

Statistics and Statistical Data Mining

Coursework 1

You are required to conduct the analyses comprised in Tasks 1 and Task 2 below, and submit a report in pdf format, comprising your name and student number, a title, followed by your analyses - in particular the R code, results, comments, a conclusion for each task, and a list of References comprising the material and learning sources you used to produce your work (starting with the main text book of the module and the module's VLE url). For producing the report you may want to use R Markdown (which runs with RStudio).

Each task is worth equally, 50 marks.

Task 1: This regression task involves a subset of the Boston data set, that is provided here for the purpose of this coursework.

- (a) You are required to split the dataset in a training dataset comprising 2 thirds of the data, and 1 test dataset containing the remaining one third of the dataset, respectively. Prior to splitting the data, set the random seed to 35.
- (b) Using the training dataset, look for correlation in the data and remove correlation over 0.5 among predictors if present, then fit a multiple regression model to predict per capita crime rate, called *crim*.
- (c) For which of the predictors can you reject the null hypothesis $H_0 : \beta_j = 0$ at a significance level of 0.05?
- (d) On the basis of your response to the previous question (e), fit a smaller model on the training set that only uses the predictors for which there is evidence of association with the outcome. Continue eliminating predictors following the “Backward elimination” method presented in class, until no more predictors can be eliminated. The result of this step is twofold: 1. the final model fitting the data and using a smaller set of predictors, and 2. its set of predictors which are also called the selected predictors (selected features).
- (e) Provide an interpretation of each coefficient in the final model of (d).
- (f) Write out the final model of (d) in equation form.
- (g) How well do the model in (b) and the final model in (d) fit the training data?
- (h) Using the final model from (d), obtain 95 % confidence intervals for the coefficient(s).
- (i) Is there evidence of outliers or high leverage observations in the final model from (d)?
- (j) For each $k=1, 2$ and 3 , fit a k -nearest neighbour model on the training set with the predictors selected on point (d). Hint: Use *knn.reg* function from *FNN* (the Fast Nearest Neighbour) R package; see the documentation for this R package at <https://cran.r-project.org/web/packages/FNN/FNN.pdf>
- (k) Evaluate the models from (b), (d) and (j) on the test set, and decide which model is the most accurate based on test MSE.

Task 2: This classification task involves a subset of the Boston data set, provided here for the purpose of this coursework.

(a) Compute a new categorical variable called class having three values: “high”, “medium”, and “low”, defined using the per capita crime rate variable as follows. If the per capita crime rate is:

- above the 0.75 percentile, then class is defined as “high”;
- below the 0.25 percentile, then class is defined as “low”;
- between the 0.25 and 0.75 percentiles, then the class is defined as “medium”.

Remove per capita rate variable from the dataset and add the class variable.

(b) Split the dataset in a training set and test set comprising 2 thirds and 1 third of the data, respectively. The “high”, “medium” and “low” classes should be represented proportionally in the training and test set, as in the original dataset (that is, a stratified sampling). *Hint: Use createDataPartition function of caret R package <https://topepo.github.io/caret/data-splitting.html>*

(c) Fit a logistic regression model on the training set to predict class, after removing correlation over 0.6 among predictors, if correlation is present. Compute accuracy on the training and test sets.

(d) Redefine class variable on the whole data by collapsing the “medium” and “low” classes into “normal”, and keeping the “high” class. That is, you get 2 classes: “high” and “normal”.

(e) Do a stratified sampling with respect to the “high” and “normal” classes, based on 2 thirds and 1 third for the training and test sets, respectively.

(f) Using the training dataset, fit a logistic regression model to predict class variable using the other variables in this data set as input variables/ predictors.

(g) Which predictors have a significant contribution to the model, at 0.05 significance level?

(h) At your choice, perform either a backward elimination or forward selection of predictors as general methods explained in class, to get a reduced set of predictors, and a final (logistic regression) model on this set.

(i) Write out the final model of (h) in equation form that show the probability for an instance to belong to the “high” class.

(j) Interpret your final model of (h) from the standpoint of which predictors and values tend to be associated with a larger risk for the “high” crime level?

(k) Evaluate the final model of (h) on the training and test sets, computing ROC AUC, accuracy, sensitivity (recall, or true positive rate) and specificity (or true negative rate) with respect to class “high”. Draw the ROC curve on the training and test sets of the final model of (h). *Hint: You can install and use the pROC R package, see online documentation of this package <https://cran.r-project.org/web/packages/pROC/pROC.pdf>.*

(l) The predictions above are by default made using the probability threshold 0.5 to delimitate the 2 classes. You are required to find a better probability threshold for a balanced detection of the two classes “high” and “normal” which are imbalanced and may bias the model: in particular find the optimal probability threshold using the ROC curve, then re-do predictions and re-evaluate accuracy, sensitivity (recall) and specificity. *Hint: Optimal probability thresholds can be computed with pROC package using youden or topleft methods, see documentation of this package for guidance.*