Compulsory exercise 3: Group 27

TMA4268 Statistical Learning V2018

Huglen, Huso and Myklebust

02 May, 2018

1a) Full classification tree

- Q1. The tree is contructed by using recursive binary splitting. This is a top-down approach because it begins at the root of the tree, successively splitting predictor space into two and two new branches as it moves down the tree. This method is also greedy because it takes the locally best choice by making the best split at each level. The split is made based on minimizing the deviance function for the relevant nodes. The deviance is a scaled version of the cross entropy criterion, which is minimized when the probability of an observation belonging to a class is either zero or one. Thus minimizing this expression ensures that a node mostly has observations from one class. The terminal nodes, the nodes at the bottom of the tree, are also known as leaves. Here, the predicted class for the observation is decided. The fact that the leaves are class labels is what makes the tree a classification tree.

b) Pruned classification tree

- Q2. The model includes alot of splits. That means that the observations from a training set may be partitioned into almost as many small regions as there are observations, and the probability of having overfitting increases. Pruning helps avoiding this by reducing the depth (and consequently the amount of splits) of the tree and thus reducing the amount of regions, yielding a higher probability of having more observations in each node. Thus we lower the probability of overfitting.
- Q3. The amount of pruning is decided by using cross-validation on the full tree model, in this case with 5-fold-CV. The classification error rate is used as evaluation for the cross-validation procedure. We pick the number of nodes in the model with the lowest error rate. This is the number of nodes in the pruned model.
- Q4. As the complexity of the model decreases with the pruned tree, it also becomes more interpretable. However, the AUC is slgithly lower for the pruned tree than for the full tree, and the misclassification rate has also increased a bit. However, one might still prefer the pruned tree if one wishes to have an easily interpreted model (since the two models perform very similarly).

c) Bagged trees

- Q5. The main motivation behind bagging is to decrease the amount of variance in the decision tree by looking at the results from many different tree models.
- Q6. Variable importance plots visualize the relative importance of each of the predictors in the model. The higher the variables are ranked, the higher is their importance. The importance, e.g. for the Gini index, is interpreted as average decrease in node impurity over splits for the predictors. Impurity here means the presence of observations from more than a single class. For this data set we observe that "checkaccount" and "amount" are important predictors, dependent on how we calculate the importance. The least important predictor is in both cases the "foreign". We see that the ranking depends slightly on whether we use the Gini Index or mean decrease in accuracy for calculating the importance.
- Q7. For bagging the AUC-value is significantly higher than for the full model. On the other hand, the interpretability is worse since bagging includes many full trees. And you have to e.g. compute the majority vote to classify an observation, while for a full tree you just have a single classification.

d) Random forest

- Q8. The parameter mtry specifies the number of variables to be considered at each split. These are picked randomly. The motivation for using 4 as the number of variables here is that it is roughly equal to the square root of the number of predictors, which is a standard choice for classification trees.
- Q9. A common problem with bagging is that when there are some very strong predictors, they are often chosen as top splits in the bagging trees. This results in many similar trees. Thus there is a high correlation between them. Therefore one does not achieve the desired reduction in variance. To avoid this, random forest is used. Here one only considers a randomly picked subset, m, of the predictors at each split, thus forcing the trees to become less correlated. For classification trees m is usually set to be approximately equal to the square root of the total number of predictors.
- Q10. We see that the AUC is a bit higher for the random forest than for the bagging. On the other hand, the misclassification rate is slightly higher for random forest. Since the models perform relatively equally on these data we prefer the random forest since this model in theory gives less correlated trees, and therefore more accurate predictions. If we had tried to create the models based on several different seeds, the random forrest would probably outperform bagging on average.

Problem 2 - Nonlinear class boundaries and support vector machine

a) Bayes decision boundary

- Q11. A Bayes classifier is a classification rule that classifies an observation to the class k for which $Pr(Y = k|X = x)$ is the greatest. That is, we classify to the class for which the probability is the greatest, given the observed values for x. A Bayes decision boundary consists of the points where there is an equal chance of an observation being in either of the classes. In a two class setting this corresponds to the points for which there is a 50% chance of an observation being in either class. In a classification setting the test error rate is defined as the average number of misclassifications on a test set. The Bayes classifier achieves the minimum of this error rate, and this is called the Bayes error rate. Since the Bayes classifier assigns an observation to the class with the highest probability, we can compute the Bayes error rate as $1 - E(max_k Pr(Y = k|X))$. Here the expectation is taken over all values of X. The Bayes error rate minimizes the test error rate, and is therefore analogous to irreducible error.
- Q12. If the Bayes decision boundary is known, we do not need a test set since we already know how to classify the observations in order to minimize test error. Thus we already have the best classification method. However, it might still be nice to have a test set in order to get a better picture of the spread in the data, and the real misclassification rate.

