Measuring Performance: Evaluating Variable Reduction Techniques

Spiro Stilianoudakis

Contents

Loading Packages

```r
library(caret)
```

```
## Warning: package 'caret' was built under R version 3.4.4
```

```
## Loading required package: lattice
```

```
## Loading required package: ggplot2
```

```
## Warning: package 'ggplot2' was built under R version 3.4.4
```

```r
#library(data.table)
library(gbm)
```

```
## Loading required package: survival
```

```
##
## Attaching package: 'survival'
```

```
## The following object is masked from 'package:caret':
##
##     cluster
```

```
## Loading required package: splines
```

```
## Loading required package: parallel
```

```
## Loaded gbm 2.1.3
```

```r
library(pROC)
```

```
## Warning: package 'pROC' was built under R version 3.4.4
## Type 'citation("pROC")' for a citation.
```

```
##
## Attaching package: 'pROC'
```

```
## The following objects are masked from 'package:stats':
##
##     cov, smooth, var
```

```r
library(plyr)
library(dplyr)
```

```
## Warning: package 'dplyr' was built under R version 3.4.4
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:plyr':
##
##     arrange, count, desc, failwith, id, mutate, rename, summarise,
##     summarize
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```r
#library(DMwR)
library(gridExtra)
```

```
##
## Attaching package: 'gridExtra'
```

```
## The following object is masked from 'package:dplyr':
##
##     combine
```

```r
library(ggplot2)
library(leaps)
library(limma)
#library(DT)
library(knitr)
```

```
## Warning: package 'knitr' was built under R version 3.4.4
```

Setting Working directory

```r
#reading in dataset
setwd("C:/Users/Spiro Stilianoudakis/Documents/TAD_data/RData/GM12878")
chr1_gm12878_f <- readRDS("C:/Users/Spiro Stilianoudakis/Documents/TAD_data/RData/GM12878/chr1_gm12878_

#set directory for selection techinques
setwd("C:/Users/Spiro Stilianoudakis/Documents/TAD_data/RData/GM12878/variable_reduction")
```

Variable Selection Techniques

Forward Selection

```
#rds objects for dataset
auc.model.fwd <- readRDS("C:/Users/Spiro Stilianoudakis/Documents/TAD_data/RData/GM12878/variable_reduct
cv.preds.fwd <- readRDS("C:/Users/Spiro Stilianoudakis/Documents/TAD_data/RData/GM12878/variable_reducti

#rds object for roc and aucs
enetlst_fwd <- readRDS("C:/Users/Spiro Stilianoudakis/Documents/TAD_data/RData/GM12878/variable_reducti

auc.fwd <- round(mean(enetlst_fwd[[3]]),3)
auc.fwd
```

```
## [1] 0.804
```

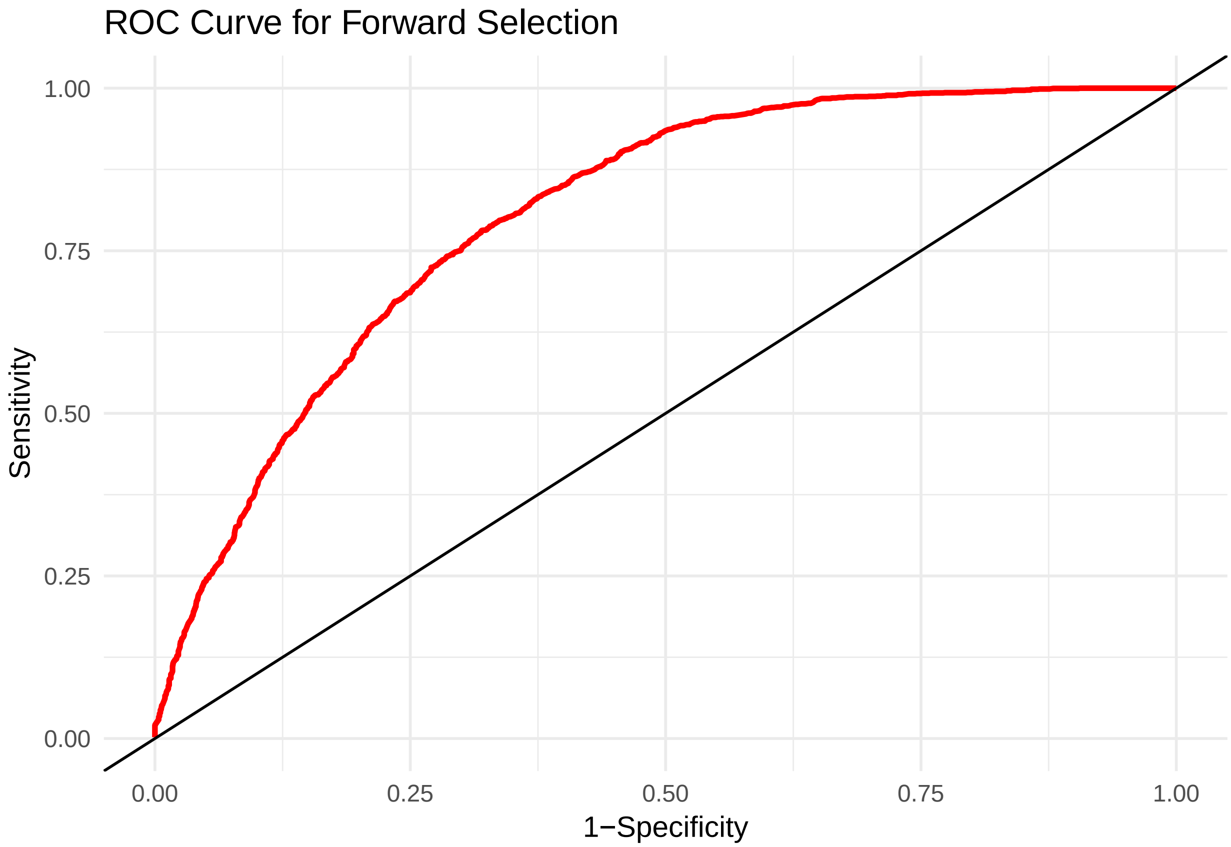
```
#0.810

vars.fwd <- na.omit(cv.preds.fwd[,which.max(auc.model.fwd)])
vars.fwd[grep("_dist",vars.fwd,invert = TRUE)] <- unlist(lapply(vars.fwd[grep("_dist",vars.fwd,invert =

chr1_gm12878_fwd <- chr1_gm12878_f[,which((names(chr1_gm12878_f) %in% vars.fwd) | names(chr1_gm12878_f)=

dim(chr1_gm12878_fwd)
```

```
## [1] 247632     23
```

```
#247632   29

names(chr1_gm12878_fwd)
```

```
##  [1] "y"                               "A_dist"
##  [3] "duplication"                     "complex_dist"
##  [5] "mobile_element_insertion_dist" "sequence_alteration_dist"
##  [7] "tandem_duplication_dist"       "DNA_dist"
##  [9] "low_complexity_dist"           "se_GM12878_dist"
## [11] "Gm12878_Heterochromlo"         "Gm12878_WeakTxn_dist"
## [13] "Gm12878_WeakPromoter_dist"     "Gm12878_WeakEnhancer6_dist"
## [15] "Gm12878_WeakEnhancer7_dist"    "Gm12878_Insulator_dist"
## [17] "Gm12878_CTCF_dist"             "Gm12878_TSS_dist"
## [19] "Gm12878_DNaseI"                "Gm12878_DNaseI_dist"
## [21] "Gm12878_H3k36me3"              "Gm12878_H3k9me3"
## [23] "Gm12878_H2az_dist"
```

```
setwd("C:/Users/Spiro Stilianoudakis/Documents/TAD_data/RData/GM12878")
saveRDS(chr1_gm12878_fwd, "chr1_gm12878_fwd.rds")

rocdat.fwd <- data.frame(sensitivity=rowMeans(enetlst_fwd[[1]]), specificity=rowMeans(enetlst_fwd[[2]])
rocdat.fwd$Selection <- "fwd"

ggplot(rocdat.fwd, aes(x=specificity, y=sensitivity)) +
  geom_line(size=1, color="red") +
  xlab("1-Specificity") +
  ylab("Sensitivity") +
  xlim(0, 1) +
  ylim(0, 1) +
  geom_abline(intercept=0, slope=1) +
  theme_minimal() +
  ggtitle("ROC Curve for Forward Selection")
```



