ST4061 – Computer Intensive Statistical Analytics II

2021-2022

In-class test 2

NAME AND SURNAME: ……..

STUDENT NUMBER: ……..

PROGRAM: ……

INTRUCTIONS

- Provide your answers in this document, after each question item.
- Paste the R code you used for each question item.
- **Save your files regularly.**

Your Word document will be copied directly from your account for assessment.

Question 1

No R coding is required for this question. Figure 1 below shows the classification decision boundaries obtained by training four different models on a 2D dataset (X1,X2,Y) where Y is a 2-class categorical response variable. Indicate, in the table provided in your answer document, which of the scenarios represented in Figure 1 the models proposed in column A correspond to. (Indicate N/A if a model proposed in column A does not match any of the scenarios of Figure 1.) In column C, provide a brief explanation for your answer about each of the proposed models.

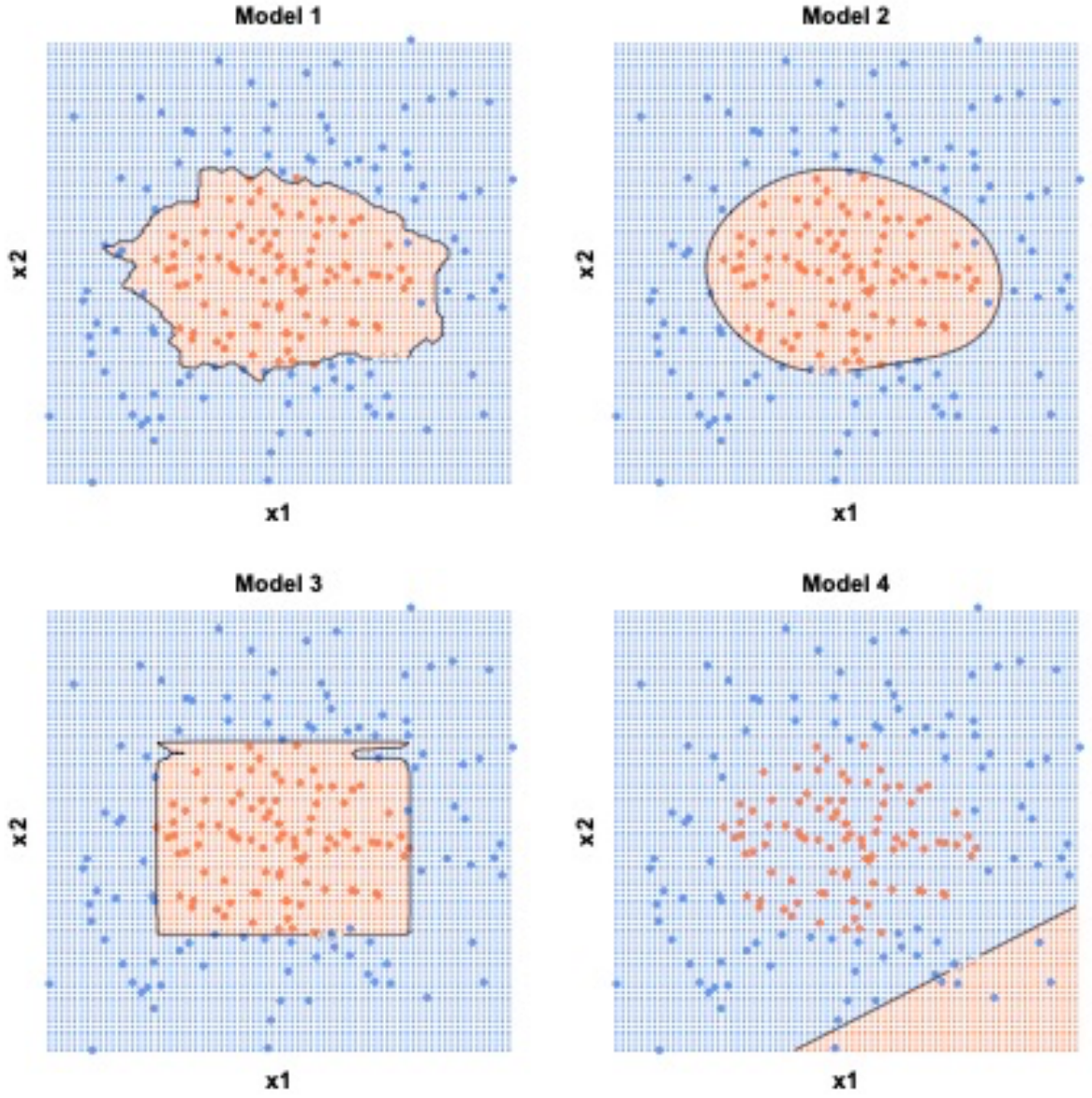

Figure 1 - Classification boundaries from 4 models on the same 2-class dataset (X1, X2, Y); blue and red dots depict the labelled data, and the areas correspond to the decision boundaries of each classifier.

Answers to Question 1

Table 1 - Match the models proposed in Column A to the models depicted in Figure 1 of Question 1. Indicate N/A when a model is not represented in Figure 1. Provide a brief explanation for each of your answers in column C.

(A) Proposed model	(B) Model of Figure 1	(C) Explanation for your answer in column B
A random forest		
A logistic regression model		
A Quadratic Discriminant Analysis with 3 Gaussian components		
A Support Vector Machine using a radial basis function		
A lasso classifier with an extremely large shrinkage parameter		
A kNN classifier (with k=5)		

Question 2

Run the following R instructions to load the required dataset and libraries:

```
library(randomForest)
library(gbm)
dat = read.csv(file="CA2_2021-22.dat", stringsAsFactors=TRUE)
```

Here the response variable of interest is `Sale_Price`, which corresponds to the selling price of a sample of US dwellings, in $1,000's. All other variables in the dataset are used as potential predictors. Do **not** perform any action on the predictors unless instructed to do so. Do **not** use any other package (such as `caret` or `tidymodels`) for this question. Provide your answers in the table below.

(1) Name which numerical features in the dataset have Pearson correlation of over 90% in absolute value, if any.

(2) Implement a simple (i.e. not repeated) 10-fold cross-validation framework, to train and test 3 models, namely a GLM, a random forest and a gradient boosting model, using all variables to predict sale prices. Set the random seed to 4061 before running the cross-validation code.

   (a) Report the mean cross-validated RMSEs for the 3 models.
   (b) Provide a boxplot of test-set RMSEs for the 3 models (within one figure).

(3) Perform the same cross-validation as in (2) but training the models on `log(Sale_Price)`.

   (a) Report the mean cross-validated RMSEs for the 3 models, in the scale of the original `Sale_Price` variable.
   (b) Provide a boxplot of test-set RMSEs for the 3 models (within one figure) , in the scale of the original `Sale_Price` variable.

(4) Explain any differences you may find between your results in (2) and (3). If you could not obtain a complete set of results in previous steps, describe what you would have expected to see.

Answers to Question 2

Question	Your answer
1	
2(a)	
2(b)	
3(a)	

3(b)	
4	

R code for Question 2: