Ridge & Lasso regression

3조 김해인 이윤정 정유정 현윤후

Ridge & Lasso regression

INDEX

1. Ridge & Lasso regression 개념

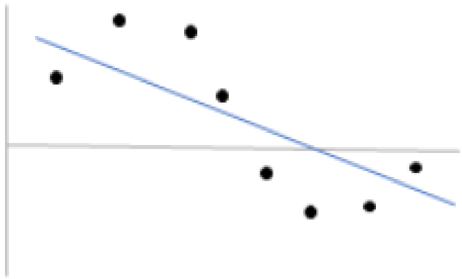
Ridge & Lasso 회귀를 왜 이용하는가?

단순/다중 선형회귀 모델의 문제점을 해결하기 위해

문제점

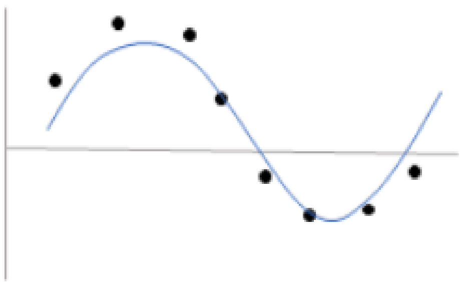
1. 과적합 (overfitting) 문제
2. 과소적합 (underfitted) 문제

과적합 & 과소적합 개념



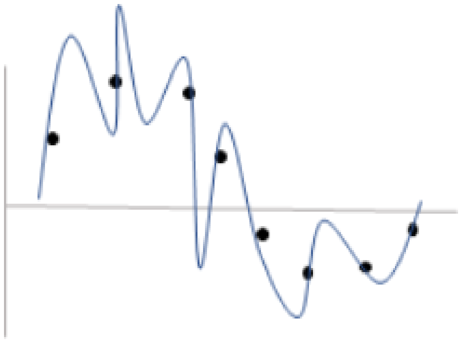
“Underfitting”

High Bias
Low Variance



“best fitting”

Middle Bias
Middle Variance



“overfitting”

Low Bias
High Variance

» 과적합, 과소적합 문제는 bias, variance와 관련되어 있음.

Bias & Variance

$$Bias = (E[f^{pred}(x)] - f(x))^2$$

Bias

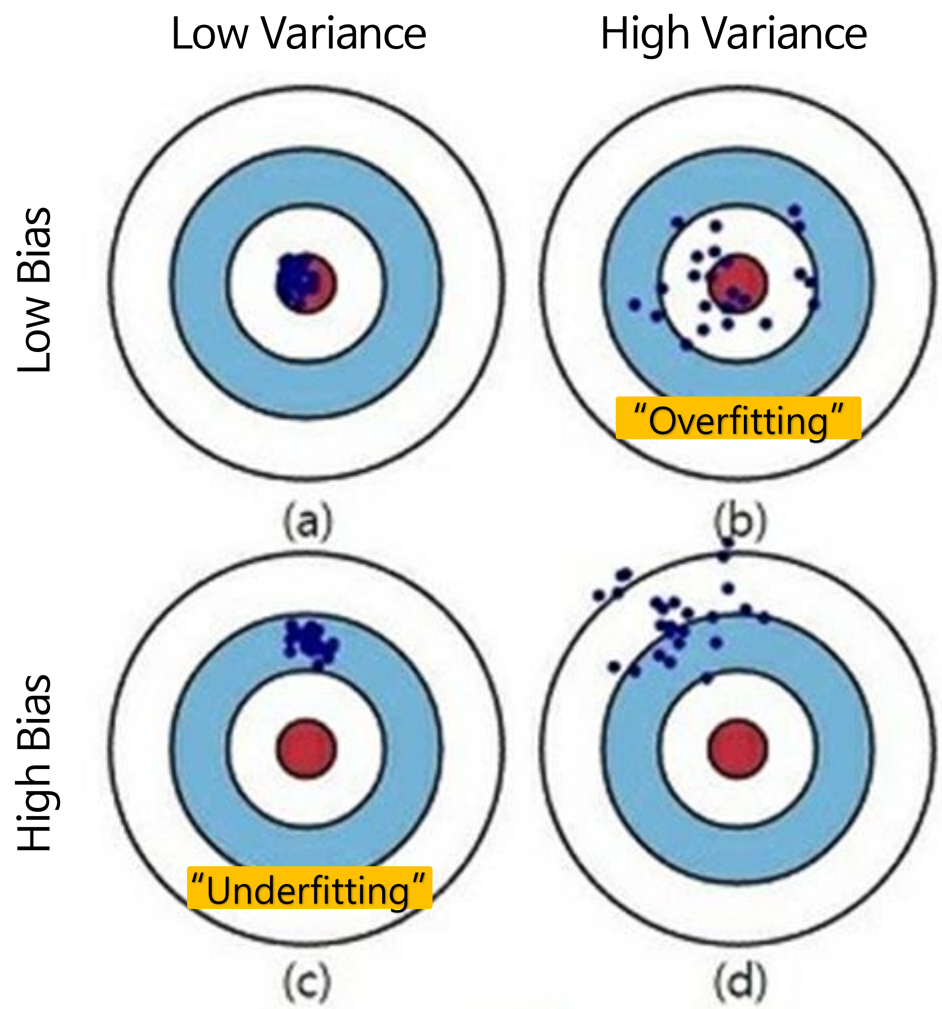
- '편향성'에 대한 개념
- 데이터 내의 정보를 충분히 고려하지 않아서 발생하는 문제점
- Bias가 높을 수록 머신러닝에서 편향된 정보를 학습하는 경향이 있음
- 예측 값과 실제 값의 차이가 클수록 '편향이 높다.'

