Breast Cancer Prediction with R

Code

Hide

```r
data <- read.csv("breast cancer.csv",header=FALSE, sep=",")

str(data)
```

```
'data.frame':   699 obs. of  11 variables:
 $ V1 : int  1000025 1002945 1015425 1016277 1017023 1017122 1018099 1018561 1033078 1033078 ...
 $ V2 : int  5 5 3 6 4 8 1 2 2 4 ...
 $ V3 : int  1 4 1 8 1 10 1 1 1 2 ...
 $ V4 : int  1 4 1 8 1 10 1 2 1 1 ...
 $ V5 : int  1 5 1 1 3 8 1 1 1 1 ...
 $ V6 : int  2 7 2 3 2 7 2 2 2 2 ...
 $ V7 : chr  "1" "10" "2" "4" ...
 $ V8 : int  3 3 3 3 3 9 3 3 1 2 ...
 $ V9 : int  1 2 1 7 1 7 1 1 1 1 ...
 $ V10: int  1 1 1 1 1 1 1 1 5 1 ...
 $
```

```
Error in gregexpr(calltext, singleline, fixed = TRUE) :
  regular expression is invalid UTF-8
Error in gregexpr(calltext, singleline, fixed = TRUE) :
  regular expression is invalid UTF-8
Error in gregexpr(calltext, singleline, fixed = TRUE) :
  regular expression is invalid UTF-8
```

```
V11: int  2 2 2 2 2 4 2 2 2 2 ...
```

Hide

```r
head(data)
```

	V1 <int>	V2 <int>	V3 <int>	V4 <int>	V5 <int>	V6 <int>	V7 <chr>	V8 <int>	V9 <int>
1	1000025	5	1	1	1	2	1	3	1
2	1002945	5	4	4	5	7	10	3	2
3	1015425	3	1	1	1	2	2	3	1
4	1016277	6	8	8	1	3	4	3	7
5	1017023	4	1	1	3	2	1	3	1
6	1017122	8	10	10	8	7	10	9	7

6 rows | 1-10 of 11 columns

Hide

```r
summary(data)
```

```
       V1                  V2               V3               V4
 Min.   :   61634   Min.   : 1.000   Min.   : 1.000   Min.   : 1.000
 1st Qu.:  870688   1st Qu.: 2.000   1st Qu.: 1.000   1st Qu.: 1.000
 Median : 1171710   Median : 4.000   Median : 1.000   Median : 1.000
 Mean   : 1071704   Mean   : 4.418   Mean   : 3.134   Mean   : 3.207
 3rd Qu.: 1238298   3rd Qu.: 6.000   3rd Qu.: 5.000   3rd Qu.: 5.000
 Max.   :13454352   Max.   :10.000   Max.   :10.000   Max.   :10.000
       V5               V6              V7                  V8
 Min.   : 1.000   Min.   : 1.000   Length:699         Min.   : 1.000
 1st Qu.: 1.000   1st Qu.: 2.000   Class :character   1st Qu.: 2.000
 Median : 1.000   Median : 2.000   Mode  :character   Median : 3.000
 Mean   : 2.807   Mean   : 3.216                      Mean   : 3.438
 3rd Qu.: 4.000   3rd Qu.: 4.000                      3rd Qu.: 5.000
 Max.   :10.000   Max.   :10.000                      Max.   :10.000
       V9              V10              V11
 Min.   : 1.000   Min.   : 1.000   Min.   :2.00
 1st Qu.: 1.000   1st Qu.: 1.000   1st Qu.:2.00
 Median : 1.000   Median : 1.000   Median :2.00
 Mean   : 2.867   Mean   : 1.589   Mean   :2.69
 3rd Qu.: 4.000   3rd Qu.: 1.000   3rd Qu.:4.00
 Max.   :10.000   Max.   :10.000   Max.   :4.00
```

Hide

```r
names(data) <- c('Id','Cl_thickness','Cell_size','Cell_shape','Marg_adhesion','Epith_c_size','Bare_nuclei','Bl_cr
omat','Normal_nucleoli','Mitoses','Class')
data$Class[data$Class == 2] = 0
data$Class[data$Class == 4] = 1
```

