
 Max.   :7.900   Max.   :4.400   Max.   :6.900   Max.   :2.500
>
> sd(iris$Sepal.Length)
[1] 0.8280661
> sd(iris$Sepal.Width)
[1] 0.4358663
> sd(iris$Petal.Length)
[1] 1.765298
> sd(iris$Petal.Width)
[1] 0.7622377
```

c)
-> After plotting *quantitive variables sepal length and sepal width and qualitative variable species* in a scatter plot, where the x and y aesthetic are sepal length and sepal width.
-> The *scatter plot has a negative direction; signifying the independent variable(sepal length) is increasing and the dependent variable(sepal width) is decreasing.*
-> The *strength of the scatter plot is moderate; which signifies there is a noticeable but not extremely strong relationship between sepal length and sepal width. Also the negative correlation coefficient solidifies the fact that weak or very weak linear relationship between sepal length and sepal width.*
-> Also, there are some outliers present for all the three species.

```r
# (c) Plot the first two features of the data. Classes must be discriminated
# by using different symbols. Please label the figure.

library(tidyverse)
library(ggthemes)

ggplot(iris, aes(x = Sepal.Length, y = Sepal.Width))+
  geom_point(aes(color = Species, shape = Species))+
  geom_smooth(method = "lm", se= FALSE)+
  labs(
    title = "Relation between Sepal length and Sepal Width",
    subtitle = "Dimensions for setosa, versicolor,virginica",
    x = "Sepal Length(cm)", y = "Sepal Width(cm)",
    color = "Species", shape = "Species"
  )+
  scale_color_colorblind()
```

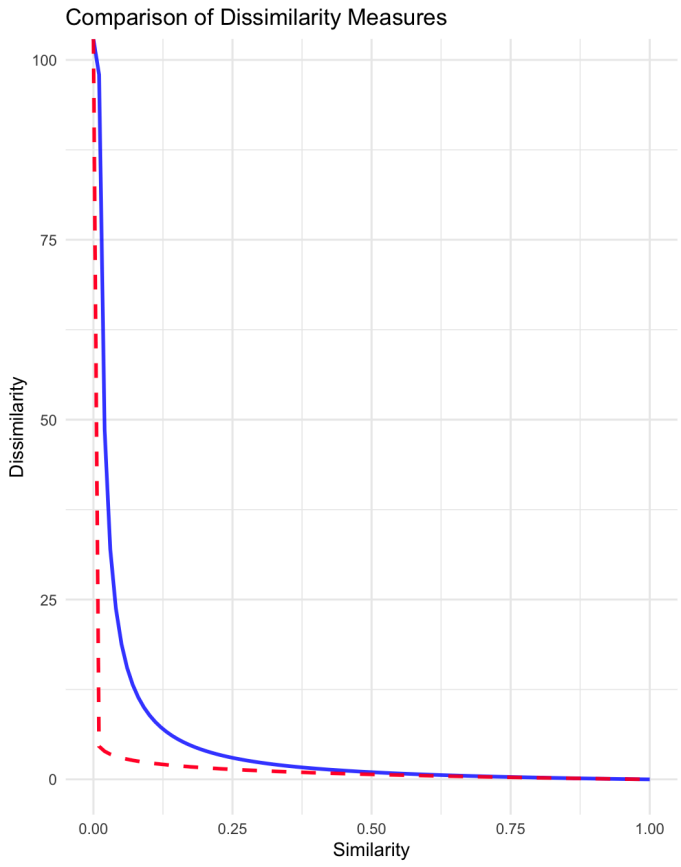
```r
cor(iris$Sepal.Length, iris$Sepal.Width)
```

