Compulsory exercise 3: Group XYZ

TMA4268 Statistical Learning V2018

Huglen, Huso and Myklebust

27 April, 2018

1a) Full classification tree

- Q1. Explain briefly how `fulltree` is constructed. The explanation should include the words: greedy, binary, deviance, root, leaves.

b) Pruned classification tree

- Q2. Why do we want to prune the full tree?
- Q3. How is amount of pruning decided in the code?
- Q4. Compare the the full and pruned tree classification method with focus on interpretability and the ROC curves (AUC).

c) Bagged trees

- Q5. What is the main motivation behind bagging?
- Q6. Explain what the importance plots show, and give your interpretation for the data set.
- Q7. Compare the performance of bagging with the best of the full and pruned tree model above with focus on interpretability and the ROC curves (AUC).

d) Random forest

- Q8. The parameter `mtry=4` is used. What does this parameter mean, and what is the motivation behind choosing exactly this value?
- Q9. The value of the parameter `mtry` is the only difference between bagging and random forest. What is the effect of choosing `mtry` to be a value less than the number of covariates?
- Q10. Would you prefer to use bagging or random forest to classify the credit risk data?

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 wich 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 vallet 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.
- Q12. When the Bayes decision boundary is known, do we then need a test set?

b) Support vector machine

- Q13. What is the difference between a support vector classifier and a support vector machine?
- Q14. What are parameters for the support vector classifier and the support vector machine? How are these chosen above?
- Q15. How would you evaluate the support vector machine decision boundary compared to the Bayes decision boundary?

Problem 3 - Unsupervised methods

a) Principal component analysis

- Q16. Explain what you see in the `biplot` in relation to the loadings for the first two principal components.
- Q17. Does this analysis give you any insight into the consumption of beverages and similarities between countries?

b) Hierarchical clustering

- Q18. Describe how the distance between *clusters* are defined for single, complete and average linkage.
- Q19. Identify which of the three dendrograms (A, B, C) correspond to the three methods single, complete and average linkage. Justify your solution.

Problem 4 - Neural networks

- Q20. What is the advantage of using a non-linear activation function such as `relu`?
- Q21. Why do we need to use a different activation function (`sigmoid`) in the output layer instead of using `relu` again?
- Q22. Plot the training and validation loss and accuracy for the simpler and more complex model mentioned above. How do they compare with the model with 16 hidden units?
- Q23. Besides reducing the network's size, what other methods can be used to avoid overfitting with neural network models? Briefly describe the intuition behind each one.