Hide

```r
summary(data)
```

```
       Id            Cl_thickness      Cell_size        Cell_shape
 Min.   :   61634   Min.   : 1.000   Min.   : 1.000   Min.   : 1.000
 1st Qu.:  870688   1st Qu.: 2.000   1st Qu.: 1.000   1st Qu.: 1.000
 Median : 1171710   Median : 4.000   Median : 1.000   Median : 1.000
 Mean   : 1071704   Mean   : 4.418   Mean   : 3.134   Mean   : 3.207
 3rd Qu.: 1238298   3rd Qu.: 6.000   3rd Qu.: 5.000   3rd Qu.: 5.000
 Max.   :13454352   Max.   :10.000   Max.   :10.000   Max.   :10.000
 Marg_adhesion     Epith_c_size    Bare_nuclei          Bl_cromat
 Min.   : 1.000   Min.   : 1.000   Length:699         Min.   : 1.000
 1st Qu.: 1.000   1st Qu.: 2.000   Class :character   1st Qu.: 2.000
 Median : 1.000   Median : 2.000   Mode  :character   Median : 3.000
 Mean   : 2.807   Mean   : 3.216                      Mean   : 3.438
 3rd Qu.: 4.000   3rd Qu.: 4.000                      3rd Qu.: 5.000
 Max.   :10.000   Max.   :10.000                      Max.   :10.000
 Normal_nucleoli     Mitoses           Class
 Min.   : 1.000   Min.   : 1.000   Min.   :0.0000
 1st Qu.: 1.000   1st Qu.: 1.000   1st Qu.:0.0000
 Median : 1.000   Median : 1.000   Median :0.0000
 Mean   : 2.867   Mean   : 1.589   Mean   :0.3448
 3rd Qu.: 4.000   3rd Qu.: 1.000   3rd Qu.:1.0000
 Max.   :10.000   Max.   :10.000   Max.   :1.0000
```

Hide

```r
sum(data == '?')
```

```
[1] 16
```

Hide

```r
data[data == '?'] <- 0
data$Bare_nuclei <- as.numeric(data$Bare_nuclei)
data$Bare_nuclei[data$Bare_nuclei == 0 ] <- mean(data$Bare_nuclei, na.rm =TRUE)
```

Hide

```r
data$Id <- NULL
str(data)
```

```
'data.frame':   699 obs. of  10 variables:
 $ Cl_thickness   : int  5 5 3 6 4 8 1 2 2 4 ...
 $ Cell_size      : int  1 4 1 8 1 10 1 1 1 2 ...
 $ Cell_shape     : int  1 4 1 8 1 10 1 2 1 1 ...
 $ Marg_adhesion  : int  1 5 1 1 3 8 1 1 1 1 ...
 $ Epith_c_size   : int  2 7 2 3 2 7 2 2 2 2 ...
 $ Bare_nuclei    : num  1 10 2 4 1 10 10 1 1 1 ...
 $ Bl_cromat      : int  3 3 3 3 3 9 3 3 1 2 ...
 $ Normal_nucleoli: int  1 2 1 7 1 7 1 1 1 1 ...
 $ Mitoses        : int  1 1 1 1 1 1 1 1 5 1 ...
 $ Class          : num  0 0 0 0 0 1 0 0 0 0 ...
```

Hide

```r
summary(data)
```

```
  Cl_thickness      Cell_size        Cell_shape     Marg_adhesion
 Min.   : 1.000   Min.   : 1.000   Min.   : 1.000   Min.   : 1.000
 1st Qu.: 2.000   1st Qu.: 1.000   1st Qu.: 1.000   1st Qu.: 1.000
 Median : 4.000   Median : 1.000   Median : 1.000   Median : 1.000
 Mean   : 4.418   Mean   : 3.134   Mean   : 3.207   Mean   : 2.807
 3rd Qu.: 6.000   3rd Qu.: 5.000   3rd Qu.: 5.000   3rd Qu.: 4.000
 Max.   :10.000   Max.   :10.000   Max.   :10.000   Max.   :10.000
  Epith_c_size     Bare_nuclei       Bl_cromat      Normal_nucleoli
 Min.   : 1.000   Min.   : 1.000   Min.   : 1.000   Min.   : 1.000
 1st Qu.: 2.000   1st Qu.: 1.000   1st Qu.: 2.000   1st Qu.: 1.000
 Median : 2.000   Median : 1.000   Median : 3.000   Median : 1.000
 Mean   : 3.216   Mean   : 3.543   Mean   : 3.438   Mean   : 2.867
 3rd Qu.: 4.000   3rd Qu.: 5.000   3rd Qu.: 5.000   3rd Qu.: 4.000
 Max.   :10.000   Max.   :10.000   Max.   :10.000   Max.   :10.000
    Mitoses           Class
 Min.   : 1.000   Min.   :0.0000
 1st Qu.: 1.000   1st Qu.:0.0000
 Median : 1.000   Median :0.0000
 Mean   : 1.589   Mean   :0.3448
 3rd Qu.: 1.000   3rd Qu.:1.0000
 Max.   :10.000   Max.   :1.0000
```