Backward Selection

```
#rds object for dataset
auc.model.bwd <- readRDS("C:/Users/Spiro Stilianoudakis/Documents/TAD_data/RData/GM12878/variable_reduc
cv.preds.bwd <- readRDS("C:/Users/Spiro Stilianoudakis/Documents/TAD_data/RData/GM12878/variable_reducti

#rds object for roc and aucs
enetlst_bwd <- readRDS("C:/Users/Spiro Stilianoudakis/Documents/TAD_data/RData/GM12878/variable_reducti

auc.bwd <- round(mean(enetlst_bwd[[3]]),3)
auc.bwd
```

```
## [1] 0.807
```

```
#0.81

vars.bwd <- na.omit(cv.preds.bwd[,which.max(auc.model.bwd)])
vars.bwd[grep("_dist",vars.bwd,invert = TRUE)] <- unlist(lapply(vars.bwd[grep("_dist",vars.bwd,invert =

chr1_gm12878_bwd <- chr1_gm12878_f[,which((names(chr1_gm12878_f) %in% vars.bwd) | names(chr1_gm12878_f)
```

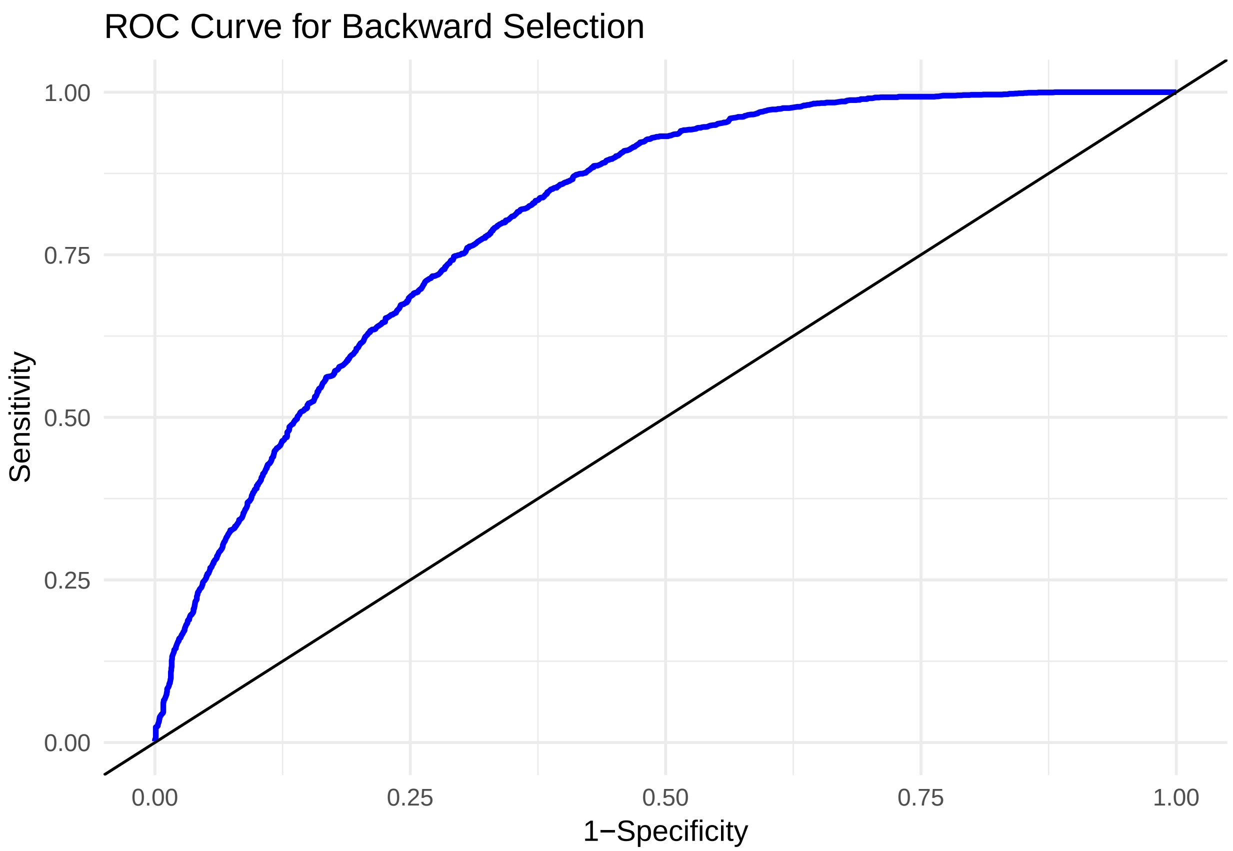
```r
dim(chr1_gm12878_bwd)
```

```
## [1] 247632     28
```

```r
#247632     35

names(chr1_gm12878_bwd)
```

```
##  [1] "y"                             "A"
##  [3] "B"                             "A_dist"
##  [5] "complex_dist"                  "mobile_element_insertion_dist"
##  [7] "sequence_alteration_dist"      "LTR"
##  [9] "se_GM12878_dist"               "Gm12878_WeakTxn"
## [11] "Gm12878_Repressed"             "Gm12878_WeakTxn_dist"
## [13] "Gm12878_Repressed_dist"        "Gm12878_Heterochromlo_dist"
## [15] "Gm12878_WeakPromoter_dist"     "Gm12878_WeakEnhancer7_dist"
## [17] "Gm12878_Insulator_dist"        "Gm12878_T"
## [19] "Gm12878_CTCF_dist"             "Gm12878_TSS_dist"
## [21] "Gm12878_DNaseI"                "Gm12878_DNaseI_dist"
## [23] "Gm12878_H3k27ac"               "Gm12878_H3k27me3"
## [25] "Gm12878_H3k36me3"              "Gm12878_H3k9ac"
## [27] "Gm12878_H3k9me3"               "Gm12878_H2az_dist"
```

```r
setwd("C:/Users/Spiro Stilianoudakis/Documents/TAD_data/RData/GM12878")
saveRDS(chr1_gm12878_bwd, "chr1_gm12878_bwd.rds")

rocdat.bwd <- data.frame(sensitivity=rowMeans(enetlst_bwd[[1]]), specificity=rowMeans(enetlst_bwd[[2]])
rocdat.bwd$Selection <- "bwd"

ggplot(rocdat.bwd, aes(x=specificity, y=sensitivity)) +
  geom_line(size=1, color="blue") +
  xlab("1-Specificity") +
  ylab("Sensitivity") +
  xlim(0, 1) +
  ylim(0, 1) +
  geom_abline(intercept=0, slope=1) +
  theme_minimal() +
  ggtitle("ROC Curve for Backward Selection")
```



