19101185 강동현 데이터분석 Assignment 08

1. forward stepwise과 backward stepwise selection을 single data set에 적용한다. P+1개의 모델을 얻고, 0, 1, 2… p개의 predictors를 포함한다.

(a). k predictors를 가지고 있는 3개의 모델 들 중 가장 작은 training RSS를 가지는 모델은 무엇인지 묻는 문제이다. 이 경우 모델들중 Best subset selection이 가장 낮은 training RSS를 가진다. Best subset selection은 가장 낮은 training RSS를 가질 수 있지만 , 모델의 경우를 모두 확인하는 과정에서 시간이 오래 걸릴 수 있다는 단점이 있다.

(b). k predictors를 가지고 있는 3개의 모델 들 중 가장 작은 test RSS를 가지는 모델은 무엇인지 묻는 문제이다. test RSS에 대해서는 3가지 모델들 중 어느 하나가 확실히 좋다고 말할 수 없다. 왜냐하면 모델을 적용할 Data set들의 특성에 따라 각기 다른 결과가 나올 수 있기 때문이다. Test RSS가 좀 더 좋게 나올 가능성이 높은 것을 뽑자면 전체 경우를 모두 확인하는 Best subset selection가 그나마 나쁘지 않을 것이다.

(c). 참 , 거짓을 고르는 문제이다.

(i). forward stepwise selection 에서 k-variable model는 (k+1)-variable model의

부분 집합인지 묻는 문제이다. forward stepwise selection은 아무것도 없는 상태에서 하나씩 채워나가는 방법이므로 k -> k+1로 넘어가게 된다. 따라서 k+1은 k개의 경우에 1개의 요소가 포함된 상태일 것이고 따라서 부분집합 관계가 성립함을 알 수 있다. 따라서 정답은 True 이다.

(ii). backward stepwise selection 에서 k-variable model는 (k+1)-variable model의

부분 집합인지 묻는 문제이다. backward stepwise selection는 모두 채워진 상태에서 하나씩 뺴가는 방법이다. 따라서 이경우도 k+1의 경우는 k개의 경우에 1가지 요소가 포함된 상태일 것이다. 따라서 부분집합 관계가 성립한다. 정답은 True이다.

(iii). backward stepwise selection 의 k-variable model은 forward stepwise selection의 (k + 1)-variable model의 부분집합인지 묻는 문제이다. backward stepwise selection과 forward stepwise selection은 방법 자체가 다르므로 경우를 확인해야 한다. 따라서 모든 경우에 대해서 부분집합이라고 말할 수 없다. 따라서 정답은 False이다.

(iv). forward stepwise selection의 k-variable model은 backward stepwise selection 의 (k + 1)-variable model의 부분집합인지 묻는 문제이다. 이 역시 iii번과 같다. backward stepwise selection과 forward stepwise selection은 방법 자체가 다르므로 경우를 확인해야 한다. 따라서 모든 경우에 대해서 부분집합이라고 말할 수 없다. 따라서 정답은 False이다.

(v). best subset selection으로 나온 best subset set인 k-variable model은 best subset selection 의 (k + 1)-variable model의 부분집합인지 묻는 문제이다. 이 경우 결론은 False이다. 왜냐하면 best subset selection은 k개를 선택하는 경우에 있어서 이전의 set이 전혀 영향을 미치지 않기 때문이다. 경우에 따라서 이전의 경우에 요소가 더해진 형태로 나올 수도 있지만 , k개를 뽑는 경우는 다른 k+1이나 k-1과는 각각 독립적이다. 따라서 정답은 False이다.

2. 문제에서 옳은 번호를 고르고 근거를 제시하는 문제이다.

(a). Lasso의 경우이다. (a)번의 경우 정답은 iii 이다. iii은 flexible이 낮아지면 prediction accuracy가 증가하고 이는 bias가 증가하는 정도가 variance가 감소하는 정도 보다 낮기 때문에 일어난다고 이야기 한다. prediction accuracy는 bias와 variance를 더한 것이 작을수록 높아진다. Bias가 증가하고 variance가 감소하는데 이 증감의 정도에 있어서 bias가 더 적다면 prediction accuracy는 증가할 것이다. 따라서 답은 iii 이다.

(b). ridge regression의 경우에 대해서도 (a)번과 같이 문제를 풀면 된다. ridge regression의 경우에도 Lasso와 같다. 따라서 답은 iii 이다. ridge regression의 경우에도 prediction accuracy는 bias와 variance를 더한 것이 작을수록 높아진다. Bias가 증가하고 variance가 감소하는데 이 증감의 정도에 있어서 bias가 더 적다면 prediction accuracy는 증가할 것이다. 따라서 답은 iii 이다.

