머신러닝

홍유경 / 이원지희 /한명수

1 손실함수

○손실함수는 머신러닝이나 딥러닝 모델이 예측한 값과 실제 값 사이의 차이를 측정하는 함수입니다. 이를 통해 모델의 성능을 평가하고, 모델이 어떤 방향으로 개선되어야 할지 알려주는 역할을 합니다. 손실함수의 값을 최소화하는 것이 모델 학습의 목표라고 할 수 있습니다.

종류:평균 제곱 오차(MSE, Mean Squared Error),크로스 엔트로피(Cross-Entropy)

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}(\hat{Y}_i - Y_i)^2$$

과소적합

과대적합

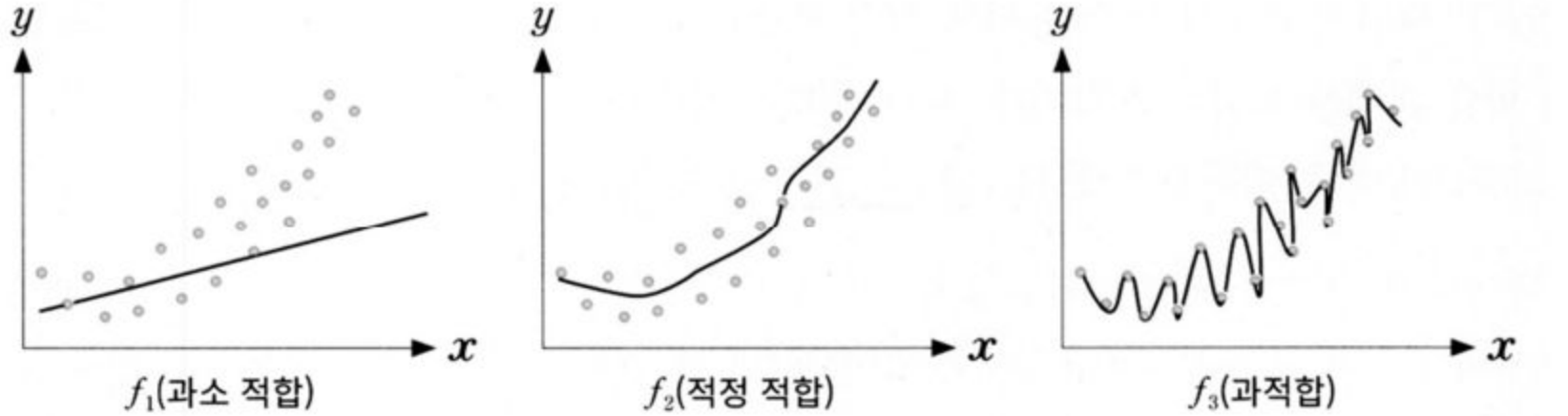


그림 1-2 과소 적합, 적정 적합, 과대 적합

x
f_1(과소 적합)
x
f_3(과적합)

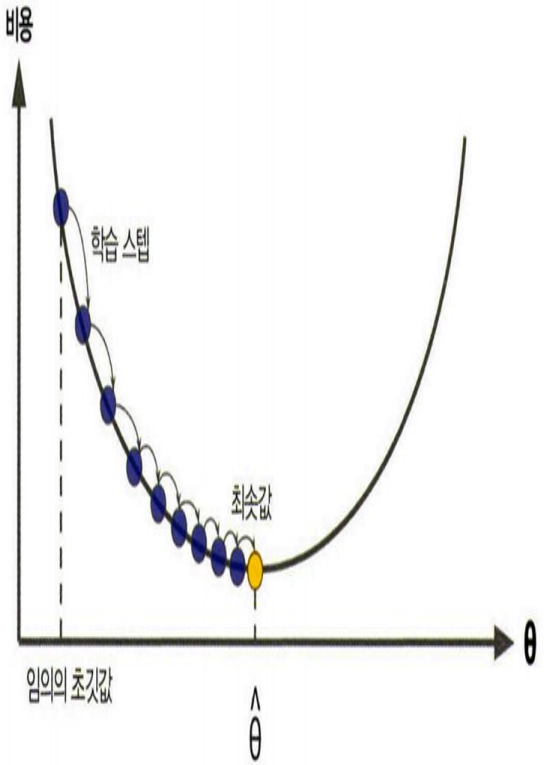
경사하강법

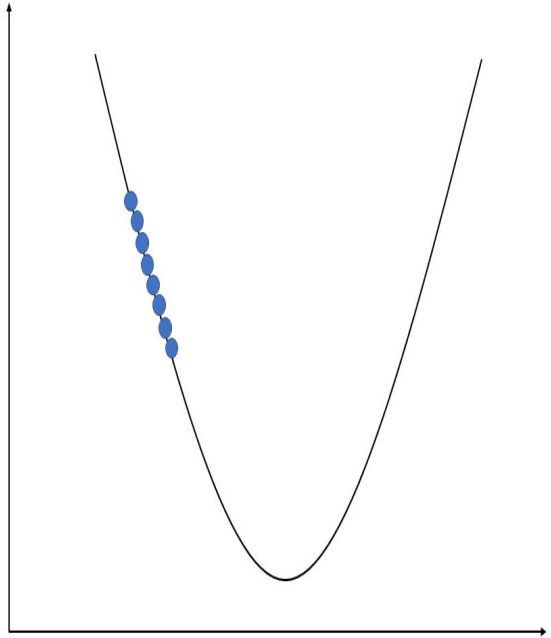
1.초기 파라메터 설정

2.함수의 경사도 계산

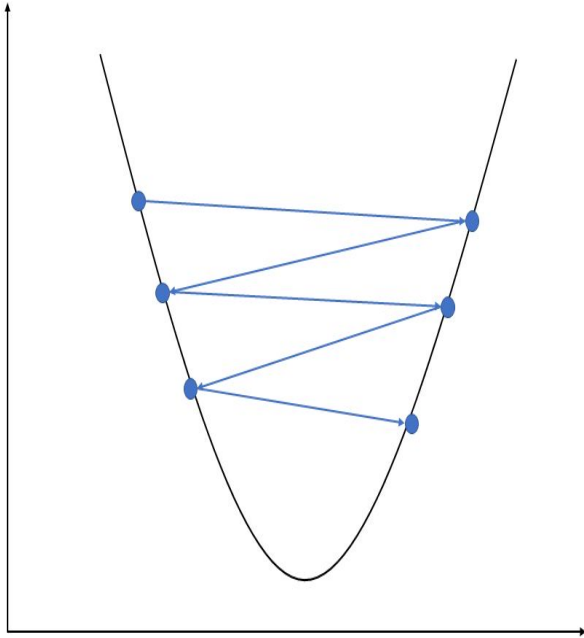
3.경사방향으로 이동

4.반복(1~3과정을 반복해서 최소값이 되는 변수값을 찾음)





learning rate 작을때



learning rate 클때

하이퍼파라메타

모델이 학습을 시작하기 전에 사람이 직접 설정해야 하는 값.

학습률 (Learning Rate)

에포크 (Epoch)

배치 크기 (Batch Size)

정규화 파라미터

2 모델의 과적합의 이해

	과소적합	과대적합
예시	학생이 시험 문제의 핵심 개념을 이해하지 못해 문제를 풀지 못하는 것	학생이 시험 문제를 외워서 풀거나 너무 세세한 부분까지 기억하려다보니, 조금만 문제가 바뀌면 풀지 못하는 것

구분	과소적합 (Underfitting)	과대적합 (Overfitting)
의미	모델이 학습 데이터를 충분히 학습하지 못함	모델이 학습 데이터에 너무 맞춰짐
발생원인	모델이 너무 단순, 학습 데이터 부족, 학습률이 낮음(이터레이션 수 적음)	모델이 너무 복잡(샘플/특징 多) 학습 데이터 부족, 학습률이 높음

