Exercise 2

Jonas Brunvoll Larsson
Svein Jakob Høie

Task 1

Compute the Euclidean distance between the following sets of points

a) (2, 5) and (8, 4)

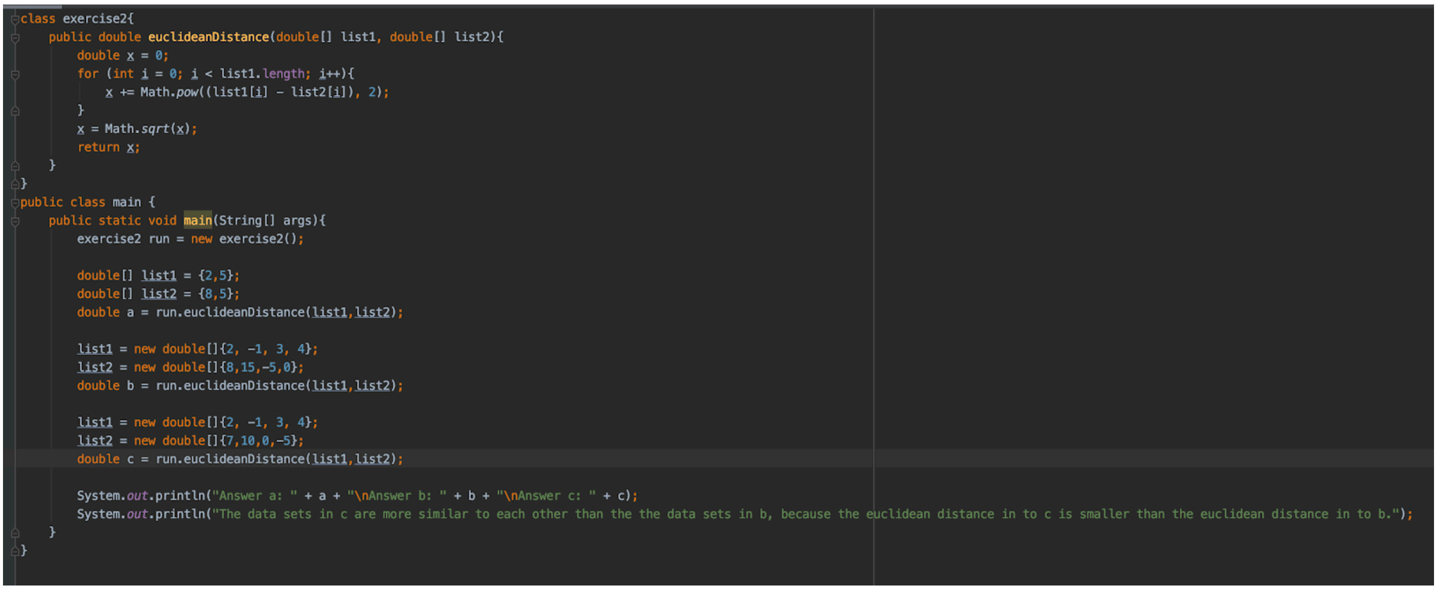
b) (2, -1, 3, 4) and (8, 15, -5, 0)

c) (2, -1, 3, 4) and (7, 10, 0, -5)

d) Compare b) and c). Which of the two sets of data points are more "similar" to each other

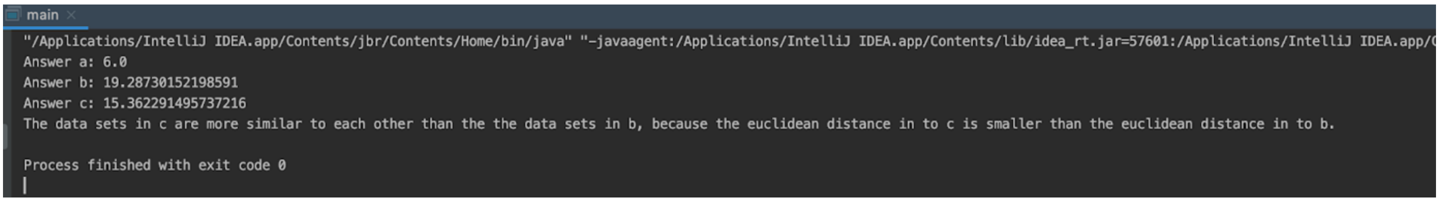
Answer:

