This is my first course uding this program and i am having trouble with it. I am asked to create an HTML file in R Markdown with 3 exercises:

Exercise 1

Develop a function to obtain a complete and customised assessment of a GLM model.

a GLM model. The function should therefore obtain:

- Exploratory analysis, using tables and graphs, of the relationship between the independent and dependent variables.

independent variables and the dependent variable. Use correlation plots and graphs,

Note: distinguish, according to the type of the dependent variable, the type of model, the type of dependent variable, the type of model, regression or classification (functions is.factor or is.numeric).

- A suitably formatted output of the coefficients estimated by the model, standard deviations, standard deviations, statistic,...

model, standard deviations, t-statistics and their p-values, both in their classical version (under model assumptions) and in their p-values.

(under the model assumptions) and Bootstrap.

- Ditto for the goodness-of-fit measures. Consider the R2, the adjusted R2 and the

ECM for regression and classification error and the pseudo R2 for classification.

- Results of the diagnostic tests of the model: Shapiro-Wilk, Lilliefors,

Levene/Brown-Forsythe, White, Breush-Godfrey, Durbin-Watson...

determining whether the residuals are normal, homoscedastic and random. Note:

using plot(model) you already get the graphical tests, but you can customise

the presentation using ggplot2 or any other graphical library (lattice, plotrix,

plotly,...). Given the results obtained, the function will return a message

advising whether or not to use the classical or Bootstrap estimators.

- Include an additional parameter, indicating whether or not to perform a selection of variables and, if so, a selection of variables.

of variables and, if so, what type: forward, backward, stepwise. Use

the BIC statistic as a criterion.

Note: to obtain the Bootstrap statistics you can rely on the three packages

implemented in R: rsample, bootstrap or boot, or make your own function for this purpose.

For classification problems, perform the bootstrap for each class separately. Use the data frame **BostonHousing** from the mlbench package to test the function in the case of regression.

*Exercise 2*

In the document on Introduction to Machine Learning an example of

Data Mining using the german credit database balanced with 750 records, 375 of each class.

records, 375 of each class. One of the estimated models has been a Logit model,

obtaining an area under the ROC curve of 0.771, very close to that of the best model, the Random Forest, with a value of 0.771.

Random Forest, with a value of 0.796.

It is to be expected that a GAM model will improve the GLM fit, but will it outperform the Random Forest?

Forest? To do this, develop a GAM by considering which type of regression to use (polynomial or curvilinear

(polynomial or curvilinear fit, Loess, regression spline, or smoothing spline) with each of the 3 quantitative independent variables

of the 3 quantitative independent variables: duration, amount and age. Note: the

implementation of caret does not currently allow this type of model to be properly fitted.

models.

As in the rest of the examples, use a Bootstrap estimation with 30

repetitions to obtain the 5 metrics proposed in the paper. Compare

statistically compare the results. Note: you can calculate CI of AUC with the pROC library,

auc() and ci.auc() functions. The rest of the metrics can be obtained through the confusion matrix obtained from the replications.

confusion matrix obtained from the Bootstrap replicates (note, the document shows the confusion matrix for the

the confusion matrix of the 750 data, not of the replicates).

Exercise 3

The **environmental** database of the lattice library presents the daily measurements of

ozone concentration, wind speed, temperature, and solar radiation in New York City from May to September of

New York City from May to September 1973. Specifically, 111

observations with the following information:

- Ozone. Average ozone concentration (hourly measurements) in parts per billion.

- Radiation. Solar radiation (from 08:00 to 12:00) in langleys.

- Temperature. Maximum daily temperature in degrees Fahrenheit.

- Wind. Average wind speed (at 07:00 and 10:00) in miles per hour.

Based on other studies, the beta coefficients or, in other words, the increase of the

ozone concentration as climatic conditions vary by one unit,

can be estimated as follows (at approximately 95% confidence):

radiation 0.05 to 0.15, temperature 1.3 to 1.7 and wind -3.7 to -3.3. We will assume the non-existence of of covariances.

For the accuracy (inverse of the residual variance), consider as parameters of the function

gamma function as the following two parameters: c0 = 56 and d0 = 24649.

Estimate a second model without including a priori estimates of the beta coefficients and their variances, considering c0 = d0 = 24649.

variances, considering c0 = d0 = 100.

Estimate a third model without incorporating a priori information. Comment on the default parameterisation and explain what this implies.

Compare the results with a classical or frequentist linear regression model. Note: The expectation and variance of the residual error in classical linear regression is:

Texto

Descripción generada automáticamente

(Ignore the spanish, just read the formula). Finally, estimate a Loess model, compare it with the rest and draw conclusions. Determine, with reasoning, which of them would be the best performer.