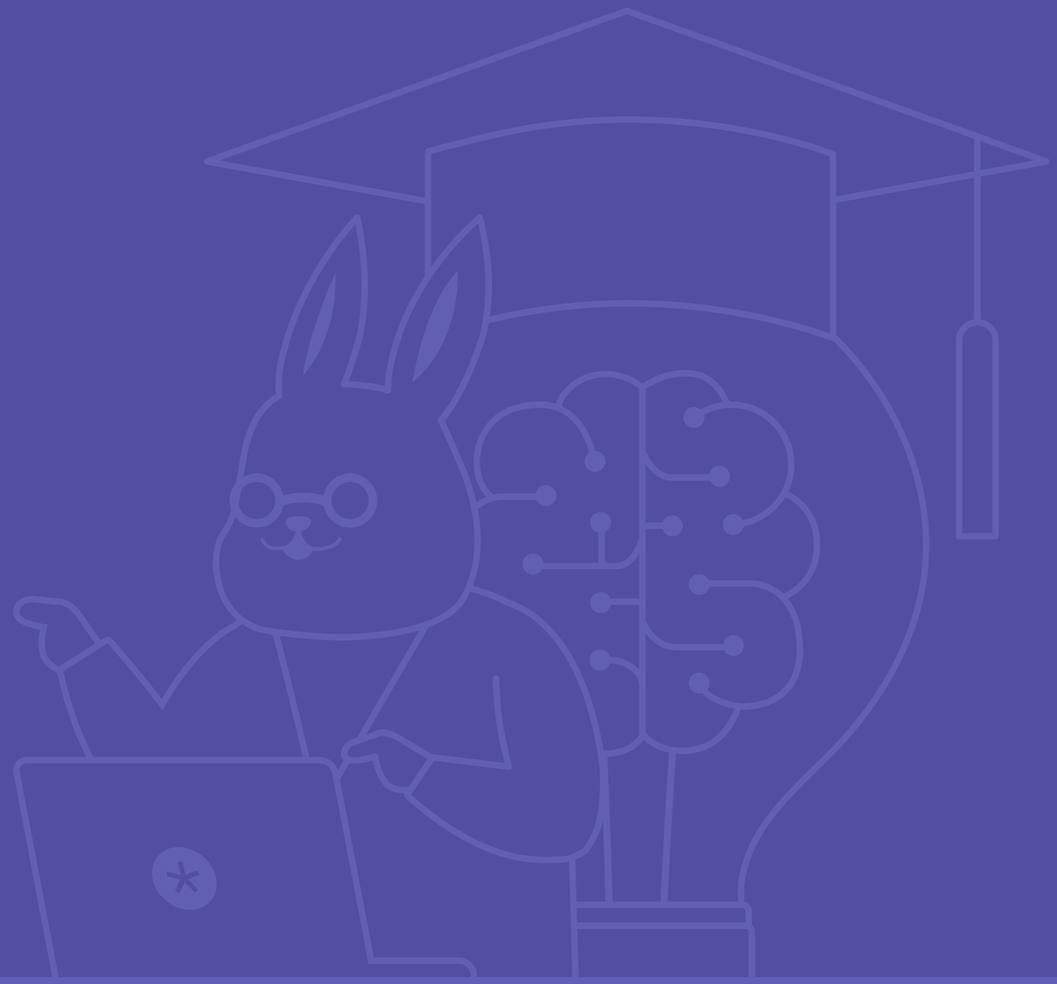


머신러닝 심화

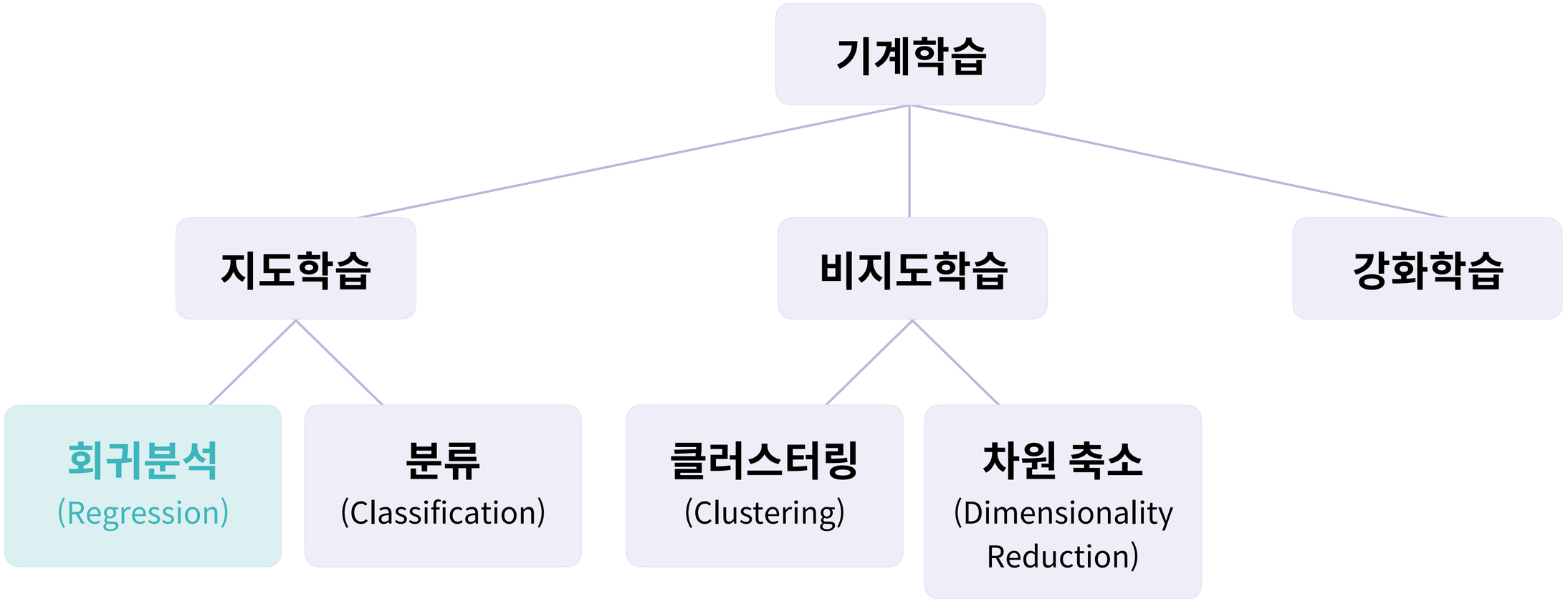
1장 회귀

Contents

01

회귀 개념 알아보기

머신러닝(기계 학습) 분야


기계학습
지도학습
비지도학습
강화학습
회귀분석
(Regression)
분류
(Classification)
클러스터링
(Clustering)
차원 축소
(Dimensionality Reduction)

가정해보기

아이스크림 가게를
운영하는 주인이라고 가정해보자.