Both

```
#rds objects for datasets
auc.model.both <- readRDS("C:/Users/Spiro Stilianoudakis/Documents/TAD_data/RData/GM12878/variable_reduc
cv.preds.both <- readRDS("C:/Users/Spiro Stilianoudakis/Documents/TAD_data/RData/GM12878/variable_reduct

#rds object for roc and aucs
enetlst_both <- readRDS("C:/Users/Spiro Stilianoudakis/Documents/TAD_data/RData/GM12878/variable_reducti

auc.both <- round(mean(enetlst_both[[3]]),3)
auc.both
```

```
## [1] 0.807
```

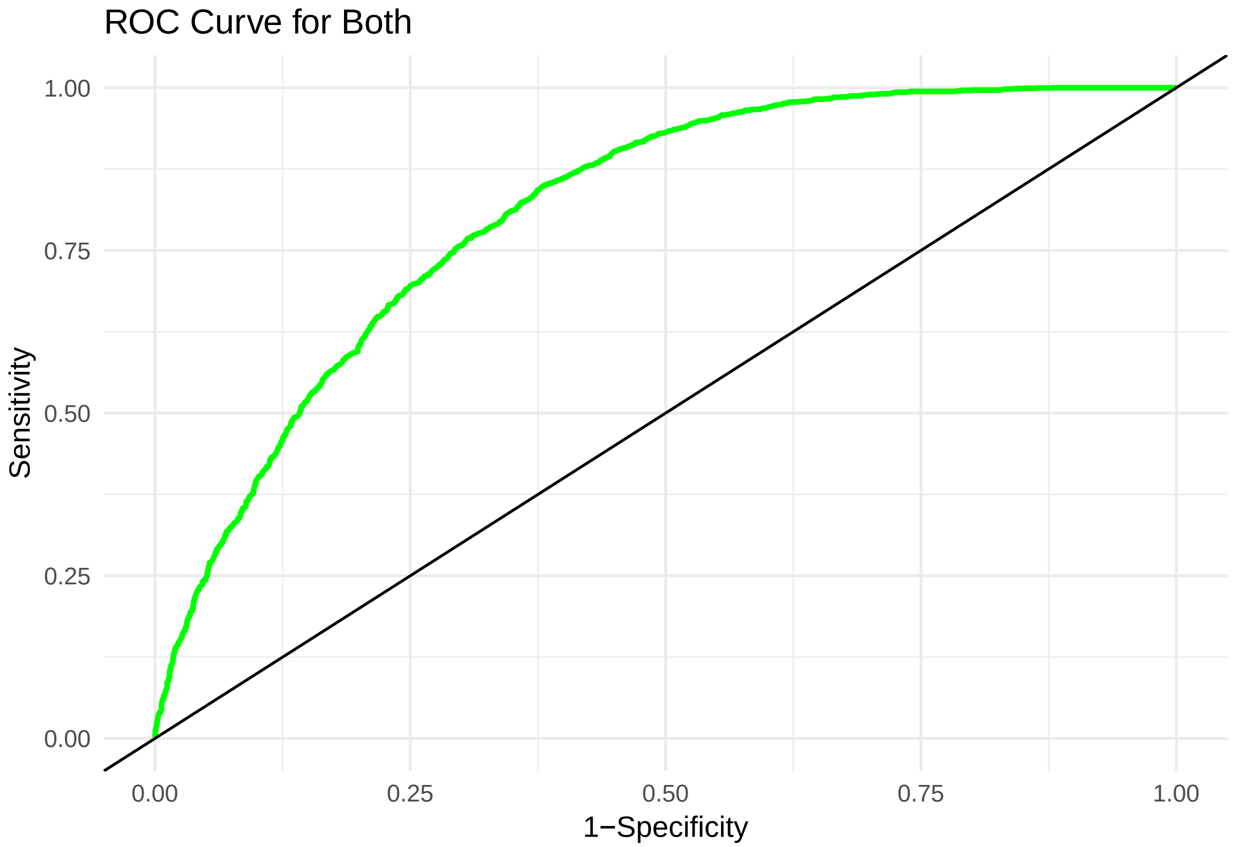
```
#0.809

vars.both <- na.omit(cv.preds.both[,which.max(auc.model.both)])
vars.both[grep("_dist",vars.both,invert = TRUE)] <- unlist(lapply(vars.both[grep("_dist",vars.both,inver

chr1_gm12878_both <- chr1_gm12878_f[,which((names(chr1_gm12878_f) %in% vars.both) | names(chr1_gm12878_

dim(chr1_gm12878_both)
```

```
## [1] 247632     26
```

```
#247632     27

names(chr1_gm12878_both)
```