(c). non-linear methods의 경우에 대해서 (a)번과 같이 문제를 푸는 문제이다. non-linear methods은 선형적이지 않으므로 좀더 flexible 하다. 따라서 답은 ii 이다. Ii은 좀더 flexible해 variance의 증가가 bias의 감소보다 작을 때 prediction accuracy가 증가한다고 이야기 한다. non-linear methods은 비 선형적이므로 좀더 flexible 하므로 ii번의 경우에 속한다. 따라서 답은 ii 이다.

3. linear regression model의 regression coefficients을 최소화 하여 추정한다. 아래 식을 보고 각 경우에 대해서 옳은 번호를 고르고 근거를 제시하는 문제이다.

(a). s가 0에서 증가할 때 , training RSS는 어떻게 될지 묻는 문제이다. 이 경우 s가 증가함에 따라서 추정되는 베타값은 증가하고 training RSS는 감소하게 된다. 이 경우 해당되는 번호는 iv번 Steadily decrease 이다. 따라서 답은 iv 이다.

(b). s가 0에서 증가할 때 , Test RSS는 어떻게 될지 묻는 문제이다. 이 경우 test RSS는 점차 감소하다가 데이터 셋에 피팅되면서 점점 줄어들 것이다. 이후 최소값을 지난후에는 test RSS는 오버 피팅떄문에 점차 증가할 것이다. 따라서 감소하다가 최저점 이후 증가하는 U모양을 가지게 될 것이다. 따라서 정답은 ii번 이다.

(c). s가 0에서 증가할 때 , variance는 어떻게 될지 묻는 문제이다. S가 증가할 수록 모델에는 베타가 더 많이 포함될 것이다. 또한 베타값 역시 s가 증가함에 따라 커질 수 있다. 따라서 다양한 수의 베타들이 넓은 범위 안에서 각기 다른 값을 가질 것이므로 variance는 커진다고 볼 수 있다. 따라서 정답은 꾸준히 증가한다는 iii번 이다.

(d). s가 0에서 증가할 때 , bias는 어떻게 될지 묻는 문제이다. 위쪽에서 다룬 variance는 s가 증가함에 따라 같이 증가한다고 보았다. Bias는 variance와의 trade-off 관계가 존재하므로 bias는 s가 증가함에 따라 감소한다고 볼 수 있다. 따라서 답은 iv번 꾸준히 감소한다 이다.

(e). s가 0에서 증가할 때 , irreducible error는 어떻게 될지 묻는 문제이다. irreducible error는 축소가 불가능한 우리가 컨트롤 하지 못하는 에러를 의미한다. 이러한 컨트롤이 불가한 에러는 우리가 모델을 가지고 조정할 수 있는 부분이 아니기에 어떠한 모델을 사용하던 혹은 s값이 변하던 그 값을 조정할 수 없다. 따라서 답은 v번 일정하게 유지됨 이다.

4. linear regression model의 regression coefficients을 최소화 하여 추정한다. 아래 식을 보고 각 경우에 대해서 옳은 번호를 고르고 근거를 제시하는 문제이다.

(a). 람다가 0에서 증가할 때 , training RSS는 어떻게 될지 묻는 문제이다. 각 베타들은 람다가 증가함에 따라서 점점 0으로 감소하게 된다. 이미 training된 set에서 베타들이 0으로 감소하기 때문에 training RSS는 증가할 것이다. 따라서 답은 iii번 점점 증가한다 이다.

(b). 람다가 0에서 증가할 때 , Test RSS 는 어떻게 될지 묻는 문제이다. 람다가 증가함에 따라서 베타가 0에 가까워지므로 점점 traing set에 대한 영향을 벗어난다. 따라서 test RSS는 점점 감소한다. 하지만 하나 둘 베타가 0이 되어 모델이 점점 단순해 질수록 test RSS는 증가할 것이다. 따라서 Test RSS는 감소하다가 최소값 이후 다시 증가하는 U모양을 띄게 될 것이다. 따라서 답은 ii번 이다.

(c). 람다가 0에서 증가할 때 , variance 는 어떻게 될지 묻는 문제이다. 람다가 커지면서 베타는 0에 가까워지고 그에 따라 모델은 점점 더 단순해 진다. 따라서 모델이 단순해짐에 따라 variance는 점점 감소할 것이다. 답은 iv 점점 감소한다 이다.

(d). 람다가 0에서 증가할 때 , bias 는 어떻게 될지 묻는 문제이다. Variance는 람다가 증가함에 따라 감소한다고 보았다. Variance는 bias와 trade-off 관계에 있으므로 bias는 점점 증가할 것이다. 따라서 답은 iii 점점 증가한다 이다.

(e). 람다가 0에서 증가할 때 , irreducible error 는 어떻게 될지 묻는 문제이다. irreducible error는 컨트롤 할 수 없는 에러 이기 떄문에 모델 혹은 람다값에 의해 변하지 않는다. 따라서 정답은 v 일정하게 유지된다 이다.