



부산대학교

PUSAN NATIONAL UNIVERSITY

Regularization

Paying to Fit Less

Bias, Variance, and Shrinking Coefficients

BX3600087-140 / DS3600087-140 / AB3600087-140

기계학습개론

2026학년도 2학기

오용경

데이터사이언스전공

yongkyung.oh@pusan.ac.kr

Section 1 • Regularization

Decomposing the Error

오차를 세 조각으로 나눈다

모델 복잡도와 예측 오차

복잡도	학습 오차	새 자료의 오차
낮음	모델이 단순해 클 수 있음	과소적합으로 클 수 있음
적절함	자료의 규칙을 반영	낮은 오차를 기대
높음	작아질 수 있음	과적합으로 커질 수 있음

전형적인 경향을 요약한 것이며 모든 자료에서 U자 곡선이 나타나는 것은 아님

편향-분산 분해(bias-variance decomposition)

기대 제곱 예측 오차는 아래 조건에서 편향의 제곱, 분산, 잡음 분산으로 나뉜다

$$\underbrace{\mathbb{E}[(y_0 - \hat{f}(x_0))^2]}_{\text{기대 제곱 예측 오차}} = \underbrace{(\mathbb{E}[\hat{f}(x_0)] - f(x_0))^2}_{\text{편향의 제곱}} + \underbrace{\text{Var}(\hat{f}(x_0))}_{\text{분산}} + \underbrace{\sigma^2}_{\text{잡음 분산(줄일 수 없는 오차)}}$$

고정된 x_0 에서 $y_0 = f(x_0) + \varepsilon$, $\mathbb{E}[\varepsilon] = 0$, $\text{Var}(\varepsilon) = \sigma^2$ 로 두며 새 잡음은 학습 자료와 독립

같은 자료 생성 과정에서 학습 자료와 새 잡음을 반복해서 뽑을 때의 평균 오차다

편향과 분산

편향과 분산은 학습 자료를 바꿔 가며 얻은 예측의 기댓값으로 정의한다

DEFINITION

편향(bias)

학습 자료를 바꾸며 얻은 예측의 평균이
참함수값과 다른 정도

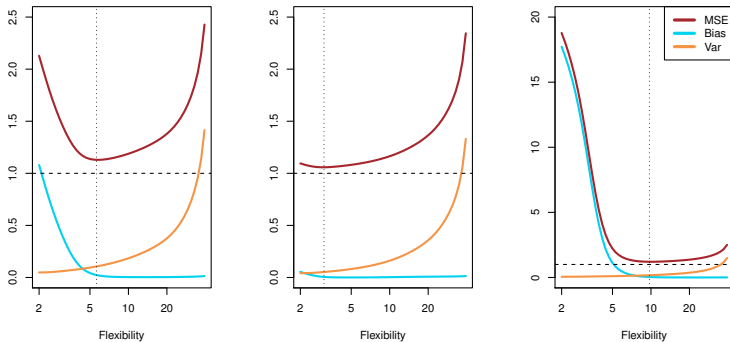
DEFINITION

분산(variance)

학습 자료를 바꿀 때 예측이 평균 예측 주위에서
달라지는 정도

같은 입력에서 평균 예측의 오차와 예측의 변동을 구분한다

복잡도에 따른 편향·분산·오차



파랑은 편향의 제곱, 주황은 분산, 붉은 곡선은 두 값에 점선의 잡음 분산을 더한 테스트 MSE

Source: James et al. (2023), Figure 2.12.

학습·검증 오차는 진단의 단서다

관찰	가능한 원인	함께 확인할 것
두 오차가 모두 큼	과소적합, 특성 부족, 최적화 실패, 잡음	모델 표현력, 자료, 학습 진행
학습보다 검증 오차가 큼	과적합 또는 자료 분포 차이	복잡도와 분할 방식
두 오차가 모두 작고 비슷함	후보가 유용할 가능성	선택 뒤 테스트 성능

분해의 각 항을 직접 구할 수 있는 경우

상황	가능한 분석
실제 자료	참함수 f 와 반복 표본을 몰라 편향·분산을 직접 구하기 어려움
생성 과정을 아는 모의실험	f 를 고정하고 학습 자료를 반복 생성해 각 항을 추정할 수 있음