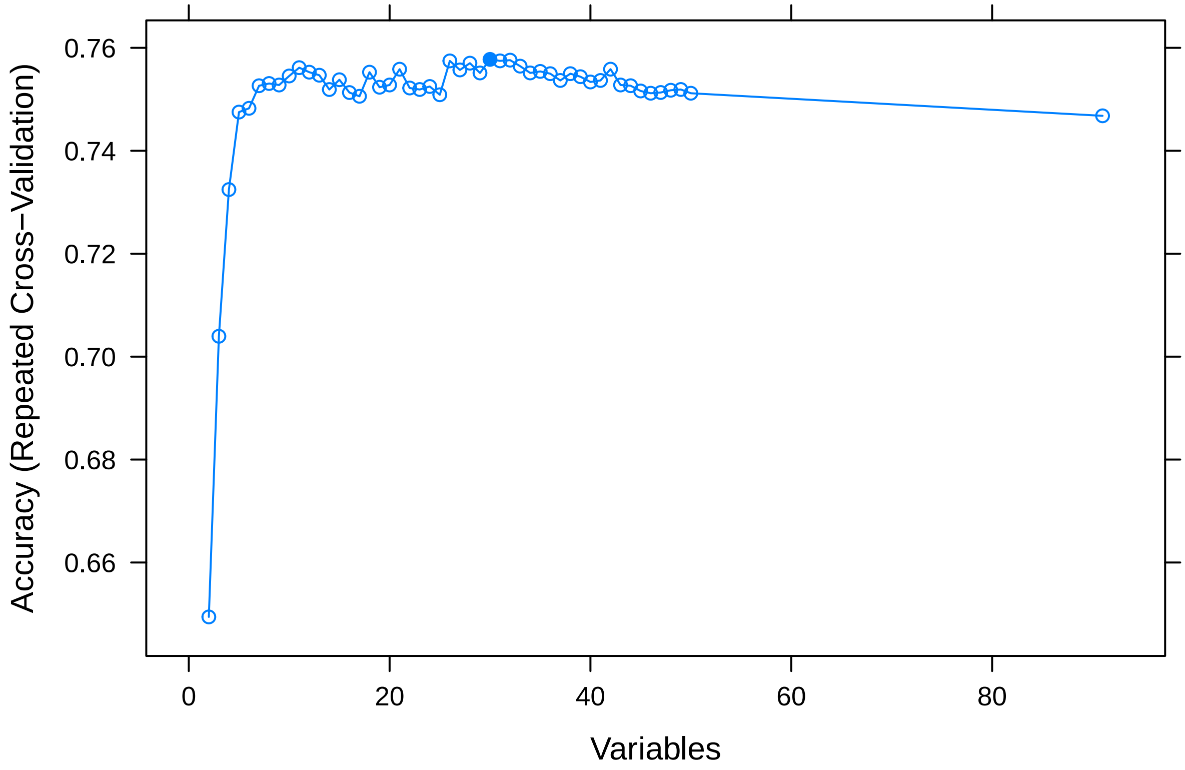
```
##  [1] "y"                              "A"
##  [3] "B"                              "A_dist"
##  [5] "complex_dist"                   "mobile_element_insertion_dist"
##  [7] "sequence_alteration_dist"       "LTR"
##  [9] "se_GM12878_dist"                "Gm12878_WeakTxn"
## [11] "Gm12878_WeakTxn_dist"           "Gm12878_Heterochromlo_dist"
## [13] "Gm12878_WeakPromoter_dist"      "Gm12878_WeakEnhancer7_dist"
## [15] "Gm12878_Insulator_dist"         "Gm12878_T"
## [17] "Gm12878_CTCF_dist"              "Gm12878_TSS_dist"
## [19] "Gm12878_DNaseI"                 "Gm12878_DNaseI_dist"
## [21] "Gm12878_H3k27ac"                "Gm12878_H3k27me3"
## [23] "Gm12878_H3k36me3"               "Gm12878_H3k9ac"
## [25] "Gm12878_H3k9me3"                "Gm12878_H2az_dist"
```

```
setwd("C:/Users/Spiro Stilianoudakis/Documents/TAD_data/RData/GM12878")
saveRDS(chr1_gm12878_both, "chr1_gm12878_both.rds")


rocdat.both <- data.frame(sensitivity=rowMeans(enetlst_both[[1]]), specificity=rowMeans(enetlst_both[[2
rocdat.both$Selection <- "both"

ggplot(rocdat.both, aes(x=specificity, y=sensitivity)) +
  geom_line(size=1, color="green") +
  xlab("1-Specificity") +
  ylab("Sensitivity") +
  xlim(0, 1) +
  ylim(0, 1) +
  geom_abline(intercept=0, slope=1) +
  theme_minimal() +
  ggtitle("ROC Curve for Both")
```



RFE

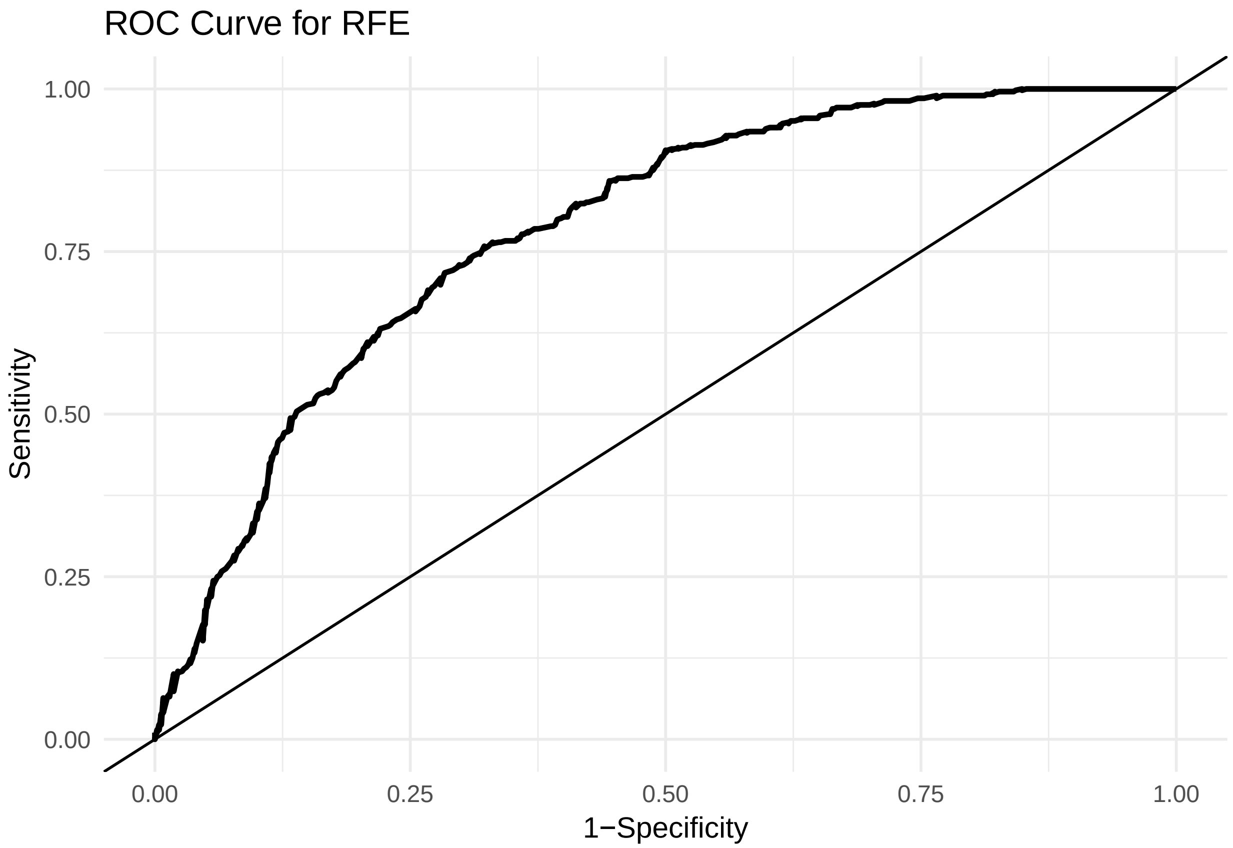
```
rfeModel <- readRDS("C:/Users/Spiro Stilianoudakis/Documents/TAD_data/RData/GM12878/variable_reduction/
roc.rfeModel <- readRDS("C:/Users/Spiro Stilianoudakis/Documents/TAD_data/RData/GM12878/variable_reducti

auc.rfe <- round(pROC::auc(roc.rfeModel),3)
#0.7986

plot(rfeModel, type="b")
```



```
accdat <- rfeModel$results
accdat <- accdat[order(accdat$Accuracy, decreasing = TRUE),]

predictors(rfeModel)
```

```
##  [1] "Gm12878_Insulator_dist"        "Gm12878_CTCF_dist"
##  [3] "Gm12878_DNaseI_dist"           "Gm12878_H3k9ac_dist"
##  [5] "Gm12878_TSS_dist"              "Gm12878_H2az_dist"
##  [7] "Gm12878_WeakPromoter_dist"     "Gm12878_PF_dist"
##  [9] "Gm12878_H3k36me3_dist"         "Gm12878_ActivePromoter_dist"
## [11] "Gm12878_H3k79me2_dist"         "Gm12878_H3k4me1_dist"
## [13] "Gm12878_WeakTxn_dist"          "Gm12878_TxnElongation_dist"
## [15] "se_GM12878_dist"               "Gm12878_StrongEnhancer5_dist"
## [17] "Gm12878_H3k4me2_dist"          "Gm12878_H3k27ac_dist"
## [19] "Gm12878_StrongEnhancer4_dist"  "Gm12878_Heterochromlo_dist"
## [21] "VMR_dist"                      "Gm12878_WE_dist"
## [23] "Gm12878_H3k4me3_dist"          "A_dist"
## [25] "Gm12878_TxnTransition_dist"    "Gm12878_WeakEnhancer6_dist"
## [27] "Gm12878_E_dist"                "Gm12878_H3k9me3_dist"
## [29] "novel_sequence_insertion_dist" "Gm12878_H4k20me1_dist"
```

```
rocdat.rfe <- data.frame(sensitivity=roc.rfeModel$sensitivities, specificity=1-roc.rfeModel$specificiti
rocdat.rfe$Selection <- "rfe"

ggplot(rocdat.rfe, aes(x=specificity, y=sensitivity)) +
  geom_line(size=1, color="black") +
  xlab("1-Specificity") +
  ylab("Sensitivity") +
  xlim(0, 1) +
  ylim(0, 1) +
  geom_abline(intercept=0, slope=1) +
  theme_minimal() +
  ggtitle("ROC Curve for RFE")
```