모델의 과적합 해결방안

구분	과소적합 (Underfitting)	과대적합 (Overfitting)
해결방안	모델 복잡도 높이기, 학습 데이터 늘리기, 학습률 높이기	규제, 모델 복잡도 낮추기, 정규화, 데이터 증강, 드롭아웃

예시

다양한 유형의 문제를 더 많이 풀어보자

문제 푸는 양을 늘리자

문제를 더 꼼꼼하게 읽고 기출변형으로 문제의 의도를 파악하게끔 도와주자
(K-FOLD Cross Validation)

공식 그만 외우고 문제 위주로 풀어보자

문제에서 요구하는 핵심 정보만 집중해서 문제를 풀게하자

기출 문제로 다양한 문제 유형을 더 풀어보자

다른 방법으로 풀 수도 있는지 생각해보자

K-FOLD Cross Validation

더욱 객관적인 평가가 필요할 때 쓰는 기법으로 k-겹 교차 검증(k-fold cross validation)이 있습니다. k-겹 교차 검증은 데이터를 서로 겹치지 않는 k개의 작은 데이터인 폴드(fold)로 분할한 뒤, 각각의 폴드를 평가 데이터로 사용하고 나머지 폴드의 합집합을 학습 데이터로 사용하는 방식입니다. k를 5로 했을 때의 예시를 다음 그림을 통해 살펴보겠습니다.

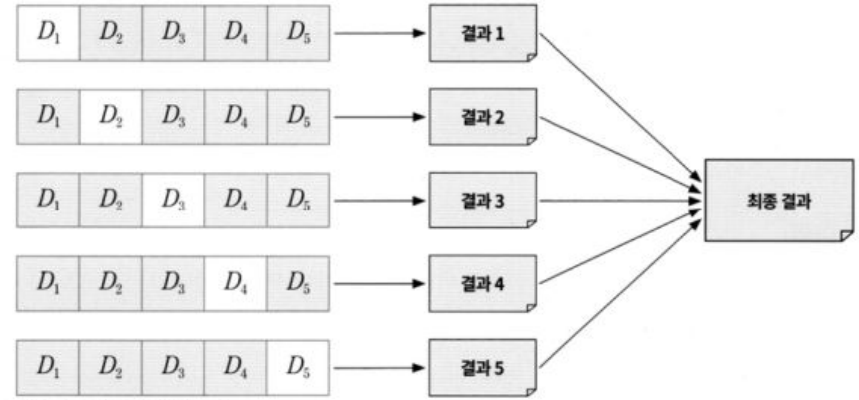
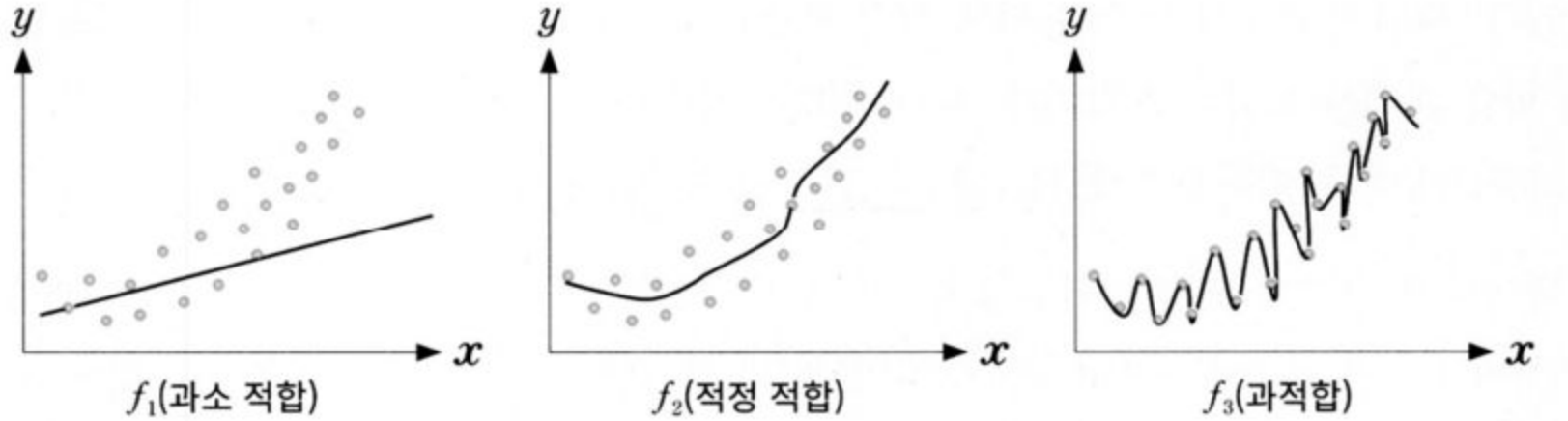