Code:



```java
class exercise2{
    public double euclideanDistance(double[] list1, double[] list2){
        double x = 0;
        for (int i = 0; i < list1.length; i++){
            x += Math.pow((list1[i] - list2[i]), 2);
        }
        x = Math.sqrt(x);
        return x;
    }
}
public class main {
    public static void main(String[] args){
        exercise2 run = new exercise2();

        double[] list1 = {2,5};
        double[] list2 = {8,5};
        double a = run.euclideanDistance(list1,list2);

        list1 = new double[]{2, -1, 3, 4};
        list2 = new double[]{8,15,-5,0};
        double b = run.euclideanDistance(list1,list2);

        list1 = new double[]{2, -1, 3, 4};
        list2 = new double[]{7,10,0,-5};
        double c = run.euclideanDistance(list1,list2);

        System.out.println("Answer a: " + a + "\nAnswer b: " + b + "\nAnswer c: " + c);
        System.out.println("The data sets in c are more similar to each other than the the data sets in b, because the euclidean distance in to c is smaller than the euclidean distance in to b.");
    }
}
```

Output:



```
"/Applications/IntelliJ IDEA.app/Contents/jbr/Contents/Home/bin/java" "-javaagent:/Applications/IntelliJ IDEA.app/Contents/lib/idea_rt.jar=57601:/Applications/IntelliJ IDEA.app/C
Answer a: 6.0
Answer b: 19.28730152198591
Answer c: 15.362291495737216
The data sets in c are more similar to each other than the the data sets in b, because the euclidean distance in to c is smaller than the euclidean distance in to b.

Process finished with exit code 0
```

Task 2

Look at the Excel file Colleagues and Universities. The characteristics of these institutions differ quite widely. Suppose that we wish to cluster them into more homogeneous groups based on the median SAT, acceptance rate, expenditures/student, percentage of students in the top 10% of their high school, and graduation rate. Our task is to cluster the university into four groups. Answer the following questions.

a) What technique(s) will you use for this task? Are the techniques supervised or unsupervised learning? Describe your techniques and a possible result from your proposed solution.

b) We might need to code categorical data into numerical data before we can apply any algorithm. What does it mean?

c) We also need to normalize all features. What does it mean?

Answer:

a) There are two different data mining techniques we can use to solve this task. We can either use classification or clustering. Classification is a supervised learning technique. In this example we would use classification to classify each university to the correct classes. Because the classes in this example are pre decided, classification is probably the most useful technique. With clustering we would try to group the different universities by their similarities. Clustering is unsupervised, which is much more explorativ. A possible result of clustering the universities are new classes based on similarities we have not thought of beforehand.

b. An algorithm does not understand words as we do. Therefore we need to rewrite the input to a format the algorithm can make use of. A great example of categorical data translation is the translation of male and female. We can for example translate male = 0 and female = 1 or even male = TRUE and female =FALSE. These are values algorithms understand and can use for different data mining techniques.

c. Normalizing all features means rescaling features to avoid some features becoming too decisive. If not, unites with a wide range have a much bigger impact than unites with a small range. There is a large difference between the salaries of two persons who earn 10 000$ a year and 1 000 000$. Here we can normalize the features and decide that if you earn a salary less 50 000$ a year = 0, and earning a salary of 50 000$ or greater = 1.

Task 3

Look at the Excel file Credit Approval Decisions. In this case, we have a historical data set in the worksheet Data. For each record, we have a categorical variable of interest (Decision) and a number of predictor variables (Homeowner, Credit Score, Years of Credit History, Revolving Balance, Revolving utilization). The task here is to classify a set of new data in the worksheet Additional Data to decide if we should Reject or Approve these credit applications. Answer the following question.