분해는 정확한 수식이지만 실제 자료에서는 각 항을 바로 관측할 수 없다

Checkpoint

Decomposing the Error

- 세 조각 중 우리가 줄일 수 없는 것은 무엇인가
- 편향과 분산은 각각 무엇을 재는가
- 복잡도를 늘리면 편향과 분산에 어떤 변화가 생길 수 있는가

Section 2 • Regularization

Ridge Regression

계수의 크기에 벌점을 매긴다

정규화로 과적합을 줄인다

DEFINITION

정규화(regularization)

모델의 복잡도에 제약이나 벌점을 주어 새 자료에서의 예측 오차를 줄이려는 방법

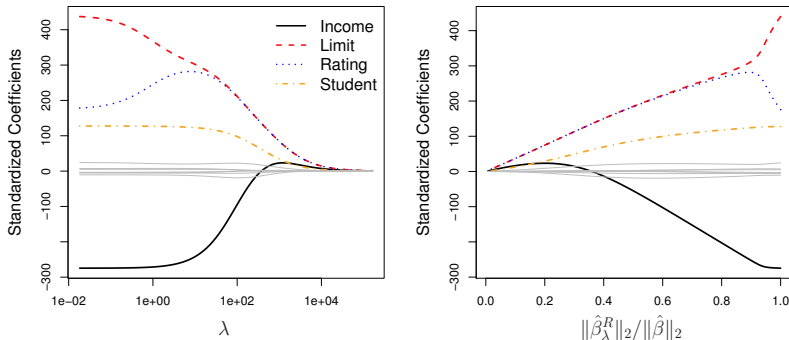
ridge regression과 lasso에서는 계수의 크기에 벌점을 준다

제곱합에 벌점을 더한다

$$\underbrace{\sum_{i=1}^n \left(y_i - \beta_0 - \sum_j \beta_j x_{ij} \right)^2}_{\text{잔차제곱합 RSS}} + \underbrace{\lambda \sum_{j=1}^p \beta_j^2}_{\text{계수 크기에 대한 벌점}}$$

ridge regression은 $\lambda \geq 0$ 을 고정한 뒤 이 목적함수를 최소화하는 계수를 구한다

정규화 강도와 계수의 변화



Credit 자료. $\hat{\beta}$ 는 벌점 없는 최소제곱 계수, $\|\beta\|_2 = \sqrt{\sum_j \beta_j^2}$ (벡터의 크기)이다. 왼쪽은 λ , 오른쪽은 최소제곱 대비 계수 벡터 크기의 비에 따른 경로. 오른쪽 가로축은 λ 가 커질수록 0 쪽이라 왼쪽과 방향이 반대. 일반적으로 계수를 정확히 0으로 만드는 방법은 아님

Source: James et al. (2023), Figure 6.4.

절편은 벌점에서 제외한다

매개변수	벌점	이 설정의 의미
특성의 계수 β_j	$\lambda\beta_j^2$	특성에 따른 예측 변화의 크기를 제한
절편 β_0	없음	y 의 평균 수준을 0 쪽으로 축소하지 않음

ridge regression과 lasso의 목적함수는 절편에 벌점을 주지 않으며, scikit-learn의 Ridge와 Lasso도 같다

같은 정보도 단위에 따라 벌점이 달라진다

단위와 특성값 x	계수 β	βx	β^2
$x = 5,000,000$ (단위: 원)	2×10^{-5}	100	4×10^{-10}
$x = 500$ (단위: 만원)	0.2	100	0.04

Caution 같은 예측도 단위만 바꾸면 벌점이 달라지므로 학습 자료 기준으로 특성을 표준화한다

정규화 강도의 양 끝

강도	계수와 예측
$\lambda = 0$	벌점이 없는 최소제곱 목적함수
$0 < \lambda < \infty$	자료 적합과 계수 크기를 함께 고려
$\lambda \rightarrow \infty$	계수는 0으로 수렴하고 절편은 학습 자료 y 의 평균으로 수렴

