Prediction of Protein localisation sites

ZhangChi Qiu

29/05/2018

Task 1 Data Mining

Build a model to predict protein localisation site

a)

use a 70-30 split to create traning and test data

```
#use a seed of 1234

set.seed(1234)
yeast <-read.table("yeast.data")
ind <- sample(2, nrow(yeast), replace = TRUE, prob=c(0.7, 0.3))
train_data <- yeast[ind == 1,]
test_data <- yeast[ind == 2,]
```

b)

use training data to train a model.

```
library(party)
```

```
## Loading required package: grid
```

```
## Loading required package: mvtnorm
```

```
## Loading required package: modeltools
```

```
## Loading required package: stats4
```

```
## Loading required package: strucchange
```

```
## Loading required package: zoo
```

```
##
## Attaching package: 'zoo'
```

```
## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
```

```
## Loading required package: sandwich
```

```
formula <- V10 ~ V2 + V3 + V4 + V5 + V6 + V7 + V8 + V9
yeast_ctree <-ctree(formula, data = train_data)
```

c)

use model to predict previoulsy unseen data using the test data

```
predict <-table(predict(yeast_ctree, newdata = test_data), test_data$V10)
predict
```

```
##
##       CYT ERL EXC ME1 ME2 ME3 MIT NUC POX VAC
##   CYT  69   0   0   0   2   1  16  23   2   3
##   ERL   0   0   0   0   0   0   0   0   0   0
##   EXC   1   3   9   1   2   0   1   0   1   0
##   ME1   0   0   1   7   2   0   1   0   0   0
##   ME2   0   1   0   0   7   0   3   0   0   0
##   ME3   2   0   0   1   2  42   7   4   0   4
##   MIT  10   0   0   0   0   0  52  10   1   0
##   NUC  56   1   1   0   0   3   8  67   0   2
##   POX   0   0   0   0   0   0   0   0   3   0
##   VAC   0   0   0   0   0   0   0   0   0   0
```

d)

Produce a confustion matrix

```
library(caret)
```

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

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

```
confusionMatrix(predict(yeast_ctree, newdata = test_data), test_data$V10)
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction CYT ERL EXC ME1 ME2 ME3 MIT NUC POX VAC
```