판매용 아이스크림 주문 시,
예상되는 실제 판매량만큼만 주문을 원한다.

이 때 만약 **평균 기온**을 활용하여
미래 판매량을 예측할 수 있다면?

문제 정의와 해결 방안

- 문제 정의

데이터 : **과거 평균 기온** (X) 과 그에 따른 **아이스크림 판매량** (Y)

가정 : 평균 기온과 판매량은 선형적인 관계를 가지고 있음

목표 : 평균 기온에 따른 아이스크림 판매량 예측하기

- 해결 방안

회귀 분석 알고리즘

X	Y
평균 기온(°C)	아이스크림 판매량(만개)
10	40
13	52.3
20	60.5
25	80

01 회귀 개념 알아보기

회귀 분석이란?

데이터를 가장 **잘 설명하는 선**을 찾아

입력값에 따른 미래 결과값을 예측하는 알고리즘

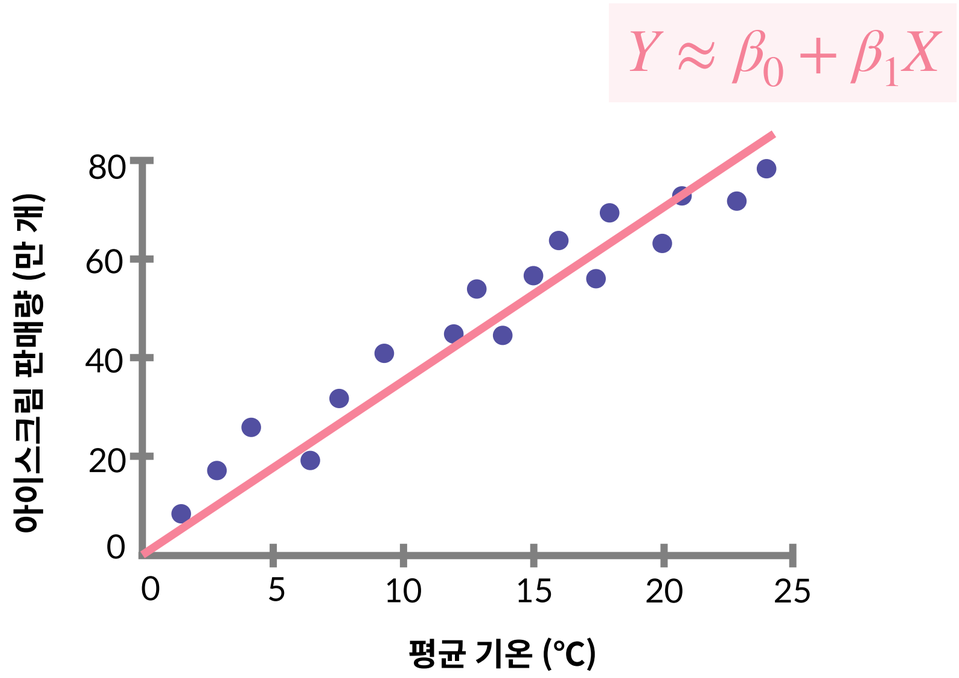
예시

$Y \approx \beta_0 + \beta_1 X$

X = 평균 기온 , Y = 아이스크림 판매량

= 적절한 β_0(y절편)과 β_1(기울기)찾기

잘 설명하는 선



적절한 β_0(y절편)과 β_1(기울기) 찾기

아이디어 : 완벽한 예측은 불가능함!

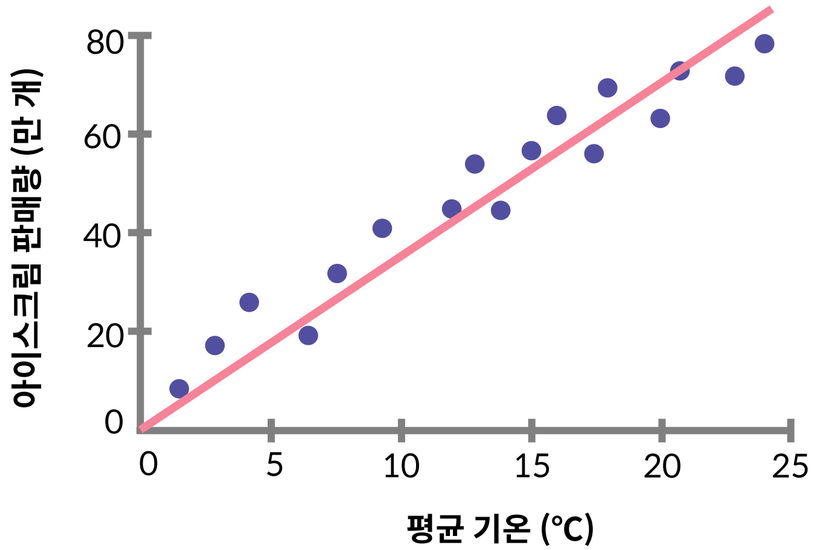
$$Y \approx \beta_0 + \beta_1 X$$

각 데이터의 실제 값과 모델이 예측하는 값의

차이를 최소한으로 하는 선을 찾자

전체 모델의 차이

즉, **Loss function**을 최소로 만드는 β_0, β_1 구하기

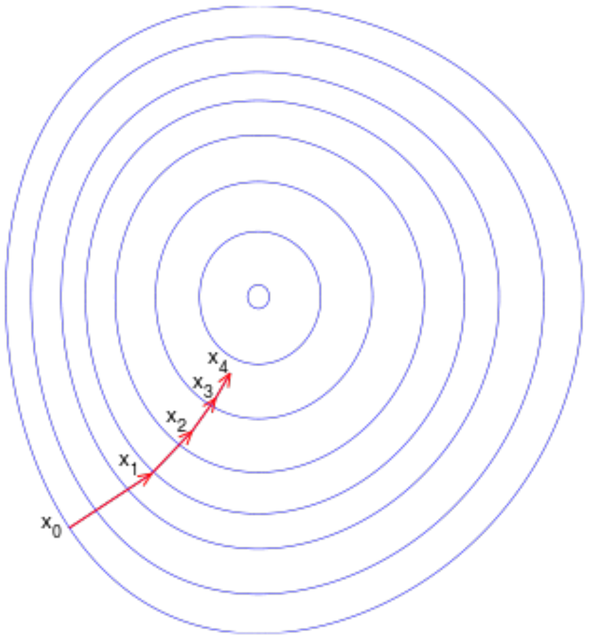


어떻게 찾을까?

