FIT5197_ass3_wk9

Chuangfu Xie, 27771539

28/05/2018

1. Loading data and inspect

```
data(iris)
summary(iris)
```

```
##   Sepal.Length    Sepal.Width     Petal.Length    Petal.Width
##  Min.   :4.300   Min.   :2.000   Min.   :1.000   Min.   :0.100
##  1st Qu.:5.100   1st Qu.:2.800   1st Qu.:1.600   1st Qu.:0.300
##  Median :5.800   Median :3.000   Median :4.350   Median :1.300
##  Mean   :5.843   Mean   :3.057   Mean   :3.758   Mean   :1.199
##  3rd Qu.:6.400   3rd Qu.:3.300   3rd Qu.:5.100   3rd Qu.:1.800
##  Max.   :7.900   Max.   :4.400   Max.   :6.900   Max.   :2.500
##        Species
##  setosa    :50
##  versicolor:50
##  virginica :50
##
##
##
```

2. Transform to boolean feature

```
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
```

3. Using `table()` to build the 4 pairwise tables

```
table(iris$SLC,iris$Species, dnn = c("SLC=Sepal.Length<6","Species"))
```

```
##                   Species
## SLC=Sepal.Length<6 setosa versicolor virginica
##               FALSE      0         24        43
##               TRUE      50         26         7
```

```
table(iris$SWC,iris$Species, dnn = c("SWC=Sepal.Width<3","Species"))
```

```
##                  Species
## SWC=Sepal.Width<3 setosa versicolor virginica
##             FALSE     48         16        29
##             TRUE       2         34        21
```

```
table(iris$PLC,iris$Species, dnn = c("PLC=Petal.Length<5","Species"))
```

```
##                   Species
## PLC=Petal.Length<5 setosa versicolor virginica
##               FALSE      0          2        44
##               TRUE      50         48         6
```

```
table(iris$PWC,iris$Species, dnn = c("PWC=Petal.Width<6","Species"))
```

```
##                  Species
## PWC=Petal.Width<6 setosa versicolor virginica
##             FALSE      0          5        47
##             TRUE      50         45         3
```

4. Extract data for fitting model

```
train_data = iris[,5:9]
model <- glm(Species~., family = binomial, data=train_data)
```

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

```
summary(model)
```

```
##
## Call:
## glm(formula = Species ~ ., family = binomial, data = train_data)
##
## 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
```

5. Naive Bayes Classifier

```
sum(iris$SWC==T & iris$SWC==T & iris$PLC==T & iris$PWC==T)
```

```
## [1] 35
```

```
##                  Species
## SLC=Sepal.Length<6 setosa versicolor virginica
##             FALSE      0         24        43
##             TRUE      50         26         7
```

```
##                  Species
## SWC=Sepal.Width<3 setosa versicolor virginica
##             FALSE     48         16        29
##             TRUE       2         34        21
```

```
##                  Species
## PLC=Petal.Length<5 setosa versicolor virginica
##             FALSE      0          2        44
##             TRUE      50         48         6
```

```
##                  Species
## PWC=Petal.Width<6 setosa versicolor virginica
##             FALSE      0          5        47
##             TRUE      50         45         3
```

By given formula:

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

$$p(setosa|SLC,SWC,PLC,PWC)=p(setosa)\frac{p(SLC|setosa)\cdot p(SWC|setosa)\cdot p(PLC|setosa)\cdot p(PWC|setosa)}{p(SLC,SWC,PLC,PWC)}=\frac{\frac{1}{3}\cdot 1\cdot\frac{2}{50}\cdot 1\cdot 1}{\frac{35}{150}}=\frac{2}{35}\approx 0.057$$

$$p(versicolor|SLC,SWC,PLC,PWC)=p(versicolor)\frac{p(SLC|versicolor)\cdot p(SWC|versicolor)\cdot p(PLC|versicolor)\cdot p(PWC|versicolor)}{p(SLC,SWC,PLC,PWC)}=\frac{\frac{1}{3}\cdot\frac{26}{50}\cdot\frac{34}{50}\cdot\frac{48}{50}\cdot\frac{45}{50}}{\frac{35}{150}}\approx 0.436$$

$$p(virginica|SLC,SWC,PLC,PWC)=p(virginica)\frac{p(SLC|virginica)\cdot p(SWC|virginica)\cdot p(PLC|virginica)\cdot p(PWC|virginica)}{p(SLC,SWC,PLC,PWC)}=\frac{\frac{1}{3}\cdot\frac{7}{50}\cdot\frac{21}{50}\cdot\frac{6}{50}\cdot\frac{3}{50}}{\frac{35}{150}}\approx 0.0006$$