4과목.빅데이터 결과 해석

(Ch_01. 분석 모형 평가 및 개선 - SEC 01. 분석 모형 평가
SEC 02. 분석 모형 개선)

빅데이터 분석 기사(4과목. 빅데이터 결과 해석)

CHAPTER 1. 분석 모형 평가 및 개선

CHAPTER 2. 분석 모형 개선

분석 모형 평가 및 개선

분석 모형 평가 및 개선 챕터는 총 2개의 작은 섹션으로 구성된다.

1. 분석 모형 평가
2. 분석 모형 개선

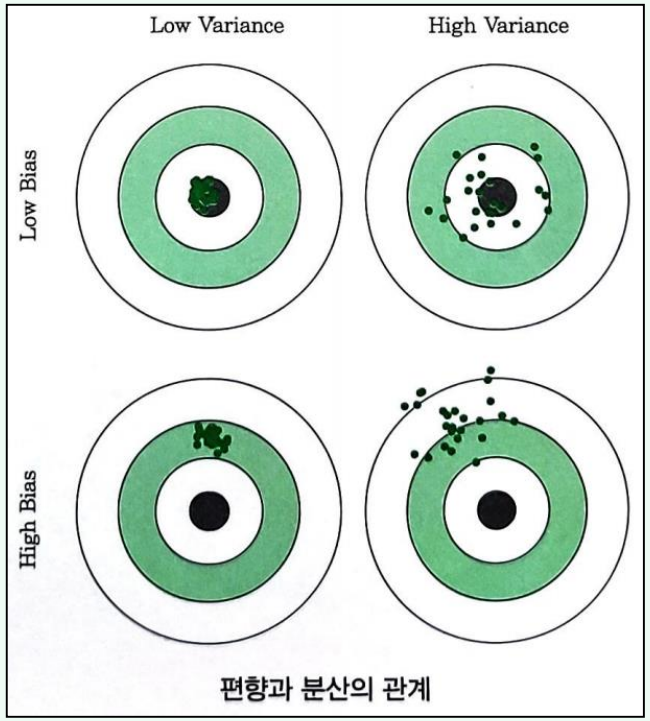
4. 분석 모형 평가 및 개선 – 분석 모형 평가

01 평가 지표

- **빅데이터 결과를 해석할 때 분석 모형의 종류(분류 모형, 회귀 모형)에 따라 다른 평가 지표를 사용한다.**
- **보통의 경우 예측 모형을 평가할 때 평가 지표의 정확도를 95% 수준으로 설정하여 평가한다.**

1) 분석 모형 설정

- 편향(Bias)은 학습 알고리즘에서 잘못된 가정을 했을 때 발생하는 오차로 예측값과 실제값의 차이고, 분산(Variance)은 훈련 데이터(Training Set)의 내재된 작은 변동으로 발생하는 오차로 데이터의 흩어진 정도이다.



편향과 분산의 관계

4. 분석 모형 평가 및 개선 – 분석 모형 평가

1) 분석 모형 설정

- 이상적인 분석 모형은 낮은 편향(Bias)과 낮은 분산(Variance)으로 설정되어야 한다.
- 예를 들어 예측값들과 정답이 멀리 떨어져 있는 경우 결과의 편향이 높다고 해석하고, 예측값들이 정답과 멀리 떨어져 흩어져 있는 경우 분산이 높다고 해석한다.

2) 분석 모형 평가 방법

- 분석 모형 평가 방법은 종속변수 유형에 따라 다르다.

종속변수 유형에 따른 모형 평가 방법

종속변수 유형	주요 분석 모형 평가 방법
범주형	혼동행렬(Confusion Matrix)
연속형	RMSE(평균제곱근오차, Root Mean Squared Error)

혼동행렬 : 분석 모델에서 구한 분류의 예측 범주와 데이터의 실제 분류 범주를 교차표 형태로 정리한 것이다.
평균제곱근오차(RMSE: Root Mean Squared Error) : 평균제곱오차(MSE)에 제곱근을 씌운 값을 의미한다.

4. 분석 모형 평가 및 개선 – 분석 모형 평가

3) 회귀모형 평가 지표

● 회귀모형의 예측 결과는 예측값과 실젯값 사이의 차이를 나타내는 오차 수치로 표현되고, 오차 수치가 작을수록 예측 모형의 정확도가 높다고 할 수 있다.

① 평가 지표

▶ 회귀 모형의 성능을 평가할 때 다음 <회귀 모형 평가 지표> 오차 수치를 활용한다.

▶ 지표 계산 시 오차 값이 상쇄되지 않게 하기 위해 오차를 제곱하거나 절댓값을 취해 계산한다.

회귀 모형 평가 지표

명칭	설명	수식
평균절대오차 (MAE: Mean Absolute Error)	모델의 실젯값과 예측값 차이에 절댓값을 취하여 평균한 값	$\frac{1}{n}\sum_{i=1}^{n}\lvert y_i-\hat{y}_i\rvert$
평균제곱오차 (MSE: Mean Squared Error)	모델의 실젯값과 예측값 차이를 제곱하여 평균한 값	$\frac{1}{n}\sum_{i=1}^{n}(y_i-\hat{y}_i)^2$
평균제곱근오차 (RMSE: Roof Mean Squared Error)	• 평균제곱오차(MSE)에 제곱근을 씌운 값 • MSE는 값이 커지는 경향이 있으므로 제곱근을 씌운 RMSE를 실무에서 일반적으로 사용한다.	$\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}$
평균절대백분율오차 (MAPE: Mean Absolute Percentage Error)	• 평균절대오차(MAE)를 퍼센트로 변환한 값 • 다른 변수 사이의 오차를 비교할 수 있다.	$\frac{100}{n}\times\sum_{i=1}^{n}\left\lvert\frac{y_i-\hat{y}}{y_i}\right\rvert$

4. 분석 모형 평가 및 개선 – 분석 모형 평가

3) 회귀모형 평가 지표

② 결정계수(Coefficient of determination, R^2)

▶ 결정계수는 선형회귀 모형의 성능 검증 지표로 많이 사용되고, 회귀 모형의 예측값이 실제값과 얼마나 유사한지를 나타내는 지표이다.

▶ 결정계수는 0~1의 범위를 갖고, 결정계수 값이 1에 가까울수록 모형의 설명력이 높다고 할 수 있다.

▶ 결정계수의 수식

$$R^2 = \frac{\text{회귀제곱합}}{\text{전체제곱합}} = \frac{SSR}{SST} = \frac{SSR}{SSR+SSE} = \left(1 - \frac{SSE}{SST}\right) \quad (0 \le R^2 \le 1)$$

구분	설명	수식
SST(Total Sum of Squares)	•전체 제곱합 •실제 관측값(y_i)과 표본의 평균값($\bar{y}$)과의 차이(편차)를 제곱하여 더한 값	$\sum_{i=1}^{n}(y_i - \bar{y})^2$
SSR(Regression Sum of Squares)	•회귀 제곱합 •예측값($\hat{y}_i$)과 평균값($\bar{y}$)의 차이를 제곱하여 더한 값	$\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$
SSE(Error Sum of Squares)	•오차 제곱합 •실젯값(y_i)과 예측값($\hat{y}_i$)의 차이를 제곱하여 더한 값	$\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$

4. 분석 모형 평가 및 개선 – 분석 모형 평가

4) 분류 모형 평가 지표

● 분류 모형의 예측 결과는 참(True) 혹은 거짓(False)으로 나타내므로 분류 모형의 예측값이 실젯값과 많이 일치할수록 예측 모델의 설명력이 높다고 할 수 있다.

① 혼동행렬(Confusion Matrix)

▶ 혼동행렬은 분석 모델에서 구한 분류의 예측 범주와 데이터의 실제 분류 범주를 교차표 형태로 정리한 것이다.

혼동행렬

		예측 범줏값	
		Predicted Positive	Predicted Negative
실제 범줏값	Actual Positive	True Positive (TP)	False Negative (FN)
	Actual Negative	False Positive (FP)	True Negative (TN)