Comparing additional performance metrics across all variable selection techniques

```
options(scipen = 999)

enetperf_fwd <- readRDS("C:/Users/Spiro Stilianoudakis/Documents/TAD_data/RData/GM12878/variable_reducti

enetperf_bwd <- readRDS("C:/Users/Spiro Stilianoudakis/Documents/TAD_data/RData/GM12878/variable_reducti

enetperf_both <- readRDS("C:/Users/Spiro Stilianoudakis/Documents/TAD_data/RData/GM12878/variable_reduct

enetperf_rfe <- readRDS("C:/Users/Spiro Stilianoudakis/Documents/TAD_data/RData/GM12878/variable_reducti

fwdtab <- round(as.matrix(rowMeans(enetperf_fwd)),2)
bwdtab <- round(as.matrix(rowMeans(enetperf_bwd)),2)
bothtab <- round(as.matrix(rowMeans(enetperf_both)),2)
```

```
enetperf_rfe <- round(enetperf_rfe,2)

fwdtab[1:5,1] <- round(fwdtab[1:5,1],0)
bwdtab[1:5,1] <- round(bwdtab[1:5,1],0)
bothtab[1:5,1] <- round(bothtab[1:5,1],0)
enetperf_rfe[1:5,] <- round(enetperf_rfe[1:5,],0)

perfdat <- cbind.data.frame(rownames(enetperf_fwd),
                            fwdtab,
                            bwdtab,
                            bothtab,
                            enetperf_rfe)
rownames(perfdat) <- NULL
colnames(perfdat) <- c("Metric", "Forward", "Backward", "Both", "RFE")

kable(perfdat)
```

Metric	Forward	Backward	Both	RFE
TN	336.00	334.00	334.00	312.00
FN	112.00	113.00	112.00	108.00
FP	152.00	154.00	154.00	178.00
TP	378.00	377.00	378.00	380.00
Total	978.00	978.00	978.00	978.00
Sensitivity	0.77	0.77	0.77	0.78
Specificity	0.69	0.68	0.68	0.64
Kappa	0.46	0.45	0.46	0.42
Accuracy	0.73	0.73	0.73	0.71
Precision	0.71	0.71	0.71	0.68
FPR	0.31	0.32	0.32	0.36
FNR	0.23	0.23	0.23	0.22
FOR	0.25	0.25	0.25	0.26
NPV	0.75	0.75	0.75	0.74
MCC	0.46	0.46	0.46	0.42
F1	0.87	0.87	0.87	0.88

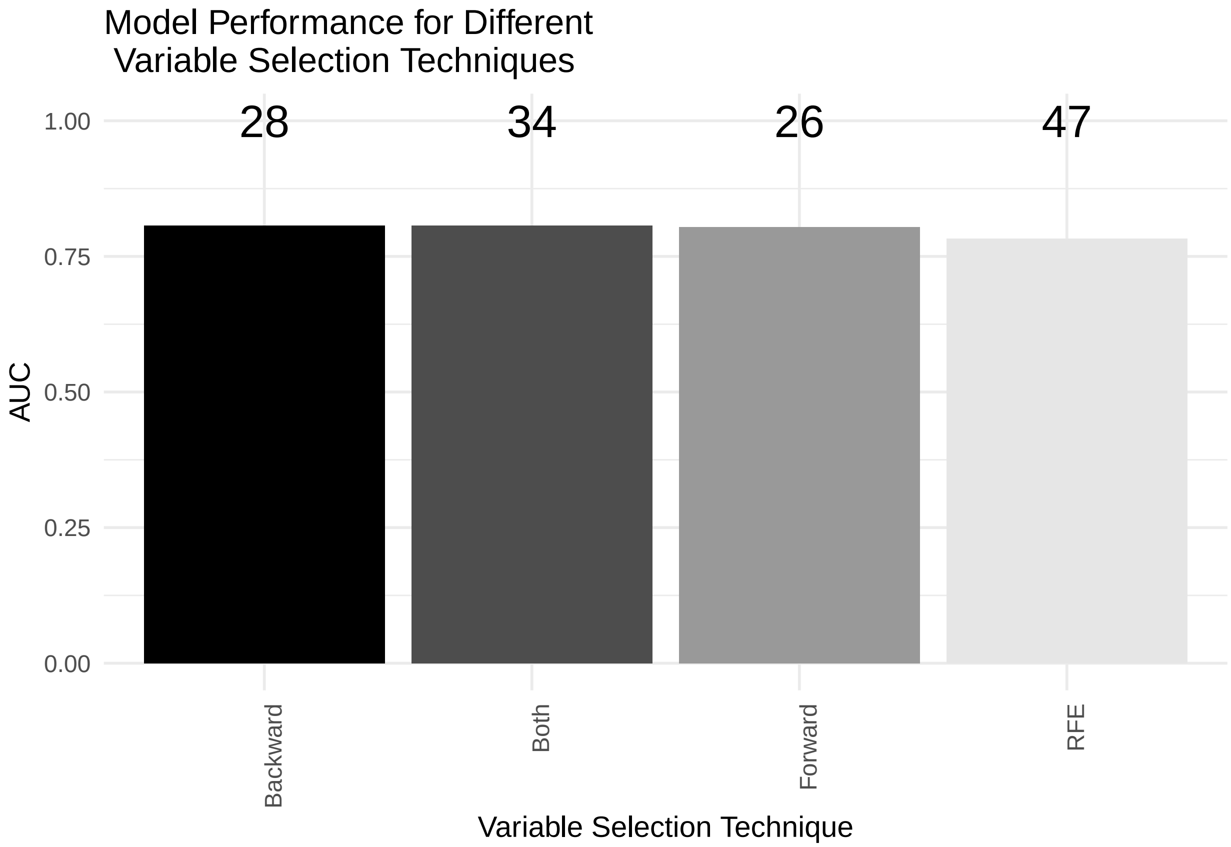
Comparing Variable Selection Techniques

```
auc.plot <- data.frame(Selection=c("Forward",
                                       "Backward",
                                       "Both",
                                       "RFE"),
                       auc=c(auc.fwd,
                             auc.bwd,
                             auc.both,
                             auc.rfe))

auc.plot <- auc.plot[order(auc.plot$auc, decreasing=TRUE),]

auc.plot$Selection <-factor(auc.plot$Selection,
                                     levels=auc.plot$Selection)
```

```r
#datatable(auc.plot)
kable(auc.plot)
```