크지만 유한한 λ 에서 계수가 모두 정확히 0이 된다는 뜻은 아니다

Checkpoint

Ridge Regression

- 목적함수에 더해지는 항은 무엇을 재는가
- 절편에 벌점을 매기지 않는 까닭은 무엇인가
- 정규화 전에 척도를 맞춰야 하는 까닭은 무엇인가

Why Shrinking Helps

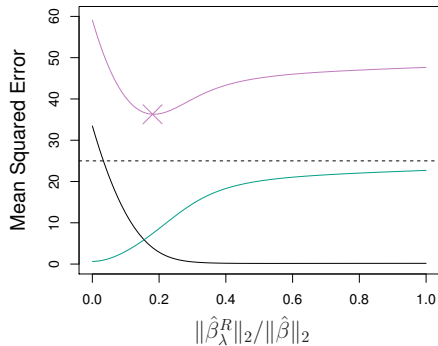
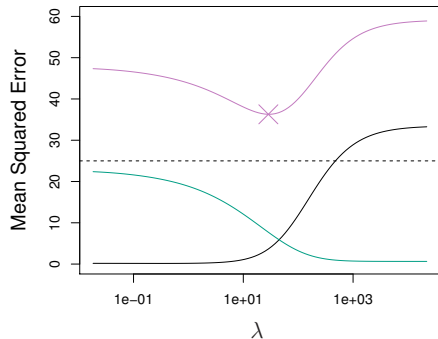
학습 자료에 덜 맞추어도 예측 오차가 줄어드는 조건을 본다

분산 감소가 편향 증가보다 크면 이득이다

정규화의 영향	기대할 수 있는 변화
계수 크기에 벌점	학습 자료에 대한 적합을 제한
자료 변화에 대한 민감도 감소	분산 감소 가능
참함수와 평균 예측의 차이	편향의 제곱 증가 가능

총 예측 오차가 줄어드는지는 두 변화의 크기에 달려 있음

λ 에 따른 편향·분산·테스트 MSE



검정은 편향의 제곱, 초록은 분산, 보라는 잡음까지 포함한 테스트 MSE. 점선은 잡음 분산, 십자 표시는 최소 MSE. 오른쪽 가로축은 λ 가 커질수록 0 쪽이라 왼쪽과 방향이 반대

Source: James et al. (2023), Figure 6.5.

계수 추정이 불안정한 상황

상황	최소제곱에서의 문제	ridge regression
특성 수 p 가 관측 수 n 보다 큼	최소제곱 해가 유일하지 않음	$\lambda > 0$ 이면 해가 유일함
특성 간 상관이 큼	자료가 바뀌면 계수가 크게 달라질 수 있음	계수 추정의 불안정성을 완화
충분한 자료와 약한 상관	이미 추정이 안정적일 수 있음	개선이 작거나 오히려 손해일 수 있음

잡음 분산은 기대 예측 오차의 하한이다

$$\mathbb{E}[(y_0 - \hat{f}(x_0))^2] = \text{Bias}^2 + \text{Var} + \sigma^2 \geq \sigma^2$$

같은 입력과 자료 생성 과정에서 편향의 제곱과 분산은 음수가 될 수 없음

Caution 이 하한은 기대 오차에 대한 것이며 특정 테스트 자료에서 계산한 MSE의 하한은 아니다

Checkpoint

Why Shrinking Helps

- 계수를 축소하면 어떤 조건에서 기대 예측 오차가 줄어드는가
- 정규화가 특히 이득인 자료는 어떤 모양인가
- 잡음 분산은 어떤 오차의 하한인가

