Mini-Project 2
Evaluating Bagging

Alexander Reinthal 880726-4851 reinthal@student.chalmers.se
Magnus Lindström 910511-1752 guslinmagc@student.gu.se

April 2017

1 Datasets

For this assignment, two datasets were used to test the different classification methods: `BudgetUK` and `Computers`, both from the `Ecdat` library.

`BudgetUK` $n = 1519, p = 10$. Contains information about the expenditures of households in the UK within different areas such as food, fuel, clothing etc as a fraction. Each row sums to 1. It also has includes the age of the household head, the income of the household and the number of children. The variable chosen for classification was the number of children, one or two.
A correlation plot of the dataset is shown in Fig. 1.

`Computers` $n = 6259, p = 10$. Contains information about different personal computers such as price, screen size, RAM size, hard drive size and whether or not a CD-ROM is present. The variable to predict was chosen to be whether or not the computer has a CD-ROM. However, every case of computers not having a CD-ROM also did not have a multimedia kit included(the `multi` variable was "no"), which caused some of the methods (QDA and NB) to not work. Therefore, the `multi` feature was not taken into account which reduced p to 9.
A correlation plot of the dataset is shown in Fig. 2.

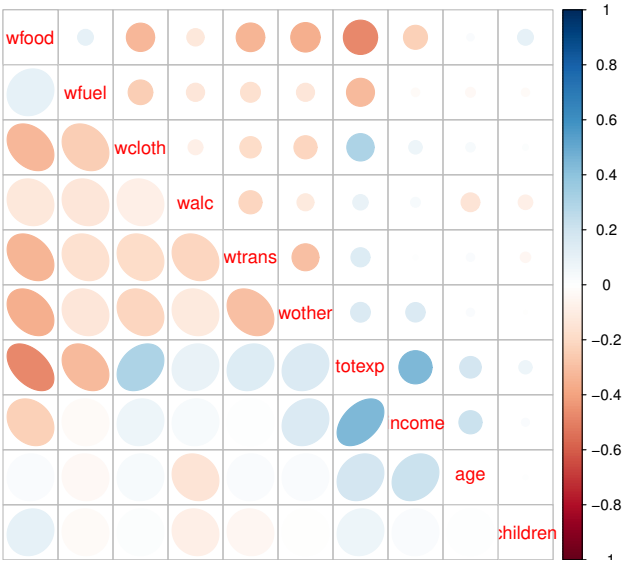

Figure 1: `BudgetUK` correlation plot.

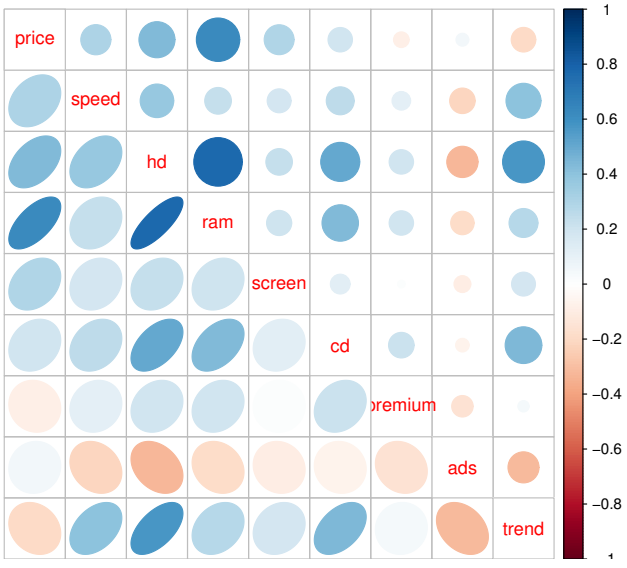

Figure 2: `Computers` correlation plot.

2 Analysis of methods

Following methods have been run on the datasets

- CART
- Random forest
- kNN
- LDA
- QDA
- PDA
- Naive Bayes
- MDA

2.1 Without bagging

Shown in Figs. 3 and 4 are the cross-validation error rates of the different methods after having trained on 75% of the data (50 data points per method). As is clear from the plots, the best methods at this point are CART, RF and NB when it comes to the `BudgetUK` dataset, and RF by far when it comes to the `Computers` dataset.

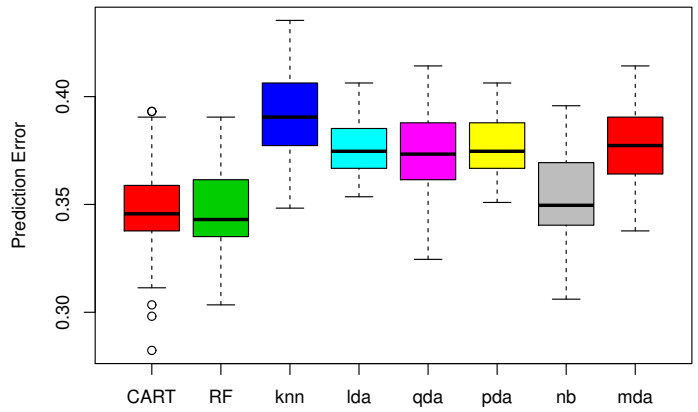

Figure 3: Boxplot of the cross-validation error rates of different methods when classifying on the `BudgetUK` dataset.

