Candy Lab

Minh Tam Hoang

10/27/2019

```
library(readr)
candy_data <- read_csv("CopyOfcandy-power-ranking/candy-data.csv")
```

```
## Parsed with column specification:
## cols(
##   competitorname = col_character(),
##   chocolate = col_double(),
##   fruity = col_double(),
##   caramel = col_double(),
##   peanutyalmondy = col_double(),
##   nougat = col_double(),
##   crispedricewafer = col_double(),
##   hard = col_double(),
##   bar = col_double(),
##   pluribus = col_double(),
##   sugarpercent = col_double(),
##   pricepercent = col_double(),
##   winpercent = col_double()
## )
```

```
head(candy_data)
```

```
## # A tibble: 6 x 13
##   competitorname chocolate fruity caramel peanutyalmondy nougat
##   <chr>              <dbl>  <dbl>   <dbl>          <dbl>  <dbl>
## 1 100 Grand              1      0       1              0      0
## 2 3 Musketeers           1      0       0              0      1
## 3 One dime               0      0       0              0      0
## 4 One quarter            0      0       0              0      0
## 5 Air Heads              0      1       0              0      0
## 6 Almond Joy             1      0       0              1      0
## # ... with 7 more variables: crispedricewafer <dbl>, hard <dbl>,
## #   bar <dbl>, pluribus <dbl>, sugarpercent <dbl>, pricepercent <dbl>,
## #   winpercent <dbl>
```

Choose a candy characteristic and build a logistic regression model.

```
model <- glm(candy_data$chocolate~candy_data$caramel + candy_data$sugarpercent + candy_data$hard)
print(summary(model))
```

```
##
## Call:
## glm(formula = candy_data$chocolate ~ candy_data$caramel + candy_data$sugarpercent +
##     candy_data$hard)
##
## Deviance Residuals:
```

```
##      Min        1Q    Median        3Q       Max
## -0.80131  -0.42713  -0.06117   0.51025   0.93883
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)                0.3912     0.1004   3.895 0.000201 ***
## candy_data$caramel         0.2527     0.1403   1.801 0.075366 .
## candy_data$sugarpercent    0.1631     0.1845   0.884 0.379249
## candy_data$hard           -0.4286     0.1336  -3.207 0.001921 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 0.2140263)
##
##     Null deviance: 20.894  on 84  degrees of freedom
## Residual deviance: 17.336  on 81  degrees of freedom
## AIC: 116.08
##
## Number of Fisher Scoring iterations: 2
```

Description of target variables and covariates

In this model, we are trying to classify candy into "chocolate" group and "non-chocalate" group. Three covariates are used in this model, which are: - Caramel: whether or not the candy has caramel flavour.

- sugar percentage: the percentage of sugar in the candy.
- hard: whether or not the candy is hard.

Logistic Regression Model is used to analyze the candy dataset.

Description of log-odds

The intercept is 0.3912

'Whether or not the candy has caramel flavour' is 0.2527 (positive log-odd)==> Candy that has caramel flavour indicates that the candy is likely to have chocolate flavour.

The sugar percentage is 0.1631(positive log-odd)==>Candy with higher sugar percentage suggests that it falls into the chocolate group.

'Whether or not the candy is hard' is -0.4286(negative log-odd) ==> Candy that is hard suggests that it is likely to have non-chocolate flavour.

Description of odds

```
print(exp(0.3912))
```

```
## [1] 1.478754
```

```
print(exp(0.2527))
```

```
## [1] 1.287497
```

```
print(exp(0.1631))
```

```
## [1] 1.177154
```

```
print(exp( -0.4286))
```

```
## [1] 0.6514204
```

The intercept is 1.478754

'Whether or not the candy has caramel flavour' is 1.287497==> If the candy has caramel flavour, it is likely to be in chocolate group.

