[ISL] 6.1 Subset Selection

태그

6.1.2 Stepwise Selection

best subset selection은 변수의 수(p)가 매우 클 때는 적용하기 어렵다. 또한 , 변수의 수가 클때는 통계적 문제가 발생할 수 있다.

이런 문제점의 대안으로 stepwise 방법이 있다.

(1) Forward Stepwise Selection

Forward Stepwise Selection은 변수 선택 방법 중 하나로, 빈 모델에서 시작하여 한 번에 하나씩 변수를 추가해나가는 방식이다. 이는 모든 가능한 변수 조합을 고려하는 Best Subset Selection의 계산적으로 효율적인 대안으로 개발되었습니다

forward stepwise selection 알고리즘

(1) Null Model M0에서 시작

(2) k = 0,1,2... , p-1까지 하나씩 돌아가며 Mk 모델에 predictior에 추가해 p-k개의 모델을 학습하고

가장 낮은 RSS 혹은 R^2을 가는 모델을 선택한다.

(3) M0부터 Mp까지의 모델 중에서 최종적으로 하나의 최적 모델을 선택한다.

ex) 변수 a,b,c가 있다고 가정하자.

1단계 M0 = 빈모델

2단계:k = 0일때 a,b,c,를 각각 추가해서 결과를 확인하고 가장 좋은 모델을 선택한다. (예: a라고 가정)

M1(a) k = 1 일때 남은 b,c,중에 하나 선택 (예: c)

 M2(a,c)k= 2 일때 남은 b 추가

M3 (a,c,b)3단계 M0, M1(a) ,M2(a,c), M3 (a,c,b) 중에서 최종 모델 선택

Forward Stepwise는 비교적 계산이 효율적인 면이 있지만, 최적의 모델을 보장하진 않음

# Variables	Best subset	Forward stepwise
One	rating	rating
Two	rating, income	rating, income
Three	rating, income, student	rating, income, student
Four	cards, income	rating, income,
	student, limit	student, limit

Best subset 방법에서 Four variables의 cards 변수가 들어가지만

Forward stepwise는 앞에서 선택된 변수가 계속 포함되기 때문에 차이가 발생한다

- Best subset selection은 매번 모든 가능한 조합을 새로 검토할 수 있어서, rating을 제거하고 cards로 대체하는 것이 가능
- 반면 Forward stepwise는 한 번 선택된 변수(rating)를 제거할 수 없고, 계속 유지한 채로 새로운 변수만 추가할 수 있다.

장점과 한계

장점

1. 계산적으로 매우 효율적
2. 구현이 간단하고 이해하기 쉬움
3. 고차원 데이터(n < p)에서도 적용 가능
4. 실제 분석에서 대체로 좋은 성능을 보임

한계

1. 최적의 모델을 보장하지 않음
2. 한 번 선택된 변수는 제거할 수 없음
3. 변수들 간의 복잡한 상호작용을 고려하지 못할 수 있음

(2) Backward Stepwise Selection

forward와 달리, 모든 p개의 예측변수를 포함한 전체 모델에서 시작한다.

그리고 가장 덜 유용한 예측변수를 하나씩 제거해 나간다.

backward stepwise selection 알고리즘

(1) Full model Mp에서 시작(p개의 predictor)

(2) k = p, p-1, ..., 1 까지 하나씩 돌아가며 k-1개의 predictor를 갖도록 Mk에서 predictor 하나를 제거한 뒤

k개의 모델을 fitting하고 가장 낮은 RSS 혹은 높은 R2를 갖는 모델 Mk-1을 선택

(3) M0, ..., Mp 의 모델에서 가장 좋은 모델 선택 (선택 방법은 뒤에 설명)

full model에서 시작하여 매 스탭마다 k개의 변수를 제거해가며, 어떤 변수를 제거했을때 성능이 좋아지는 지를 체크하는 방법

- 평가해야 할 총 모델 수: 1 + p(p+1)/2
- Best Subset Selection(2^p)에 비해 매우 효율적
- 예: 변수 20개일 때
  - Best Subset: 1,048,576개 모델
  - Backward Stepwise: 211개 모델

장점과 한계

장점

1. 계산이 빠르고 효율적
2. 구현이 상대적으로 간단
3. 변수 간 관계를 점진적으로 평가 가능
4. 각 단계에서 변수의 중요도를 평가 가능

한계점

1. 최적의 모델을 보장하지 않음
2. n < p인 경우 사용 불가 (표본 수가 변수보다 적을 때)
3. 한 번 제거된 변수는 다시 고려할 수 없음
4. 전체 모델에서 시작해야 하므로 초기 계산 부담이 있음

backward stepwise는 처음에 full model을 fitting 해야 하기 때문에 forward stepwise 와는 달리 **p > n 일 때 적용이 불가능**하다

Forward stepwise와 Backward stepwise 방법의 공통적인 단점은

한번 포함한 변수 혹은 한번 제거한 변수를 다시 고려하지 않기 때문에 global minimum을 찾지 못한다는 점이다

그래서 적절하게 forward와 backward의 혼합 버전인 Hybrid Approaches방법을 사용한다.

(1) 빈 모델에서 시작

(2) 가장 좋은 변수 추가 (Forward)

(3) 추가 후, 기존 변수들 중 제거할 만한 것이 있는지 검토 (Backward)

(4) 이 과정을 반복