산 정상 오르기

산 정상이 되는 지점을 찾고 싶다.

아무 곳에서나 시작했을 때,
가장 정상을 빠르게 찾아가는 방법은?



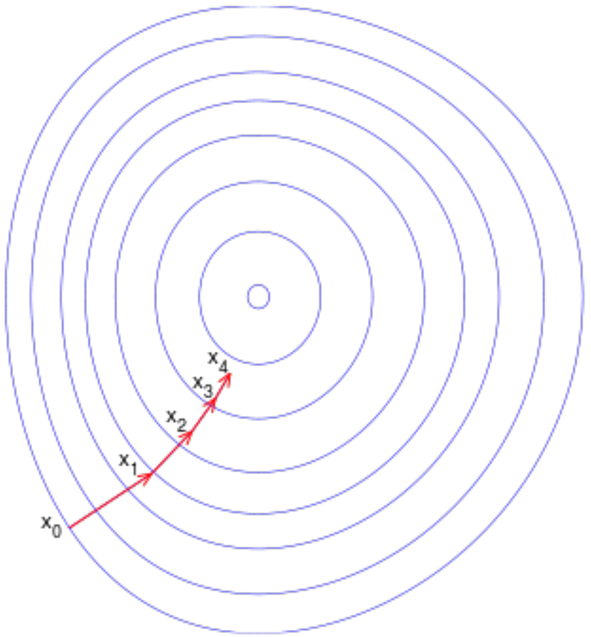
어떻게 찾을까?

가정

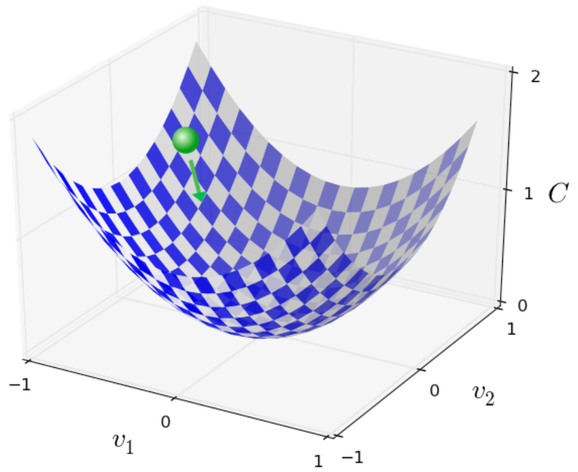
- 정상의 위치는 알 수 없다.
- 현재 나의 위치와 높이를 알 수 있다.
- 내 위치에서 일정 수준 이동할 수 있다.

방법

- 현재 위치에서 가장 **경사가 높은 쪽**을 찾는다.
- 오르막 방향으로 일정 수준 이동한다.
- 더 이상 높이의 변화가 없을 때까지 반복!



거꾸로 된 산을 내려가기



Gradient Descent(경사하강법)

- 최적의 값을 찾기 위한 거꾸로 된 산을 내려가는 방법
- 전체 모델의 차이 즉, **Loss function**을 최소로 만드는 β_0, β_1 을 선정함

회귀 분석 개념 정리하기

Loss Function(실제 값과 모델이 예측하는 값의 오차)를 최소화하는
Gradient Descent(최적의 β_0, β_1 를 찾는 알고리즘)을 통해

데이터를 가장 잘 설명할 수 있는 선을 찾는 방법

02

단순 선형회귀

문제에 적용하기

평균 기온 (X) 에 따른 **아이스크림 판매량** (Y) 을 예측하고자 함

예시

$$Y \approx \beta_0 + \beta_1 X$$

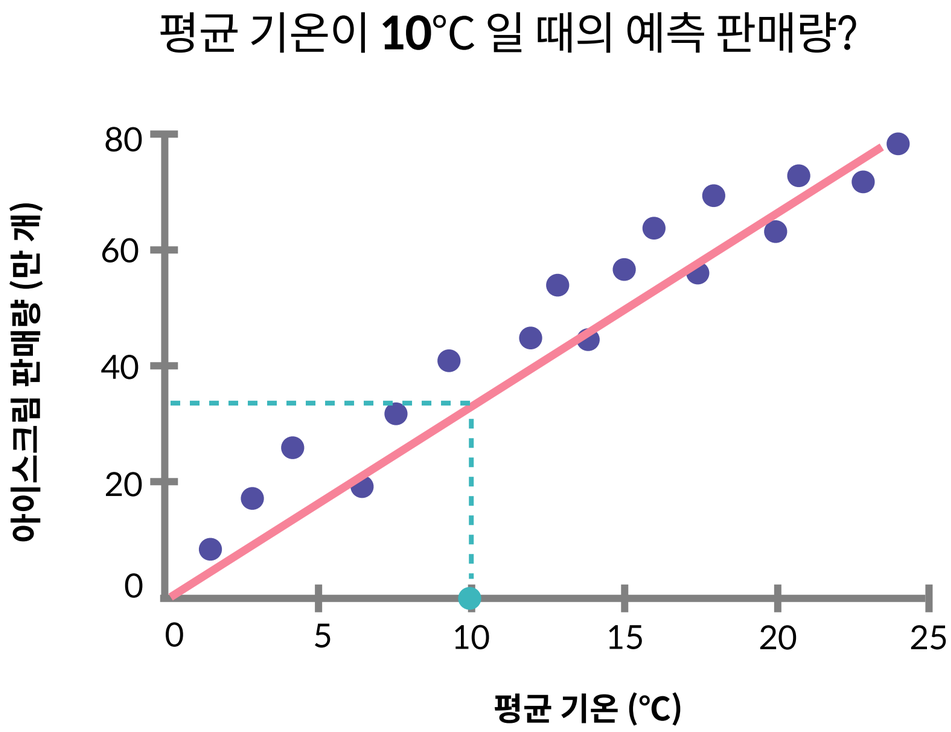
아이스크림 판매량 = β_0 + β_1 * 평균기온

적용 결과

데이터를 **잘 설명할 수 있는 선**을 찾아
새로운 데이터에 대한 결과값 확인 가능

이러한 가장 단순한 회귀 모델을

단순 선형회귀
(Simple Linear Regression)라고 한다.



