Clara

Anqi Chi SID:460204008

22/10/2018

Data Clean

```
## [1] 12153   145
```

```
##     BDYMSQ04           BMR             SEX             EIBMR1      
##  Min.   :0.0000   Min.   : 4275   Min.   :0.0000   Min.   :0.0000  
##  1st Qu.:0.0000   1st Qu.: 5734   1st Qu.:0.0000   1st Qu.:0.9381  
##  Median :0.0000   Median : 6504   Median :0.0000   Median :1.2371  
##  Mean   :0.1382   Mean   : 6710   Mean   :0.4757   Mean   :1.3064  
##  3rd Qu.:0.0000   3rd Qu.: 7595   3rd Qu.:1.0000   3rd Qu.:1.6033  
##  Max.   :1.0000   Max.   :12759   Max.   :1.0000   Max.   :5.8661  
##     ADTOTSE        CHOPER1         FATPER1         PROPER1     
##  Min.   :   0   Min.   : 0.00   Min.   : 0.00   Min.   : 0.00  
##  1st Qu.:1410   1st Qu.:36.11   1st Qu.:24.99   1st Qu.:14.14  
##  Median :2165   Median :43.51   Median :30.88   Median :17.40  
##  Mean   :2358   Mean   :43.22   Mean   :30.83   Mean   :18.34  
##  3rd Qu.:3180   3rd Qu.:50.48   3rd Qu.:36.53   3rd Qu.:21.38  
##  Max.   :9180   Max.   :98.73   Max.   :82.87   Max.   :63.65  
```

LDA

Load the functions from cvTools into memory.

Fit the LDA method

```r
n = length(y)
dat = data.frame(y,X)
X1 = model.matrix(~-1+BMR+EIBMR1+ADTOTSE+SEX+BDYMSQ04+CHOPER1+FATPER1+PROPER1,data=dat)
X1 = data.frame(X1)
summary(X1)
```

```
##       BMR            EIBMR1          ADTOTSE          SEX        
##  Min.   : 4275   Min.   :0.0000   Min.   :   0   Min.   :0.0000  
##  1st Qu.: 5734   1st Qu.:0.9381   1st Qu.:1410   1st Qu.:0.0000  
##  Median : 6504   Median :1.2371   Median :2165   Median :0.0000  
##  Mean   : 6710   Mean   :1.3064   Mean   :2358   Mean   :0.4757  
##  3rd Qu.: 7595   3rd Qu.:1.6033   3rd Qu.:3180   3rd Qu.:1.0000  
##  Max.   :12759   Max.   :5.8661   Max.   :9180   Max.   :1.0000  
##     BDYMSQ04         CHOPER1         FATPER1         PROPER1     
##  Min.   :0.0000   Min.   : 0.00   Min.   : 0.00   Min.   : 0.00  
##  1st Qu.:0.0000   1st Qu.:36.11   1st Qu.:24.99   1st Qu.:14.14  
##  Median :0.0000   Median :43.51   Median :30.88   Median :17.40  
##  Mean   :0.1382   Mean   :43.22   Mean   :30.83   Mean   :18.34  
##  3rd Qu.:0.0000   3rd Qu.:50.48   3rd Qu.:36.53   3rd Qu.:21.38  
##  Max.   :1.0000   Max.   :98.73   Max.   :82.87   Max.   :63.65  
```

```r
res <- lda(y~., data=X1,subset=1:n)
res
```

```
## Call:
## lda(y ~ ., data = X1, subset = 1:n)
##
## Prior probabilities of groups:
##         0         1
## 0.7236162 0.2763838
##
## Group means:
##        BMR   EIBMR1  ADTOTSE       SEX  BDYMSQ04  CHOPER1  FATPER1
## 0 6426.343 1.380460 2313.456 0.4791776 0.1073358 43.38162 30.74535
## 1 7452.157 1.112322 2475.082 0.4666971 0.2189781 42.79542 31.06969
##    PROPER1
## 0 18.05556
## 1 19.08146
##
## Coefficients of linear discriminants:
##                    LD1
## BMR       1.248732e-03
## EIBMR1   -4.342037e-01
## ADTOTSE  -4.590553e-05
## SEX      -2.093857e+00
## BDYMSQ04  3.632406e-01
## CHOPER1  -5.031900e-03
## FATPER1   1.558290e-03
## PROPER1   2.123950e-03
```

Comment:

The above output suggests the following interpretations for each of the variables. * People who have higher BMR are more likely to obese * Lower EIBMR1(Energy intake) increases obese probability (since the coefficient -0.434 is negative). * Lower ADTOTSE(Total mins spent sitting or lying down) increases obese probability (since the coefficient -4.590553e-05 is negative).

- Sex = 1 for males. So females are more likely to obese
- People on a diet are more likely to obese compared with people not on a diet
- People who have high fat and high protein diet type are more likely to obese
- People who have high carbon diet type decrease the probability of obesity

CV error

```r
#cross validation error for LDA.
res.lda = cv.da(X1,y,method="lda",V,seed=1)
res.lda
```

```
## [1] 0.1823225
```