$$Variance = E[f^{pred}(x) - E[f^{pred}(x)]]^2$$

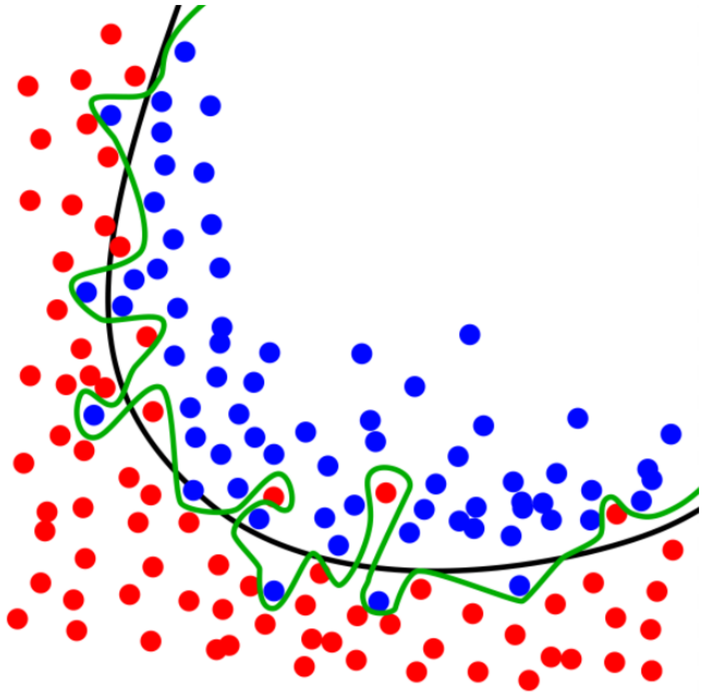
Varience

- '편차'에 대한 개념
- 데이터 내의 정보를 과도하게 고려하여 발생하는 문제점
- Varience가 높을 수록 머신러닝에서 ramdom하게 일어나는 데이터까지 학습하는 경향이 있음
- 예측 값과 예측 값의 평균의 차이가 클수록 '편차가 크다.'

Bias & Variance



Overfitting



Overfitting(과적합)

- Variance가 높은 경우 발생
- 학습모델이 trainset의 noise까지 학습하여 trainset에서는 정확도가 높지만 testset에서는 정확도가 낮은 문제가 발생함
- 과적합은 데이터의 요소를 많이 고려할수록 발생하기 쉬움
- 과적합을 방지하는 방법으로는 요소 수를 줄이는 방법과 정규화 방법이 있음

Overfitting 해결법

1. Parameter 수를 줄인다.

주요 특징을 직접 선택하고, 나머지는 버린다.