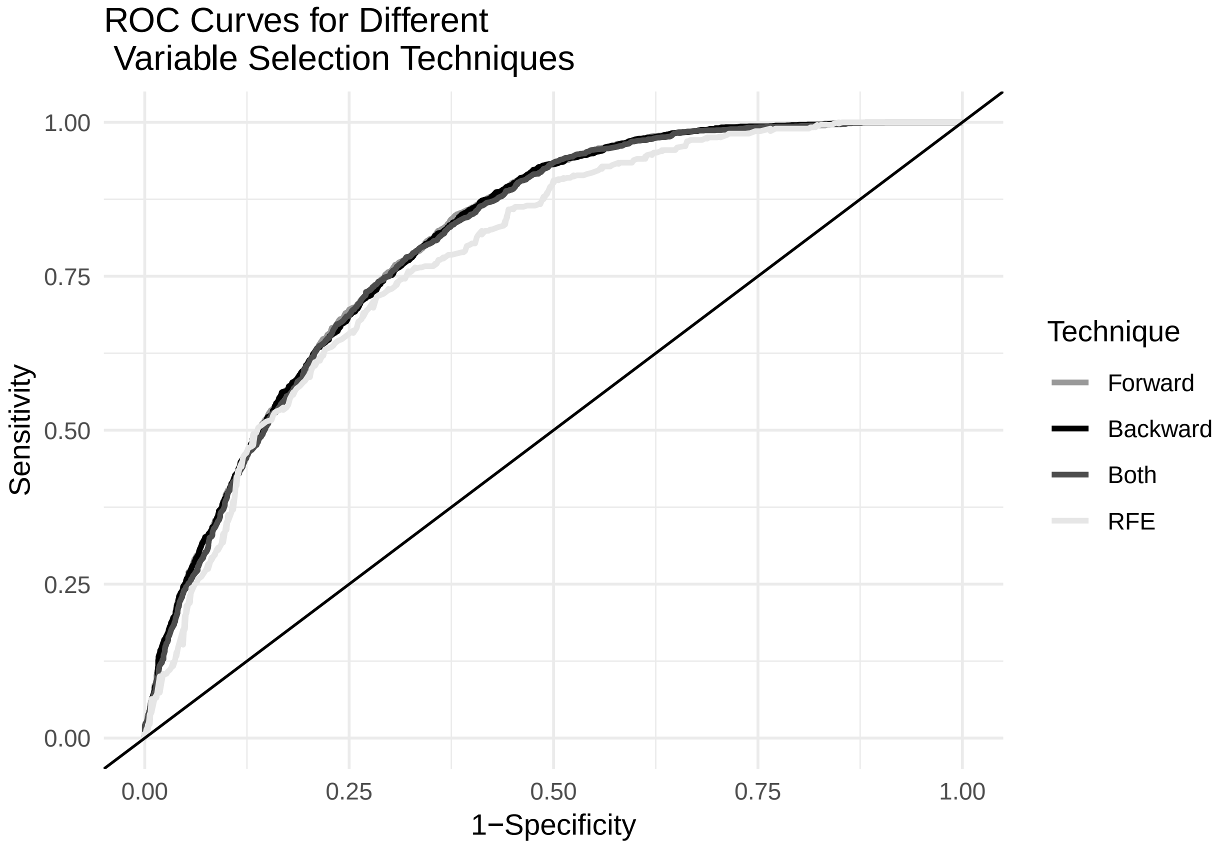
	Selection	auc
2	Backward	0.807
3	Both	0.807
1	Forward	0.804
4	RFE	0.783

```r
p<-ggplot(data=auc.plot, aes(x=Selection, y=auc, fill=Selection)) +
  xlab("Variable Selection Technique") + ylab("AUC") +
  geom_bar(stat="identity") + ylim(0,1) +
  scale_fill_manual(values=gray(c(0,.3,.6,.9)), guide=FALSE) +
  annotate("text", x=1, y=1, label= "28", size=6) +
  annotate("text", x=2, y=1, label= "34", size=6) +
  annotate("text", x=3, y=1, label= "26", size=6) +
  annotate("text", x=4, y=1, label= "47", size=6) +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 90, hjust = 1)) +
  ggtitle("Model Performance for Different \n Variable Selection Techniques")
p
```



```r
allrocdat <- rbind.data.frame(rocdat.fwd, rocdat.bwd, rocdat.both, rocdat.rfe)
```

```
ggplot(data=allrocdat, aes(x=specificity, y=sensitivity, color=Selection)) +
  geom_line(size=1) +
  scale_colour_manual(name="Technique",
    labels=c("Forward",
             "Backward",
             "Both",
              "RFE"),
    values=c("#999999", "#000000", "#4D4D4D", "#E6E6E6")) +
  xlab("1-Specificity") +
  ylab("Sensitivity") +
  xlim(0, 1) +
  ylim(0, 1) +
  geom_abline(intercept=0, slope=1) +
  theme_minimal() +
  ggtitle("ROC Curves for Different \n Variable Selection Techniques")
```



```
intersect(vars.fwd,intersect(vars.bwd,vars.both))
```

```
##  [1] "Gm12878_DNaseI_dist"            "Gm12878_TSS_dist"
##  [3] "Gm12878_Insulator_dist"         "Gm12878_WeakTxn_dist"
##  [5] "Gm12878_H3k36me3"               "Gm12878_H2az_dist"
##  [7] "Gm12878_DNaseI"                 "Gm12878_WeakEnhancer7_dist"
##  [9] "sequence_alteration_dist"       "Gm12878_H3k9me3"
## [11] "A_dist"                         "Gm12878_CTCF_dist"
## [13] "complex_dist"                   "se_GM12878_dist"
## [15] "mobile_element_insertion_dist" "Gm12878_WeakPromoter_dist"
```