단순 선형회귀(Simple Linear Regression)

가장 기본적이고 간단한 방법의 회귀 알고리즘
입력값 X와 결과값 Y의 관계를 설명할 때 가장 많이 사용되는 단순한 모델

회귀 알고리즘의 기초로 이를 응용한 다수 알고리즘 존재

단순 선형회귀의 함수식

$$Y \approx \beta_0 + \beta_1 X$$

단순 선형회귀 특징

- 가장 기초적이나 여전히 많이 사용되는 알고리즘
- 입력값(X)이 1개인 경우에만 적용이 가능함
- 입력값과 결과값의 관계를 알아보는 데 용이함
- 입력값이 결과값에 얼마나 영향을 미치는지 알 수 있음
- 두 변수 간의 관계를 **직관적으로 해석**하고자 하는 경우 활용

가정해보기

만약, 평균 기온 뿐만 아니라
강수량이라는 새로운 입력값을 추가하여
판매량을 예측하고자 한다면 어떻게 해야 할까?

데이터 유형 및 예측 결과에 따라
다양한 회귀 알고리즘 필요

03

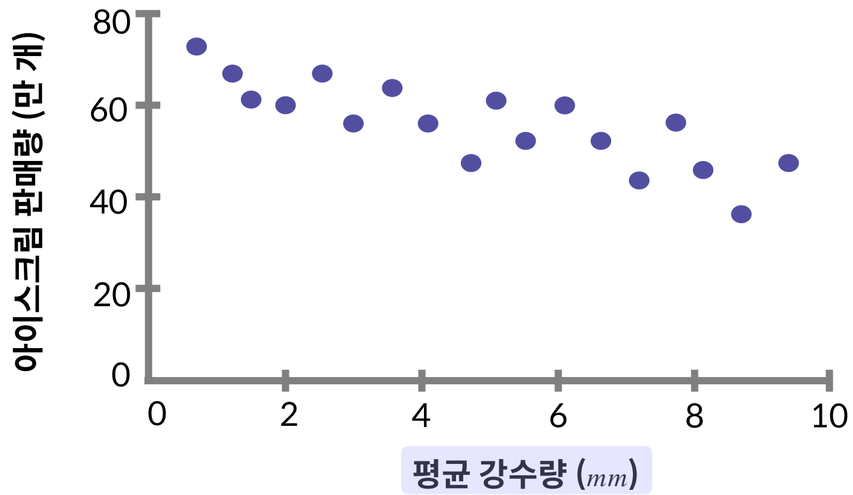
다중 선형회귀와 다항 회귀

문제

만약, 입력값 X 에 **강수량**이 추가된다면?

즉, **평균 기온**과 **평균 강수량**에 따른 **아이스크림 판매량**을 예측하고자 할 때

X_0	X_1	Y
평균 기온(℃)	평균 강수량(mm)	아이스크림 판매량(만개)
10	10	40
13	7.5	52.3
20	2	60.5
25	0	80



문제 해결하기

평균 기온 (X_1) 과 **평균 강수량** (X_2)에 따른 **아이스크림 판매량** (Y)을 예측하고자 함

여러 개의 입력값(X_i)으로 결과값(Y)을 예측하고자 하는 경우 → **다중 선형회귀 (Multiple Linear Regression)**

다중 선형 회귀(Multiple Linear Regression)

입력값 X가 여러 개(2개 이상)인 경우 활용할 수 있는 회귀 알고리즘
각 **개별** X_i에 해당하는 **최적의** β_i를 찾아야 함

$$Y \approx \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_i X_i$$

문제 : **평균 기온**과 **평균 강수량**에 따른 **아이스크림 판매량**을 예측하고자 할 때

아이스크림 판매량 = β_0 + β_1 * 평균기온 + β_2 * **강수량**

03 다중 선형회귀와 다항 회귀

다중 선형 회귀 특징

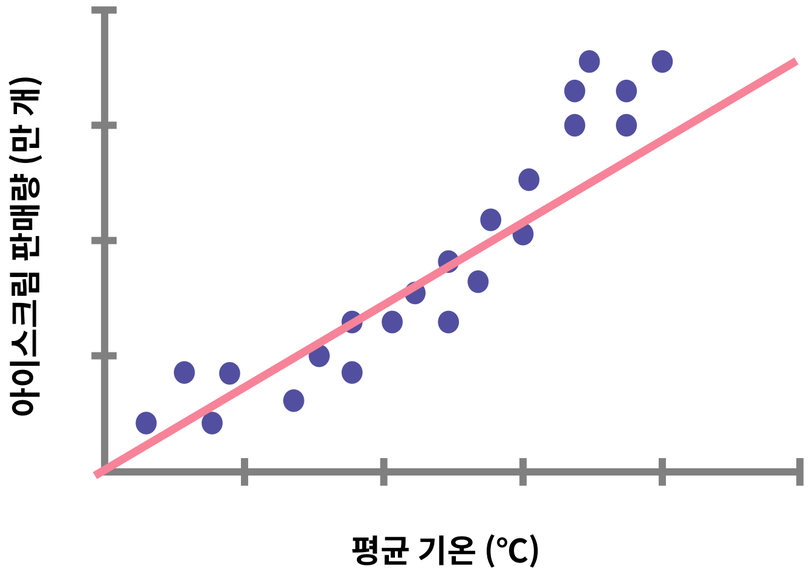
- 여러 개의 입력값과 결과값 간의 관계 확인 가능
- 어떤 입력값이 결과값에 어떠한 영향을 미치는지 알 수 있음.
- 여러 개의 입력값 사이 간의 **상관 관계***가 높을 경우 결과에 대한 신뢰성을 잃을 가능성이 있음.

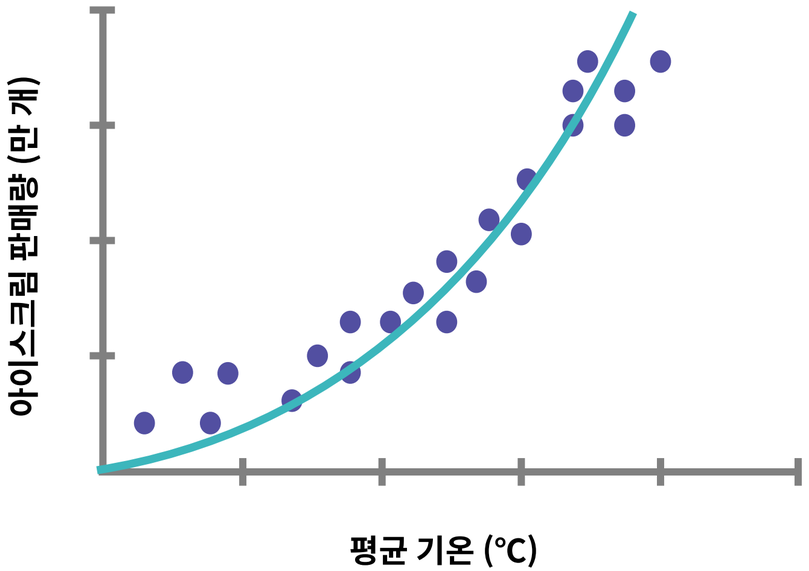
상관 관계

두 가지 것의 한쪽이 변화하면 다른 한쪽도 따라서 변화하는 관계

더 생각해보기

다중 선형회귀 알고리즘 적용 시 예측 결과값이 좋지 않은 경우
만약, **평균기온**과 **예측 판매량** 간의 관계가 **선형적**이지 않다면?





문제 해결하기

다중 선형회귀 적용 시 예측 결과값이 좋지 않고,
데이터들간의 관계가 선형적이지 않은 경우

→

다항 회귀
(Polynomial Linear Regression)

다항 회귀(Polynomial Regression)

1차 함수 선형식으로 표현하기 어려운 분포의 데이터를 위한 회귀
복잡한 분포의 데이터의 경우 일반 선형 회귀 알고리즘 적용 시 낮은 성능의 결과가 도출됨
데이터의 분포에 더 잘 맞는 모델이 필요함

$$Y \approx \beta_0 + \beta_1 X_1 + \beta_2 X_2^2 + \ldots + \beta_i X_i^i$$

문제 : 아이스크림 판매량이 평균 기온의 **제곱**과 관련이 있는 경우

$$\text{아이스크림 판매량} = \beta_0 + \beta_1 * \text{평균기온} + \beta_2 * \text{강수량}^2$$

03 다중 선형회귀와 다항 회귀

다항 회귀 원리

기존 **입력값** X_i를 전처리 한 새로운 변수를 추가시켜 선형 회귀 모델로 예측할 수 있도록 함.

$$Y \approx \beta_0 + \beta_1 X_1 + \beta_2 X_2^2$$

사실은 다중 선형 회귀와 동일한 원리

$$Y \approx \beta_0 + \beta_1 X_1 + \beta_2 X_2$$

: 각 **개별 입력값** X_i에 해당하는 **최적의** β_i를 찾아야 함.

03 다중 선형회귀와 다항 회귀

다항 회귀 특징

- 일차 함수 식으로 표현할 수 없는 복잡한 데이터 분포에도 적용 가능
- 극단적으로 높은 차수의 모델을 구현할 경우 과도하게 학습 데이터에 맞춰지는 **과적합** 현상 발생
- 데이터 관계를 선형으로 표현하기 어려운 경우 사용

04

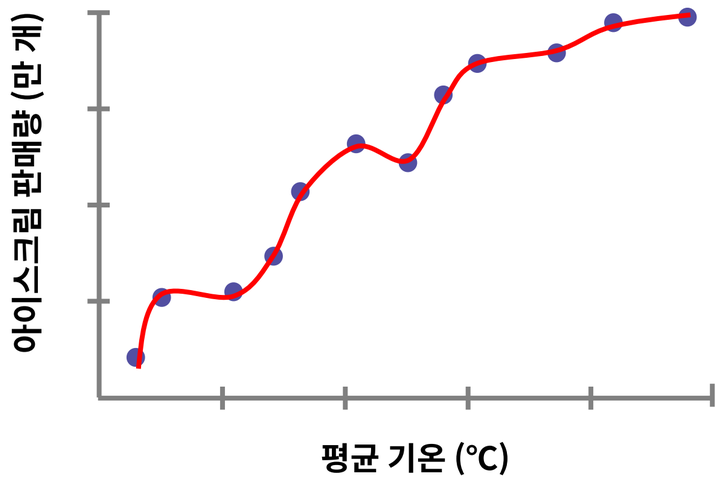
과적합과 정규화

04 과적합과 정규화

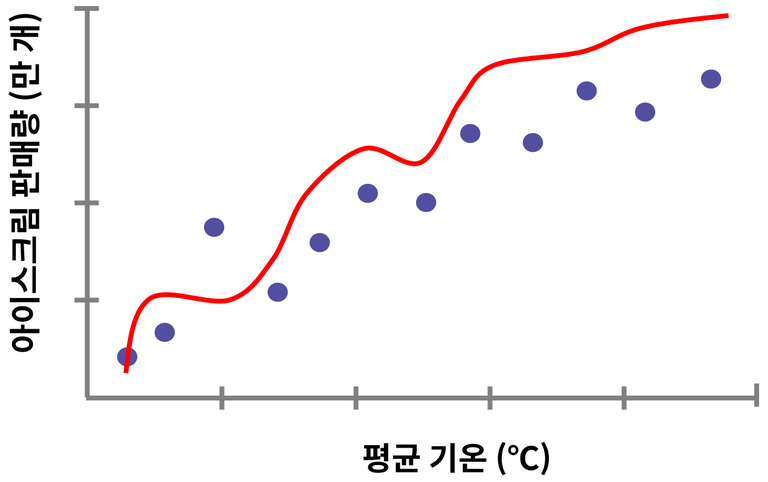
문제

만약, 과거 데이터를 학습한 다항 회귀 모델을 활용하여
새로운 데이터에 대한 예측 결과가 매우 낮다면?
= **과적합(Overfitting)** 발생

- 2019년 데이터로 학습한 다항 회귀 모델 A



- 학습된 다항 회귀 모델 A에 새로운 데이터 입력

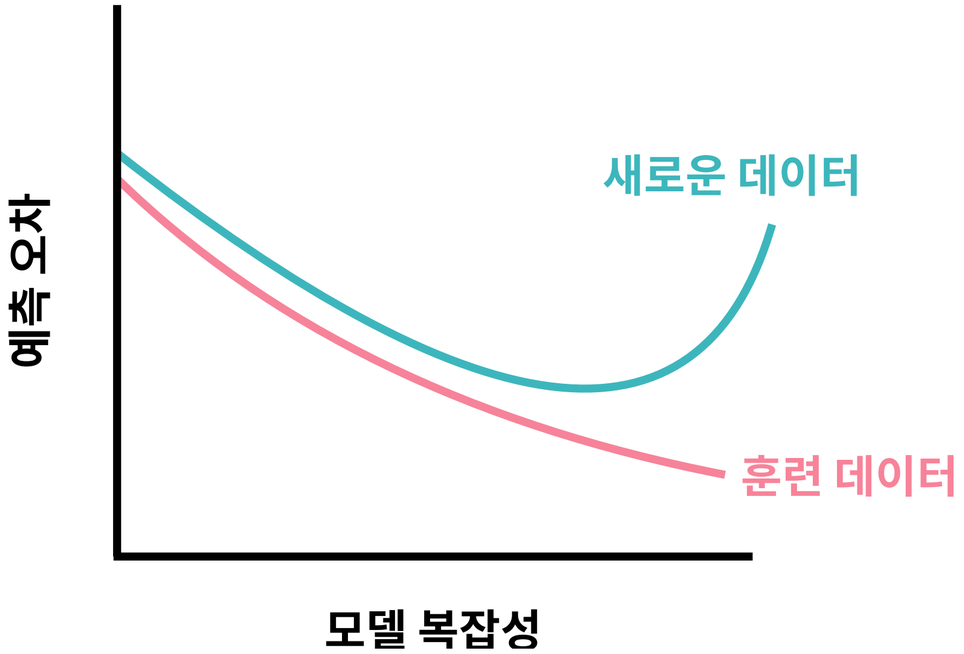


/* elice */

과적합(Overfitting)