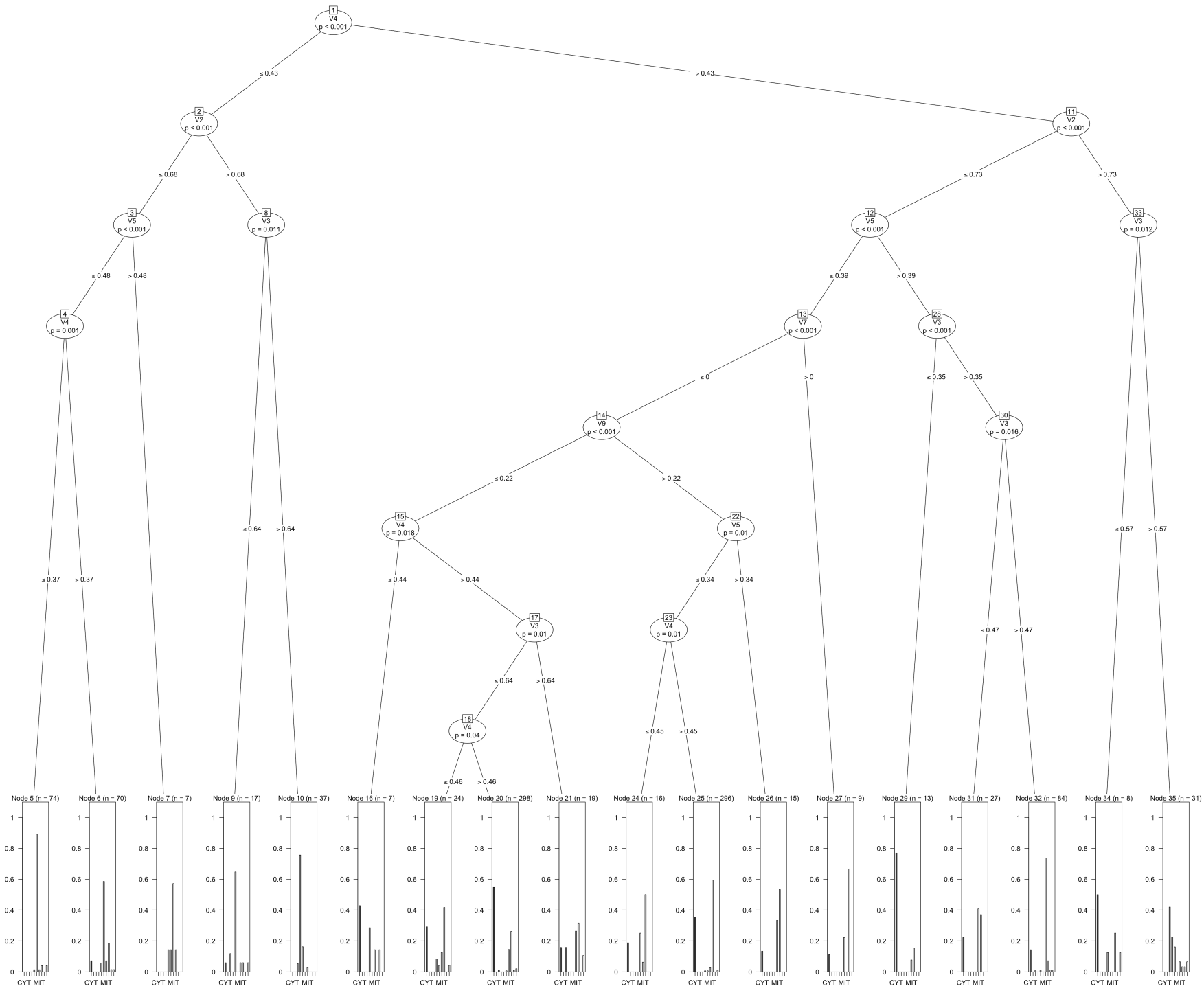
```
##        CYT  69    0    0    0    2    1   16   23    2    3
##        ERL   0    0    0    0    0    0    0    0    0    0
##        EXC   1    3    9    1    2    0    1    0    1    0
##        ME1   0    0    1    7    2    0    1    0    0    0
##        ME2   0    1    0    0    7    0    3    0    0    0
##        ME3   2    0    0    1    2   42    7    4    0    4
##        MIT  10    0    0    0    0    0   52   10    1    0
##        NUC  56    1    1    0    0    3    8   67    0    2
##        POX   0    0    0    0    0    0    0    0    3    0
##        VAC   0    0    0    0    0    0    0    0    0    0
##
## Overall Statistics
##
##                Accuracy : 0.5926
##                  95% CI : (0.5446, 0.6393)
##     No Information Rate : 0.3194
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 0.481
##  Mcnemar's Test P-Value : NA
##
## Statistics by Class:
##
##                      Class: CYT Class: ERL Class: EXC Class: ME1
## Sensitivity              0.5000    0.00000    0.81818    0.77778
## Specificity              0.8401    1.00000    0.97862    0.99054
## Pos Pred Value           0.5948        NaN    0.50000    0.63636
## Neg Pred Value           0.7816    0.98843    0.99517    0.99525
## Prevalence               0.3194    0.01157    0.02546    0.02083
## Detection Rate           0.1597    0.00000    0.02083    0.01620
## Detection Prevalence     0.2685    0.00000    0.04167    0.02546
## Balanced Accuracy        0.6701    0.50000    0.89840    0.88416
##                      Class: ME2 Class: ME3 Class: MIT Class: NUC
## Sensitivity             0.46667    0.91304     0.5909     0.6442
## Specificity             0.99041    0.94819     0.9390     0.7835
## Pos Pred Value          0.63636    0.67742     0.7123     0.4855
## Neg Pred Value          0.98100    0.98919     0.8997     0.8741
## Prevalence              0.03472    0.10648     0.2037     0.2407
## Detection Rate          0.01620    0.09722     0.1204     0.1551
## Detection Prevalence    0.02546    0.14352     0.1690     0.3194
## Balanced Accuracy       0.72854    0.93062     0.7649     0.7139
##                      Class: POX Class: VAC
## Sensitivity            0.428571    0.00000
## Specificity            1.000000    1.00000
## Pos Pred Value         1.000000        NaN
## Neg Pred Value         0.990676    0.97917
## Prevalence             0.016204    0.02083
## Detection Rate         0.006944    0.00000
## Detection Prevalence   0.006944    0.00000
## Balanced Accuracy      0.714286    0.50000
```

Task 2

Visualization of classification and results

a)

```
plot(yeast_ctree)
```



b)

```
library(ggplot2)
#normalize predictions between 0 and 1
normalized <- (predict-min(predict))/(max(predict)-min(predict))
normalzied_data<-data.frame(normalized)

#use ggplot to prodcue the heatmap visualziation
ggplot(normalzied_data, aes(Var1,Var2 )) +
  geom_tile(aes(fill = Freq), color = "yellow") +
  scale_fill_gradient(low = "white", high = "red") +


  theme(legend.title = element_text(size = 10),
        legend.text = element_text(size = 10),

        axis.title=element_text(size=10,face="bold"),
        axis.text.x = element_text(angle = 0, hjust = 1)) +
  labs(fill = "Frequency")
```