```
> cor(iris$Sepal.Length, iris$Sepal.Width)
[1] -0.1175698
```

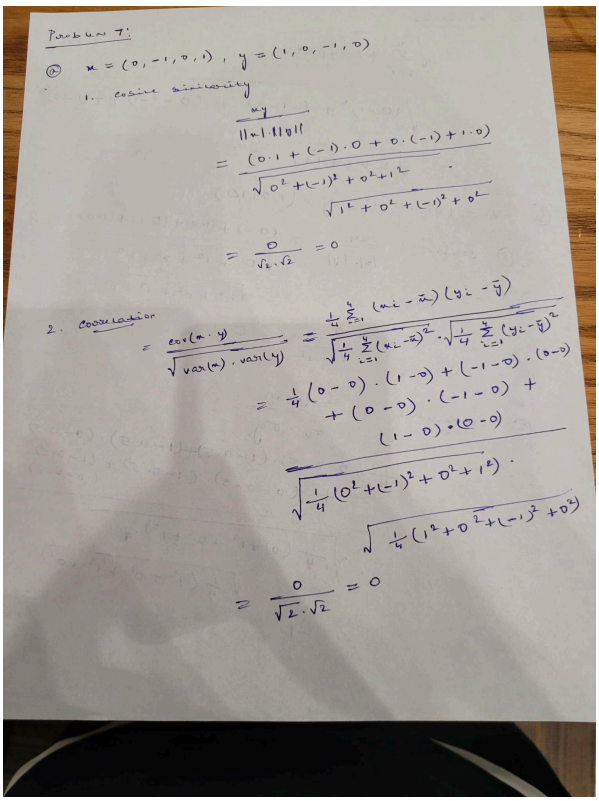

Relation between Sepal length and Sepal Width
Dimensions for setosa, versicolor,virginica
Sepal Width(cm)
Sepal Length(cm)
4.5
4.0
3.5
3.0
2.5
2.0
5
6
7
8
Species
setosa
versicolor
virginica

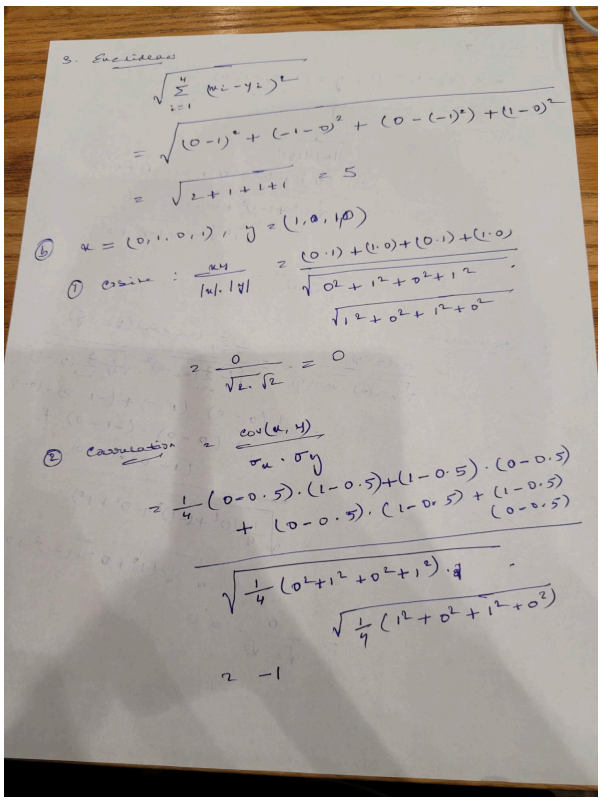
Problem 4:

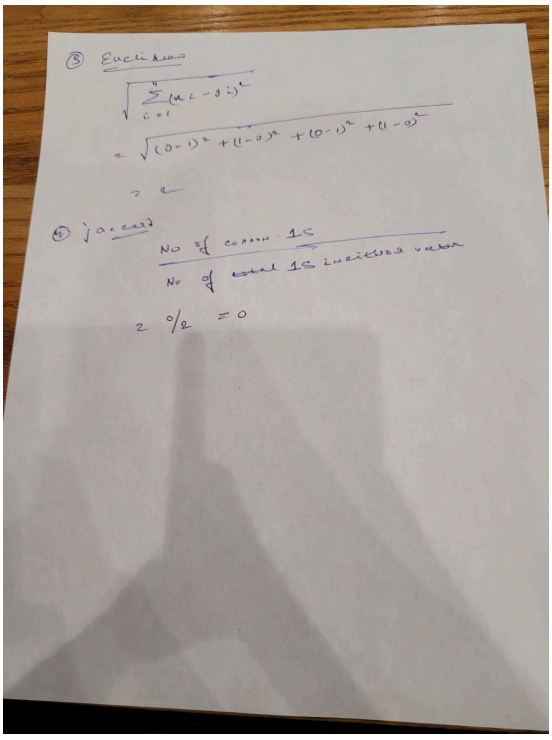
```r
# Problem 4

# Generate a sequence of similarity values
similarity_values <- seq(0, 1, length.out = 100)

# Calculate dissimilarity values for d1 and d2
d1_values <- (1 - similarity_values) / similarity_values
d2_values <- -log(similarity_values)

# Create a data frame for plotting
plot_data <- data.frame(similarity = similarity_values, d1 = d1_values, d2 = d2_values)

