An introduction to statistical learning - Ch4 - Ex13

Thalles Quinaglia Liduares

23/03/2023

Using the Boston data set from ISLR package, fit classification models in order to predict whether a given suburb has a crime rate above or below the median. Explore logistic regression, LDA, and KNN models using various subsets of the predictors.

Upload package

```
library(ISLR2)
```

Upload data

```
Boston<-ISLR2::Boston
```

Binary variable highcrime indicates whether the crime rate is above or below the median

```
Boston$highcrime <- ifelse(Boston$crim > median(Boston$crim), 1, 0)
table(Boston$highcrime)
```

```
##
##   0   1
## 253 253
```

Logistic regression

```
fit1 <- glm(highcrime ~ ., data = Boston, family = "binomial")
```

```
## Warning: glm.fit: algoritmo não convergiu
```

```
## Warning: glm.fit: probabilidades ajustadas numericamente 0 ou 1 ocorreu
```

```
summary(fit1)
```

```
##
## Call:
## glm(formula = highcrime ~ ., family = "binomial", data = Boston)
##
## Deviance Residuals:
##        Min          1Q      Median          3Q         Max
## -2.638e-03  -2.000e-08   0.000e+00   2.000e-08   2.689e-03
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -3.270e+02  2.030e+05  -0.002    0.999
## crim         1.056e+03  2.021e+04   0.052    0.958
## zn           2.251e+00  6.284e+01   0.036    0.971
## indus       -3.859e+00  1.542e+03  -0.003    0.998
## chas        -5.407e+00  1.089e+04   0.000    1.000
## nox          1.467e+02  2.190e+05   0.001    0.999
## rm          -4.152e+01  1.990e+03  -0.021    0.983
## age          4.756e-01  8.017e+01   0.006    0.995
## dis         -1.335e+01  2.827e+03  -0.005    0.996
## rad         -4.353e+00  3.454e+03  -0.001    0.999
## tax         -1.346e-01  1.581e+02  -0.001    0.999
## ptratio      1.464e+01  6.733e+03   0.002    0.998
## lstat       -9.119e-01  5.204e+02  -0.002    0.999
## medv         3.491e+00  7.710e+02   0.005    0.996
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 7.0146e+02  on 505  degrees of freedom
## Residual deviance: 2.8134e-05  on 492  degrees of freedom
## AIC: 28
##
## Number of Fisher Scoring iterations: 25
```

We can see that all variables are significant except `age` and `dis` .

LDA model

```
library(MASS)
```

```
##
## Attaching package: 'MASS'
```

```
## The following object is masked _by_ '.GlobalEnv':
##
##     Boston
```

```
## The following object is masked from 'package:ISLR2':
##
##     Boston
```

```
fit2 <- lda(highcrime ~ zn + indus + nox + rm + tax + ptratio  + lstat, data = Boston)
fit2
```

```
## Call:
## lda(highcrime ~ zn + indus + nox + rm + tax + ptratio + lstat,
##     data = Boston)
##
## Prior probabilities of groups:
##   0   1
## 0.5 0.5
##
## Group means:
##          zn     indus       nox       rm      tax  ptratio     lstat
## 0 21.525692  7.002292 0.4709711 6.394395 305.7431 17.90711  9.419486
## 1  1.201581 15.271265 0.6384190 6.174874 510.7312 19.00395 15.886640
##
## Coefficients of linear discriminants:
##                  LD1
## zn      -0.008662315
## indus   -0.003523247
## nox      9.389656234
## rm       0.424228460
## tax      0.002590662
## ptratio  0.045979026
## lstat    0.015640517
```

```
library(class)
x <- subset(Boston, select = c(zn, indus, nox, rm, tax, ptratio, lstat))
y <- Boston$highcrime
set.seed(123)
train <- sample(1:nrow(Boston), nrow(Boston)/2)
test <- setdiff(1:nrow(Boston), train)
yhat <- knn(x[train,], x[test,], y[train], k = 5)
mean(yhat == y[test])
```

```
## [1] 0.9486166
```

We can see that the KNN model has the highest accuracy among the three models (85.2%).

Overall, we found that the KNN model performed the best in predicting whether a given suburb has a crime rate above or below the median, using a subset of significant variables