- *a) What technique(s) will you use for this task? Are the techniques supervised or unsupervised learning? Describe your techniques and a possible result from your proposed solution.*
- *b) Use this task and task 2) to elaborate the difference between supervised and unsupervised learning.*
- *c) Suppose your algorithm resulted in the following confusion matrix for the test data. What is the True Positive Rate, True Negative Rate and Accuracy of your algorithm? Why is it useful to know these numbers?*

Answer:

a) For this task would classification absolutely be the best technique to choose. A good classification algorithm would be K Nearest Neighbor (KNN). This algorithm maps the applicants and compares them against old applicants. If a new applicant is more similar to older approved applicants, we can conclude that we probably also should approve the new applicant. Same goes for rejected applicants. The result of the additional data would be two classes. One we probably should accept, and one we probably should reject.

b) A supervised learning algorithm takes input and maps it to an output where the different output solutions are pre decided. This is useful when we want to classify something. Example: classify a customer. An unsupervised learning algorithm takes input and maps it to an output where the different output solutions are not pre decided. This is useful when we want to cluster groups based on similarities. Supervised learning is more strict and useful when we know what output we want, while unsupervised is more exploratory and useful for finding new classes based on similarities.

c. True positiv = TP

False Negativ = FN

False Positiv = FP

True negativ = TN
The true positive rate = TP / (TP + FN)
= 3 / (3+0) = 100%

The true negative rate = TN / (TN+FP)
= 2 /(2+1) = 66%

Accuracy = TP+TN / (TP+TN+FP+FN)
= 3+2 / (3+2+1+0) = 5 / 6 = 83%

It is useful to know these numbers for two reasons. Firstly, they give a good impression of how exact the classification model is and therefore how much we should trust this model. Secondly, the numbers are useful when choosing which classification model one should use.

Task 4

a. What technique(s) will you use for this task? Are the techniques supervised or unsupervised learning? Describe your techniques and a possible result from your proposed solution.

i.For this task I would use Association rule mining. This technique is an unsupervised learning algorithm that tries to identify insights between objects, often in retail sales as we see in this task.

b. How can your result be used to improve businesses?

.The result of the Association rule mining can be used in order to suggest users what they could buy along with what they are already buying. It can also be used in order to create deals and packages that contain products that the customer is more likely to find an interest in.

c. Consider that your algorithm discovered the following rules. Explain support, confidence and lift in the context of this example. How can they be used to choose which rules are most useful?

.**Support** identifies how often the combination appears. In this case, the combination of {15_inch_screen, 320_GB} has the highest support of the three entries, and we can therefore conclude that this combination is the most likely one to find. **Confidence** is an indication of how often the rule has been found to be true. In all three entries we have a confidence of 1. That means that for every combination LHS, the RHS is found to also be true. In our case, **lift** is greater than 1 in these entries. How much larger the lift is compared to 1, tells us something about the degree to which these occurrences are dependent on one another. A lift value of 3.9 tells us that this rule is potentially very useful for predicting consequent combinations in the future. The confidence attribute is especially useful to predict future occurrences.

Task 5

Business objective:

The proposal contains a clear objective of what they want to achieve. They will target 20,000 customers and attempt to migrate those customers over to the new application.

A project plan, though simple, is also provided. It says that High Arch Consulting will use data to build a model of whether or not a customer will migrate given the incentive. They will use a set of attributes related to the customers in order to find the most appropriate ones for the task.

The business success criteria is lacking in the proposal. It does not say how many of the customers should migrate for the project to be a success, whether it should be 90% or the customers, or a lower / higher percentage.

Data understanding:

The proposal states that High Arch Consulting will gather relevant data from the customers in order to predict whether they would migrate or not. Data such as usage of the app, location of the customer, tenure with the firm, and other loyalty indicators are some of the data they plan on using. The proposal could be improved on this topic, with adding summaries about the proposed data to be used, its quality and if there is any problem that could occur.

Data preparation:

The proposal does not contain information about the data preparation that has to be done, which could mean that High Arch Consulting believes that the data they will get from Big Blue will already be ready for usage.

Modelling:

The proposal states that High Arch Consulting will use a linear regression model to estimate target variables.If the estimate is greater than 0.5, they will predict that the customer will in fact migrate. HAC wants to ensure that the accuracy of the model is substantially greater than a random guess of customers.

Evaluation:

Evaluation is not specifically mentioned in the proposal, other than what was previously mentioned; they want to achieve greater accuracy than random guesses.

Deployment:

Based on the proposal it seems that Higher Arch Consulting plans on applying the regression model for each of the customers in order to estimate the target variable.

The proposal presented by Higher Arch Consulting could certainly be more detailed and informative, but it is still a proposal that gives the customer, Big Blue, an overview over what will happen and how. Specifically, the proposal should have a success criteria included.