**Lab01: MSE, Cluster Detection, Scaling, Binary Target Prediction and ROC**

**Handed out:** Wednesday, February 08, 2023

**Return date:** Wednesday, February 22, 2023, at the eLearning link **Lab01Submit** in the **Lab01** folder.

**Grades:** This lab counts 16 % towards your final grade

**Format of answer:** Your answers (statistical figures and verbal description) should be submitted electronically as Word document. Add a running title with the following information: Lab01, your name and page numbers. Use this document as template: add your answers underneath each subtask in a red color as well as any requested statistical figures. Trial and error answers will lead to a deduction of points. You are expected to hand in professionally formatted answers: use a fixed pitch font, like **Courier New**, for any  code and output.

## Task 1: MSE, Variance and Bias (3 points)

*This task does not require that you show  code.*

The **-script **MSEVarBias.R** generates a **nSim** samples of length **nLength** to evaluate the mean-square-error, variance and bias of a spline function estimator for an underlying population function **popFct**. The smoothness of the spline function estimator is controlled by the flexibility degree parameter **iFlex**.

[a] Describe how the sample datasets and the test dataset are calculated. What distinguishes the each of the sample datasets and the test dataset and what do they all have in common? (1 point)

[b] How are the variance and square bias calculated at a given flexibility **iFlex**? Do the calculation in the script match the theoretical equations and ? (1 point)

[c] Discuss the joint plot of variance, squared bias und mean-square error in relation to the flexibility of the spline smoother. What is the variance-bias tradeoff? (1 point)

## Task 2: Cluster Detection and Axis Scaling (4 points)

*This task does not require that you show  code.*

The **findClusters.r** script generates three equidistant clusters in the 2-dimensional space.

[a] Through experimentation find the smallest **offset** value in line 24 within the range , which allows the **mclust( )** estimator to identify these 3 clusters. Discuss why the clusters need to be sufficiently separated. (2 points)

[b] Activate the normalization of the - axis in just line 50 . How does the configuration of the clusters change? Why is the **mclust( )** estimator no longer capable of distinguishing the clusters? (1 point)

[c] Discuss in general terms why for *distance-based* machine learning methods the scaling of the individual variables is important to distinguish between objects? Use the 2-dimensional distance function with the - and -axes for your arguments. (1 point)

## Task 3: Classification of Binary Outcomes (6 points)

*Show the relevant  code underneath each sub-task. You can use the code of* ***kNNVersusLogistic.R*** *as template.*

Open the **Default** dataset in the package **ISLR2**. The binary variable **default** is your target variable.

[a] Prepare the data for a nearest neighbors prediction and generate **default**-stratified 30% test and 70% training samples. (1 point)

[b] Use nearest neighbors to identify the optimal accuracy -value. (1.5 points)

[c] For the optimal -value report the confusion matrix and the ROC-curve. Discuss both. (1 point)

[d] For the same training and test samples estimate a logistic regression model. (1 point)

[e] Report and discuss for the logistics regression model the confusion matrix and the ROC-curve. (1 point)

[f] Which estimation method performs better for the **Default** dataset? (0.5 points)

## Task 4: Receiver Operating Curve (3 points)

*Show the relevant  code underneath each sub-task.*

[a] Calculate the *sensitivity* and *specificity* at cut-off values for predicted *positive* probabilities Also draw the associated ROC diagram. The predicted probabilities and the true observed values are given in the table below:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | *True Observed* | *Predicted* |  | *True Observed* | *Predicted* |
| *1:* | *negative* | *0.00* | ***11:*** | *positive* | *0.55* |
| *2:* | *negative* | *0.05* | ***12:*** | *positive* | *0.60* |
| *3:* | *negative* | *0.10* | ***13:*** | *positive* | *0.65* |
| *4:* | *negative* | *0.15* | ***14:*** | *positive* | *0.70* |
| *5:* | *negative* | *0.20* | ***15:*** | *positive* | *0.75* |
| *6:* | *negative* | *0.25* | ***16:*** | *positive* | *0.80* |
| *7:* | *negative* | *0.30* | ***17:*** | *positive* | *0.85* |
| *8:* | *negative* | *0.35* | ***18:*** | *positive* | *0.90* |
| *9:* | *negative* | *0.40* | ***19:*** | *positive* | *0.95* |
| *10:* | *negative* | *0.45* | ***20:*** | *positive* | *1.00* |

*Table 1*

[b] Calculate the *sensitivity* and *specificity* at cut-off values for predicted *positive* probabilities . Also draw the associated ROC diagram. The predicted probabilities and true observed values are given in the table below:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | *True Observed* | *Predicted* |  | *True Observed* | *Predicted* |
| *1:* | *negative* | *0.55* | ***11:*** | *positive* | *0.00* |
| *2:* | *negative* | *0.05* | ***12:*** | *positive* | *0.60* |
| *3:* | *negative* | *0.65* | ***13:*** | *positive* | *0.10* |
| *4:* | *negative* | *0.15* | ***14:*** | *positive* | *0.70* |
| *5:* | *negative* | *0.75* | ***15:*** | *positive* | *0.20* |
| *6:* | *negative* | *0.25* | ***16:*** | *positive* | *0.80* |
| *7:* | *negative* | *0.85* | ***17:*** | *positive* | *0.30* |
| *8:* | *negative* | *0.35* | ***18:*** | *positive* | *0.90* |
| *9:* | *negative* | *0.95* | ***19:*** | *positive* | *0.40* |
| *10:* | *negative* | *0.45* | ***20:*** | *positive* | *1.00* |

*Table 2*

[c] Interpret and compare both ROC diagrams with respect to their underlying data in tasks 4 [a] and [b].