연세-네이버클라우드 데이터사이언스 교육과정

과목명: 선형회귀분석

데이터 파일: 보스턴 주택가격 데이터 (파이썬 패키지에서 제공)

관찰치수 (n), 변수의 수 (p): n=506, p=14

과제 설명:

sklearn에서 제공하는 보스턴 주택가격 데이터에는 반응변수인 주택의 중앙가격 (MEDV) 과 이를 설명하는 13개의 설명변수들이 존재한다. (해당 데이터를 불러읽는데는 첨부된 주피터노트북을 참고)

1. 설명변수들 중 LSTAT (하위 소득수준. 예를 들어 4는 하위 4%를 의미) 만을 이용해 단순회귀모형을 적합한 후 모형을 평가하시오. (결정계수, 잔차진단 등)
2. 모든 설명변수를 다 포함하는 다중회귀모형을 적합한 후 모형을 평가 하시오.
3. AIC를 기준으로 한 전진선택법을 이용해 필요한 설명변수를 모형에 차례대로 포함하시오. 반응변수를 설명하는데 중요한 변수들은 무엇이라 생각하는가?
4. Lasso를 이용해 필요없는 설명변수를 모형에서 제거하시오. 튜닝파라미터의 변화에 따른 회귀계수의 변화를 볼 때 가장 중요한 변수들은 무엇이라 생각하는가?