Model Selection Algorithm 사용

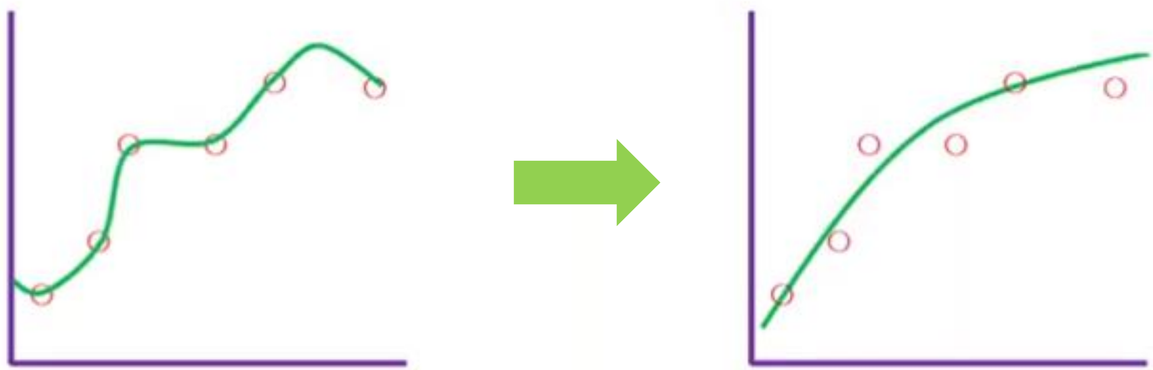
2. 정규화를 수행한다.

선형회귀 계수(weight)에 제약조건을 추가한다.

1. Ridge Regression Model
2. Lasso Regression Model
3. Elastic Net Model

정규화

Overfitting model General model



$\beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \beta_4 x^4$ $\beta_0 + \beta_1 x + \beta_2 x^2$

현재 데이터에 대한 예측력도 중요하지만, 미래에 예측할 데이터도 중요하므로 일반화 필요

⇒ 제약조건 추가

▶▶ 학습 데이터에 대한 설명력을 다소 포기 + 미래 데이터 변화에 상대적으로 안정적인 결과 도출

정규화

일반화 식

페널티 항

$$L(\beta) = \min_{\beta} \underbrace{\sum_{i=1} (y_i - \hat{y}_i)^2}_{\text{(1) Training accuracy}} + \underbrace{\lambda \sum_{j=1}^{p} \beta_j^2}_{\text{(2) Generalization accuracy}}$$

$$\beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \beta_4 x^4 \quad \Rightarrow \quad \beta_0 + \beta_1 x + \beta_2 x^2$$

(1) Training accuracy : 최소제곱법

(2) Generalization accuracy : 베타 값에 제약 → 정규화

정규화를 통해 계수 추정지를 줄여주는 정규화 방법을 shrinkage method 라고 함.

정규화

정규화 — L1 정규화 / L2 정규화

Ridge 회귀는 L2정규화, Lasso 회귀는 L1 정규화 이용.

- L1 정규화 : 벡터 p, q의 각 원소들의 차의 절댓값의 합

$$\|\mathbf{p}-\mathbf{q}\|_1 = \sum_{i=1}^{n} |p_i - q_i|$$

→ Lasso 페널티 항 $|\beta|$ 꼴로 나타남

- L2 정규화 : 벡터 p, q의 유클리드 거리(직선거리)

$$\|\boldsymbol{x}\|_2 := \sqrt{x_1^2 + \cdots + x_n^2}$$

→ Ridge 페널티 항 β^2 꼴로 나타남

정규화

(3) λ 역할

큰 λ 값

적은 변수
간단한 모델
해석 쉬움
높은 학습오차(Underfitting 위험 증가)

작은 λ 값

많은 변수

복잡한 모델

해석 어려움

낮은 학습오차(Overfitting 위험 증가)

2. 알고리즘

Ridge regression (L2 Regression)

ridge regression? 모형의 설명력에 기여하지 못하는 독립변수의 회귀계수 크기를 0에 근접하도록 축소하여 회귀계수가 과다 추정 되는 것을 방지하는 방법.

