Data Mining HW 2

Due: 2022.05.02 24:00

Exercises for Cassification

1. 주식에 대한 배당금을 지급할지 여부('Yes', 'No')를 예측하기 위해, 작년 한 해 동안의 수익률을 설명변수 X로 사용하고자 한다. 많은 수의 회사를 조사해보니 배당금을 지급한 회사의 주식 수익률의 평균 $\bar{X}$=10 이고, 지급하지 않은 회사의 주식 수익률의 평균 $\bar{X}$=0이다. 두 그룹에서 수익률의 분산은 두 그룹 모두 $\hat{\sigma}^2$=36이다. X가 정규분포를 따른다고 가정하자. 어떤 회사의 작년 수익률 $X = 4$였다면, 이 회사가 배당금을 지급할 확률은 얼마인가?

2. 'Auto.csv' 데이터를 이용하여 자동차의 연비가 높을지 낮을지에 대해 예측을 하고자 한다. 다음 물음에 답하여라. (NA 처리 후 분석)

   (a) mpg01 변수를 생성하여라. 이 변수는 mpg가 mpg의 중앙값보다 크면 1의 값을 갖고 아니면 0의 값을 갖는 변수이다. (중앙값 : median())

   (b) 시각화를 통하여 mpg01과 다른 변수들 사이의 관계를 확인하고 설명하여라. 어떤 변수가 mp01을 예측하는 데 가장 유용할 것으로 생각되는가?

   (c) 데이터를 training data (60%)와 testing data(40%)로 나누어라.

   (d) (b)에서 연관이 있다고 생각되는 변수들을 이용하여, mpg01을 예측하기 위한 LDA를 수행하여라. test 오분류율은 얼마인가?

   (e) (b)에서 연관이 있다고 생각되는 변수들을 이용하여, mpg01을 예측하기 위한 QDA를 수행하여라. test 오분류율은 얼마인가?

   (f) (b)에서 연관이 있다고 생각되는 변수들을 이용하여, mpg01을 예측하기 위한 logistic regressiond을 수행하여라. test 오분류율은 얼마인가?

   (g) (b)에서 연관이 있다고 생각되는 변수들을 이용하여, mpg01을 예측하기 위한 KNN을 수행하여라. KNN을 수행할 때 몇개의 k 값을 선택하여 분석하여라. test 오분류율은 얼마인가? 어떤 k를 선택했을 때 결과가 가장 좋았는가?

3. 'Boston.csv' 데이터를 이용하여, 어떤 지역의 범죄율이 중앙값 이상인지 아닌지를 예측하기 위한 분류 모형(logistic regression, LDA, KNN)을 적합하여라.

   - 변수설명

     crim : per capita crime rate by town.

zn : proportion of residential land zoned for lots over 25,000 sq.ft.

indus : proportion of non-retail business acres per town.

chas : Charles River dummy variable (= 1 if tract bounds river; 0 otherwise).

nox : nitrogen oxides concentration (parts per 10 million).

rm : average number of rooms per dwelling.

age : proportion of owner-occupied units built prior to 1940.

dis : weighted mean of distances to five Boston employment centres.

rad : index of accessibility to radial highways.

tax : full-value property-tax rate per $10,000.

ptratio : pupil-teacher ratio by town.

lstat : lower status of the population (percent).

medv : median value of owner-occupied homes in $1000s.