Diamonds II

02450 Introduction to Machine Learning & Data Mining

Oriade Simpson (s172084) Pietro Lombardo (s231756)

From: 2023-10-20 To: 2023-10-31

Contents

Contribution Table

Task	Oriade	Pietro
Student ID	s172084	s231756
Question A.1	x	
Question A.2	x	
Question B.1	x	
Question B.2	x	
Question B.3	x	
Question C.1		x
Question C.2		x
Question C.3		x
Question C.4		x
Question C.5		x
Exam Problem 1	x	
Exam Problem 2		x
Exam Problem 3	x	
Exam Problem 4		x
Exam Problem 5	x	
Exam Problem 6		x

LINEAR REGRESSION

Section A

Question 1

Feature Transformation

Here, the price is converted from United States Dollars ($) to Danish Kroner (DKK) , Euro and Pound Sterling (£). The length, width and depth was converted from millimetres (mm) to micrometers (um). In addition to this, carat was converted to milligrams. The original columns for carat, length, width and depth were removed.

Outliers

The outliers are any values that lie above the upper boundary or below the lower boundary. There are 20 diamonds with the value of depth listed at 0 in the dataset. The smallest depth is 1,070 micrometers. Two diamonds have a width of 3,730 micrometers and 7 diamonds have the width listed as 0. There are also 8 diamonds that have a length of 0 in the dataset. It is important to deal with outliers because they may distort the statistical model.

Regression

The regression problem looks at the analysis of attributes in order to predict the carat of a diamond. In the multiple linear regression analysis, the price, table, length, width and depth of a diamond is used to compute the weight of a diamond.

Question 2

Section B

Question 1

Question 2

Question 3

CLASSIFICATION

Question 1

Question 2

Question 3

Exam Problems

Question 1

The answer

Question 2

Answer **D**: we have a dataset of $N = 135$ elements, divided in 4 classes as follows:

37-31-33-34 (R)

By considering a tree made of two branches based on the value of x_7, we obtain the two following sub-groups:

$x_7 = 2$ 0-1-0-0 (A) with $N_2 = 1$

$x_7 \neq 2$ 37-30-33-34 (B) with $N_2 = 134$

By computing the *classification error impurity measure* for each branch, we obtain:

$I_R = 1 - 37/135 = 0.726$; $I_A = 1 - 1 = 0$; $I_B = 1 - 37/134 = 0.724$

And finally we can calculate the purity gain based on the rule $x_7 = 2$:

$\Delta_2 = 0.726 - \frac{134}{135} \cdot 0.724 = 0.0074$

Question 3

The answer

Question 4

Answer **D**: we concentrate on the class 4 and we notice that it is the only one dependent only on b_1 (Fig. 4). We see from Fig. 3 that rules A and C lead to class 4, so those rules must regard conditions on b_1. By looking at the four possible answers, only answer **D** shows both A and C rules regarding b_1.

Question 5

The answer

Question 6

The answer

References