모델이 주어진 훈련 데이터에 과도하게 맞춰져
새로운 데이터가 입력 되었을 때 잘 예측하지 못하는 현상

즉, 모델이 과도하게 복잡해져
일반성이 떨어진 경우를 의미함



과적합 방지 방법

모델이 잘 적합되어 실제 데이터와 유사한 예측 결과를 얻을 수 있도록
과적합 방지를 위해 다양한 방법을 사용함

예시

교차 검증(Cross Validation), 정규화(Regularization)

교차 검증(Cross Validation)

모델이 잘 적합되었는지 알아보기 위해
훈련용 데이터와 별개의 **테스트 데이터,** 그리고 **검증 데이터로** 나누어 성능 평가하는 방법

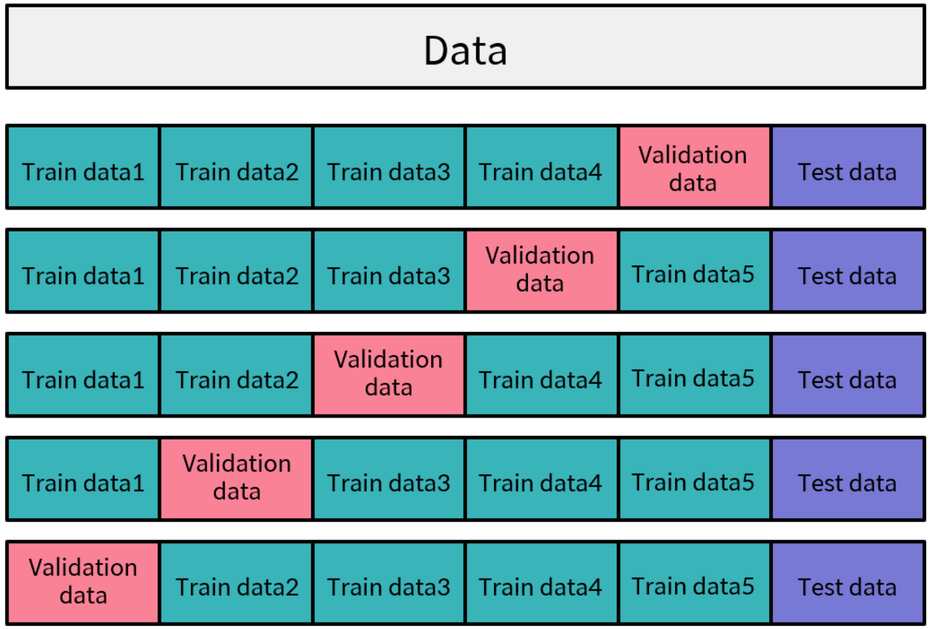
다양한 방법들이 있지만, 일반적으로 **k-fold 교차 검증**을 많이 사용함

Train (훈련용 데이터)	Validation (검증용 데이터)	Test (테스트용 데이터)

K-fold 교차 검증

훈련 데이터를 계속 변경하며 모델을 훈련시킴
: 데이터를 K등분으로 나누고 K번 훈련시킴

1. K를 설정하여 데이터 셋을 K개로 나눔
2. K개 중 한 개를 검증용, 나머지를 훈련용으로 사용
3. K개 모델의 평균 성능이 최종 모델 성능



정규화(Regularization)

모델의 복잡성을 줄여
일반화된 모델 구현을 위한 방법

→

모델 β_i에 **패널티를 부여함***

*선형 회귀를 위한 정규화 : L1, L2 정규화

선형 회귀를 위한 정규화 방법

- **L1 정규화(Lasso)**

불필요한 입력값에 대응되는 β_i를 **정확히 0**으로 만든다.

- **L2 정규화(Ridge)**

아주 큰 값이나 작은 값을 가지는 이상치에 대한 β_i를 **0에 가까운 값**으로 만든다.

정규화 방법을 적용한 회귀 알고리즘

회귀 알고리즘에 정규화 방법을 적용하면?

