Test Topic DM 1B 2023-2024 - RESIT

Total 40 points (final version after deleting question 4: 34 points)

Q1: [4 points]; each question 2 points max, 1 point if for question a only one of the two is mentioned, -1 for each wrong item chosen

Consider a case study in which patients are at high risk of postoperative death after a hip fracture based on preoperative characteristics. Data is available from different hospitals on thousands of hip fracture patients. Typically, from each patient, several characteristics are known such as symptoms of dementia, age, living situation, risk of malnutrition, fracture type, Hb level, mobility score in the pre-fracture situation, length of stay in the hospital after surgery, post-operative wound infections, and survival after 30 days following hip fracture surgery.

We are interested in predicting for each patient the probability of postoperative death within 30 days after hip fracture surgery with logistic regression based on preoperative characteristics to support the decision for surgical treatment compared to nonoperative treatment.

a. What kind of problem is it? Select all answers that apply.

a. a supervised problem
b. an unsupervised problem
c. a classification problem
d. a regression problem

b. What might be the features (attributes / predictors)? Select all answers that apply.

a. the length of stay at the hospital after surgery
b. symptoms of dementia
c. age of the patient
d. Type of surgery

Remarks on a: you want to classify based on the probability of mortality the patients in order to decide whether they should have a surgery. The outcome variable in this case is death yes / no, so a qualitative outcome. This implies we are dealing with a classification problem. And since we want a model based on ***preoperative*** *characteristics only 'symptoms of dementia' and 'age of the patient' are correct.*

Q2 [5 points]:

Assume you have a dataset which you randomly split into a training and test dataset. Fill in below the correct answers:

If you have a model that has been fit on the training dataset, the training error is expected to be *[select the correct answer] larger than / smaller than / similar to* the test error.

In case of underfitting, both the training and test error are [select the correct answer] large / small.

In case of overfitting, the training error is [select the correct answer] larger / smaller than the test error.

A model suffering from underfitting will most likely be having [select the correct answer] high bias / high variance.

If you fit a model using K-nearest neighbours which of the following choices of the parameter K: [select the correct answer] K = 2 / K = 10, K = 20 will result in overfitting of your data?

Q3 [2 points]:

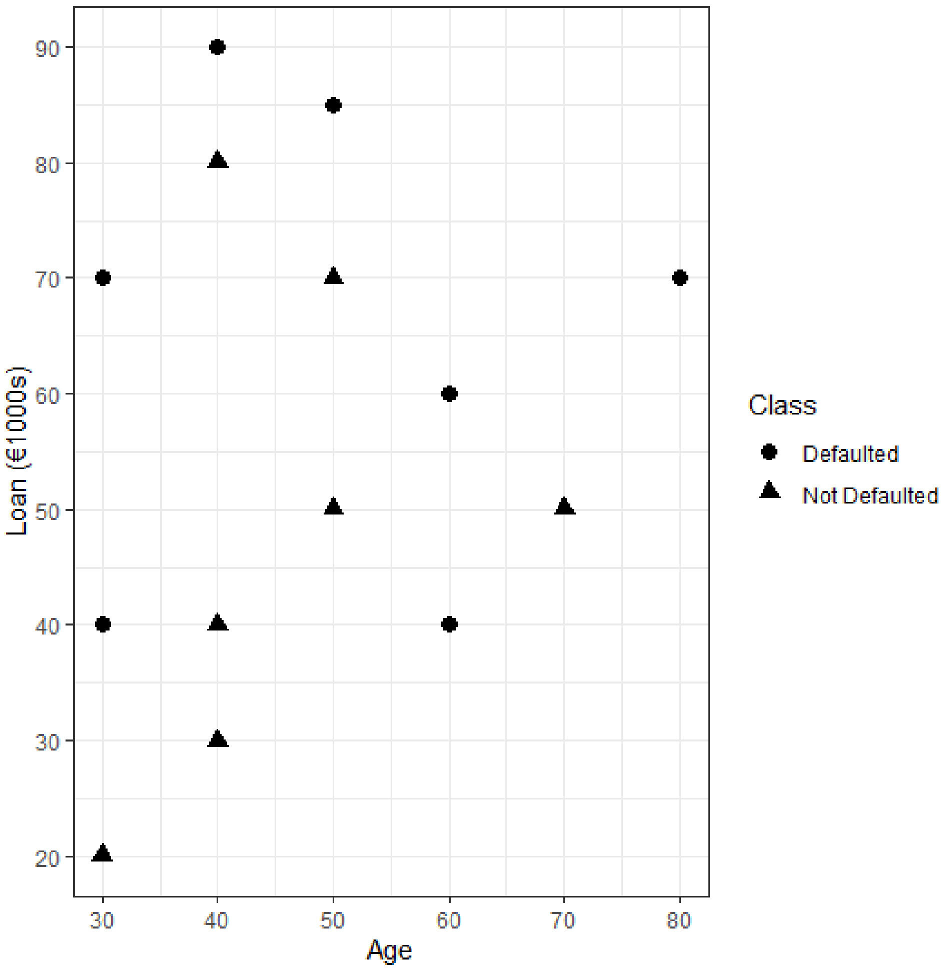
Assume you have a fixed training dataset (and test dataset) and you want to develop a classification model for that data. By sequentially adding parameters to give more flexibility to your classification model, you are more likely to observe (select all answers that apply):