페널티 항

$$\sum_{i=1}^{n}\left(y_i-\beta_0-\sum_{j=1}^{p}\beta_j x_{ij}\right)^2+\lambda\sum_{j=1}^{p}\beta_j^2=\mathrm{RSS}+\lambda\sum_{j=1}^{p}\beta_j^2$$

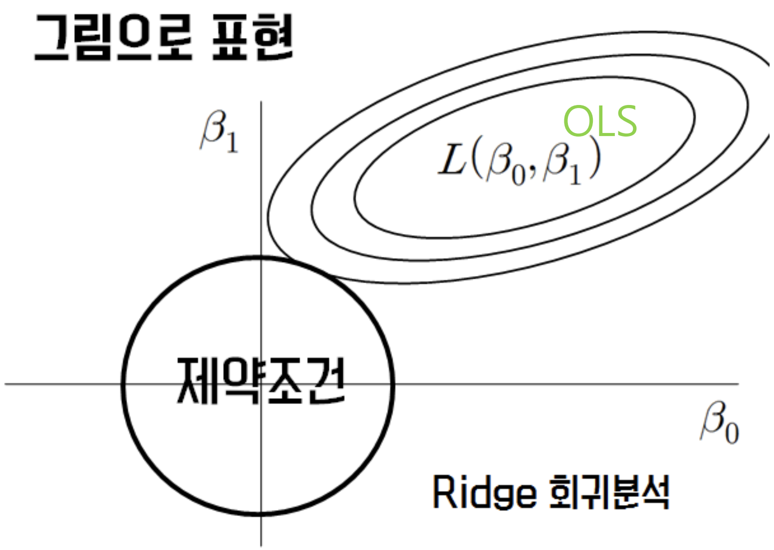
잔차제곱합(RSS : residual sum of squares)+페널티 항

λ에 따라 페널티를 얼마나 부과할지 조절

λ가 0에 가까워지면 → 페널티항 효과 X
Linear Regression에 가까워 짐

λ가 커지면 → β^2의 크기↓ 0에 가까워짐
모델의 복잡도, 다중공선성 문제의 영향↓

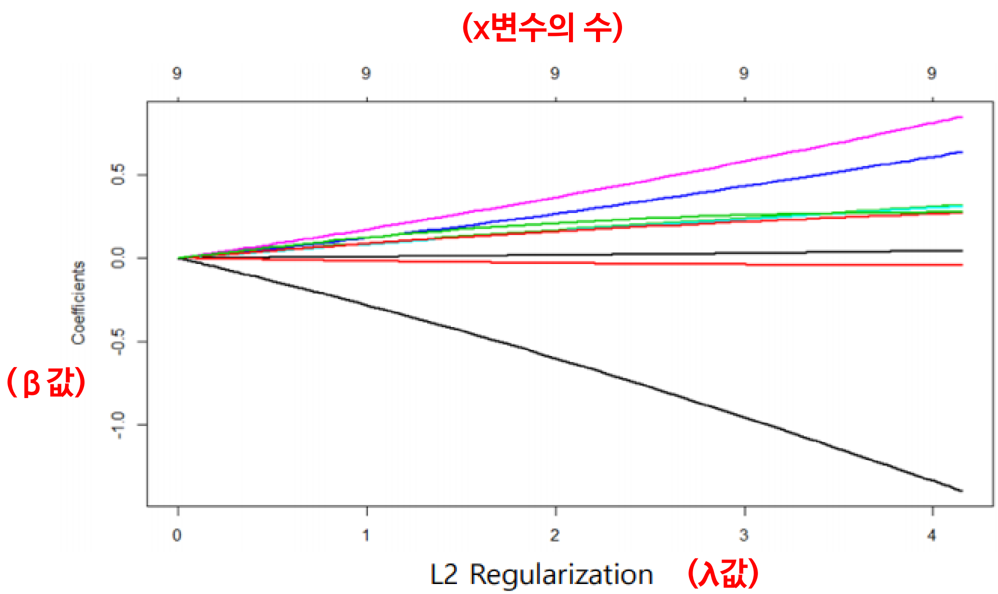
Ridge regression (L2 Regression)