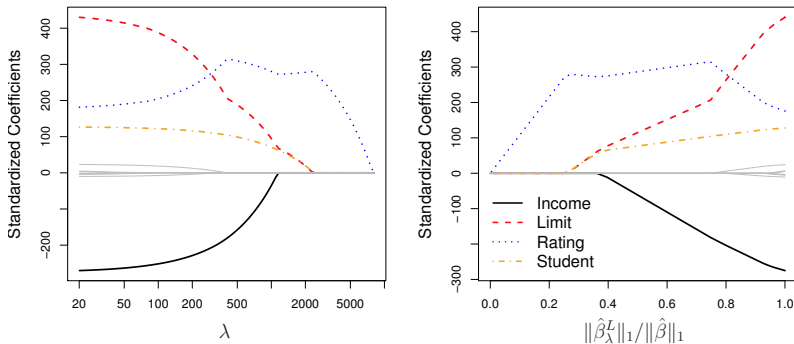
The Lasso

벌점의 모양을 바꾸면 계수가 정확히 0이 될 수 있다

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

lasso는 계수 절댓값의 합에 벌점을 주어 일부 계수를 정확히 0으로 만들 수 있다

lasso의 계수 경로



$\|\beta\|_1 = \sum_j |\beta_j|$. 왼쪽은 λ , 오른쪽은 최소제곱 대비 계수 절댓값 합의 비에 따른 경로. 오른쪽 가로축은 λ 가 커질수록 0 쪽이라 왼쪽과 방향이 반대.
일부 계수가 0이 되는 구간이 나타남

Source: James et al. (2023), Figure 6.6.

계수 축소와 특성 선택

정규화 방법	계수에 미치는 영향
ridge regression	계수를 0 쪽으로 줄이지만, 일반적으로 정확히 0으로 만들지는 않는다
lasso	강도에 따라 일부 계수를 정확히 0으로 만들 수 있다

lasso에서 계수가 0이 된 특성은 예측식에 기여하지 않는다

두 방법은 검증 성능으로 비교한다

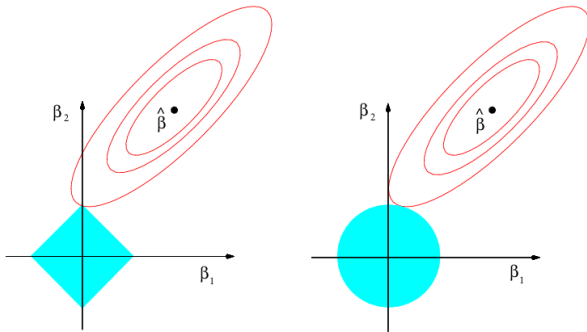
예상하는 구조	우선 검토할 방법	이유
여러 특성이 조금씩 기여	ridge regression	계수를 함께 축소
일부 특성만 예측에 기여	lasso	일부 계수를 0까지 축소 가능
구조를 알 수 없음	두 방법 모두	같은 분할과 지표로 교차검증

서로 상관된 특성이 있을 때

방법	계수 선택의 경향
ridge regression	상관된 특성들에 계수가 나뉘어 나타날 수 있음
lasso	일부 특성만 남길 수 있지만 항상 하나만 남기는 것은 아님

Caution lasso가 선택한 특성이 유일하게 중요하거나 인과적인 특성이라는 뜻은 아니다

제약 영역의 모양과 0인 계수



각 λ 마다 같은 계수를 주는 예산 s 가 있어, 벌점 목적함수는 RSS를 최소화하되 계수 크기를 s 이하로 제한하는 문제와 같다(λ 가 클수록 s 가 작다). $\hat{\beta}$ 는 최소제곱 추정값. 파란 영역은 제약 $|\beta_1| + |\beta_2| \leq s$ (왼쪽 lasso)와 $\beta_1^2 + \beta_2^2 \leq s$ (오른쪽 ridge), 붉은 타원은 RSS 등고선이다. 등고선이 축 위에 있는 lasso 제약의 모서리에 닿으면 한 계수가 0이 된다

Source: James et al. (2023), Section 6.2.2, Figure 6.7.

Checkpoint

The Lasso

- 두 벌점의 식은 무엇이 다른가
- lasso에서 계수가 정확히 0이 되면 무슨 일이 일어나는가
- 비슷한 특성이 여럿일 때 두 방법은 어떻게 다르게 행동하는가