b) Support vector machine

- Q13. A support vector classifier (SVC) is a classification method that uses linear decision boundaries, but allows some degree of misclassification. Some of the observations are allowed to be on the wrong side of the margin. This is to improve the robustness of the method and improve the classification of the majority of the data. A support vector machine (SVM) is an extension of the SVC that allows for non-linear decision boundaries as well.
- Q14. For the SVC the relevant parameters are the data points as well as the tuning parameter C, which can be interpreted as a budget for how much and much many of the data points are allowed to cross the margin. In the `svm` library we rather specify a cost, which represents the penalty for crossing the margin. The SVM has the same parameters as SVC, and additionally one has to specify which kernel to use, e.g. radial or polynomial. Furthermore, there might be extra parameters that need to be specified for the kernels, e.g. the polynomial degree or the γ constant for the radial kernel. In the code above, these parameters are chosen using 10-fold cross-validation.

- Q15. It looks like the SVM boundary is better than the Bayes boundary since the SVM correctly classifies several more of the black points while only misclassifying a couple more of the red points than the Bayes boundary does. This might seem a bit strange, since the Bayes decision boundary (when known) minimizes the test error rate. Therefore it appears that the SVM model has overfitted to the training data. In reality the SVM boundary cannot outperform the Bayes boundary on the real data.

Problem 3 - Unsupervised methods

a) Principal component analysis

- Q16. The loading of the first principal component describe the direction in the 6-dimensional space of which the linear combination of the variables has the highest variance. Similarly, the loading of the second principal component describes the direction of which the remaining variance (the variance not accounted for by the first principal component) is largest. The loading of the first principal component places most weight on tea and wine, though with opposite signs. This implies that the variables have negative correlation - a country which consumes high amounts of tea consumes low amounts of wine, and vice versa. Liquor, on the other hand, has weight close to zero, and has therefore almost no correlation with consumption of tea or wine. The loading of the second principal component places most weight on coffee and cocoa, and some on beer and liquor. Here, tea and wine has weights close to zero. We see that that the consumption of coffee and cocoa has high correlation, and that liquor has negative correlation in regard to both coffee, cocoa and beer.
- Q17. The biplot also visualizes the scores for the different countries. Countries whose names are overlapping have similar scores in regard to both first principal components, and indicate that their drinking patterns are very similar. We se here that Norway and Sweden, Italy and France, and Great Britain and Ireland are examples of countries which seems to have similar consumption. These countries are located near one another, and have similar cultures, so the results are as expected. One can also note that countries such as Schweitz and Portugal have almost the same score on the first principal component but quite different scores on the second principal component, implying that they have similar consumption in regard to tea and wine, and different consumption in regard to coffee, cocoa, beer and liquor.