4. 분석 모형 평가 및 개선 – 분석 모형 평가

4) 분류 모형 평가 지표

② 혼동행렬을 통한 분류 모형의 평가 지표

분류 모형 평가 지표

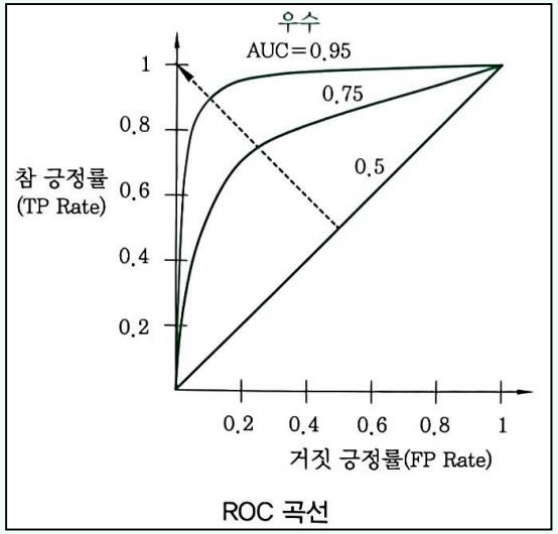
명칭	설명	수식
정확도(Accuracy) =정 분류율	전체 범주 중 정확히 예측한 비율	$\frac{TP+TN}{TP+TN+FP+FN}$
오차 비율(Error Rate)	전체 범주 중 잘못 예측한 비율	$\frac{FP+FN}{TP+TN+FP+FN}$
참 긍정률(TP Rate) =재현율(Recall) =민감도(Sensitivity)	실제 '긍정' 범주 중 '긍정'의 비율	$\frac{TP}{TP+FN}$
특이도(Specificity)	실제 '부정' 범주 중 '부정'의 비율	$\frac{TN}{TN+FP}$
거짓 긍정률(FP Rate)	실제 '부정' 범주 중 '긍정'의 비율	$\frac{FP}{TN+FP}$
정밀도(Precision)	예측 '긍정' 범주 중 '긍정'의 비율	$\frac{TP}{TP+FP}$
F-Measure (F1-Score)	0~1 사이의 범위를 가짐	$2\times\frac{\text{Precision}\times\text{Recall}}{\text{Precision}+\text{Recall}}$

4. 분석 모형 평가 및 개선 – 분석 모형 평가

4) 분류 모형 평가 지표

③ ROC(Receiver Operating Characteristic Curve)

- ROC 곡선은 가로축(x)을 혼동행렬의 거짓 긍정률(FP Rate)로 두고, 세로축(y)을 참 긍정률(TP Rate)로 두어 시각화한 그래프이다.
- ROC 곡선은 가능한 모든 임계값(threshold)에 대한 거짓 긍정률(FPR)과 참 긍정률(TPR)의 비율을 표현한다.
- 그래프가 왼쪽 꼭대기에 가까울수록 분류 성능이 우수하다고 할 수 있다.
- AUC(Area Under the ROC Curve)는 진단의 정확도를 측정할 때 사용하는 것으로 ROC 곡선 아래의 면적을 모형의 평가 지표로 삼는다.
- AUC 값은 항상 0.5~1의 값을 가지며, 1에 가까울수록 좋은 모형이라고 평가한다.



ROC 곡선

4. 분석 모형 평가 및 개선 – 분석 모형 평가

개념 체크

01 다음 중 회귀 모형 평가 지표에 속하지 않는 것은?

① MAE ② MSE

③ MAPE ④ TMSE

TMSE란 평가지표는 없다.

회귀 모형 평가 지표

1. 평균절대오차(MAE; Mean Absolute Error)

● 모델의 실젯값과 예측값 차이에 절댓값을 취하여 평균한 값을 의미한다.

● 수식 : $\frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$

2. 평균제곱오차(MSE; Mean Squared Error)

● 모델의 실젯값과 예측값 차이를 제곱하여 평균한 값이다.

● 수식 : $\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$

3. 평균제곱근오차(RMSE; Root Mean Squared of Error)

● 평균제곱오차(MSE)에 제곱근을 씌운 값

● MSE는 값이 커지는 경향이 있으므로 제곱근을 씌운 RMSE를 실무에서 일반적으로 사용한다.

02 다음 결정계수(R^2)에 대한 설명 중 옳지 않은 것은?

① 결정계수의 범위는 -1~1이다.

② 결정계수 검정 요소에는 SST, SSR, SSE가 포함된다.

③ SST는 전체제곱합으로 수식은 $\sum_{i=1}^{n}(y_i - \bar{y})^2$ 와 같다.

④ SSR은 회귀제곱합으로 수식은 $\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$ 와 같다.

결정계수(Coefficient of Determination, R^2)

● 결정계수는 선형회귀 모형의 성능 검증 지표로 많이 사용되고, 회귀 모형의 예측값이 실제값과 얼마나 유사한지를 나타내는 지표다.

● 결정계수는 0~1의 범위를 갖고, 결정계수 값이 1에 가까울수록 모형의 설명력이 높다고 할 수 있다.