# Plotting
ggplot(plot_data, aes(x = similarity)) +
  geom_line(aes(y = d1), color = "blue", linetype = "solid", size = 1, alpha = 0.8, label = "d1") +
  geom_line(aes(y = d2), color = "red", linetype = "dashed", size = 1, alpha = 0.8, label = "d2") +
  labs(title = "Comparison of Dissimilarity Measures",
       x = "Similarity",
       y = "Dissimilarity",
       color = "Measure") +
  theme_minimal()
```



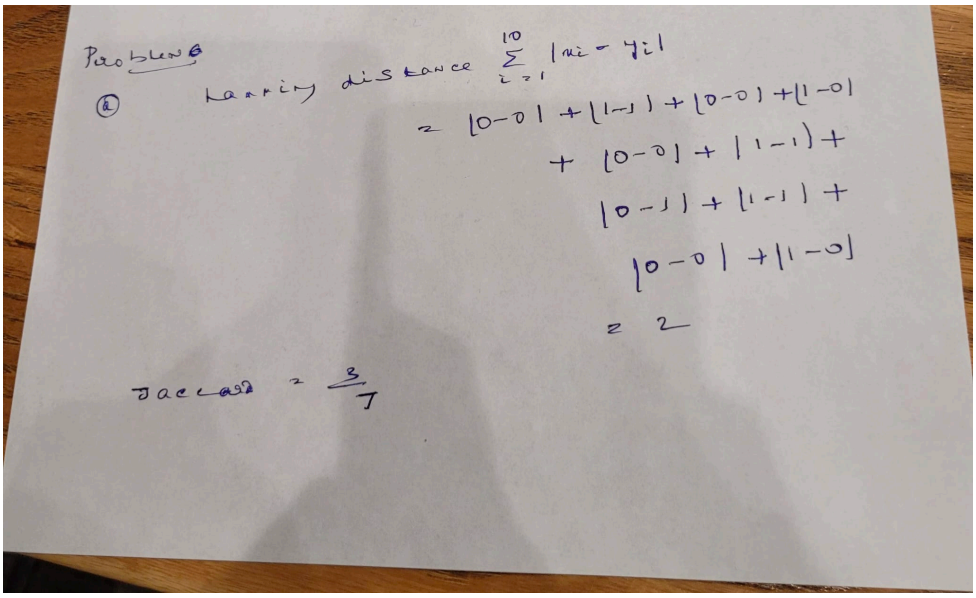
Problem 7:

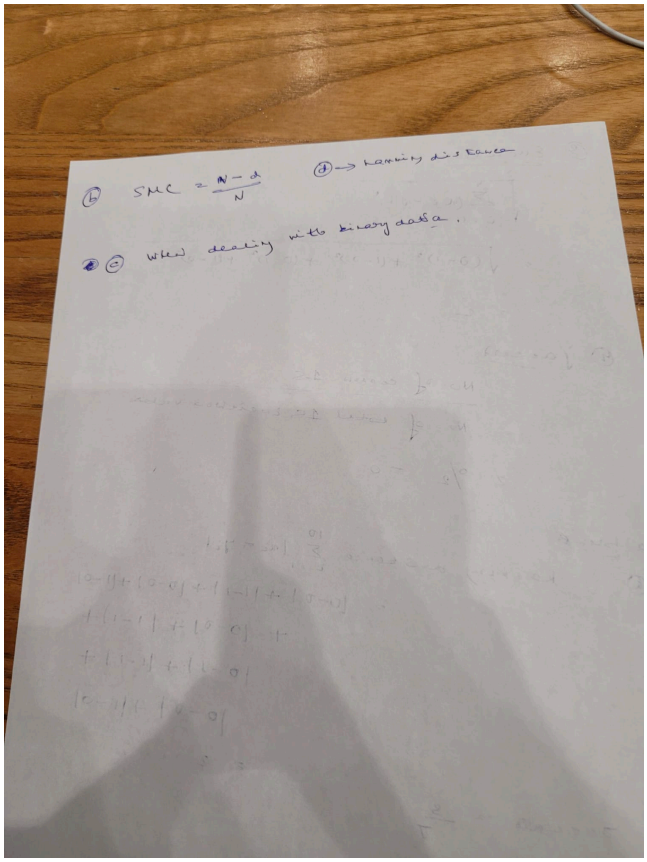

Problem 7:
a) x = (0, -1, 0, 1), y = (1, 0, -1, 0)
1. cosine similarity
2. Correlation


3. Euclidean
b) x = (0, 1, 0, 1), y = (1, 0, 1, 0)
1. cosine
2. Correlation


3. Euclidean
4. Jaccard

Problem 6:


Problem 6
(a) Hamming distance $\sum_{i=1}^{10} |x_i - y_i|$
$= |0-0| + |1-1| + |0-0| + |1-0| + |0-0| + |1-1| + |0-1| + |1-1| + |0-0| + |1-0|$
$= 2$
Jaccard $= \frac{3}{7}$


(b) SMC $= \frac{M-d}{N}$ (d) → Hamming distance
(c) When dealing with binary data.

Problem 5:

a) Effect of the Transformation:

- The effect is a scaling of the term frequency by the logarithm of the total number of documents (m).
- This scaling reduces the impact of terms that are present in very few documents, emphasizing the importance of terms that occur in a more substantial portion of the corpus.