Hide

```r
library(caTools)
set.seed(123)
split=sample.split(data, SplitRatio = 0.8)
training_set=subset(data,split==TRUE)
test_set=subset(data,split==FALSE)
dim(training_set)
```

```
[1] 559  10
```

Hide

```r
training_set[,1:9] = scale(training_set[,1:9])
test_set[,1:9] = scale(test_set[,1:9])
```

Hide

```r
Classifier = glm(formula = Class ~ .,
                 family = binomial,
                 data = training_set)
```

Hide

```r
prob_pred = predict(Classifier, type = 'response', newdata = test_set[1:9])
prob_pred
```

```
          5           8          15          18          25          28
0.038800843 0.010011047 0.999875910 0.022122579 0.004326489 0.026918754
         35          38          45          48          55          58
0.012214239 0.222147240 0.999019664 0.003056325 0.999349755 0.869301407
         65          68          75          78          85          88
0.003056325 0.962959991 0.962447877 0.043209554 0.999821609 0.996978797
         95          98         105         108         115         118
0.007474883 0.037730478 0.999991536 0.995008064 0.048630123 0.999909901
        125         128         135         138         145         148
0.997359324 0.012884720 0.009682919 0.006455280 0.005285325 0.005003077
        155         158         165         168         175         178
0.002158245 0.007474883 0.102711084 0.999937986 0.997478333 0.976448474
        185         188         195         198         205         208
0.997731560 0.999999255 0.012884720 0.085469008 0.004326489 0.004075962
        215         218         225         228         235         238
0.999996563 0.004326489 0.999892682 0.998523034 0.035920270 0.997810358
        245         248         255         258         265         268
0.004326489 0.963479042 0.998047967 0.009125070 0.994939724 0.936990556
        275         278         285         288         295         298
0.014202821 0.003056325 0.999083117 0.009682919 0.008870513 0.193330979
        305         308         315         318         325         328
0.997485950 0.004326489 0.002879135 0.999512587 0.004326489 0.003056325
        335         338         345         348         355         358
0.997457016 0.004326489 0.999881243 0.002475756 0.003056325 0.999997191
        365         368         375         378         385         388
0.007474883 0.999941252 0.012214239 0.002879135 0.003734723 0.109154654
        395         398         405         408         415         418
0.006073169 0.011135473 0.004687212 0.003056325 0.997738218 0.003056325
        425         428         435         438         445         448
0.006455280 0.970863924 0.991087299 0.011135473 0.111043267 0.019143511
        455         458         465         468         475         478
0.004840716 0.999362988 0.011135473 0.998897357 0.019143511 0.011135473
        485         488         495         498         505         508
0.025536879 0.999998387 0.763531146 0.012295840 0.002158245 0.007910529
        515         518         525         528         535         538
0.999902456 0.002879135 0.009125070 0.022122579 0.005285325 0.050014819
        545         548         555         558         565         568
0.012634610 0.003518342 0.006455280 0.033653140 0.024362386 0.036691925
        575         578         585         588         595         598
0.994529946 0.003056325 0.080991836 0.029629437 0.987811123 0.066024291
        605         608         615         618         625         628
0.996177046 0.002158245 0.004979554 0.005912059 0.033633440 0.020902428
        635         638         645         648         655         658
0.006455280 0.105338463 0.003734723 0.003844525 0.012884720 0.287808475
        665         668         675         678         685         688
0.016166478 0.012884720 0.003056325 0.019143511 0.002158245 0.011095114
        695         698
0.010544126 0.972222443
```