Comment: The CV error for LDA is 18.232 percent

CART

```
library(rpart)
```

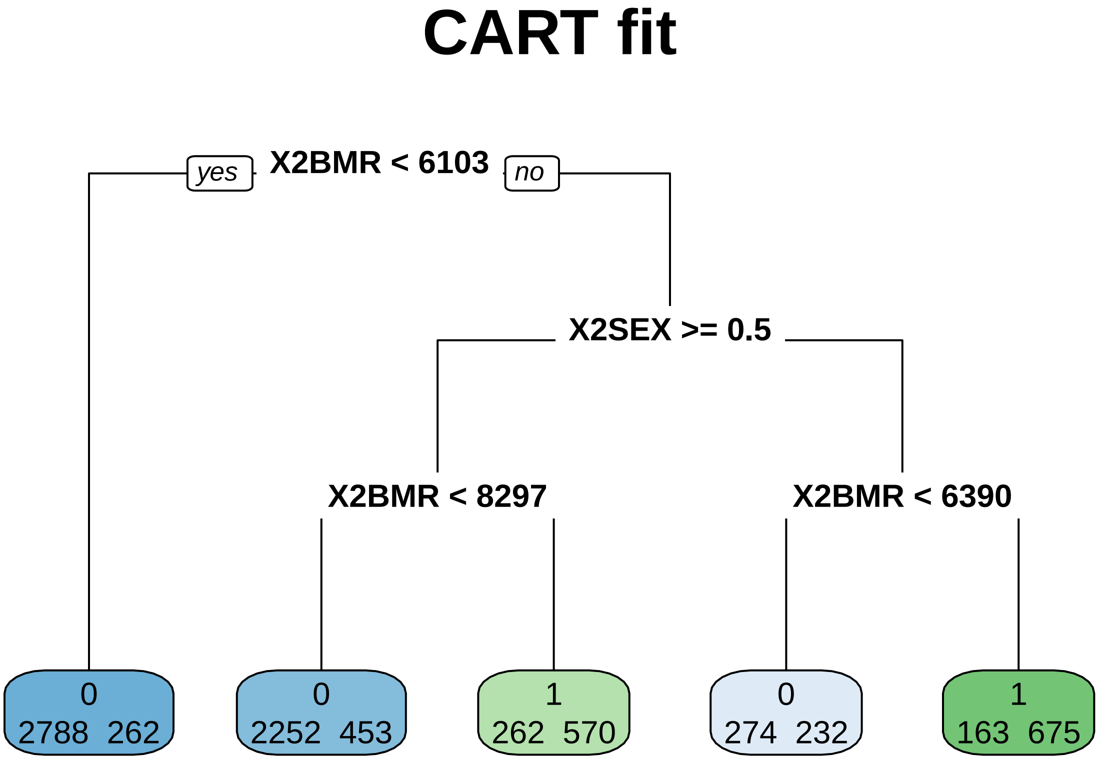
```
## Warning: package 'rpart' was built under R version 3.4.3
```

```
X2 = model.matrix(~-1+BMR+EIBMR1+ADTOTSE+SEX+BDYMSQ04+CHOPER1+FATPER1+PROPER1,data=dat)
# Be careful as coding y as a factor here, Otherwise R will do a regression tree rather than a classifi
res.rpart <- rpart(as.factor(y) ~ X2, data=dat)

library(rpart.plot)
```

```
## Warning: package 'rpart.plot' was built under R version 3.4.4
```

```
rpart.plot(res.rpart,type=0,extra=1,main="CART fit",cex.main=2,cex=1)
```



##Comment: Interpret??

cross-validation for rpart.

CV error

```
res.rpart = cv.rpart(X2,y,V,seed=1)
res.rpart
```

```
## [1] 0.1755138
```

Comment: The CV error for CART is 17.551 percent

Logistic regression

Fit a logistic regression model on most of the data.

```
res.glm = glm(y~.,family=binomial,data=X1)
summary(res.glm)
```

```
##
## Call:
## glm(formula = y ~ ., family = binomial, data = X1)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.7799  -0.6546  -0.3851   0.4044   3.2858
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.093e+01  4.700e-01 -23.246  < 2e-16 ***
## BMR          1.879e-03  5.203e-05  36.109  < 2e-16 ***
## EIBMR1      -7.113e-01  7.220e-02  -9.852  < 2e-16 ***
## ADTOTSE     -5.952e-05  2.449e-05  -2.430   0.0151 *
## SEX         -3.423e+00  1.174e-01 -29.159  < 2e-16 ***
## BDYMSQ04     3.802e-01  8.342e-02   4.557 5.18e-06 ***
## CHOPER1     -1.014e-02  3.948e-03  -2.567   0.0103 *
## FATPER1      1.507e-03  4.357e-03   0.346   0.7295
## PROPER1     -9.982e-04  5.854e-03  -0.171   0.8646
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 9350.7  on 7930  degrees of freedom
## Residual deviance: 6587.0  on 7922  degrees of freedom
## AIC: 6605
##
## Number of Fisher Scoring iterations: 5
```

Comment:

The full model has the coefficeints for BMR, EIBMR1, ADTOTSE, SEX, BDYMSQ04, CHOPER1 as statisticlly signifficantly different from zero at the 0.05 level. The fitted model is

$$\text{logit}(p) = -0.1093 + 0.001879 \cdot \text{BMR} - 3.423 \cdot \text{SEX} - 0.7113 \cdot \text{EIBMR1} - 0.00005952 \cdot \text{ADTOTSE} + 0.3802 \cdot \text{BDYMSQ04} - 0.01014 \cdot \text{CH}$$

where p is the probability of obesity.

The effect of the significant variables on surival are: * Larger BMR reduces the probability of obesity. * Males have reduced the probability of obesity compared to women.

CV error for glms.

```
## [1] 0.1777834
```

Comment: The CV error for glm is 17.78 percent

summary

We now display the results nicely using the package huxtable.

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

Methods	**Errors**
LDA	18.2
rpart	17.6
glm	17.8

The effect of different predictors on obesity are relatively consistent. In summary: