

06.05 정규화 선형회귀

정규화 선형회귀 방법

선형회귀 계수(weight)에 대한 제약조건 추가한 최적화 -> 과최적화 막음 (모델을 조금 더 부드럽게)

- 1) Ridge 회귀모형
- 2) Lasso 회귀모형
- 3) Elastic Net 회귀모형

각각 하이퍼모수 '람다'(페널티)로 두고, **w**에 제한을 둬

*페널티가 클 수록, 정규화정도가 커짐 (1page)

1. statsmodels의 정규화 회귀모형

- 1) Lasso : `L1_wt = 1`
- 2) Ridge : `L1_wt = 0`
- 3) Elastic net : `L1_wt = 0`과 `1`사이

4-6page

3가지 모델의

공통점 : 회귀 선의 추세는 비슷한 모습

차이점 : 회귀계수 추정값이 `Lasso > Elastic net > Ridge` 순으로 '0'이 많음

*Lasso 는 정규화 정도가 클수록(람다 클수록) 0의 숫자가 많아짐 (Lasso path)

2. sklearn의 정규화 회귀모형

- Ridge, Lasso, ElasticNet 이라는 별도의 클래스 사용
- 9-11 page

3. 정규화 모형의 장점, 의미, Lasso - Ridge 차이

- 1) 정규화 모형 장점 : 안정적 모형 \Leftrightarrow 다중공선성 \rightarrow 조건수 커짐 \rightarrow 데이터가 조금만 커져도 추정값 변화 큼(불안정성)
- 2) 정규화의 의미 : 정규화가 없는 최적화 문제에 부등식 제한 조건을 추가하는 것
- 3) Ridge - Lasso 차이 (11page - 13page)

: Lasso는 일부 가중치 계수를 먼저 0으로 수렴시키며 정규화 정도를 높여감 (ridge는 가중치 계수를 전부다 조금씩 축소하면서 정규화 진행)

4. Lasso 모형의 쓸모 : 중요하지 않은 변수 먼저 0으로 수렴한다 (Lasso Path)

- 13page

5. Validation-Curve : 그렇다면, 페널티는 얼마를 줘야 검증성능이 제일 높을까?

- 최적 정규화
- validation curve 를 그려 최적 람다 찾기
 - *사실, 람다마다 test-performance를 일일이 찾아서 그래프를 그린 것 (18page)

6. Validation-Curve : 다항회귀의 최적 차수 결정

- 정규화 : 람다가 클수록 정규화 정도가 높음(제약이 큰 것)
- 다항회귀 : 차수가 클수록 정규화 정도가 낮음(제약이 낮은 것. 사용 가능한 계수가 많아짐) (18page)