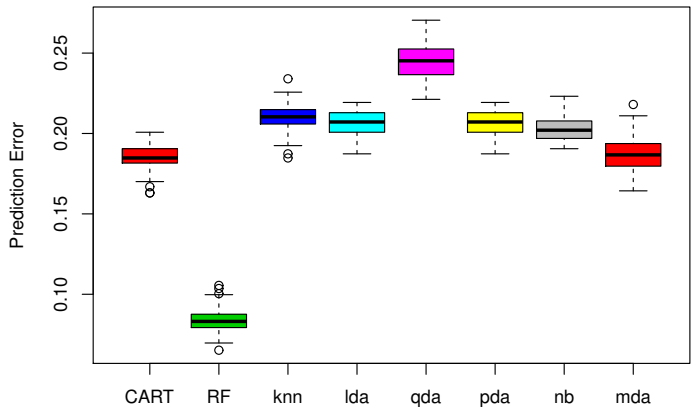

Figure 4: Boxplot of the cross-validation error rates of different methods when classifying on the `Computers` dataset.

2.2 With bagging

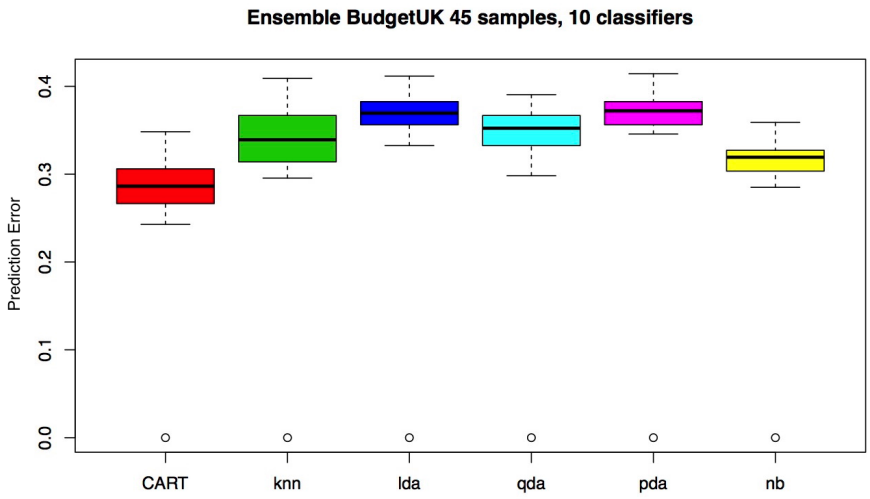

Figure 5: Above figure shows 45 runs of ensembles of size 10 on the BudgetUK dataset

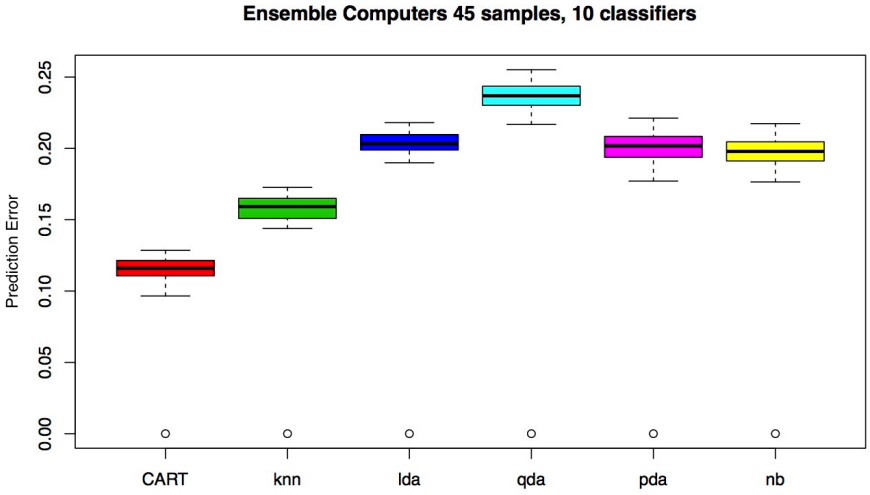

Figure 6: Above figure shows 45 runs of ensembles of size 10 on the Computers dataset

3 Concluding Remarks

The methods that significantly improved from bagging were CART and kNN. This is to be expected, since they are both local rules with high variance that can be reduced by bagging. The other methods, discriminant analysis and NB are global methods that are not greatly improved by bagging. Their prediction error consists mostly of their bias, not of their variance, and bias is not improved by bagging.

Every method is pretty bad at predicting the BudgetUK dataset, and they all seem like they are equally bad. This points to the class boundaries being rather simple, linear, but at the same time being muddy; no clear lines can be drawn between classes (if they could, RF would have been better).

The fact that RF is superior to the other methods in prediction on the Computers dataset points to the fact that the class boundaries are rather complex, and can not be captured by the simple methods. And the reason that the complex method CART has a high error rate could be that the boundaries between classes are not clear-cut, and so the variance is high (which contributes to the error).