- 제약조건: 페널티항에 따라 $\beta_0^2 + \beta_1^2$인 원
- 기존의 OLS (Ordinary Least Squares)가 제약조건에 다다랐을 때 **최적값**
- OLS가 제약조건까지 커지면 RSS도 증가↑

  제약조건까지 오는 가장 작은 RSS를 고르면 variance를 최소화

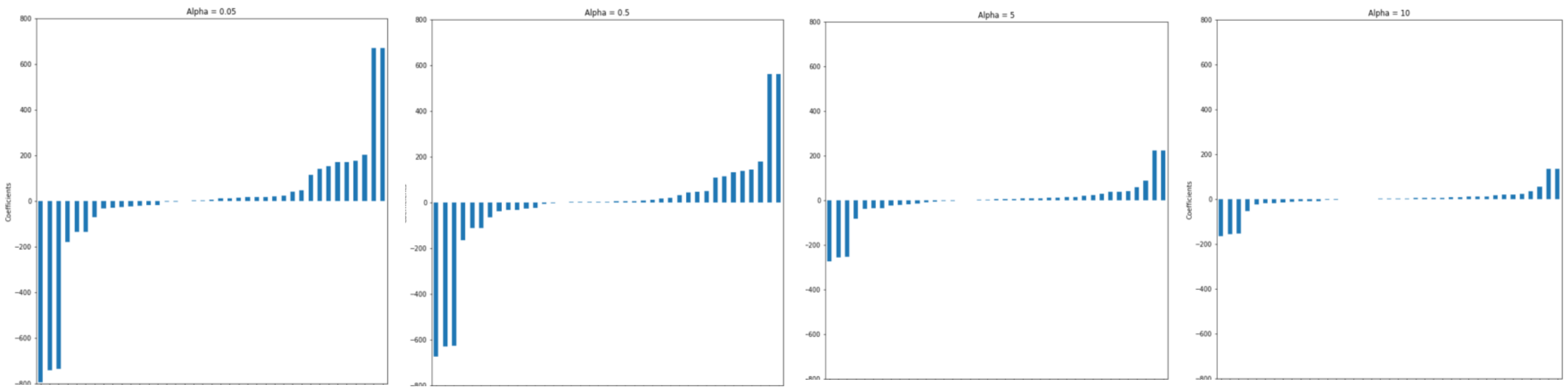
Ridge regression (L2 Regression)



- y축 : coefficient (β 값)
- 아래 x축 : λ / 위 x축 : x변수의 수
- λ값이 0이 됐을 때 모든 coefficient(β 값)이 동시에 0이 됨.
- Ridge는 coefficient shrinkage와 모델링을 동시에 수행.

Ridge regression (L2 Regression)

- coefficient shrinkage (회귀계수 값 축소)



λ=0.05 λ=0.5 λ=5 λ=10

λ가 증가할수록 회귀계수가 작아짐

Lasso regression (L1 Regression)

페널티 항

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

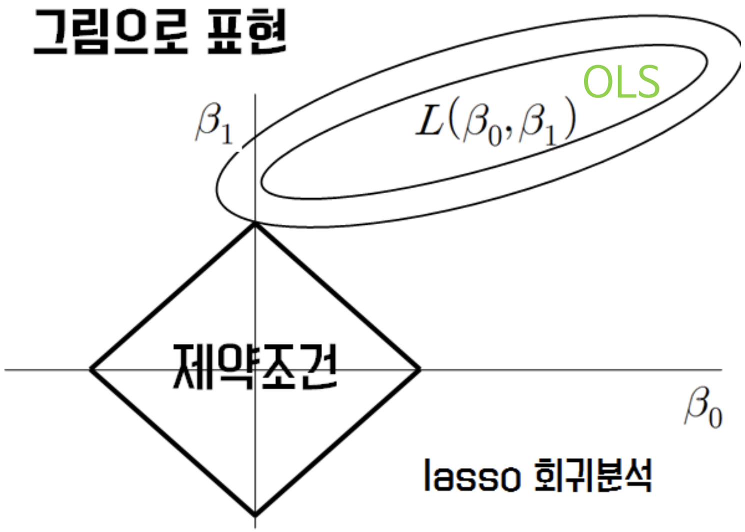
- Ridge Regression과 형태 비슷함. (RSS : residual sum of squares, 잔차제곱합 + 페널티항)
- 페널티항 | β | 형태로 나타남. → 제약조건이 마름모꼴 형태로 나타남.

