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

Regarding the classification problem, we want to train a model which labels whether a diamond has an *Ideal cut* or not. This aim appears to be feasible according to the projection of data onto the space defined by the first three Principal Components. From the top figure of page 11 of Report 1, it can be seen that by coloring data according to their *cut*, they appear to be clustered, especially the *Ideal* ones. So we aim to find a model which classifies diamonds in two binary classes: *Ideal* and *Non-Ideal* cut, according to their depth, table, price in DKK, carat in milligrams, length, width and depth in micrometers. We choose to use only continuous attributes and to ignore information coming from the color and clarity of diamonds.

Question 2 & 3

Different models can be trained to classify diamonds in *Ideal* and *Non-Ideal* cut. The simplest one is the Base-Line (BL), based only on the vector "y" of the outputs. In our case it is represented by the attribute *cut* transformed as follow:

$$y = \begin{cases} 1 & \text{if cut = "Ideal"} \\ 0 & \text{otherwise} \end{cases}$$

By computing the average of "y", we obtain the value of 0.4, meaning that the dataset contains more *Non-ideal* diamonds than *Ideal* ones. According to this information, the BL model always predicts a new diamond as *Non-Ideal*, regardless its characteristics (depth, table, etc...), with a classification error of 40%.

The other three models we want to analyse are:

- the Logistic Regression based on a linear combination of the attributes (LRL)
- the Decision Tree (DT)
- the Logistic Regression based on a quadratic combination of the first four Principal Components of the dataset (LRQ)

The choice of training a forth model is made because a quadratic combination of the 7 considered attributes is very hard to be trained (26 new columns representing the quadratic model would be added to the dataset, resulting in an "X" matrix of 33 columns): the computational time is too high and the process does not converge to a solution. So, by reducing the dimensions of the problem thanks to the Principal Component Analysis, we can consider a quadratic combination of the first four Principal Components (the "X" matrix turns out to have only 14 columns).

Each of the three above models requires a complexity parameter managing the regularization of the model. Higher values of the complexity parameter mean that large weights are penalized and data are less important in the training of the model. On the other hand, lower values of the complexity parameters allow the model to better follow data but to be less general in case of new data.

- Logistic regression models (LRL and LRQ) are regularized by the λ parameter, for which we do not know the value. Its value can be chosen by training the same model on the same data but with different values of λ. As the dataset consists of 53879 diamonds after the cleaning from the outliers, only four values of $lambda$ have been tempted in the training of the LRL and LRQ models. The tempted values are: $\lambda = \{10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}\}$. We will choose the λ associated to the lowest generalization error computed on a dataset of diamonds independent from the one used to train the model.
- Decision Tree is regularized by the c_P parameter (note that $c_P \in [0, 1]$). c_P close to zero means more complex decision trees and more importance of data, c_P close to 1 means easier decision trees and less importance to data. The selection of the best value of c_P is the same of the Logistic Regression. The tempted values are: $c_P = \{0.05, 0.01, 0.005, 0.001\}$. In addition, DT is dependent on two more parameters, that are the minimum number of data to create a new node (question) and the minimum number of data to create a leaf. The choice have been arbitrary made by taking them respectively equal to 100 and 1.

The large number of observations leads to choose a "light" cross-validation method, that is the K-fold partition of the dataset with a small number of folds. In particular, we choose 4 outer partitions and 6 inner folds. This means that each model will be trained:

$$\text{Nr. of trainings} = \text{4 outer folds } \cdot \text{6 inner folds } \cdot \text{4 complexity parameters } + \text{4 re-trainings } = \text{100 times}$$

i	E_i^{test} [%]	$olambda_i$	E_i^{test} [%]	$olambda_i$	E_i^{test} [%]	$c^*_{P,i}$	E_i^{test} [%]
1	39.73	1e-04	19.91	1e-05	12.47	0.005	11.41
2	40.44	1e-04	19.81	1e-04	13.51	0.001	11.98
3	40.36	1e-05	20.33	1e-05	12.87	0.001	11.95
4	39.41	1e-04	20.42	1e-05	13.21	0.005	11.76

Question 4

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