Section 5 • Regularization

Choosing the Strength

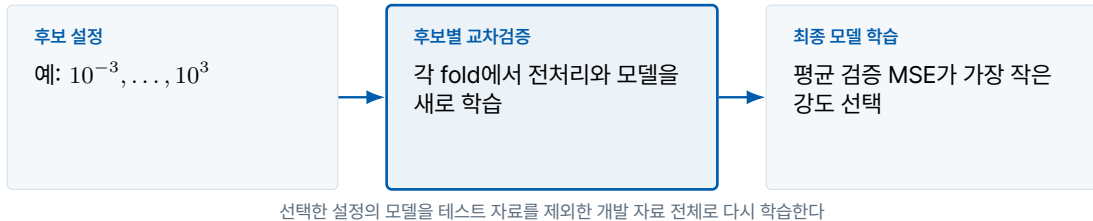
정규화 강도를 자료가 정하게 한다

계수 학습과 정규화 강도 선택

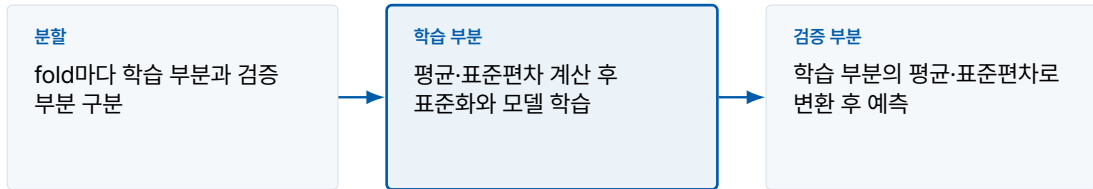
정할 값	사용하는 기준
계수	고정된 λ 에서 학습 목적함수 최소화
정규화 강도 λ	교차검증의 평균 검증 오차 비교

벌점은 0 이상이므로, 학습 목적함수를 λ 에 대해서도 최소화하면 $\lambda = 0$ 이 항상 가장 작은 값을 준다

후보마다 교차검증을 수행한다



표준화도 학습 fold 안에서 수행한다



Caution 전처리와 모델을 하나의 pipeline으로 묶어 검증 자료의 정보가 학습에 들어가지 않게 한다

코드로 정규화 강도를 고른다

Hitters 자료(선수 263명)에서 특성 19개로 연봉(천 달러)을 예측한다. 범주형 특성 3개는 0과 1로 바꿨다

```
1 import numpy as np
2 from sklearn.model_selection import GridSearchCV, KFold
3 from sklearn.pipeline import make_pipeline
4 from sklearn.preprocessing import StandardScaler
5 from sklearn.linear_model import Ridge
6
7 model = make_pipeline(StandardScaler(), Ridge())
8 grid = {"ridge__alpha": np.logspace(-3, 3, 13)}
9 cv = KFold(n_splits=5, shuffle=True, random_state=0)
10 search = GridSearchCV(model, grid, cv=cv,
11                        scoring="neg_mean_squared_error")
12 search.fit(X, y)
13 print(search.best_params_)
14 # {'ridge__alpha': 3.1622776601683795}
15 print(-search.best_score_)
16 # 115722.7986063871
```

Ridge의 alpha는 앞 목적함수의 λ 와 같다. 후보 13개마다 fold 5개에서 표준화와 ridge를 새로 학습한다. 이 예는 263명 전체를 개발 자료로 쓰며, 최종 평가까지 하려면 테스트 자료를 먼저 떼어 둔다

Adapted from James et al. (2023), Chapter 6 lab · data: James et al. (2023), Hitters.

후보별 평균 검증 MSE

alpha (λ)	평균 검증 MSE	fold 간 표준편차
0.001	122,661	18,003
0.1	120,169	15,235
3.16 (선택)	115,723	15,512
100	117,591	23,605
1000	131,993	28,678

앞 쪽 `search.cv_results_`의 후보 13개 중 5개다. 가장 작은 후보는 최소제곱 회귀(122,699)와 거의 같다. 같은 fold끼리 0.001과 3.16을 비교하면 3.16이 낮은 fold는 5개 중 2개이고, 평균 차이 약 6,900은 대부분 한 fold(약 41,000)에서 나온다. 규칙대로 3.16을 고르지만, 이 차이만으로 정규화가 일관되게 낮다고 말하지는 않는다

