Exercise 7

Advanced Methods for Regression and Classification

December 22, 2022

Load the data `OJ` from the package `ISLR`. Our goal is to find a classification model that allows to predict the grouping variable `Purchase`. Which of the remaining variables should be considered in the model? For this task we shall use Generalized Additive Models, implemented in the function `gam()` of the package `library(mgcv)`.

Select randomly a training set of about 2/3 of the observations, build the classification model, predict the group membership for the (remaining) test data and compute the misclassification rate.

(a) The smooth functions in GAMs can be defined for every variable by `s(variable)`, see also course notes. With the parameter `k` you could also set an upper bound for the degrees of freedom. It might not make sense to use smooth functions for all variables. Now compute the GAM model based on your chosen "formula".

(b) Which variables are significant in the model? How complex are the smooth functions?

(c) Plot the explanatory variables against their smoothed values as they are used in the model. You can simply use:
`plot(gam.object,page=1,shade=TRUE,shade.col="yellow")`
How can you interpret this plot?

(d) Report the misclassification error for the test set.

(e) Similar as for logistic regression we can try to improve the classifier by variable selection. A natural option would be stepwise variable selection. A look into the help file of `step.gam` says that *There is no step.gam in package mgcv.* Nice. However, you can find some hints to still improve the model. Try out one of these ideas. Which variables are not used in the model? Compare with the misclassification error from (d).