λ에 따라 페널티를 얼마나 부과할지 조절

λ가 0에 가까워지면 → 페널티항 효과 X
Linear Regression에 가까워 짐

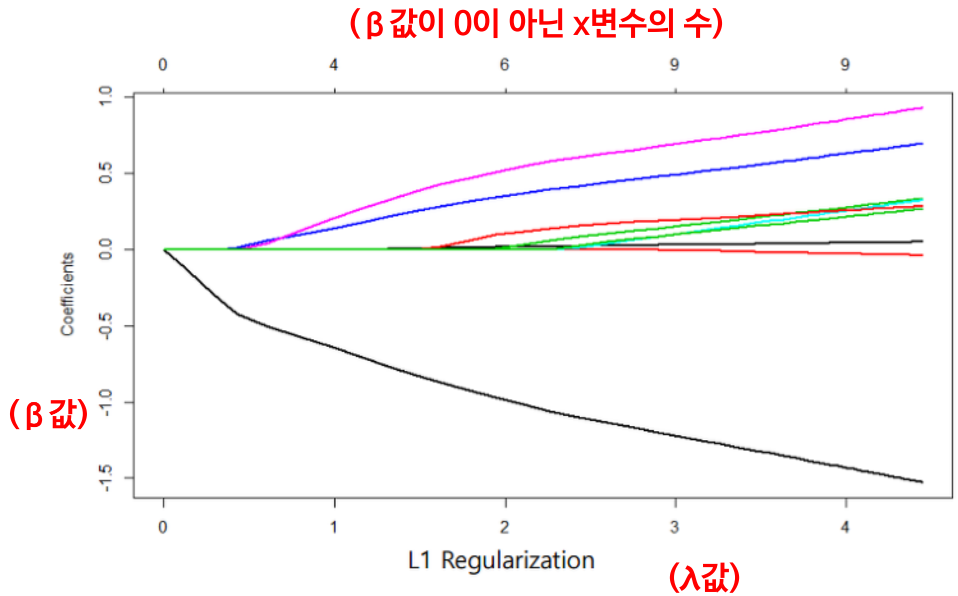
λ가 커지면 → $|\beta|$크기 ↓ 0에 가까워짐
모델의 복잡도, 다중공선성 문제의 영향 ↓

Lasso regression(L1 Regression)



- 제약조건: 페널티항($|\beta|$)에 따라 마름모 형태 → 최적값이 마름모 모서리에 나타날 확률이 높음
- OLS(Ordinary Least Squares)가 제약조건에 다다랐을 때 **최적값**
- 유의미하지 않은 변수들의 계수를 0으로 만들어 변수를 모델에서 삭제 → 모델 단순화
- Ridge Regression은 변수들의 계수(β값)을 0으로 만들지 않고,
  0에 가깝게 하여 상관성을 가지는 변수들에 대해서 적절한 가중치를 배분. **(Lasso Regression과의 차이점)**

Lasso regression(L1 Regression)



- y축 : coefficient (β값)
- 아래 x축 : λ / 위 x축 : β값이 0이 아닌 x변수의 수
- Ridge Regression에서는 λ값이 0이 됐을 때 동시에 coefficient(β값)이 0이 되지만,
  Lasso Regression에서는 λ값이 0이 되기 전에 몇몇 변수들의 coefficient(β값)이 0이 됨.
  (이를 변수 선택에 이용)
- Lasso는 변수 선택과 coefficient shrinkage를 모델링하는 과정에서 동시에 수행.

