Homework 8

Mariia Nikitash
Prepare your answers as a **single PDF file**.
Group work: You may work in groups of 1-3. Include all group member names in the PDF file. You may work with students in both sections (375-01, -02). Only one person in the group should submit to Canvas.
Due: check on Canvas.

1. Consider the toy dataset below which shows if 4 subjects have diabetes or not, along with two diagnostic measurements. (Note: do **NOT** write any code for this problem. The answers are to be computed by hand.)

Preg	**BP**	**HasDiabetes**	**Preg.Norm**	**BP.Norm**
2	74	No	0.5	1
3	58	Yes	1	0.2
2	58	Yes	0.5	0.2
1	54	No	0	0
2	70	?	0.5	0.8

a. Which variable is the "Class" variable?
   HasDiabetes

b. Normalize the Preg and BP values by scaling the minimum-maximum range of each column to 0-1. Filled in the empty columns in the table.

c. Predict whether a subject with Preg=2, BP=70 will have diabetes using the 1-NN algorithm and
   i. Using Euclidean distance on the original variables
      The nearest distance is 4, the nearest neighbor is "Row 1", prediction "No"
   ii. Using Euclidean distance on the normalized variables
      The nearest distance is 0.2, the nearest neighbor is "Row 1", prediction "No"
   iii. Using Manhattan distance on the original variables
      The nearest distance is 4, the nearest neighbor is "Row 1", prediction "No"
   iv. Using Manhattan distance on the normalized variables
      The nearest distance is 0.2, the nearest neighbor is "Row 1", prediction "No"

   For each of these cases, give the nearest distance, nearest neighbor (e.g., "Row 1" or "Row 2"), and prediction.

Euclidean Distance

$$|x - y| = \sqrt{\sum_{i=1}^{d}(x_i - y_i)^2}$$

I. Predict if Preg=2, BP = 70 will have diabetes using 1-NN algorithm

1. Use Euclidean dist. on orig. values:

1 row: $\sqrt{(2-2)^2 + (74-70)^2} = \sqrt{16} = \boxed{4}$ nearest dist Prediction: „No"

2 row: $\sqrt{(3-2)^2 + (58-70)^2} = \sqrt{1+144} \approx 12.04$

3 row: $\sqrt{(2-2)^2 + (58-70)^2} = \sqrt{144} = 12$

4 row: $\sqrt{(1-2)^2 + (54-70)^2} = \sqrt{1 + 256} = \sqrt{257} \approx 16.03$

II. Use Euclidean dist. on norm. variables :

1 row: $\sqrt{(0.5-0.5)^2 + (1-0.8)^2} = \sqrt{0.04} = \boxed{0.2}$ nearest dist Prediction "No"

2 row: $\sqrt{(1-0.5)^2 + (0.2-0.8)^2} = \sqrt{0.25 + 0.36} = \sqrt{0.61} \approx 7.8$

3 row: $\sqrt{(0.5-0.5)^2 + (0.2-0.8)^2} = \sqrt{0.36} = 0.6$

4 row: $\sqrt{(0-0.5)^2 + (0-0.8)^2} = \sqrt{0.25 + 0.64} = \sqrt{0.89} \approx 0.94$

Manhattan $\sum |x - y|$

III Use Manhattan dist. on orig. values :

1 row : $|2-2| + |74-70| = \boxed{4}$ nearest dist Prediction: „No"

2 row : $|3-2| + |58-70| = 1 + 12 = 13$

3 row : $|2-2| + |58-70| = 12$

4 row: $|1-2| + |54-70| = 1 + 16 = 17$

IV Use Manhattan dist on norm. values:

1 row: $|0.5-0.5| + |1-0.8| = \boxed{0.2}$ Nearest dist Prediction „No"

2 row: $|1-0.5| + |0.2-0.8| = 0.5 + 0.6 = 1.1$

3 row: $|0.5-0.5| + |0.2-0.8| = 0.6$

4 row: $|0.-0.5| + |0-0.8| = 0.5 + 0.8 = 1.3$

2. The `pima-indians-diabetes-resampled.csv` file on Canvas contains records indicating whether the subjects have diabetes or not, along with certain diagnostic measurements. All subjects are of Pima Indian heritage and this dataset is called the Pima Indian Diabetes Database[1]. The goal is to use the k-NN algorithm to predict if a subject has diabetes given some of the diagnostic measurements.

a. Read the data file [code]

```
pima_indians_diabetes_resampled <- read_csv("pima-indians-diabetes-resampled.csv")
```

b. What does “Preg” represent in the dataset? (2-3 sentences. Search for the Pima Indian Diabetes Database online and read up on its background.)