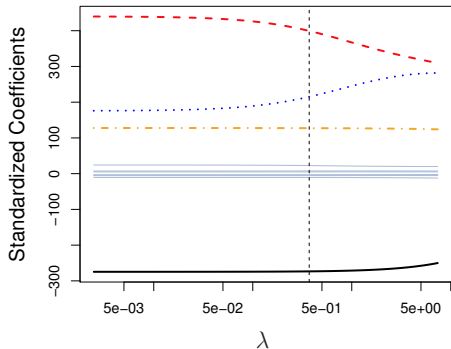
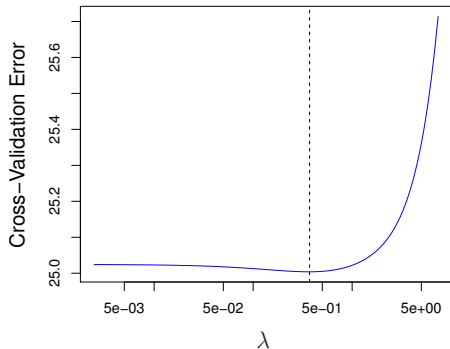
`best_score_`는 선택에 쓴 점수이므로 최종 성능은 테스트 자료로 따로 잰다

Data: James et al. (2023), Hitters.

용어	의미
편향	고정된 입력에서 평균 예측과 참함수값의 차이
분산	학습 자료를 바꿀 때 예측이 평균 주위에서 변하는 정도
ridge regression	계수 제곱의 합에 벌점을 주는 선형 회귀
lasso	계수 절댓값의 합에 벌점을 주는 선형 회귀
λ	벌점의 강도를 정하는 tuning parameter(학습 전에 정하는 하이퍼파라미터의 하나)

scikit-learn은 λ 를 alpha로 쓴다. Ridge의 alpha는 λ 와 같고, Lasso는 RSS를 $2n$ 으로 나눈 목적함수를 써서 $\alpha = \lambda / (2n)$ 이다

교차검증으로 선택한 정규화 강도



Credit 자료의 ridge regression. 왼쪽은 LOOCV 오차, 오른쪽은 계수 경로이며 수직 점선은 선택한 λ . 왼쪽 세로축 범위가 좁다. 선택한 λ 가 작고 곡선이 평평해 최소제곱과 차이가 거의 없는 경우다

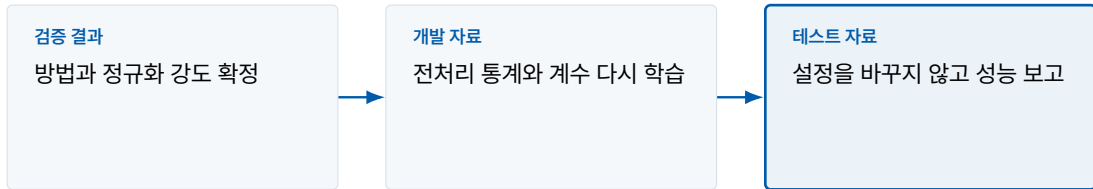
Source: James et al. (2023), Figure 6.12.

같은 선형 예측식, 다른 벌점

방법	벌점	계수의 특징
ridge regression	$\lambda \sum_{j=1}^p \beta_j^2$	일반적으로 모든 특성을 유지
lasso	$\lambda \sum_{j=1}^p \beta_j $	일부 특성의 계수가 0일 수 있음

두 방법 모두 절편은 벌점에서 제외하고 특성의 척도를 맞춘 뒤 비교

선택 이후의 최종 평가



테스트 점수를 보고 정규화 강도를 다시 고르면 독립적인 최종 평가가 아니다

선택한 모델을 해석할 때의 주의점

확인할 것	해석	주의
선택한 강도의 위치	후보 범위 안에서의 최적값	끝에 있으면 후보 범위 재검토
계수의 부호	다른 특성을 고정할 때 예측 방향	인과 효과로 해석하지 않음
계수가 0인 특성	예측식에 직접 기여하지 않음	일반적인 중요성이 없다는 뜻은 아님

Caution 후보 범위를 다시 정할 때도 테스트 자료는 사용하지 않는다

Checkpoint

Choosing the Strength

- λ 를 목적함수만으로 고를 수 없는 까닭은 무엇인가
- 표준화를 pipeline 안에 넣는 까닭은 무엇인가
- 선택한 강도가 후보 범위 끝에 있으면 무엇을 확인하는가

What If One Class Almost Never Happens?

Next Lecture: Imbalanced Data

- 높은 정확도에 가려지는 소수 범주의 오류
- 불균형 자료의 평가지표
- 임계값을 옮기고 자료와 학습을 바꾸기