그림 1-4 k-겹 교차 검증 (k = 5)

규제



f_1(과소 적합) f_2(적정 적합) f_3(과적합)

그림 1-2 과소 적합, 적정 적합, 과대 적합

f_1(과소 적합) f_3(과적합)

3 결측치란?

데이터 확인 도중 발견하는 수치. 원래 있어야 하는 데이터가 없는 것

데이터 정합성 검토

- NONE 값, NaN 값, Outlier 확인, 논리적으로 이상한 값 확인
- NONE : 없어도 괜찮은 값
- NaN : 있어야 되는데 없는 값 (결측치 처리 대상)
- Outlier : 이상치 (데이터 분포에서 많이 벗어난 값)
- 해당 값을 추정하거나, 삭제, 대체한다.

결측치 처리방법

- 입력되어야 하는데 누락된 값. 결측치가 포함된 데이터로는 학습할 수 없다.
- 반드시 결측치를 제거하거나 원래 값을 추정해야 한다.
  - 추정값 : 평균(mean), 중간값(median), 최빈값(mode) 등을 채운다

표 1-2 결측치가 포함된 데이터 예시

ID	x_1	x_2	x_3
1	10	25	5
2	8		10
3		7	4
4	5		6
5	12	3	6

? **None 과 NaN**

None : 값이 없는 것이 정상 → 새로운 값 부여(예 : 무직자, 백수, 응답 없음 등)

NaN : 값이 있어야 정상 → 제거 or 원래 값 추정

[결측치 삭제 시 장단점]

- 장점 : 쉽고 빠르다. 결측치를 잘못 추정해서 발생하는 위험이 없다
- 단점 : 샘플이 부족해진다
  - 모든 결측치를 제거할 경우, 새로운 결측에 적절히 대처할 수 없다
  - 소수의 결측 삭제로 예측에 중요한 특징이 삭제될 수 있다

결측치 제거 꿀팁

1. 샘플이 충분히 많고, 새로 입력할 데이터에 결측이 없겠다 판단하는 경우
   → 행 제거
2. 결측치가 있는 열 가운데 결측치의 비율이 너무 높아 활용이 어렵다면
   → 열 삭제

결측치 확인 함수

```
# df.isna() == .isnull()
# df에서 결측치가 있는 행을 찾아서 반환
df[df.isna( ).any(axis=1)]


# isna() : 데이터프레임 안에 모든 값들을 검사 True, False로 판단.
#          신규 데이터프레임으로 반환.
#          결측값이면 True, 정상값이면 False

# .any : 위의 불리언 값중에서 특정 조건을 만족하는 값을 확인
# (axis=1) : 조건. 각 행에서 하나라도 True가 있는지 확인해서 반환
# (axis=0) : 조건. 각 열에서 하나라도 True가 있는지 확인해서 반환


# 결측값 없는지 확인하는 함수(반대)
# df.notna() == .notnull()
# 결측값이면 False, 정상값이면 True
```

