Assignment 3 - Week 9 - Solution

Bo Dao

25 May 2018

```
# loading the iris data
data(iris)

# tranform the dataset to a new one with the four Boolean features

iris$SLC <- iris$Sepal.Length < 6
iris$SWC <- iris$Sepal.Width < 3
iris$PLC <- iris$Petal.Length < 5
iris$PWC <- iris$Petal.Width < 1.6
```

```
# #Change the labeld
# iris$SLC<-factor(iris$SLC,labels =c("SLC=Sepal.Length>=6","SLC=Sepal.Length<6"))
# iris$SWC<-as.factor(iris$SWC)
# iris$SWC<-factor(iris$SWC,labels=c("SWC=Sepal.Width>=3","SWC=Sepal.Width<3"))
# iris$PLC<-as.factor(iris$PLC)
# iris$PLC<-factor(iris$PLC,label=c("PLC=Petal.Length>=5","PLC=Petal.Length<5"))
# iris$PWC<-as.factor(iris$PWC)
# iris$PWC<-factor(iris$PWC,label=c("PWC=Petal.Width>=1.6","PLC=Petal.Width<1.6"))
```

Building 4 pairwise tables for Species vs the new Boolean variables SLC, SWC, PLC and PWC

```
t.SLC <- table(iris$SLC, iris$Species)
t.SLC
```

```
##
##         setosa versicolor virginica
##   FALSE      0         24        43
##   TRUE      50         26         7
```

```
print("FALSE: SLC=Sepal.Length>=6, TRUE: SLC=Sepal.Length<6")
```

```
## [1] "FALSE: SLC=Sepal.Length>=6, TRUE: SLC=Sepal.Length<6"
```

```
t.SWC <- table(iris$SWC, iris$Species)
t.SWC
```

```
##
##         setosa versicolor virginica
##   FALSE     48         16        29
##   TRUE       2         34        21
```

```
print("FALSE: SWC=Sepal.Width>=3, TRUE: SWC=Sepal.Width<3")
```

```
## [1] "FALSE: SWC=Sepal.Width>=3, TRUE: SWC=Sepal.Width<3"
```

```
t.PLC <- table(iris$PLC, iris$Species)
t.PLC
```

```
##
##         setosa versicolor virginica
##   FALSE      0          2        44
##   TRUE      50         48         6
```

```
print("FALSE: PLC=Petal.Length>=5, TRUE: PLC=Petal.Length<5")
```

```
## [1] "FALSE: PLC=Petal.Length>=5, TRUE: PLC=Petal.Length<5"
```

```
t.PWC <- table(iris$PWC, iris$Species)
t.PWC
```

```
##
##         setosa versicolor virginica
##   FALSE      0          5        47
##   TRUE      50         45         3
```

```
print("FALSE: PWC=Petal.Width>=1.6, TRUE: PWC=Petal.Width<1.6")
```

```
## [1] "FALSE: PWC=Petal.Width>=1.6, TRUE: PWC=Petal.Width<1.6"
```

PART 1. Naive Bayes formulas for the classifier:

$p(Species|SLC, SWC, PLC, PWC) = \frac{p(Species)p(SLC|Species)p(SWC|Species)p(PLC|Species)p(PWC|Species)}{p(SLC,SWC,PLC,PWC)}$

While ignoring the nomaliser $p(SLC, SWC, PLC, PWC)$, the formulas for each species without normaliser as follows:

1. $p(Species = "setosa"|SLC, SWC, PLC, PWC) = p(Species = "setosa")p(SLC|Species = "setosa")p(SWC|Species = "setosa")p(PLC|Species = "setosa")p(PWC|Species = "setosa")$
2. $p(Species = "versicolor"|SLC, SWC, PLC, PWC) = p(Species = "versicolor")p(SLC|Species = "versicolor")p(SWC|Species = "versicolor")p(PLC|Species = "versicolor")p(PWC|Species = "versicolor")$
3. $p(Species = "virginica"|SLC, SWC, PLC, PWC) = p(Species = "virginica")p(SLC|Species = "virginica")p(SWC|Species = "virginica")p(PLC|Species = "virginica")p(PWC|Species = "virginica")$

```
# calculating proportional tables
p.Species <- prop.table(table(iris$Species))
p.SLC <- prop.table(t.SLC)
p.SWC <- prop.table(t.SWC)
p.PLC <- prop.table(t.PLC)
p.PWC <- prop.table(t.PWC)

p.Species
```

```
##
##     setosa versicolor  virginica
##  0.3333333  0.3333333  0.3333333
```

```
print("P(Species)")
```

```
## [1] "P(Species)"
```

```
p.SLC/p.Species[1]
```

```
##
##         setosa versicolor virginica
##   FALSE   0.00       0.48      0.86
##   TRUE    1.00       0.52      0.14
```

```
print("P(SLC|Species)")
```

```
## [1] "P(SLC|Species)"
```

```
p.SWC/p.Species[1]
```

```
##
##         setosa versicolor virginica
##   FALSE   0.96       0.32      0.58
##   TRUE    0.04       0.68      0.42
```

```
print("P(SWC|Species)")
```

```
## [1] "P(SWC|Species)"
```

```
p.PLC/p.Species[1]
```

```
##
##         setosa versicolor virginica
##   FALSE   0.00       0.04      0.88
##   TRUE    1.00       0.96      0.12
```

```
print("P(PLC|Species)")
```

```
## [1] "P(PLC|Species)"
```

```
p.PWC/p.Species[1]
```

```
##
##         setosa versicolor virginica
##   FALSE   0.00       0.10      0.94
##   TRUE    1.00       0.90      0.06
```

```
print("P(PWC|Species)")
```

```
## [1] "P(PWC|Species)"
```

Based on above tables, for the case Species = setosa and all of SLC, SWC, PLC, PWC are TRUE, $p(Species = "setosa"|SLC, SWC, PLC, PWC) = p(Species = "setosa"")p(SLC = TRUE|Species = "setosa"")p(SWC = TRUE|Species = "setosa")p(PLC = TRUE|Species = "setosa")p(PWC = TRUE|Species = "setosa")$

$= 0.33x1 * 0.04 * 1 * 1 = 0.013$

PART 2. Logistic Regression

The full formula with all the features will be

$p(Species|SLC, SWC, PLC, PWC) = \frac{1}{1+\exp(-(\beta_0+\beta_1 SLC+\beta_2 SWC+\beta_3 PLC+\beta_4 PWC))}$

where $\beta_0, \beta_1, \beta_2, \beta_3, \beta_4$ are intercept and coefficients.

```
# Optional: building a logistic model for the data
model <-  glm(Species ~ SLC+SWC+PLC+PWC, family = binomial, data = iris)
```

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```
summary(model)
```

```
##
## Call:
## glm(formula = Species ~ SLC + SWC + PLC + PWC, family = binomial,
##     data = iris)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -2.18993  -0.44518   0.00000   0.00003   2.17295
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)   38.252   4144.095   0.009    0.993
## SLCTRUE      -21.026   2672.380  -0.008    0.994
## SWCTRUE        4.564      0.878   5.199 2.01e-07 ***
## PLCTRUE        1.597   7373.605   0.000    1.000
## PWCTRUE      -21.085   6839.814  -0.003    0.998
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 190.954  on 149  degrees of freedom
## Residual deviance:  46.525  on 145  degrees of freedom
## AIC: 56.525
##
## Number of Fisher Scoring iterations: 20
```