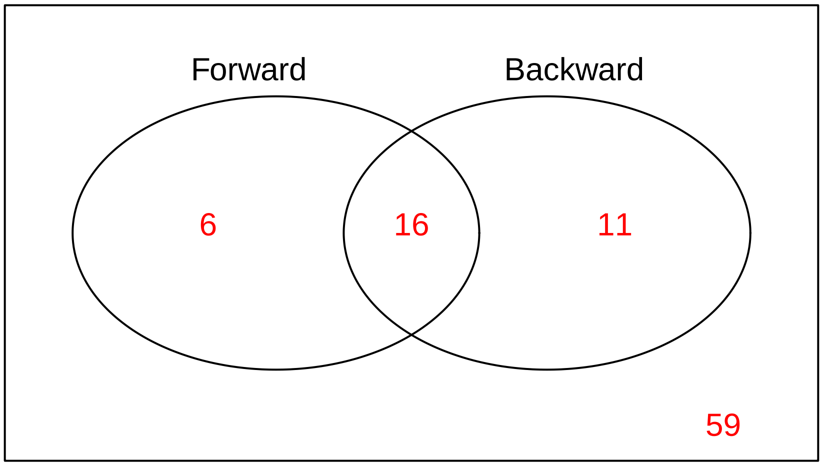
```r
intersect(vars.fwd,intersect(vars.bwd,predictors(rfeModel)))
```

```
## [1] "Gm12878_DNaseI_dist"        "Gm12878_TSS_dist"
## [3] "Gm12878_Insulator_dist"     "Gm12878_WeakTxn_dist"
## [5] "Gm12878_H2az_dist"          "A_dist"
## [7] "Gm12878_CTCF_dist"          "se_GM12878_dist"
## [9] "Gm12878_WeakPromoter_dist"
```

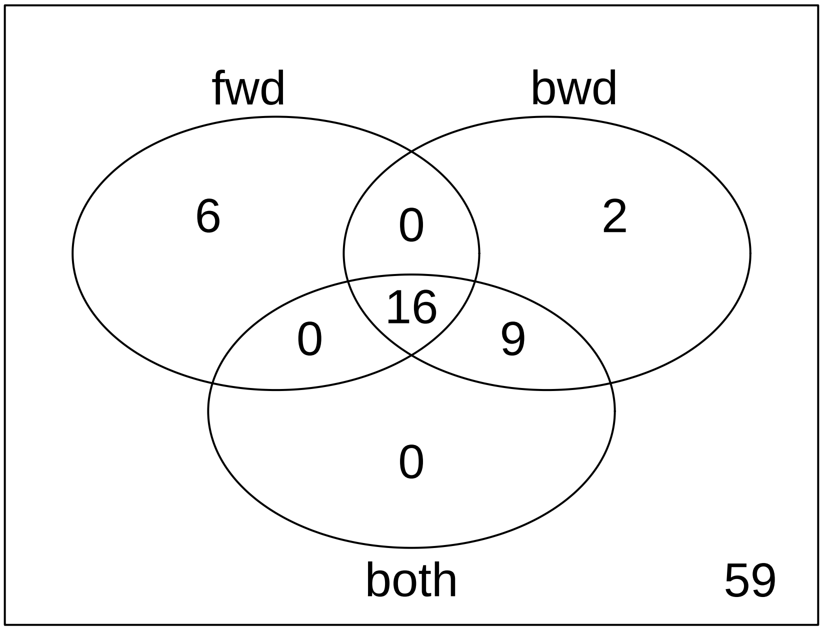
```r
intersect(intersect(vars.fwd,intersect(vars.bwd,vars.both)),
          intersect(vars.fwd,intersect(vars.bwd,predictors(rfeModel))))
```

```
## [1] "Gm12878_DNaseI_dist"        "Gm12878_TSS_dist"
## [3] "Gm12878_Insulator_dist"     "Gm12878_WeakTxn_dist"
## [5] "Gm12878_H2az_dist"          "A_dist"
## [7] "Gm12878_CTCF_dist"          "se_GM12878_dist"
## [9] "Gm12878_WeakPromoter_dist"
```

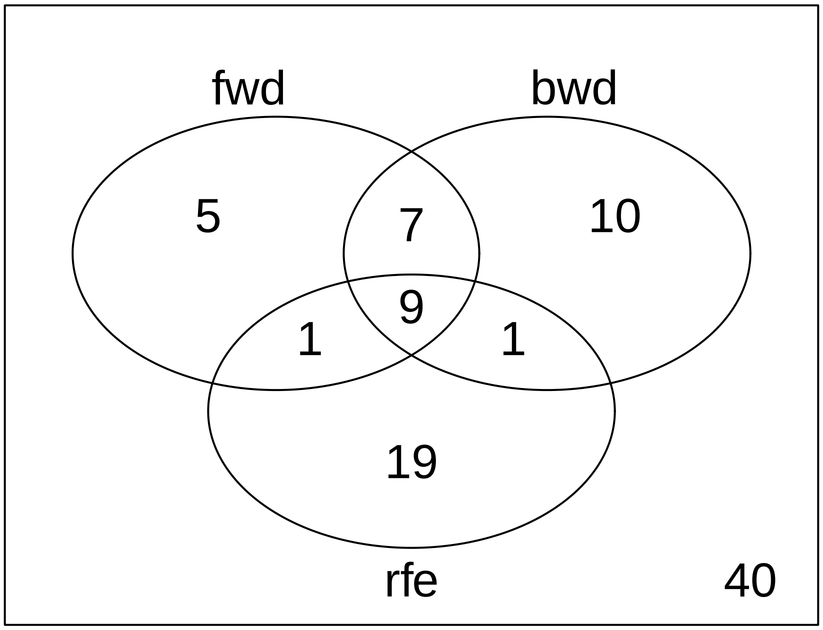
```r
fwd <- (names(chr1_gm12878_f) %in% vars.fwd)
bwd <- (names(chr1_gm12878_f) %in% vars.bwd)
both <- (names(chr1_gm12878_f) %in% vars.both)
rfe <- (names(chr1_gm12878_f) %in% predictors(rfeModel))

#fwd compared to bwd
venndatfb <- cbind(fwd,bwd)
fb <- vennCounts(venndatfb)
vennDiagram(fb, include = "both",
  names = c("Forward", "Backward"),
  cex = 1, counts.col = "red")
```