● 결정계수의 수식

$$R^2 = \frac{\text{회귀제곱합}}{\text{전체제곱합}} = \frac{SSR}{SST} = \frac{SSR}{SSR+SSE} = \left(1 - \frac{SSE}{SST}\right) \quad (0 \le R^2 \le 1)$$

SST(Total Sum of Squares)

● 전체 제곱합

● 실제 관측값과 표본의 평균값과의 차이(편차)를 제곱 하여 더한 값

4. 분석 모형 평가 및 개선 – 분석 모형 평가

03 다음과 같은 특징을 갖는 명칭은?

가로축(x)을 혼동행렬의 거짓 긍정률(FP Rate)로 두고 세로축(y)을 참 긍정률(TP Rate)로 두어 시각화한 그래프이다.

그래프가 왼쪽 꼭대기에 가까울수록 분류 성능이 우수하다고 할 수 있다.

① ROC 곡선 ② AUC

③ 이항분포 ④ 정규분포

ROC(Receiver Operating Characteristic Curve)

● 가로축(x)을 혼동행렬의 거짓 긍정률(FP Rate)로 두고 세로축(y)을 참 긍정률(TP Rate)로 두어 시각화한 그래프 이다.

● ROC 곡선은 가능한 모든 임계값에 대한 거짓 긍정률(FPR)과 참 긍정률(TPR)의 비율을 표현한다.

다음 혼동행렬을 보고 물음에 답하시오.

예측값 \ 실젯값	Positive	Negative
Positive	40	70
Negative	20	60

04 다음 혼동행렬에서 오차비율은 얼마인가?

① 10/19 ② 9/19

③ 7/19 ④ 13/19

오차비율은 전체 범주 중 잘못 예측한 비율을 의미한다.

$$\frac{FP+FN}{TP+TN+FP+FN}=\frac{70+20}{40+60+70+20}=\frac{90}{190}=\frac{9}{19}$$

4. 분석 모형 평가 및 개선 – 분석 모형 평가

05 다음 혼동행렬에서 거짓 긍정률은 얼마인가?

① 7/13 ② 6/13

③ 1/4 ④ 3/4

거짓 긍정률은 실제 부정 범주 중 긍정의 비율을 의미한다.

$$\frac{FP}{TN+FP} = \frac{70}{60+70} = \frac{70}{130} = \frac{7}{13}$$ 이 된다.

06 다음 혼동행렬에서 특이도는 얼마인가?

① 1/13 ② 3/13

③ 6/13 ④ 1/3

특이도는 실제 부정 범주 중 부정의 비율을 의미한다.

$$\frac{TN}{TN+FP} = \frac{60}{60+70} = \frac{60}{130} = \frac{6}{13}$$ 이 된다.

4. 분석 모형 평가 및 개선 – 분석 모형 평가

07 다음과 같은 혼동행렬의 빈칸에 알맞은 명칭은?

		예측 범줏값	
		Predicted Positive	Predicted Negative
실제 범줏값	Actual Positive	㉠	㉡
	Actual Negative	㉢	㉣

① ㉠ : FN ② ㉡ : TN

③ ㉢ : FP ④ ㉣ : TP

㉠은 True Positive(TP), ㉡은 False Negative(FN)

㉢은 False Positive(FP), ㉣은 True Negative(TN)

08 혼동행렬 평가 지표에서 예측 긍정 범주 중 긍정의 비율 을 나타내는 것은?

① 민감도 ② 정밀도

③ 특이도 ④ 오차 비율

분류 모형 평가 지표

1. 정확도(Accuracy, 정 분류율)

● 전체 범주 중 정확히 예측한 비율

● 수식 : $\frac{TP+TN}{TP+TN+FP+FN}$

2. 오차 비율(Error Rate)

● 전체 범주 중 잘못 예측한 비율

● 수식 : $\frac{FP+FN}{TP+TN+FP+FN}$

3. 참 긍정률(TP Rate, 재현율(Recall), 민감도(Sensitivity))

● 실제 긍정 범주 중 긍정의 비율

● 수식 : $\frac{TP}{TP+FN}$

4. 특이도(Specificity)

● 실제 부정 범주 중 부정의 비율

● 수식 : $\frac{TN}{}$

4. 분석 모형 평가 및 개선 – 분석 모형 평가

02 분석 모형 진단

1) 분석 모형 진단의 정의

- 분석 모형 진단은 분석에 사용된 데이터가 가정 및 규칙을 잘 지키고 있는지 확인하는 절차이다.
- 정확한 분석 결과를 얻기 위해서는 분석 모형에 대한 기본 가정이 제대로 이루어졌는지 사용된 분석 방법은 적합했는지에 대한 진단이 필요하다.