b) Hierarchical clustering

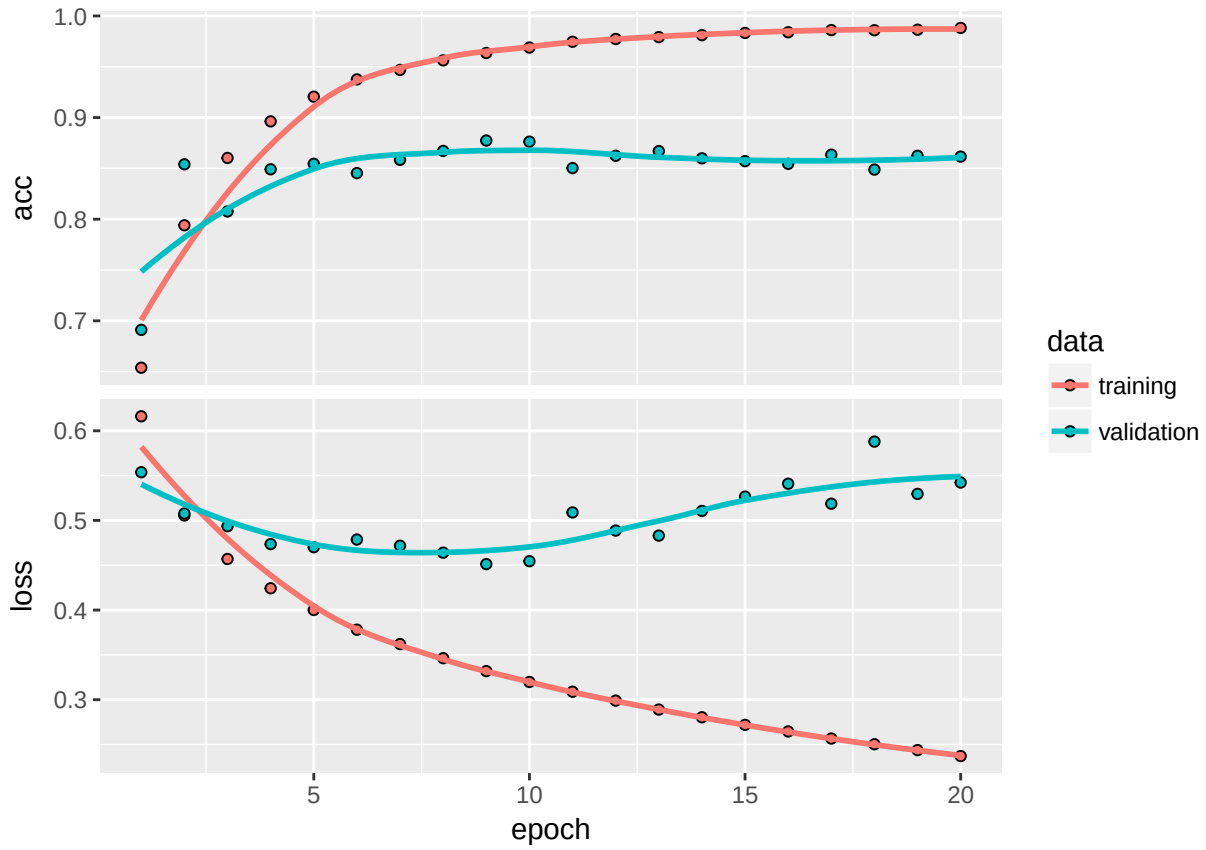
- Q18. For all three linkage methods, the distance between all pairs of observations from two. For single linkage the distance between the clusters is defined as the smallest pairwise distance, for complete linkage the distance is defined as the largest pairwise distance, and for average linkage the distance is defined as the mean of all the pairwise distances.
- Q19. In the three dendogrammes given in the assignment, the cluster fusing in C happens at the lowest heights, and the fusing in A happens at the highest heigths. Lower distance between clusters results in fusing at lower heights. Since the distance between clusters when using single linkage is always lower than or equal to the distance when using complete or average linkage, dendrogram C must correspond to single linkage. Similarily, the distance between clusters when using complete linkage is always higher than or equal to the distance when using single or average linkage, dendrogram A must correspond to complete linkage. The mean has distances in-between the other to linkages, and corresponds to dendrogram B, where the fusing heights are also in-between the heights for A and C. Another hint that dendrogram B corresponds to average linkage, is that the heights of the fuses have decimal value as opposed to the two other linkages, which is a consequence of the distances being a mean of integers.