샘플에 비해 변수가 너무 많을 때,
고전적인 방법 VS Ridge/Lasso Regression

고전적인 방법 : Feature selection step과 modeling step을 분리하여 독립적으로 수행

Ridge, Lasso 기법 : 모델링 과정에서 페널티를 줌으로써 coefficient의 값을 낮춰,
Bias값에 대해 손해를 보는 대신, Variance를 줄이기 위한 노력을 수행
(overfitting 문제 해결)

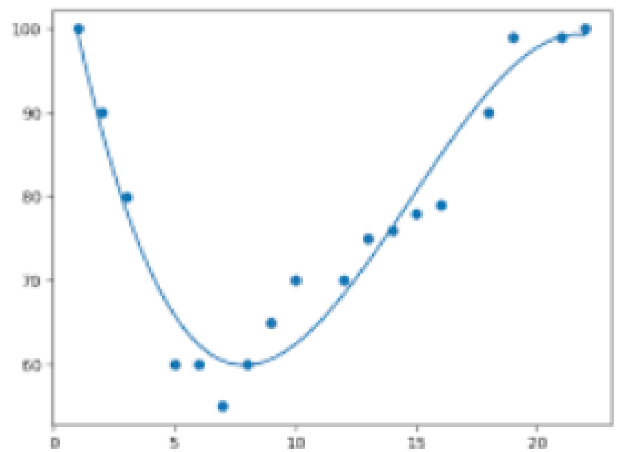
Ridge & Lasso 모형 비교

구분	Ridge	Lasso
제약식	L_2 norm	L_1 norm
변수 선택	불가능	가능
Solution	Closed Form	명시해 없음
장점	변수간 상관관계가 높아도 좋은 성능	변수간 상관관계가 높으면 성능↓
특징	크기가 큰 변수를 우선적으로 줄임	비중요 변수를 우선적으로 줄임

3. 실적용 사례

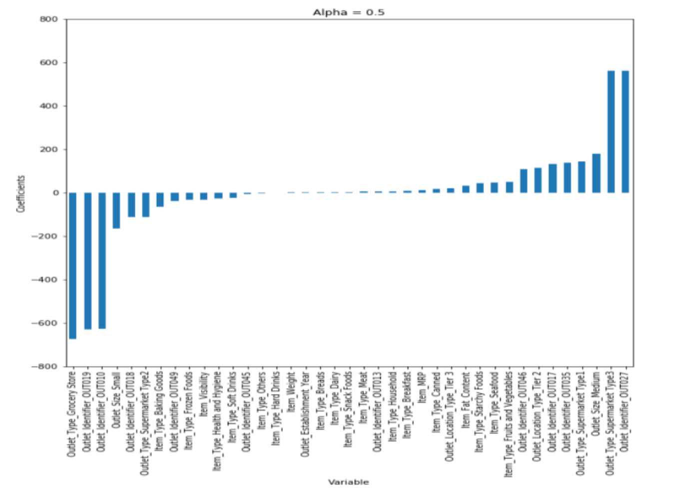
Ridge & Lasso regression 실 적용

예측 모형 연구



- 여러 개의 독립변수에 영향을 받는 하나의 종속변수에 대한 예측 가능
- 과적합 되지 않은 적절한 예측 모형 생성

유의한 설명 변수 선별



- 람다 값에 따라 독립변수의 개수나 중요도를 조절할 수 있음
- 독립변수 중에서 더 종속변수와 상관계수가 높은 중요 요인을 찾을 수 있음
  → 유의한 설명 변수를 선별함

example

- 예측 모형 연구
  - 간경변 발생 예측 모형 연구
  - 부동산 헤도닉 가격모형 연구
  - 골목상권 외식업종 점포의 월 매출액 예측 모형
- 주요 변수 선별
  - 예대 금리차 결정요인 분석
  - 입원 환자수에 영향을 미치는 날씨 변수 선택
  - 부동산 가격의 주요 설명변수 선택

4. 데이터를 통한 예제

> > Python 이동

Ridge & Lasso regression

Thank you

B.a.f | 3조