**빅데이터 예측분석 과제**

정보통신대학원 A57013 성향숙

1. auto.csv를 이용한 linear regression 문제(교재 chapter 3 exercise 9)

(a) scatterplot / (b) correlation / (c) fit a linear regression model with mpg as the response and all other variables(excl. name) as predictors : 첨부 homework.ipynb 파일

comments on the output:

* fit한 모델의 F-statistics의 p-value가2.04e-139로 매우 작으므로 null hypothesis를 거부할 수 있는 모델임(=predictors과 response간의 상관 관계가 있다고 할 수 있음)
* acceleration, horsepower, cylinders는 p-value가 크므로 null-hypothesis를 거부할 수 없음(=response와 별 관련이 없음)
* response(mpg)와 관련이 있는 predictors(p-value가 작은 predictors)는 weight, year, origin, displacement 이며, 그 중 year와 origin의 coefficient가 크므로 response에 가장 큰 영향을 주는 predictors
* year의 coefficient가 0.75이므로 year가 1 커지면(1년 늦어지면) response인 mpg는 0.75 늘어남

(e) Use the \* and : symbols to fit linear regression models with interaction effects. Do any interactions appear to be statistically significant?

* displacement:weight의 p-value가 0이므로 displacement와 weight는 서로 상관 관계가 높은 predictors임.
* displacement:year 또한 상관관계 높은 predictos

(f) Try a few different transformations of the variables, such as log(X), √X, X2. Comment on your findings.

- displacement와 weight의 square root로 모델을 fit하면(단, 이 때 acceleration, horsepower, cylinders는 제외하고 나머지 4개 variables로만 fit), 4개의 predictors가 비교적 균일한 coefficient를 가지는 모델을 얻게 된다. 즉, 4개 predictor의 근소한 변화에도 민감하게 반응하는 모델이 됨

2. boston 데이터를 이용한 classification 문제(교재 chapter 4 exercise 13)

데이터 전처리: 새로운 컬럼(crime\_above\_median)을 추가하여 crim이 중간값보다 높으면 1, 낮으면 0을 입력함

boston 의 observation이 약 500여개이므로 트레이닝 데이터 400개, 테스트 데이터 100개로 나누어 k-fold cross validation을 통해 모델 평가

모델 fitting과 퍼포먼스

* scikit learn을 이용한 classification:

crime\_above\_median을 response로 하여 모델을 fit했을 때 coefficient가 높은 predictors: nox, rad, chas, dis ... 이므로 해당 4개의 predictors만 사용하여 모델을 다시 fit함. 이 모델에 대하여 5-fold cross validation을 수행하면 0.72~0.90의 정확도를 보임.  
predictors를 nox, rad로만 사용하여 모델을 다시 fit후 k-fold cross validation해보니 정확도가 0.69~0.95로 편차가 커짐

* statsmodel의 logistic regression을 이용한 classification:

crime\_above\_median을 response로 하여 모델을 fit했을 때P값이 낮은 nox, rad, age, medv 을 사용하여 모델을 다시 fit함. 이 모델에 대하여 5-fold cross validation을 수행하면 0.73~0.89의 정확도를 보임.  
predictors를 nox, rad로만 사용하여 모델을 다시 fit후 k-fold cross validation해보니 정확도가 0.73~0.96으로 개선됨

* KNN을 이용한 classification:

crime\_above\_median을 response로 하여 모델을 fit했을 때predictors가 nox, rad인 경우 precision이 가장 높았음. 그리고 k값이 1,2,3,9일 때 precision이 높았음. K값을 3으로 두고 predictors를 nox, rad로 한 모델에 대하여 5-fold cross validation을 수행하면 0.75~0.93의 정확도를 보임.

위와 같이 3개의 모델을 만들고 트레이닝/테스트셋으로 나누어 각각 평가해 본 결과, predictor를 nox, rad로 하고 k값을 1~3으로 둔 KNN 모델의 퍼포먼스가 가장 좋았음.