Problem 4 - Neural networks

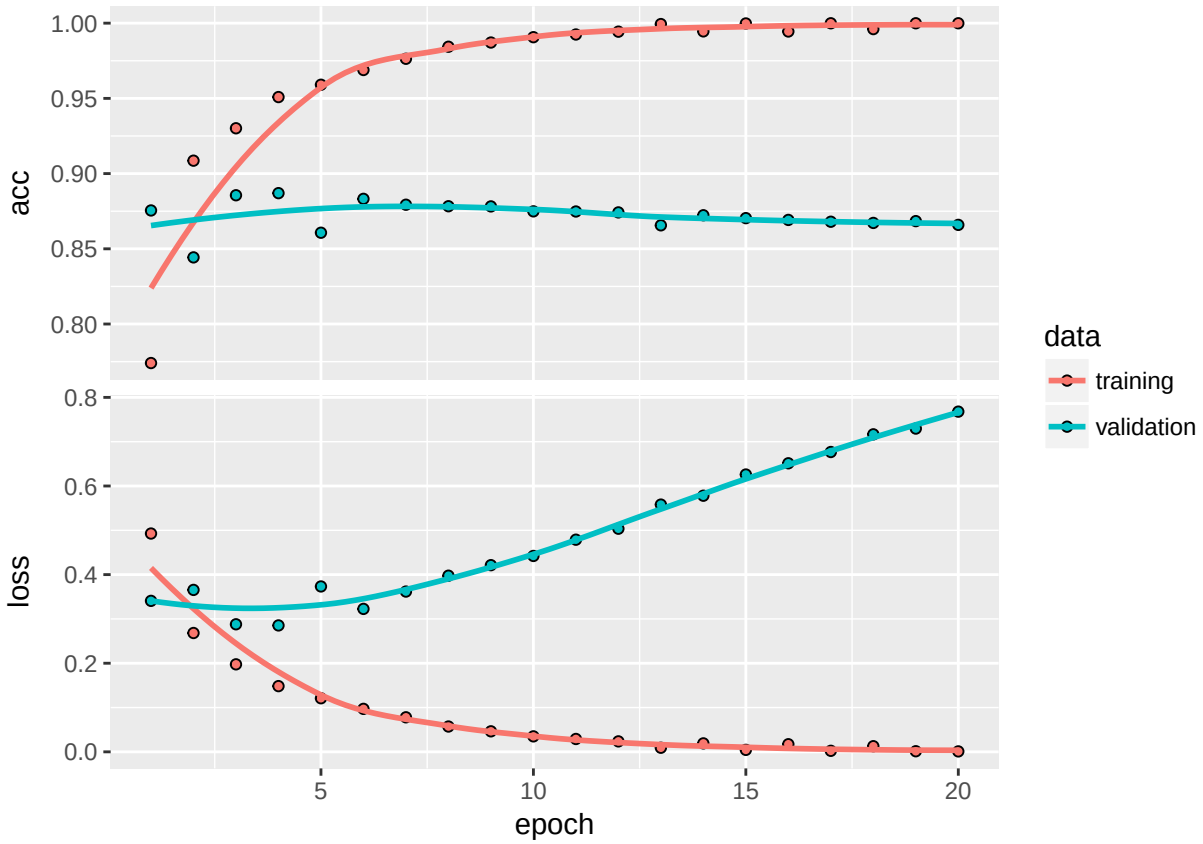
- Q20. Non-linear activation functions are used in order to capture more complex patterns in the data. If one did not use any non-linear functions the model would only be able to capture linear patterns.
- Q21. In the final layer we need a function that allows us to classify the observation into one of the two classes. The sigmoid function is an S-shaped function that gives values between 0 and 1. The relu-function does not allow us to classify the data, since it outputs values between 0 and inf.

```
library(keras)
# Collect data
imdb <- dataset_imdb(num_words = 10000)

train_data <- imdb$train$x
train_labels <- imdb$train$y
test_data <- imdb$test$x
test_labels <- imdb$test$y

# Vectorize data
vectorize_sequences <- function(sequences, dimension = 10000) {
  results <- matrix(0, nrow = length(sequences), ncol = dimension)
  for (i in 1:length(sequences))
    results[i, sequences[[i]]] <- 1
  results
}

x_train <- vectorize_sequences(train_data)
x_test <- vectorize_sequences(test_data)

y_train <- as.numeric(train_labels)
y_test <- as.numeric(test_labels)

# Defining the models (simple and complex)
model.simple <- keras_model_sequential() %>%
  layer_dense(units = 4, activation = "relu", input_shape = c(10000)) %>%
  layer_dense(units = 4, activation = "relu") %>%
  layer_dense(units = 1, activation = "sigmoid")

model.complex <- keras_model_sequential() %>%
  layer_dense(units = 32, activation = "relu", input_shape = c(10000)) %>%
  layer_dense(units = 32, activation = "relu") %>%
  layer_dense(units = 1, activation = "sigmoid")

# Compiling the models
model.simple %>% compile(
  optimizer = "rmsprop",
  loss = "binary_crossentropy",
  metrics = c("accuracy")
)

model.complex %>% compile(
  optimizer = "rmsprop",
  loss = "binary_crossentropy",
  metrics = c("accuracy")
)
```

```r
# Creating validation set
val_indices <- 1:10000

x_val <- x_train[val_indices,]
partial_x_train <- x_train[-val_indices,]
y_val <- y_train[val_indices]
partial_y_train <- y_train[-val_indices]

# Creating the fit
history.simple <- model.simple %>% fit(
  partial_x_train,
  partial_y_train,
  epochs = 20,
  batch_size = 512,
  validation_data = list(x_val, y_val)
)

history.complex <- model.complex %>% fit(
  partial_x_train,
  partial_y_train,
  epochs = 20,
  batch_size = 512,
  validation_data = list(x_val, y_val)
)


plot(history.simple)
```




```
plot(history.complex)
```



- Q22. We see from the plots that the simpler model has about the same accuracy and loss for the training data as the one with 16 units. However, the loss for the validation data is somewhat smaller for the simpler model. Conversely, the loss for the validation data increases for the complex model with 32 units. The loss and accuracy stays roughly the same for this mode as well. The fact that the loss increases for the more complex models is probably due to overfitting. The model adapts very well to the training data, but fails to give better predictions for the validation set.
- Q23. Besides reducing the network's size, there are three other main ways of preventing overfitting. The first action one can take is to get more data. This helps to prevent overfitting because more data points gives a better representation of the patterns in the data, and makes it more difficult for the network to interpret random noise as trends. The second action one can take is to add a weight regularizer. This penalizes the network for using large coefficient values in the weight matrix of each layer and thus prevents the network from making too complex models. The final option is to add a dropout, which is to set a number of random feature values from each layer equal to zero. The idea here is that by setting some of the ouput values to zero, some random noise is introduced and this prevents the network from picking up on insignificant patterns.