적용한 정규화 방법에 따라
라쏘(Lasso), 릿지(Ridge), 엘라스틱 넷(Elasticnet) 회귀로 분류

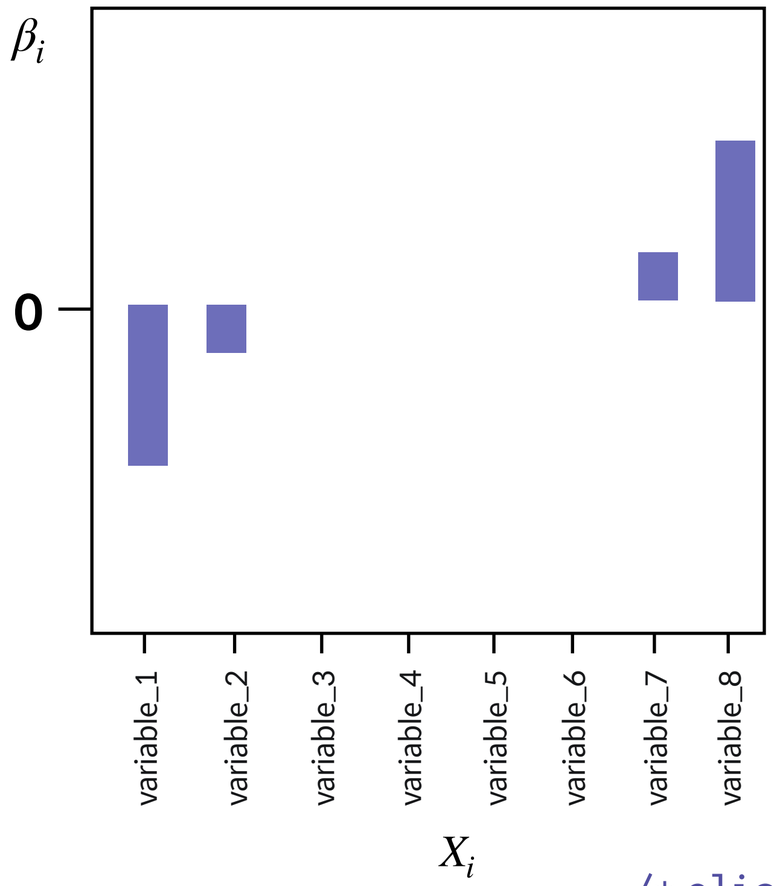
05

정규화를 적용한 회귀

정규화 기법을 적용한 회귀(1) - Lasso Regression

회귀 학습에 사용되는
Loss Function(비용 함수)에 **L1 정규화** 항을 추가

중요하지 않은 β_i를 **0으로 만들어**
모델의 복잡성을 줄일 수 있음



/* elice */

정규화 기법을 적용한 회귀(1) - Lasso Regression

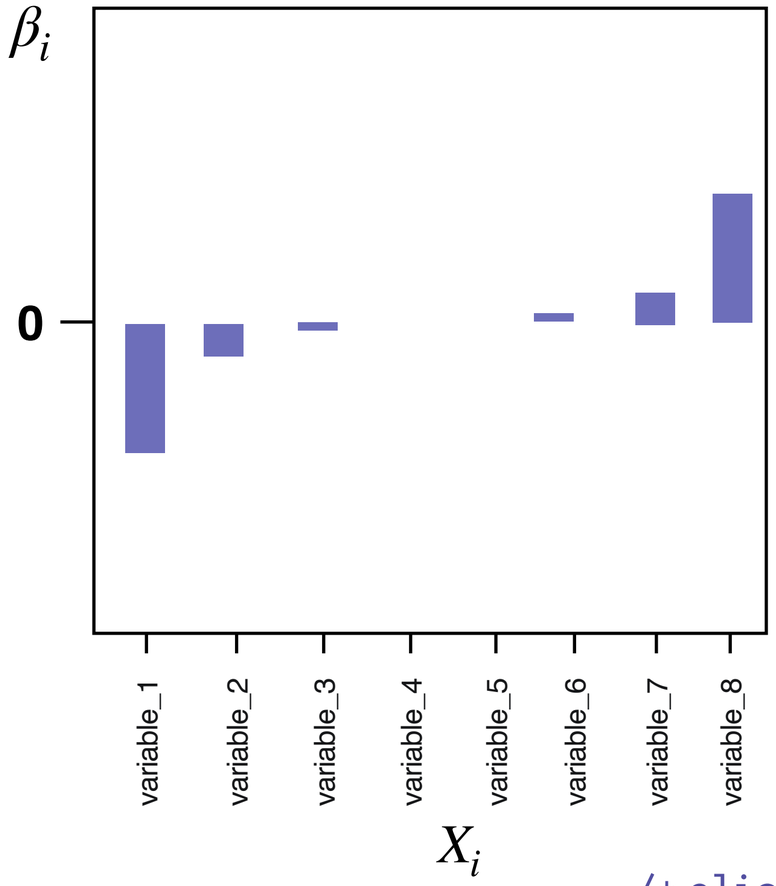
특징

- 너무 많은 β_i를 0으로 만들 수 있어 모델의 정확성이 떨어질 수 있음
- 몇 개의 중요 변수만 선택하기 때문에 정보 손실의 가능성이 있음

정규화 기법을 적용한 회귀(2) - Ridge Regression

회귀 학습에 사용되는
Loss Function(비용 함수)에 **L2 정규화** 항을 추가

중요하지 않은 β_i를 **0에 가깝게 만들어**
모델의 복잡성을 줄일 수 있음



정규화 기법을 적용한 회귀(2) - Ridge Regression

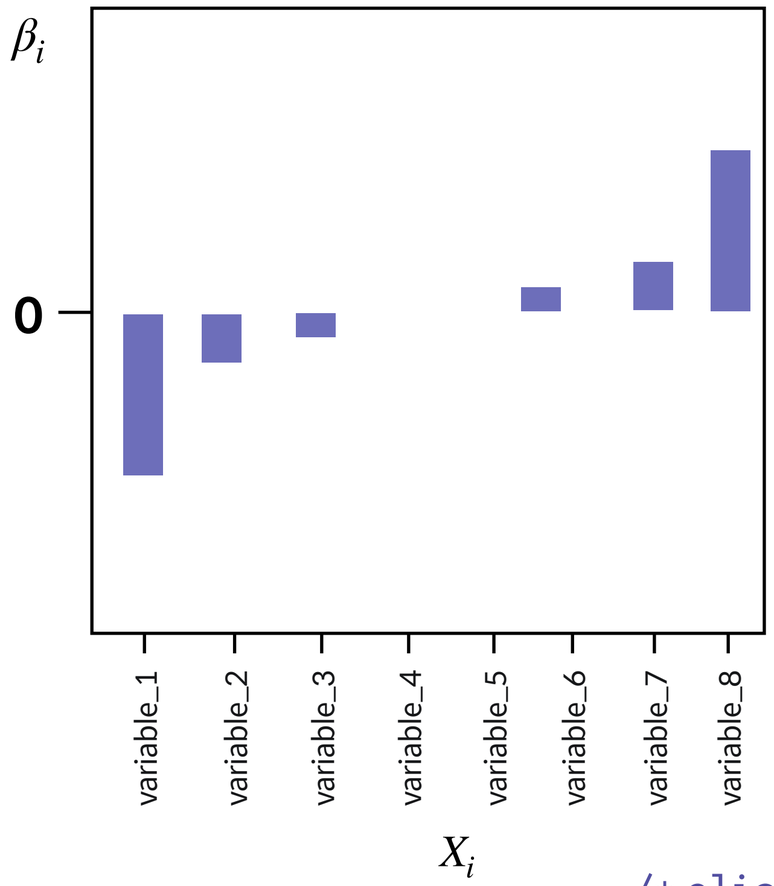
특징

- β_i를 0에 가깝게 만들지만 완전한 0은 아니기 때문에 모델이 여전히 복잡할 수 있음

정규화 기법을 적용한 회귀(3) - Elastic Net Regression

Lasso 회귀, Ridge 회귀의 단점을 보완하기 위함

Lasso 회귀의 L1 정규화와
Ridge 회귀의 L2 정규화
적용 **비율을 조정**하여 모델 구현



/* elice */

06

회귀 알고리즘 평가 지표

06 회귀 알고리즘 평가 지표

회귀 알고리즘 평가

어떤 모델이 좋은 모델인지를 어떻게 평가할 수 있을까?
목표를 얼마나 잘 달성했는지 정도를 평가해야 함.