▼ df.isna( ) 사용 결과 예시

	지점	지점명	일시	평균기온(℃)	최저기온(℃)	최고기온(℃)
15676	108	서울	1950-09-01	NaN	NaN	NaN
15677	108	서울	1950-09-02	NaN	NaN	NaN
15678	108	서울	1950-09-03	NaN	NaN	NaN
15679	108	서울	1950-09-04	NaN	NaN	NaN
15680	108	서울	1950-09-05	NaN	NaN	NaN
...	...	...	...	...	...	...
16430	108	서울	1953-11-29	NaN	NaN	NaN
16431	108	서울	1953-11-30	NaN	NaN	NaN
21260	108	서울	1967-02-19	-1.7	NaN	NaN
39758	108	서울	2017-10-12	11.4	8.8	NaN
41519	108	서울	2022-08-08	26.8	NaN	28.4

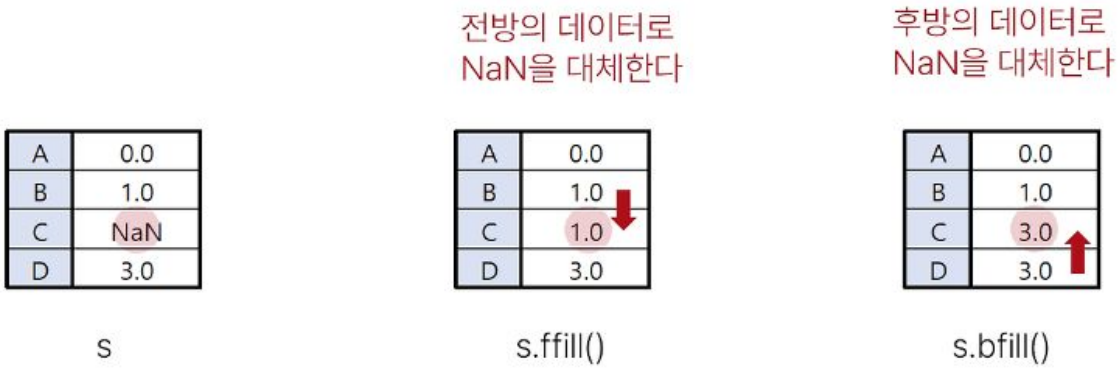
759 rows × 6 columns

결측치 채우는 함수

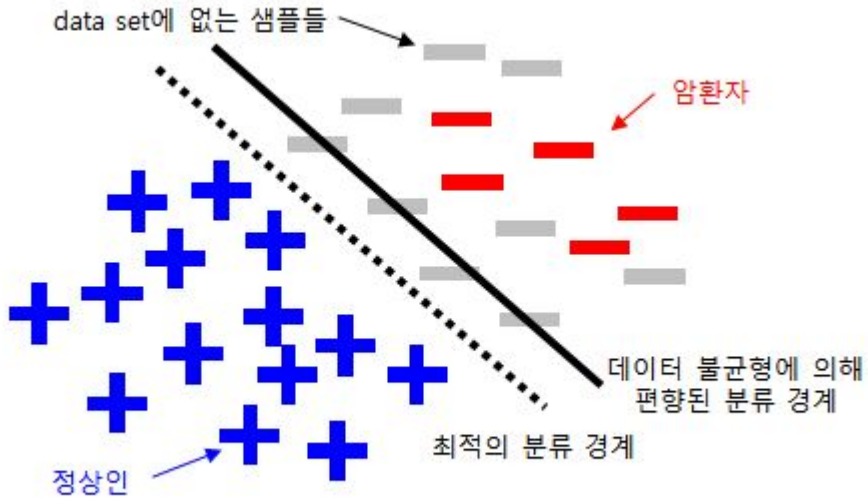
```
# df.fillna()
# df에서 결측값을 원하는 값으로 변경

df.fillna(value=None, axis=None,
          inplace=False, limit=None, downcast=None)


# value   : 결측값 대체할 값. ('') > 빈칸으로 채우기
# axis    : 0 > 행, 1 > 열
# limit   : 결측값을 변경할 횟수 입력
# downcast: 'infer'를 넣으면  float64 > int64
```