The datasets consists of several medical predictor variables and one target variable, Outcome. Predictor variables includes the number of pregnancies the patient has had, their BMI, insulin level, age, and so on. Respectively, “Preg” variable represent the number of times female patient of at least 21 years of age was pregnant.

c. 0 values in the Glucose column indicate missing values. Remove rows which contain missing values in the Glucose column. You should have 763 rows. [code]

```
pima <- pima_indians_diabetes_resampled %>% filter(Glucose != 0) %>% View()
```

d. Create three new columns/variables which are the normalized versions of Preg, Pedigree, and Glucose columns, scaling the minimum-maximum range of each column to 0-1 (you can use the code developed in class). [code]

```
normalize <- function(x) {
return ((x - min(x)) / (max(x) - min(x)))
}
```

Use this function to normalize Preg, Pedigree, and Glucose columns and assign then to new variables

```
norm_prim <- normalize(pima$Preg)
norm_pedigree <- normalize(pima$Pedigree)
norm_glucose <- normalize(pima$Glucose)

norm_pima <- pima %>% mutate(norm_prim, norm_pedigree,norm_glucose)
%>% View()
```

e. Split the dataset into train and test datasets with the *first 500 rows* for training, and the remaining rows for test. Do NOT randomly sample the data (though resampling is usually done, this hw problem does not use this step for ease of grading).

```
train <- norm_pima[1:500, ]
test <- norm_pima[501:763,]
```

[1] https://github.com/jbrownlee/Datasets/blob/master/pima-indians-diabetes.names

f. Train and test a k-nearest neighbor classifier with the dataset. *Consider only the normalized Preg and Pedigree columns*. Set k=1. What is the error rate (number of misclassifications)? [code, error rate]

```
trainfeatures <- norm_pima[1:500, c("norm_prim", "norm_pedigree") ]
testfeatures <- norm_pima[501:763, c("norm_prim", "norm_pedigree")]

trainlabels <- norm_pima[1:500, "HasDiabetes" ]
testlabels <- norm_pima[501:763, "HasDiabetes" ]

predicted <- knn(train = trainfeatures, test = testfeatures, cl=trainlabels$HasDiabetes,
k=1)

table(testlabels$HasDiabetes, predicted)
   predicted
      0   1
  0 120  50
  1  57  36
```

k=1 error rate is 107(errors) / 263(total num of predictions) == 40.68%

g. Repeat part (f) but *consider the normalized Preg, Pedigree, and Glucose columns*. Set k=1. What is the error rate? Will the error rate always decrease with a larger number of features? Why or why not: answer in 2-3 sentences? [code, error rate, answer]

```
trainfeatures <- norm_pima[1:500, c("norm_prim", "norm_pedigree", "norm_glucose") ]
 testfeatures <- norm_pima[501:763, c("norm_prim", "norm_pedigree", "norm_glucose")]

trainlabels <- norm_pima[1:500, "HasDiabetes" ]
testlabels <- norm_pima[501:763, "HasDiabetes" ]

predicted <- knn(train = trainfeatures, test = testfeatures, cl=trainlabels$HasDiabetes,
k=1)

 table(testlabels$HasDiabetes, predicted)
   predicted
      0   1
  0 128  42
  1  42  51
```

k=1 error rate is: 179(errors) / 263(total num of predictions) ==69.06%

The error rate will not always decrease with a larger number of features, since more features can lead to increased complexity and a risk of overfitting. In this example adding extra feature led to a higher error rate percent.

h. Repeat part (g) but set k=5. What is the error rate? [code, error rate]

```
predicted <- knn(train = trainfeatures, test = testfeatures, cl=trainlabels$HasDiabetes, k=5)
table(testlabels$HasDiabetes, predicted)
  predicted
    0   1
 0 149  21
 1  42  51
```

k=5 error rate is: 200(errors) / 263(total num of predictions) ==76.04%

i. Repeat part (h) but set k=11. What is the error rate? Considering your observations from (g)-(i), which is the best value for k? [code, error rate, answer]

```
predicted <- knn(train = trainfeatures, test = testfeatures, cl=trainlabels$HasDiabetes, k=11)
table(testlabels$HasDiabetes, predicted)
  predicted
    0   1
 0 154  16
 1  42  51
```

k=11 error rate is: 205(errors) / 263(total num of predictions) ==77.95%

1<5<11
The best value for k is k=1 since it has the lowest error rate.