Part. 02
회귀분석

Feature selection 정리

FASTCAMPUS
ONLINE
머신러닝과 데이터분석 A-Z

강사. 이경택

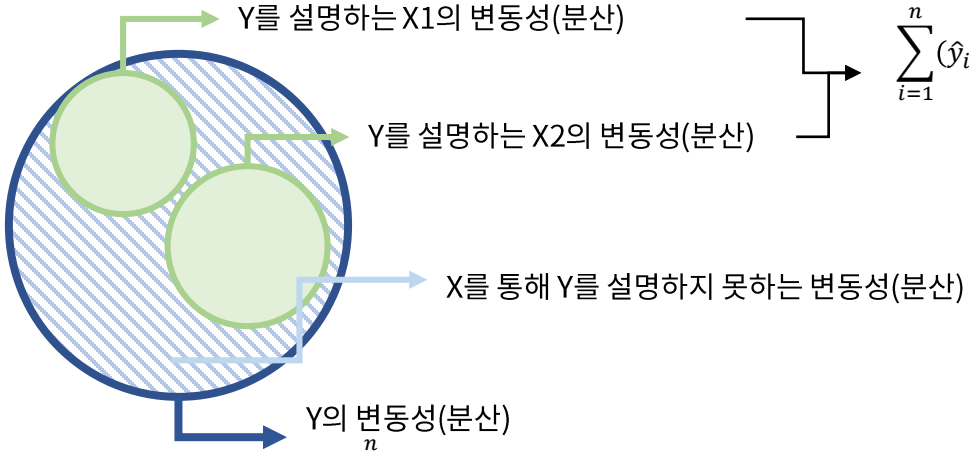
I FeatureSelection

- 다중 선형 회귀 모델 검정

귀무가설 : $B_1 = B_2 \ldots B_p = 0$ (모든 회귀계수는 0이다, 즉 변수의 설명력이 하나도 존재 하지 않는다)

대립가설 : 하나의 회귀계수라도 0이 아니다. (즉 설명력이 있는 변수가 존재 한다.)

➡ 기각 하기 너무 쉬운 가설. 변수가 추가 되면 추가 될수록 기각하기 쉬워진다.



$\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$

② 제곱합의 형태 이기 때문에 변수가 추가되면 추가될수록 자연스레 증가한다.

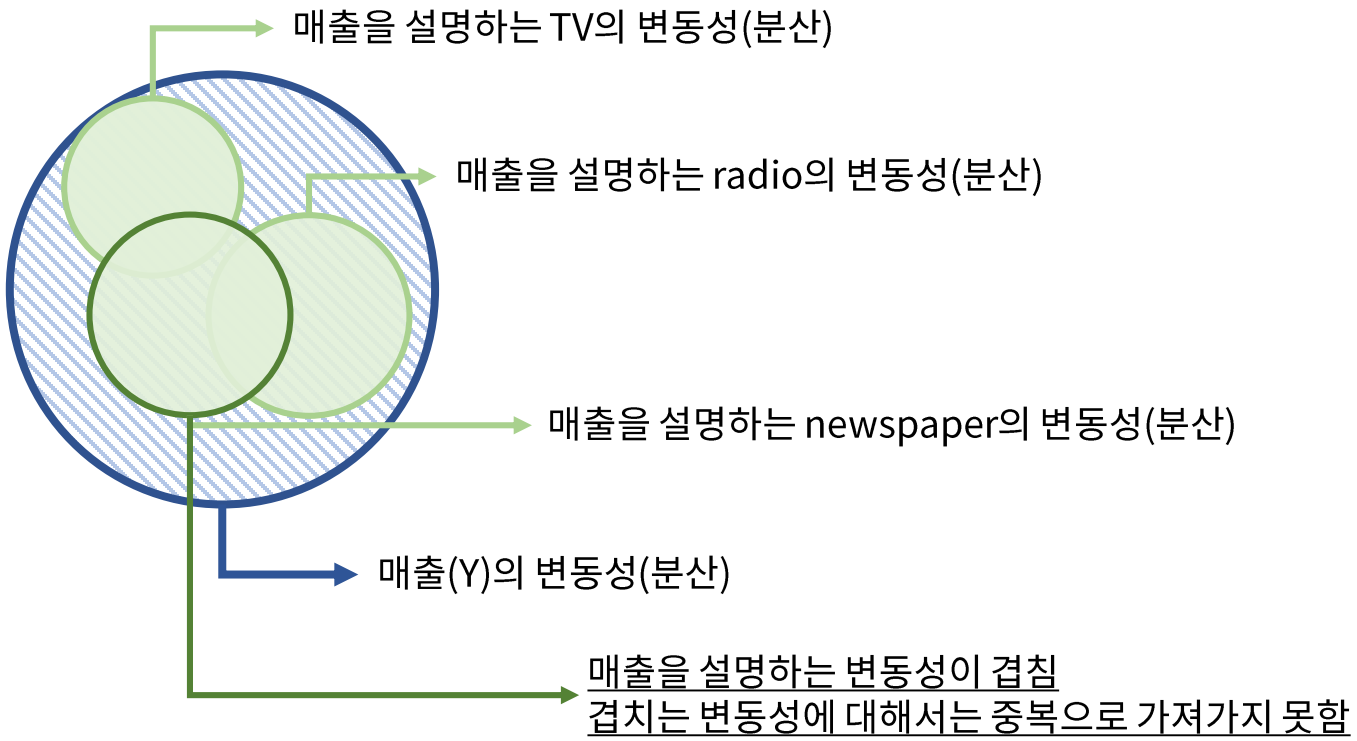
③ 제곱합의 형태로 검정을 하는 F검정의 특성상 변수가 추가되면 자연스레 기각하기 쉬워진다. (R^2도 커짐)