결측치 채우는 함수

```
df.ffill(axis=0, limit=None)
```



클래스 불균형 해결방안

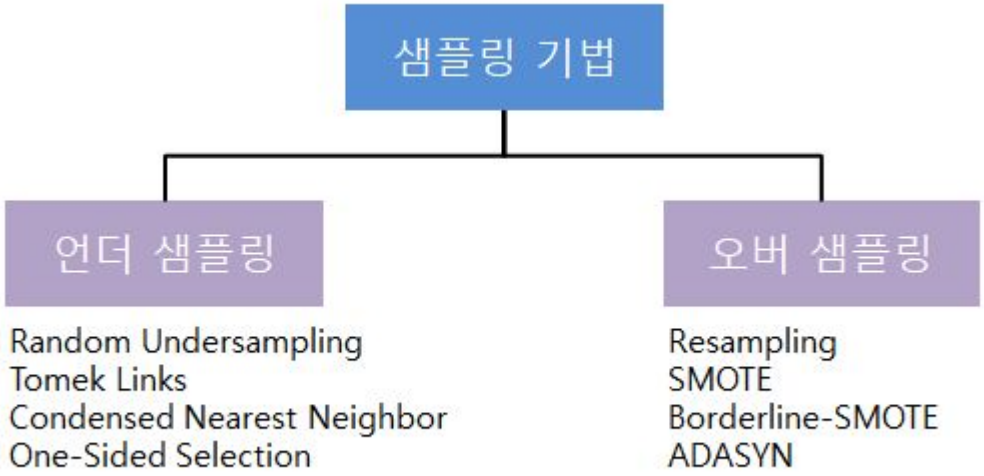


예시 : 암환자 분류

암 환자 수 < 정상인 수
샘플도 정상인이 더 多

정상인 수에 가중치를 두면
암환자를 정상인으로 진단 가능

클래스 불균형 해결 방안


샘플링 기법
언더 샘플링
오버 샘플링
Random Undersampling
Tomek Links
Condensed Nearest Neighbor
One-Sided Selection
Resampling
SMOTE
Borderline-SMOTE
ADASYN

4 데이터 스케일링

서로 다른 변수의 값 범위를 일정한 수준으로 맞추는 작업

관측개체 번호	나이 (x_1)	월별 소득 (x_2)
1	30	3,620,000
2	13	0
3	21	600,000
4	61	500,000
5	7	0
⋮	⋮	⋮

정규화(Normalization)

모든 값을 **0~1** 사이의 값으로 바꾸는 것

데이터의 최소값을 **0**, 최대값을 **1**로 설정하고, 나머지 값들을 그 사이에 맞춰 조정합니다.

ex)min-max nomalization

○서로 다른 값의 범위를 **0~1**사이로 옮겨준다

○사이킷런 **MinMaxScaler()**함수로 구현가능

$$x_{scaled} = \frac{x - x_{min}}{x_{max} - x_{min}}$$

Oli	Wax	Concentration	Width	Hyperspectral
Canola	Candelilla	5%	9.33	3307.231934
Canola	Candelilla	5%	9.38	3306.134033
Olive	Candelilla	5%	9.67	3329.056885
Olive	Candelilla	5%	9.6	3328.093506
Soybean	Carnauba	10%	9.28	4554.740723
Soybean	Carnauba	10%	9.18	4513.872559
Flaxseed	Bees	15%	9.78	4455.133789
Flaxseed	Bees	15%	9.83	4445.144531

Oli	Wax	Concentration	Width	Hyperspectral
Canola	Candelilla	5%	0.230769	0.0008793
Canola	Candelilla	5%	0.307692	0
Olive	Candelilla	5%	0.753846	0.018358745
Olive	Candelilla	5%	0.646154	0.017587182
Soybean	Carnauba	10%	0.153846	1
Soybean	Carnauba	10%	0	0.967268985
Flaxseed	Bees	15%	0.923077	0.920225533
Flaxseed	Bees	15%	1	0.912225209

