Machine Learning - Block02 Assignment 01

Agustín Valencia

1. Ensemble Methods

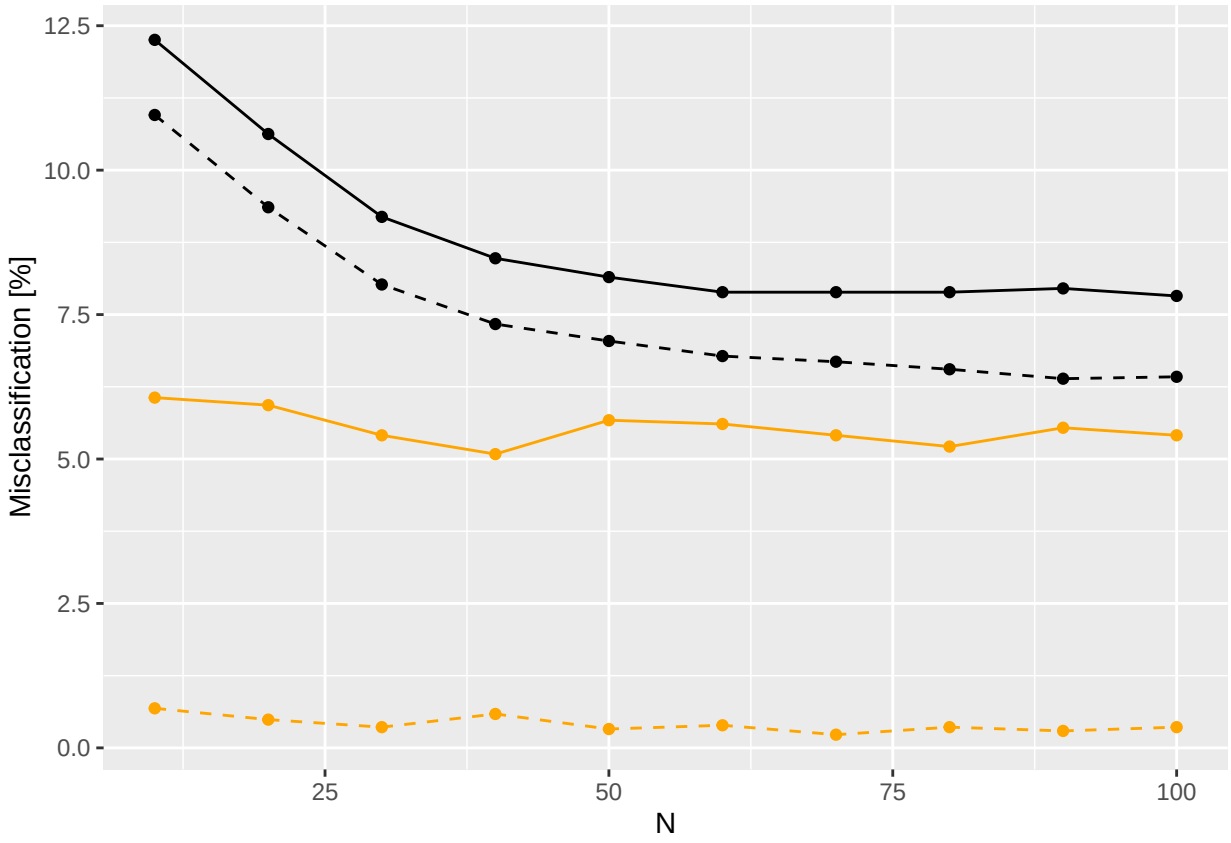
The file spambase.csv contains information about the frequency of various words, characters, etc. for a total of 4601 e-mails. Furthermore, these e-mails have been classified as spams (spam = 1) or regular e-mails (spam = 0). You can find more information about these data at https://archive.ics.uci.edu/ml/datasets/Spambase.

Your task is to evaluate the performance of Adaboost classification trees and random forests on the spam data. Specifically, provide a plot showing the error rates when the number of trees considered are 10, 20, . . . , 100. To estimate the error rates, use 2/3 of the data for training and 1/3 as hold-out test data.

To learn Adaboost classification trees, use the function blackboost() of the R package mboost. Specify the loss function corresponding to Adaboost with the parameter family. To learn random forests, use the function randomForest of the R package randomForest. To load the data, you may want to use the following code:

Solution

For trees trained using adaboost we have:



In the above graph the black lines corresponds to trees trained using adaboost with depth = N and the orange lines corresponds to random forests with N trees. In the same sense, dotted lines refers to training results and continuous to test results.

It can be seen that random forests need less effort (depth-wise) than boosted trees to get better results.

2. Mixture Models

Your task is to implement the EM algorithm for mixtures of multivariate Benoulli distributions. Please use the template in the next page to solve the assignment. Then, use your implementation to show what happens when your mixture models has too few and too many components, i.e. set $K = 2, 3, 4$ and compare results. Please provide a short explanation as well.

```
## The max likelihood for K = 2 0
```

```
## The max likelihood for K = 3 0
```

```
## The max likelihood for K = 4 0
```

In the following graph: