Module 8: Recommended Exercises

TMA4268 Statistical Learning V2020

Martina Hall, Michail Spitieris, Stefanie Muff, Department of Mathematical Sciences, NTNU

March 05, 2020

Problem 1

We recommend that you work throught he lab in the course book (Section 8.3).

Problem 2 – Theoretical

a) Provide a detailed explanation of the algorithm that is used to fit a regression tree. What is different for a classification tree?

b) What are the advantages and disadvantages of regression and classification trees?

c) What is the idea behind bagging and what is the role of bootstap? How do random forests improve that idea?

d) What is an out-of bag (OOB) error estimator and what percentage of observations are included in an OOB sample? (Hint: The result from RecEx5-Problem 4c can be used)

e) Bagging and Random Forests typically improve the prediction accuracy of a single tree, but it can be difficult to interpret, for example in terms of understanding which predictors are how relevant. How can we evaluate the importance of the different predictors for these methods?

Problem 3 – Regression (Book Ex. 8)

In the lab, a classification tree was applied to the Carseats data set after converting Sales into a qualitative response variable. Now we will seek to predict Sales using regression trees and related approaches, treating the response as a quantitative variable.

a) Split the data set into a training set and a test set (todo: say which proportions; should it be half-half, or 80:20, 2:1?).

b) Fit a regression tree to the training set. Plot the tree, and interpret the results. What test MSE do you obtain?

c) Use cross-validation in order to determine the optimal level of tree complexity. Does pruning the tree improve the test MSE?

d) Use the bagging approach in order to analyze this data. What test MSE do you obtain? Use the `importance()` function to determine which variables are most important.

e) Use random forests to analyze this data. What test MSE do you obtain? Use the `importance()` function to determine which variables are most important. Describe the effect of m, the number of variables considered at each split, on the error rate obtained.

f) What is the effect of the number of trees (`ntree`) on the test error? Plot the test MSE as a function of `ntree` for both the bagging and the random forest method.

Problem 4 – Classification

In this exercise you are going to implement a spam filter for e-mails by using tree-based methods. Data from 4601 e-mails are collected and can be uploaded from the kernlab library as follows:

```
library(kernlab)
data(spam)
```

Each e-mail is classified by `type` (`spam` or `nonspam`), and this will be the response in our model. In addition there are 57 predictors in the dataset. The predictors describe the frequency of different words in the e-mails and orthography (capitalization, spelling, punctuation and so on).

a) Study the dataset by writing `?spam` in R.

b) Create a training set and a test set for the dataset (todo: please say which proportion should be in the training/test sets).

c) Fit a tree to the training data with `type` as the response and the rest of the variables as predictors. Study the results by using the `summary()` function. Also create a plot of the tree. How many terminal nodes does it have?

d) Predict the response on the test data. What is the misclassification rate?

e) Use the `cv.tree()` function to find the optimal tree size. Prune the tree according to the optimal tree size by using the `prune.misclass()` function and plot the result. Predict the response on the test data by using the pruned tree. What is the misclassification rate in this case?

f) Create a decision tree by using the bagging approach. Use the function `randomForest()` and consider all of the predictors in each split. Predict the response on the test data and report the misclassification rate.

g) Apply the `randomForest()` function again, but this time consider only a subset of the predictors in each split. This corresponds to the random forest-algorithm. Study the importance of each variable by using the function `importance()`. Are the results as expected based on earlier results? Again, predict the response for the test data and report the misclassification rate.

h) Use `gbm()` to construct a boosted classification tree. Predict the response for the test data and report the misclassification rate.

i) Compare the misclassification rates in d-h. Which method gives the lowest misclassification rate for the test data? Are the results as expected?