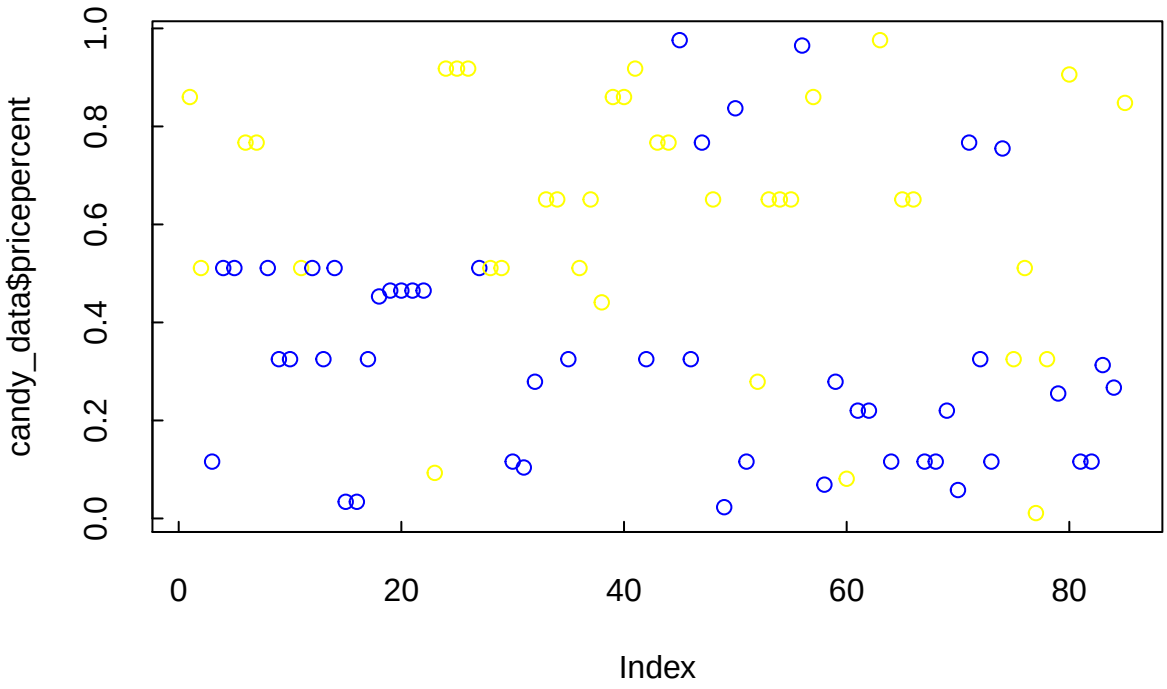
The sugar percentage is 1.177154==>For each unit increase in the sugar percentage of the candy, the odds that it has caramel flavour increase by 1.177154. 'Whether or not the candy is hard' is 0.6514204 < 1 ==> Candy that is hard is unlikely to have chocolate flavour.

Description of significant variables

'Whether or not the candy is hard' is a significant covariate. The p-value is less than 0, which indicates that the null hypothesis that the distribution is centered at log-odds = 0 is rejected.

Plot:

```
colors = ifelse( candy_data$chocolate==1,'yellow','blue')
plot(candy_data$pricepercent,tck=0.02,col=colors)
```




```
coefficients = as.matrix(coef(model))
print(coefficients)
```

```
##                                [,1]
```

```
## (Intercept)              0.3912449
## candy_data$caramel       0.2526762
## candy_data$sugarpercent  0.1630924
## candy_data$hard         -0.4285799
```

#Accuracy = TP+TN/ALL

```
trainingData  = candy_data[,c('sugarpercent','caramel','hard')]
trainingData  = cbind(1,trainingData)

predictions   = as.matrix(trainingData) %*% coefficients
probabilities = exp(predictions)/(1+predictions)
decisions = ifelse(predictions>=0, yes = 1,no = 0)

Accuracy = mean(candy_data$chocolate==decisions)
print(Accuracy)
```

```
## [1] 0.4705882
```

With the logistic regression model which include 'sugar percentile', 'whether it is caramel' and 'whether it is hard', we are able to guess correctly if the candy falls into the chocolate group approximately 47% of the time.