2) 데이터 분석 모형의 오류

① 일반화 오류(Generalization Error)

▶ 분석 모형을 만들 때 주어진 데이터의 특성이 지나치게 반영되어 발생하는 오류를 의미하고, 이를 과대적합(Over-Fitting)되었다고 표현한다.

② 학습 오류(Training Error)

▶ 분석 모형을 만들 때 주어진 데이터의 특성을 지나치게 덜 반영하여 발생하는 오류를 의미하고, 이를 과소적합(Under-Fitting)되었다고 표현한다.

4. 분석 모형 평가 및 개선 – 분석 모형 평가

03 교차 검증

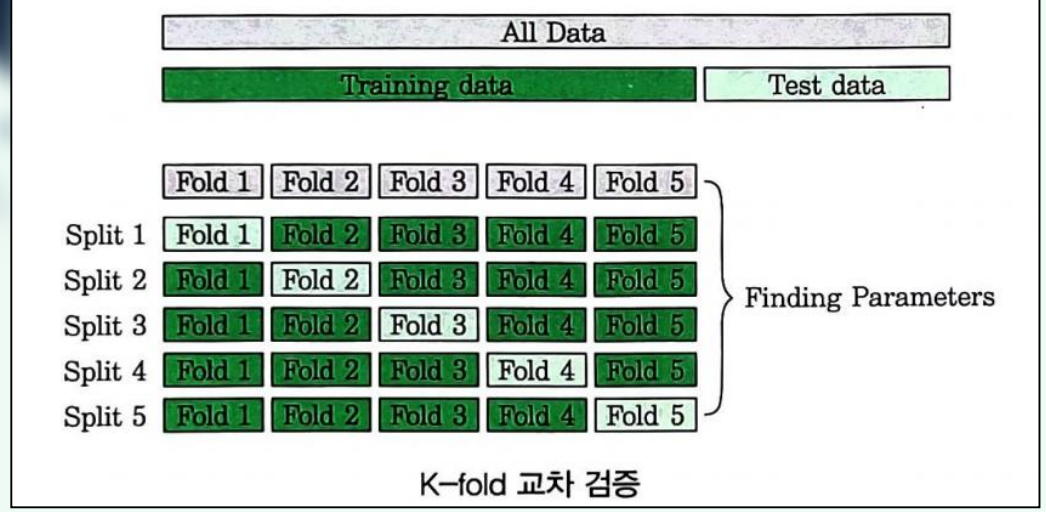
- 교차검정은 예측 모델의 정확도를 높이기 위해 데이터를 훈련, 평가 데이터로 나누어 여러 차례 검증하는 방법이다.
- 교차 검증의 목적은 과적합을 피하고, 매개변수를 튜닝하여 일반적인 모델을 만들고, 더욱 신뢰성 있는 모델 평가를 하기 위해서이다.
- 데이터를 분할하여 일부는 분석 모형 학습에 사용하고, 나머지는 모델의 검증에 사용하는 방법을 여러 차례 반복 수행하고, 이를 통해 분석 모형이 새로운 데이터에 대해 일반화된 성능을 보일 수 있을지 확인한다.

4. 분석 모형 평가 및 개선 – 분석 모형 평가

03 교차 검증

교차 검증 방법

방법	설명
K-fold 교차 검증 (K-fold cross validation)	• 학습 데이터를 K개의 그룹(fold)으로 나누어 (K-1)개는 학습에, 나머지 하나는 검증에 사용하는 방법이다. • 방법 : 테스트 데이터를 제외한 데이터를 무작위로 중복되지 않는 K개의 데이터로 분할 → K-1개의 데이터를 학습 데이터로 사용하고, 나머지 1개 데이터를 검증 데이터로 사용 → 검증 데이터를 바꾸며 K번 반복해서 분할된 데이터가 한 번씩 검증 데이터로 사용된다. • LOOCV보다 연산량이 작고, 중간 정도의 편향(Bais)과 분산(Variance)을 가진다.