표준화(standarlization)

데이터를 평균 **0**, 표준편차 **1**로 맞추는 방법

평균기준으로 데이터를 변환하기 때문에 상대적으로 오류가 적음

$$Z = \frac{X - \mu}{\sigma}$$

Oli	Wax	Concentration	Width	Hyperspectral
Canola	Candelilla	5%	9.33	3307.231934
Canola	Candelilla	5%	9.38	3306.134033
Olive	Candelilla	5%	9.67	3329.056885
Olive	Candelilla	5%	9.6	3328.093506
Soybean	Carnauba	10%	9.28	4554.740723
Soybean	Carnauba	10%	9.18	4513.872559
Flaxseed	Bees	15%	9.78	4455.133789
Flaxseed	Bees	15%	9.83	4445.144531

Oli	Wax	Concentration	Width	Hyperspectral
Canola	Candelilla	5%	-0.71955	-0.95052
Canola	Candelilla	5%	-0.51542	-0.95227
Olive	Candelilla	5%	0.668517	-0.91581
Olive	Candelilla	5%	0.382738	-0.91734
Soybean	Carnauba	10%	-0.92368	1.033408
Soybean	Carnauba	10%	-1.33193	0.968415
Flaxseed	Bees	15%	1.117596	0.875002
Flaxseed	Bees	15%	1.321724	0.859116

파이프라인

데이터 수집 및 전처리 : 데이터는 머신러닝 모델의 성능에 큰 영향을 미치므로, 적절한 데이터 수집과 전처리 작업이 필요합니다. 이 단계에서는 데이터를 수집하고, 결측치 처리, 이상치 제거, 정규화 등의 작업을 수행

특징 추출:머신러닝 모델에 입력으로 사용될 데이터의 특징을 추출하는 단계

모델 선택 및 학습:머신러닝 모델을 선택하고 학습하는 단계

모델 평가: 학습된 모델의 성능을 평가하는 단계

5. 모델 평가 지표

회귀모델 평가 지표

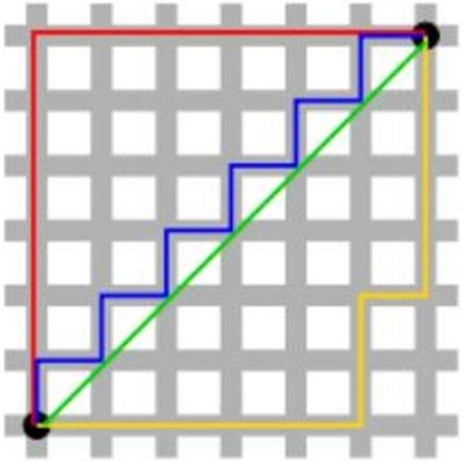
1. MAE 평균절대오차
2. MSE 평균제곱오차
3. RMSE

분류모델 평가지표 (confusion matrix)

1. accuracy
2. precision
3. recall
4. f1 score
5. precision과 recall의 관계 (분류임계점)

L1 / L2 NORM

Norm이란 두 벡터간의 거리를 측정하는 방법



검정색 두 점 사이의 거리를 측정하는데
L1 Norm은 빨간색, 파란색, 노란색 선으로
표현할 수 있고, L2 Norm은 초록색 선으로만
표현될 수 있다.

직선거리

즉, L1 Norm은 여러가지 Path를 갖지만,
L2 Norm은 하나의 Path만 갖는다.

6. L1 loss / L2 loss 와 L1 규제 / L2 규제

오차란 : 실제 데이터값 - 예측된 데이터값

손실함수의 개념 : 오차합의 평균

L1 LOSS / L2 LOSS : MAE와 MSE

이상치와 LOSS

과대적합과 규제 : 규제된 손실함수 = 기존 손실함수 + 규제항

L1 규제 / L2 규제 : LASSO와 RIDGE

L1 규제와 L2 규제의 차이(feature)

ELASTICNET(L1 RATIO)