$\sum_{i=1}^{n}(y_i - \bar{y})^2$

① ➡ 총 변동은 정해져 있다 (바뀌지 않는다)

I FeatureSelection

- 다중공선성(Multicollinearity)



이러한 현상에 대해서 변수들간의 다중공선성(Multicollinearity)이 있다고 한다.

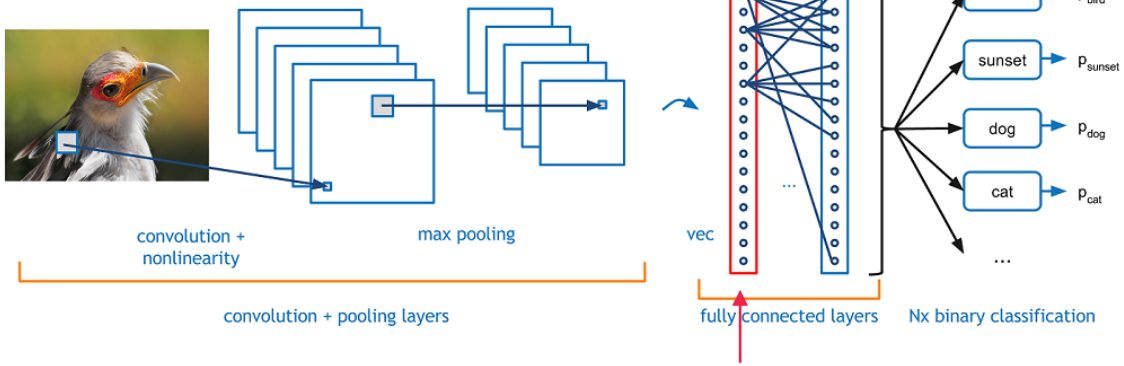
잘못된 변수해석, 예측 정학도 하락 등을 야기시킨다

I FeatureSelection

- CNN vs NN

  CNN의 해주는 역할은 이미지의 feature를 잘 뽑기 위한 것 뿐

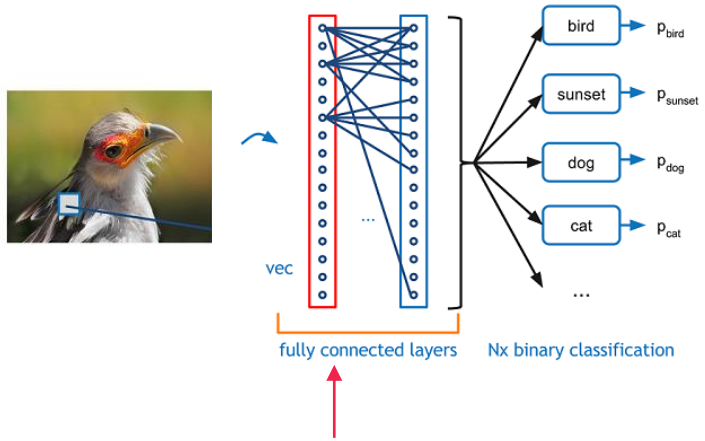
CNN



Convolution을 통해 뽑은 feature

>

NN



이미지 pixel feature

I FeatureSelection

최근의 모델(boosting)들은 중요변수를 추출해주는 알고리즘이 내장되어 있음

① 모든 변수 10,000개 vs 중요 변수 100개

② 모든 변수 10,000개 - > 모델 -> 중요변수 추출(100개) -> 모델 재학습 (성능하락가능성)

- 모델의 알고리즘이 관여치 않은 상태에서 가장 중요한 변수를 넣는것이 매우 중요!!