회귀 분석 알고리즘 목표

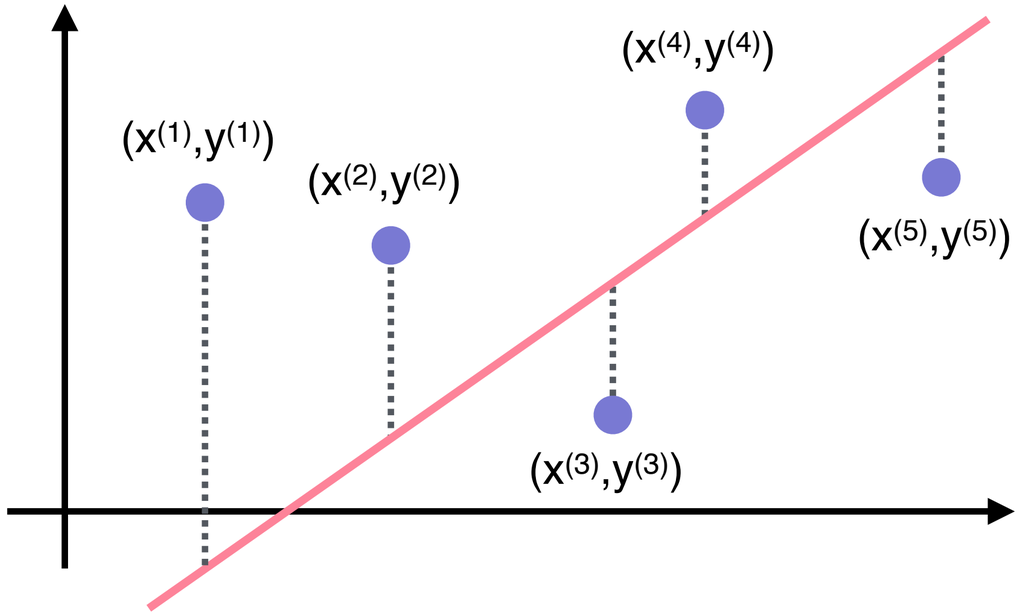
과대적합의 반대의미로,
모델이 너무 단순하여 학습된 데이터조차 잘 예측하지 못하는 현상

목표 달성 평가 방법

실제 값과 모델이 예측하는 값의 **차이**에 기반한 평가 방법 사용

예시

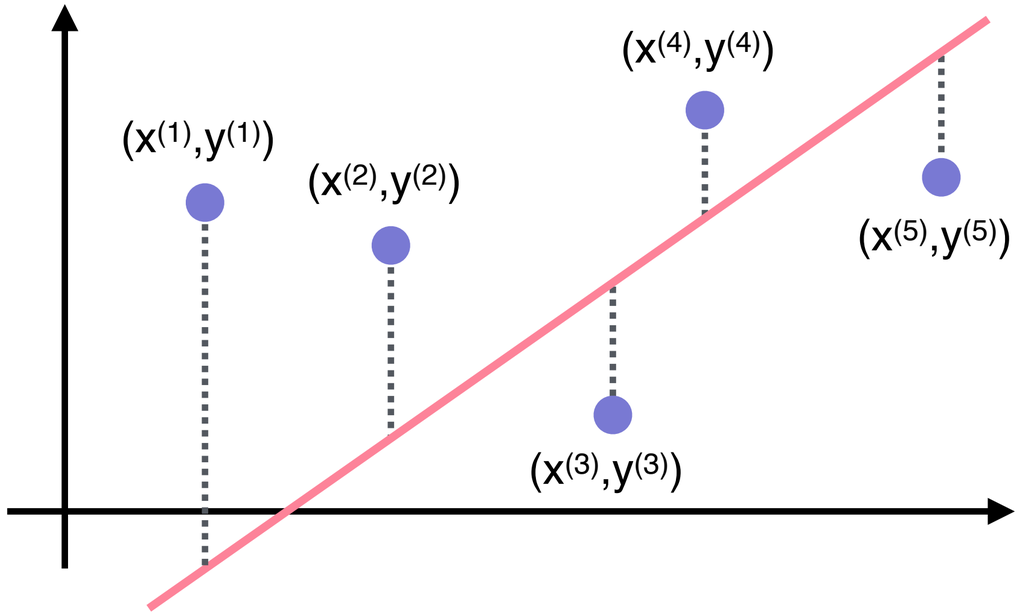
$$RSS, MSE, MAE, MAPE, R^2$$



RSS - 단순 오차

- RSS(Residual Sum of Squares)

1. 실제값과 예측값의 **단순 오차 제곱 합**
2. 값이 작을수록 모델의 성능이 높음
3. 전체 데이터에 대한 실제 값과 예측하는 값의 오차 제곱의 총합



06 회귀 알고리즘 평가 지표

RSS 특징

- 가장 간단한 평가 방법으로 직관적인 해석이 가능함
- 그러나 오차를 그대로 이용하기 때문에 입력값의 **크기에 의존적**임

06 회귀 알고리즘 평가 지표

MSE, MAE - 절대적인 크기에 의존한 지표

- MSE(Mean Squared Error)

평균 제곱 오차, RSS에서 데이터 수만큼 나눈 값
작을수록 모델의 성능이 높다고 평가할 수 있음.

- MAE(Mean Absolute Error)

평균 절대값 오차, 실제값과 예측값의 오차의 절대값의 평균
작을수록 모델의 성능이 높다고 평가할 수 있음.

06 회귀 알고리즘 평가 지표

MSE, MAE 특징

- MSE : 이상치(Outlier) 즉, 데이터들 중 크게 떨어진 값에 민감함
- MAE : 변동성이 큰 지표와 낮은 지표를 같이 예측할 시 유용
- 가장 간단한 평가 방법들로 직관적인 해석이 가능함
- 그러나 평균을 그대로 이용하기 때문에 입력값의 **크기에 의존적**임

06 회귀 알고리즘 평가 지표

R^2 (결정 계수)

회귀 모델의 설명력을 표현하는 지표

1에 가까울수록 높은 성능의 모델이라고 해석할 수 있음.

예시

$$1 - \frac{\text{SSE}}{\text{TSS}} \quad (0 \leq R^2 \leq 1)$$

06 회귀 알고리즘 평가 지표

R^2 특징

- 백분율로 표현하기 때문에 크기에 의존적이지 않음
- 실제값이 1보다 작을 경우, 무한대에 가까운 값 도출, 실제값이 0일 경우엔 계산 불가

06 회귀 알고리즘 평가 지표

평가 지표 선정 방법

절대적인 평가 지표는 존재하지 않음!
다양한 평가 지표를 적용해보고, 결과값을 비교하며 모델의 성능을 **다양한 측면에서 확인**해봐야 함

Contact

TEL

070-4633-2015

WEB

https://elice.io

E-MAIL

contact@elice.io