Homework 8 - LogitBoost

Thomas Johansen and Kyle Shaw

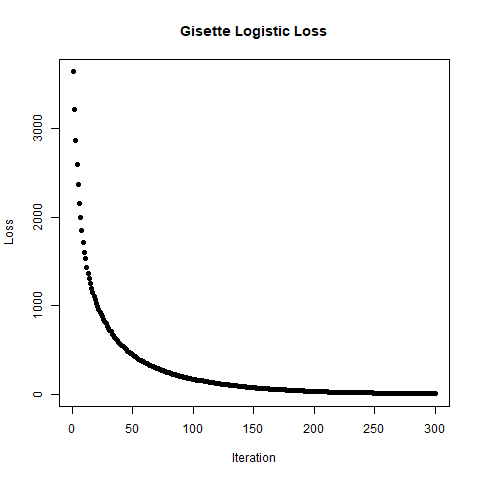
November 6, 2017

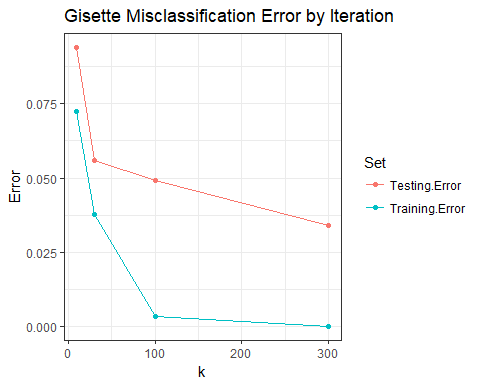
## Method

The "ada" package in R was used to run LogitBoost on four datasets. The ada function was modified to also return the loss so that it could be plotted. The plots below show the loss vs. iteration and misclassification error vs. iteration. A table of results is also included, as well as the script and a bibliography.

## Gisette

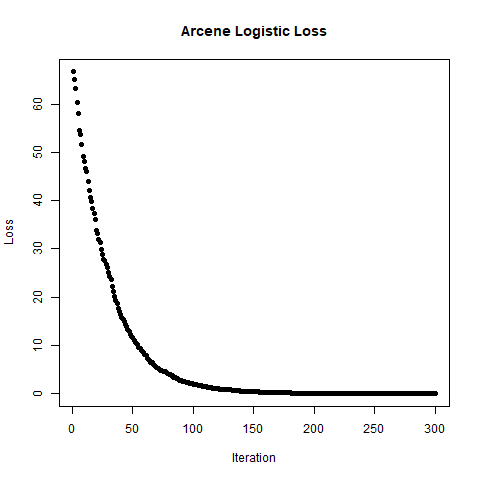
R ran out of memory on full dataset, so a subset of 500 features was used for analysis.

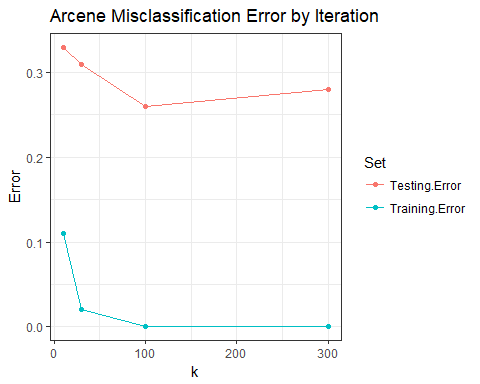




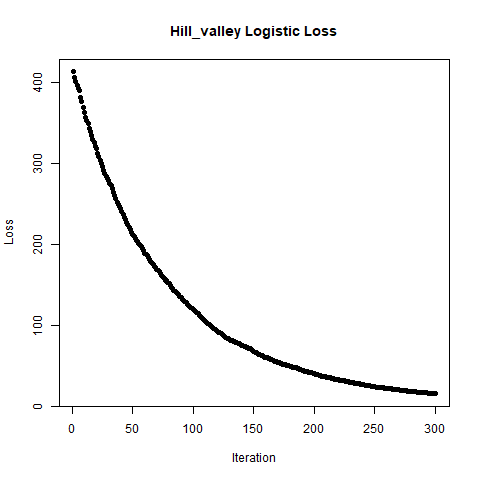
## Arcene

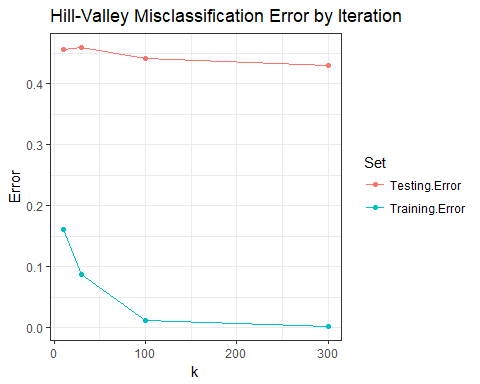
R ran out of memory on full dataset, so a subset of 2000 features was used for analysis.