I FeatureSelection

- 모델 선택(변수선택)

  - 변수가 여러 개 일 때 최적의 변수 조합을 찾아내는 기법

  - 변수의 수가 p개일 때 변수의 총 조합은 2^P으로 변수 수가 증가함에 따라 변수 조합의 수는 기하급수적으로 증가

  - 총 변수들의 조합 중 최적의 조합을 찾기 위한 차선의 방법

    (optimal은 아님, optimal한 조합을 찾는 방법은 모든 경우의 수 조합을 다 해보는 것)

    1) Feedforward Selection 방법

    2) Backward Elimination 방법

    3) Stepwise 방법

I FeatureSelection

계수축소법의 종류

- 계수축소법은 기본적으로 다중선형회귀와 유사
- 다중선형회귀에서 잔차를 최소화했다면, 계수축소법에서는 잔차와 회귀계수를 최소화
- 계수축소법에는 크게 3 가지의 방법이 있음: Ridge 회귀, Lasso 회귀, Elastic-Net 회귀
- 아래 식은 다중선형회귀의 SSE이며, 다중선형회귀에서는 RSS가 최소화되는 회귀계수를 추정

$$minimize\ SSE = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2$$

- 계수축소법에서는 위 식에 회귀계수를 축소하는 항을 추가

$$minimize\ SSE + f(\beta)$$

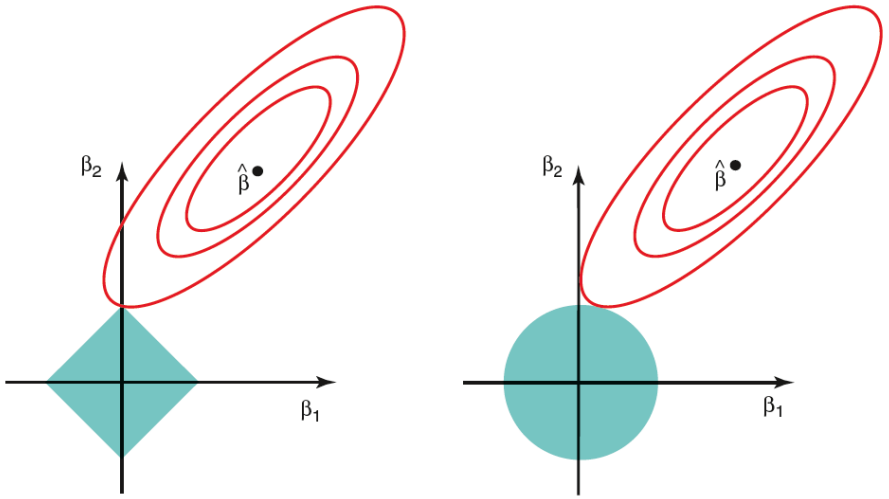
I FeatureSelection

Ridge 회귀와 Lasso의 차이점

- Ridge 회귀와 Lasso의 가장 큰 차이점은 Ridge는 계수를 축소하되 0에 가까운 수로 축소하는 반면, Lasso는 계수를 완전히 0으로 축소함
- Ridge 회귀: 입력 변수들이 전반적으로 비슷한 수준으로 출력 변수에 영향을 미치는 경우에 사용
- Lasso 회귀: 출력 변수에 미치는 입력 변수의 영향력 편차가 큰 경우에 사용

- 초록색 그림: 회귀계수가 가질 수 있는 영역(feasible region)
- 빨간색 원: SSE가 같은 지점을 연결한 그림(가운데로 갈수록 오차가 작아짐)
- Lasso 회귀의 경우 회귀계수가 0이 될 수 있지만, Ridge 회귀는 불가능

〈Lasso 회귀〉 〈Ridge 회귀〉



James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning*. New York: Springer.

I FeatureSelection

Elastic-Net 회귀

- ElasticNet 회귀는 Lasso 와 Ridge 회귀의 하이브리드(정규화) 회귀 모델
- Lasso에 적용된 회귀계수의 절대값의 합과 Ridge에 적용된 회귀계수의 제곱의 합을 모두 $f(\beta)$ 에 대입

$$minimize \sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2 + \lambda_1 \sum_{j=1}^{p}\beta_j^2 + \lambda_2 \sum_{j=1}^{p}|\beta_j|$$

- λ_1, λ_2는 Ridge와 Lasso 속성에 대한 강도를 조절
- Lasso의 변수 축소 효과로 sparse model 생성
- Ridge의 정규화 속성으로 변수의 grouping effect 유도 및 Lasso 의 sparsity를 안정화

  - Grouping effect : Lasso는 상관관계가 있는 다수의 변수들 중 하나를 무작위로 선택하여 계수를 축소하는 반면, elastic-net은 상관성이 높은 다수의 변수들을 모두 선택

- 따라서, 다수의 변수 간에 상관관계가 존재할 때 효과적

Part. 02
회귀분석

| 차원축소법

FASTCAMPUS
ONLINE
머신러닝과 데이터분석 A-Z|

강사. 이경택