Hide

```r
y_pred = ifelse(prob_pred > 0.5, 1, 0)
y_pred
```

```
  5   8  15  18  25  28  35  38  45  48  55  58  65  68  75  78  85  88  95
  0   0   1   0   0   0   0   0   1   0   1   1   0   1   1   0   1   1   0
 98 105 108 115 118 125 128 135 138 145 148 155 158 165 168 175 178 185 188
  0   1   1   0   1   1   0   0   0   0   0   0   0   0   1   1   1   1   1
195 198 205 208 215 218 225 228 235 238 245 248 255 258 265 268 275 278 285
  0   0   0   0   1   0   1   1   0   1   0   1   1   0   1   1   0   0   1
288 295 298 305 308 315 318 325 328 335 338 345 348 355 358 365 368 375 378
  0   0   0   1   0   0   1   0   0   1   0   1   0   0   1   0   1   0   0
385 388 395 398 405 408 415 418 425 428 435 438 445 448 455 458 465 468 475
  0   0   0   0   0   0   1   0   0   1   1   0   0   0   0   1   0   1   0
478 485 488 495 498 505 508 515 518 525 528 535 538 545 548 555 558 565 568
  0   0   1   1   0   0   0   1   0   0   0   0   0   0   0   0   0   0   0
575 578 585 588 595 598 605 608 615 618 625 628 635 638 645 648 655 658 665
  1   0   0   0   1   0   1   0   0   0   0   0   0   0   0   0   0   0   0
668 675 678 685 688 695 698
  0   0   0   0   0   0   1
```

Hide

```r
cm = table(test_set[,10], y_pred)
cm
```

```
   y_pred
     0  1
  0 96  2
  1  0 42
```

Hide

```r
accuracy = (cm[1,1] + cm[2,2]/(cm[1,1] + cm[2,2]+ cm[1,2] + cm[2,1]))
accuracy
```

```
[1] 96.3
```

Hide

```r
library(caret)
```

```
Loading required package: lattice
Loading required package: ggplot2
Registered S3 method overwritten by 'data.table':
  method           from
  print.data.table
```

Hide

```r
folds = createFolds(training_set$Class, k = 10)
CrossValidation = lapply(folds, function(x){
  training_fold = training_set[-x,] # taking all the training set but without the fold
  test_fold = training_set[x,]
  Classifier = glm(formula = Class ~ .,
                   family = binomial,
                   data = training_fold)
  prob_pred = predict(Classifier, type = 'response', newdata = test_fold[1:9])
  y_pred = ifelse(prob_pred > 0.5, 1, 0)
  cm = table(test_fold[,10], y_pred)
  accuracy = ((cm[1,1] + cm[2,2])/(cm[1,1] + cm[2,2]+ cm[1,2] + cm[2,1]))
  return(accuracy)
})
CrossValidation
```

```
$Fold01
[1] 0.9821429

$Fold02
[1] 0.9821429

$Fold03
[1] 0.9821429

$Fold04
[1] 0.9642857

$Fold05
[1] 0.9107143

$Fold06
[1] 0.9464286

$Fold07
[1] 0.9818182

$Fold08
[1] 0.9464286

$Fold09
[1] 0.9642857

$Fold10
[1] 0.9821429
```

Hide

```r
accuracies = mean(as.numeric(CrossValidation))
accuracies
```

```
[1] 0.9642532
```

Hide

```r
library(caret)
classifier = train(form = as.factor(Class) ~ ., data = training_set, method='glm')
classifier
```

```
Generalized Linear Model

559 samples
  9 predictor
  2 classes: '0', '1'

No pre-processing
Resampling: Bootstrapped (25 reps)
Summary of sample sizes: 559, 559, 559, 559, 559, 559, ...
Resampling results:

  Accuracy   Kappa
  0.9512439  0.8926616
```

Hide

```r
classifier$bestTune
```

	parameter <chr>
1	none

1 row

Hide

```r
summary(Classifier)
```

```
Call:
glm(formula = Class ~ ., family = binomial, data = training_set)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.1991  -0.1547  -0.0718   0.0297   2.2621

Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)      -0.9498     0.3134  -3.031 0.002439 **
Cl_thickness      1.5076     0.4069   3.705 0.000211 ***
Cell_size         0.2892     0.6459   0.448 0.654360
Cell_shape        0.8320     0.6726   1.237 0.216104
Marg_adhesion     0.8626     0.3897   2.214 0.026858 *
Epith_c_size      0.1314     0.3666   0.358 0.719992
Bare_nuclei       1.5123     0.3676   4.113  3.9e-05 ***
Bl_cromat         0.8234     0.4189   1.966 0.049319 *
Normal_nucleoli   0.2803     0.3297   0.850 0.395298
Mitoses           0.7276     0.5501   1.323 0.185983
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 727.9  on 558  degrees of freedom
Residual deviance: 103.0  on 549  degrees of freedom
AIC: 123

Number of Fisher Scoring iterations: 8
```

Hide

```r
summ = summary(Classifier)
p_values = summ$coefficients[,4]
p_values =  p_values[p_values>0.5]
imp_cols = names(p_values)
imp_cols
```

```
[1] "Cell_size"    "Epith_c_size"
```

Hide

```r
set = data
plot(set)
```