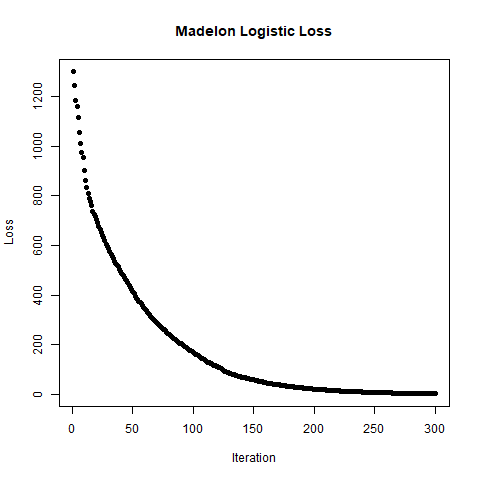


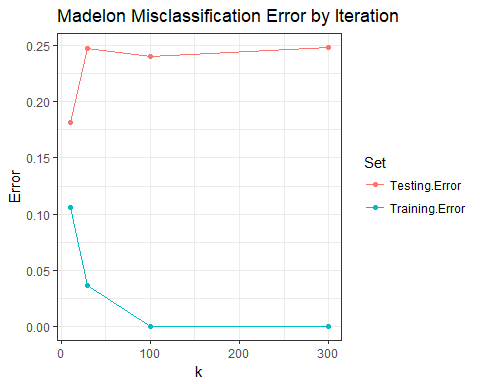
## Hill-Valley





## Madelon





## Table of Results

|  |  |  |  |
| --- | --- | --- | --- |
| Data | k | Training.Error | Testing.Error |
| gisette | 10 | 0.0725000 | 0.0940000 |
| gisette | 30 | 0.0376667 | 0.0560000 |
| gisette | 100 | 0.0035000 | 0.0490000 |
| gisette | 300 | 0.0000000 | 0.0340000 |
| arcene | 10 | 0.1100000 | 0.3300000 |
| arcene | 30 | 0.0200000 | 0.3100000 |
| arcene | 100 | 0.0000000 | 0.2600000 |
| arcene | 300 | 0.0000000 | 0.2800000 |
| hill\_valley | 10 | 0.1617162 | 0.4570957 |
| hill\_valley | 30 | 0.0874587 | 0.4603960 |
| hill\_valley | 100 | 0.0115512 | 0.4422442 |
| hill\_valley | 300 | 0.0016502 | 0.4306931 |
| madelon | 10 | 0.1060000 | 0.1816667 |
| madelon | 30 | 0.0370000 | 0.2466667 |
| madelon | 100 | 0.0000000 | 0.2400000 |
| madelon | 300 | 0.0000000 | 0.2483333 |

## Script

## run logitboost  
library(ada)  
results <- data.frame()  
set.seed(5)  
files <- c("gisette", "arcene", "hill\_valley", "madelon")  
propcase <- function(string) {paste0(toupper(substring(string, 1, 1)), substring(string, 2))}  
print(Sys.time())  
for (i in 1:4) {  
 message(paste("Begin", files[i]))  
 filelist <- read\_data(files[i])  
 for (k in c(10, 30, 100, 300)) {  
 logitBoost <- ada(filelist$X, filelist$Y$Y, filelist$Xtest, filelist$Ytest$Y,  
 loss="logistic", iter=k, verbose = FALSE)  
 if (k == 300) {  
 png(filename = paste0("C:/Users/joh10/Desktop/FSU/FA17/5635/git/hw8/", files[i], "\_loss.png"))  
 plot(logitBoost$model$plot\_loss, xlab = "Iteration", ylab = "Loss",   
 main = paste0(propcase(files[i]), " Logistic Loss"), pch=16)  
 dev.off()  
 message("Plot saved.")  
 }  
 results <- rbind(results, c(k, logitBoost$model$errs[k,1], logitBoost$model$errs[k,3]))  
 message(paste("k =", k, "finished"))  
 }  
 print(Sys.time())  
}  
  
colnames(results) <- c("k", "Training Error", "Testing Error")  
results <- data.frame(Data = rep(files, each = 4), results)  
saveRDS(results, "C:/Users/joh10/Desktop/FSU/FA17/5635/git/hw8/results.rds")

#### Modifying the ada function

The "trace" function in base R was used to modify the ada package so that the loss function could be plotted.

trace(ada:::ada.default, edit = TRUE)  
trace(ada:::ada.machine, edit = TRUE)  
  
## code added:  
# plot\_loss <- c()  
# plot\_loss[m] <- sum(log(1 + exp(-y \* fits)))  
# add plot\_loss to return obj  
  
#untrace(ada:::ada.machine)

## Bibliography

* Mark Culp, Kjell Johnson and George Michailidis (2016). ada: The R Package Ada for Stochastic Boosting. R package version 2.0-5. <https://CRAN.R-project.org/package=ada>
* H. Wickham. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York, 2009.
* Hadley Wickham and Lionel Henry (2017). tidyr: Easily Tidy Data with 'spread()' and 'gather()' Functions. R package version 0.7.2. <https://CRAN.R-project.org/package=tidyr>