a. A wider difference between train and test errors
b. A reduction in the difference between train and test errors
c. An increased or steady train error
d. A decrease in the train error

Q4 [6 points]:

THIS QUESTION WAS DELETED FROM THE TEST AS THE TEXT FOR THE FIRST QUESTION WAS MISSING

Below is a plot containing a number of data points with a known classification. The x-axis represents a person’s age, and the y-axis represents a loan amount in thousands of dollars. Each point is classified as either “Not Defaulted” (with a triangle), i.e. did pay back their loan or “Defaulted” (with a circle), i.e. did not pay back their loan.



a. [2points] If we assume Euclidean distance determines the similarity of two points, what Class will K-nearest neighbours (K-NN) with k = 3 predict for a 40 year old client with a loan of €70k?
   i. Defaulted
   ii. Not Defaulted

b. [2points] What will be the prediction for k = 5?
   i. Defaulted
   ii. Not Defaulted

c. [2points] If the loan amount was given in terms of euros instead of thousands of euros what would our K-NN model predict for the same client with k = 3?
   i. Defaulted
   ii. Not Defaulted

Q5 [2 points]:

A client has a classification problem and asks you to develop a model for this problem based on his data. You want to use regularized logistic regression (logistic regression in combination with lasso) as a method to obtain a model of the data. As a performance measure of your model, you choose the accuracy (percentage of correctly classified cases). The regularized logistic regression model has a hyperparameter lambda (the regularization parameter) that needs to be tuned. You will tune your model for this lambda parameter with a grid search. For validation of your model, you will use cross-validation. With cross-validation, you divide the dataset into 5 folds and for each split, you get an optimal value for *lambda* and an estimate of the accuracy. What is the **final model** that you will deliver to the client to use for his classification problem?

a. The regularized logistic regression model based on the whole dataset with the average value of *lambda* over the five folds.
b. A regularized logistic regression model based on one of the folds (taken randomly).
c. A regularized logistic regression model based on the whole dataset which is tuned for hyperparameter *lambda* with grid search on the whole dataset.

The purpose of validation of your model is to get a good estimate of your performance of your model. The final model is made on the whole dataset. See for a discussion on the final model: https://machinelearningmastery.com/train-final-machine-learning-model/

Q6 [2 points]:

What performance measures are most useful for a classification model?

a. Accuracy, Sensitivity, RMSE (Root mean squared error)
b. Precision, R^2, Sensitivity
c. Sensitivity, Precision, Accuracy
d. Accuracy, R^2, RMSE (Root mean squared error)

Remark: RMSE, R2 are measures for a regression model

Q7 [3 points]

Suppose that the regression equation relating systolic blood pressure (*y*, in mmHg), age (*x*, in years), and smoking (*z*, yes = 1, no = 0) is:

$$y = 98 + 0.8 * x + 5.0 * z + 0.1 * x * z$$

What is the increase in systolic blood pressure according to the model for a one-year increase in age for a smoker? Give your answer as a fraction with 1 decimal.

Answer: 0.8 + 0.1 = 0.9

Remark: the question is on the increase in systolic blood pressure if age is increased with one year. So that is something different than predicting the systolic blood pressure itself. Therefore the intercept doesn't play a role in the calculation

Q8 [16 points]:

You are working at a bank and you want to develop a classification model to discriminate whether clients who borrow money will default or not based on three attributes: Home Owner (Yes / No), Marital Status (Single / Married / Divorced), and Annual Income (High / Low). You choose to learn a Naïve Bayes classifier and a Decision Tree. You are given the following (noisy) examples:

Example	**X_1: Home Owner**	**X_2: Marital Status**	**X_3: Annual Income**	**Default**
Client 1	Yes	Single	High	No
Client 2	No	Married	High	No
Client 3	No	Single	Low	No
Client 4	Yes	Married	High	No
Client 5	No	Divorced	Low	Yes
Client 6	No	Married	Low	No
Client 7	Yes	Divorced	High	No
Client 8	No	Single	Low	Yes
Client 9	No	Married	Low	No
Client 10	No	Single	Low	Yes

Based on this data we have the following tables:

X_1: Home Owner		
	Default: Yes	Default: No
Yes	0	3
No	3	4
Total	3	7

X_2: Marital Status		
	Default: Yes	Default: No
Single	2	2
Married	0	4
Divorced	1	1
Total	3	7

X_3: Annual Income		
	Default: Yes	Default: No
High	0	4
Low	3	3
Total	3	7

Recall that Bayes' rule allows you to rewrite the conditional probability of a class given the attributes as the conditional probability of the attributes given the class:

$$\mathrm{P}(\mathrm{Default} \mid X_1, X_2, X_3) = \frac{\mathrm{P}(X_1, X_2, X_3 \mid \mathrm{Default})\ \mathrm{P}(\mathrm{Default})}{\mathrm{P}(X_1, X_2, X_3)}$$

Assume that the attributes ‘Home owner’, ‘Marital status’ and ‘Annual Income’ are **conditionally independent** given someone has (or hasn’t) defaulted.

a. [2 points] Calculate the conditional probability that a client who doesn’t default owns a house based on the available data, i.e. P(X_1 = Yes | Default = No). Give your answer as a fraction with 2 decimals. Answer: 3/7 = 0.43, correct = 0.425 – 0.431 exactly *(Remark: due to the inconsistency between the text (i.e. owns a house) and the formula (X_1 = No) in the test the answer based on the formula P(X_1 = No | Default = No) is also valid, i.e. 4/7 = 0.57; if rounding was not done appropriate 1 point is given i.e. 0.42 instead of 0.43).*

b. Consider a new example with the following attributes: Home owner = No, Marital Status = Divorced, Annual Income = Low.

a. [3 points] What is the probability that this client will default? (give your answer as a fraction with two decimals)

*P(Default= Yes| X_1 = No , X_2 = Divorced, X_3 = Low) = (3/3 * 1/3 * 3/3 * 3/10)/(3/3 * 1/3 * 3/3 * 3/10 + 4/7*1/7*3/7*7/10) = 0.80 (with more decimals:* 0.8032787*)*

b. [1 point] How would you classify this example according to the Naïve Bayes classifier?

i. Defaulted = Yes

ii. Defaulted = No

iii. Indecisive

Next, you decide to explore the use of a Decision tree to get another model.

a. [2 points] What is the classification error rate for the attribute ‘Marital Statues’? Give your answer as a fraction with 2 decimals. *3/10 = 0.3 [correct is 0.42 – 0.44]*

b. [3 points] What will be the splitting attribute in the top root of the Decision Tree if one uses the classification error rate ?

Classification error rate Home owner: 3/10

Classification error rate Marital Status: 3/10

Classification error rate Annual Income: 3/10

So all attributes give the smallest classification error rate.

i. Home owner

ii. Marital Status

iii. Annual Income

iv. The error rate of all attributes is the same, so all attributes could be chosen

c. [5 points] What is the overall Gini index for the attribute Marital Status? (give your answer as a fraction with 2 decimals). Recall that the general formula for the Gini index equals

$$G = \sum_{k=1}^{K} \hat{p}_{mk}(1 - \hat{p}_{mk})$$

Gi(Single) = 2 x 2/4 x 2/4 = 1/2

Gi(Married) = 0

Gi(Divorced) = 2 x 1/2 x 1/2 = 1/2

*Gi(Marital Status) = 4/10 * 1/2 + 0 + 2/10 * 1/2 = = 0.30*