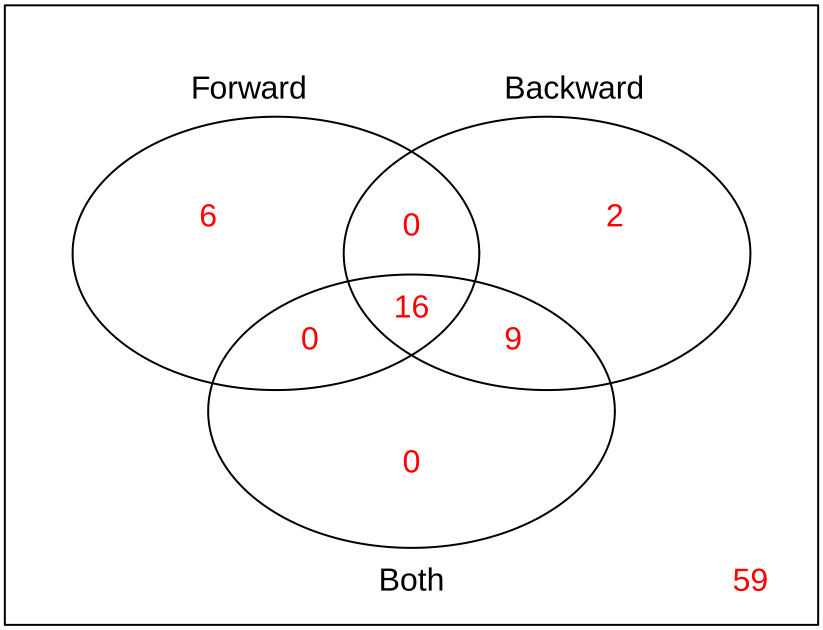
```r
venndat1 <- cbind(fwd,bwd,both)
venndat2 <- cbind(fwd,bwd,rfe)
venndat3 <- cbind(fwd,bwd,both,rfe)

a <- vennCounts(venndat1)
b <- vennCounts(venndat2)
c <- vennCounts(venndat3)

vennDiagram(a)
```



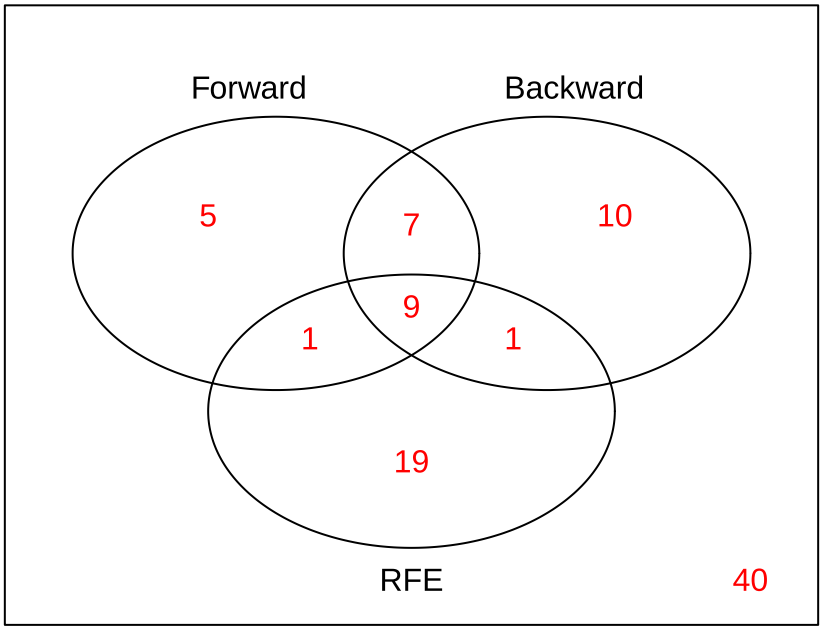
```r
vennDiagram(b)
```



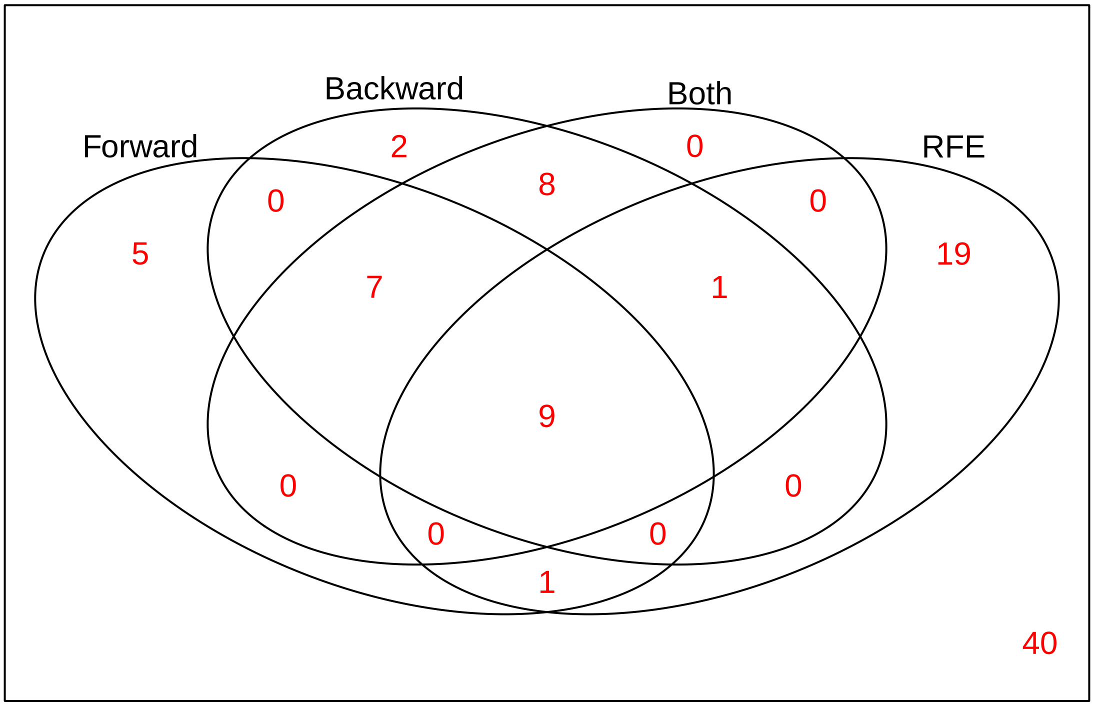
```r
vennDiagram(a, include = "both",
  names = c("Forward", "Backward", "Both"),
  cex = 1, counts.col = "red")
```



```
vennDiagram(b, include = "both",
  names = c("Forward", "Backward", "RFE"),
  cex = 1, counts.col = "red")
```



```
vennDiagram(c, include = "both",
  names = c("Forward", "Backward","Both", "RFE"),
  cex = 1, counts.col = "red")
```


Backward
Both
Forward
RFE
2
0
8
0
0
5
19
7
1
9
0
0
0
0
1
40