K-fold 교차 검증

홀드 아웃 (Hold-out) 교차 검증	• 데이터를 무작위로 7 : 3 또는 8 : 2 비율로 학습 데이터와 검증 데이터로 나누는 방법이다. • 가장 보편적으로 랜덤 추출을 통해 데이터를 분할하는 방법으로 학습 데이터와 검증 데이터가 60~80%이고, 테스트 데이터가 20~40%이다.
LOOCV (Leave-One-Out Cross Validation)	• N개 데이터 중 1개만 평가 데이터로 사용하고, 나머지 N-1개는 훈련 데이터로 사용하는 과정을 N번 반복하는 방법이다. • 많은 데이터를 훈련 데이터로 활용할 수 있지만 계산량이 많아 실행 시간이 오래 걸린다. • 낮은 편향(Bais)과 높은 분산(Variance)을 가진다.
LpOCV (Leave-p-Out Cross Validation)	데이터 중 p개의 관측치를 검증 데이터로 사용하고, 나머지는 학습 데이터로 사용하는 방법이다.
부트스트랩 (Bootstrap)	주어진 자료에서 단순 랜덤 복원추출 방법을 활용해 동일한 크기의 표본을 여러 개 생성하는 방법이다.

편향(Bias) : 학습 알고리즘에서 잘못된 가정을 했을 때 발생하는 오차로 예측값과 실제값의 차이
분산(Variance) : 훈련 데이터(Training Set)의 내재된 작은 변동 으로 발생하는 오차로 데이터의 흩어진 정도

4. 분석 모형 평가 및 개선 – 분석 모형 평가

04 모수 유의성 검정

1) 모집단과 모수

- 모집단(Population)은 연구자가 연구를 통해 실제로 알고 싶은 전체 집단을 의미한다. 예를 들어 초등학생 6학년 평균 신장을 조사하고자 하는 경우 국내 전체 초등학교 6학년 학생들의 키가 모집단이 된다.
- 모수(Population Parameter)란 모집단을 조사하여 얻을 수 있는 통계적인 특성 수치를 의미하고, 모집단 분포의 특성을 규정짓는 척도가 된다. 예를 들어 모평균, 모분산, 모표준편차 등이 있다.
- 표본(Sample)이란 모집단에 대한 분석을 위해 표집되는 부분 집합을 의미한다.

2) 모집단에 대한 유의성 검정

- 모집단에 대한 유의성 검정 방법에는 Z-검정, T-검정, 분산 분석, 카이제곱 검정, F-검정 방법이 있다.

4. 분석 모형 평가 및 개선 – 분석 모형 평가

2) 모집단에 대한 유의성 검정

모집단에 대한 유의성 검정 방법

검정 방법	설명
Z-검정 (Z-Test)	• 정규분포를 가정하고, 추출된 표본이 동일 모집단에 속하는지 가설을 검증하기 위해 사용된다. • 분산 또는 표준편차를 알고 있는 경우 사용된다. $Z_0 = \left\| \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \right\|$ Z_0 : Z통계량 $\bar{X}$: 표본 평균 μ : 기대 평균 σ : 표준편차 n : 표본 개수
T-검정 (T-Test)	모집단의 분산이나 표준편차를 알지 못할 때, 표본으로부터 추정된 분산이나 표준편차를 이용하여 두 모집단의 평균의 차이를 알아보는 검정 방법이다.
분산 분석 (ANOVA, Analysis of Variance)	• 두 개 이상의 집단 평균 차이를 비교할 때 사용하는 가설 검정 방법이다. • T-검정에서 집단이 두 개 이상인 경우 분산 분석 방법을 사용한다.
공분산 분석 (ANCOVA, Analysis of Covariance)	분산 분석과 회귀 분석을 결합한 모형으로 독립변수가 범주형이고, 종속변수가 연속형일 경우 사용하는 분석 방법이다.
카이제곱 검정 (Chi-Squared Test)	• 어떤 그룹이 서로 독립인지 아닌지 확인하는 방법으로 범주형 데이터에서 사용된다. • 데이터가 예상 분포에 얼마나 잘 맞는지 확인하는 방법으로 모집단이 정규분포를 따르며, 분산을 알고 있는 경우에 사용된다.
F-검정 (F-Test)	두 모집단의 분산의 차이가 있는지를 검정하는 방법으로, F-값이 클수록 두 집단 간의 분산 차이가 존재하는 것을 의미한다.

범주형 데이터 : 카테고리 A, B, C와 같이 종류를 표시하는 데이터

4. 분석 모형 평가 및 개선 – 분석 모형 평가

개념 체크

01 다음 중 주어진 데이터에서 p개의 관측치를 검증 데이터로 사용하고, 나머지는 학습 데이터로 사용하는 방법은?

① LOOCV

② LpOCV

③ Hold-out 교차 검증

④ K-fold 교차 검증

교차검증 방법

1. K-fold 교차검증

● 학습 데이터를 K개의 그룹(fold)으로 나누어 (K-1)개는 학습에, 나머지 하나는 검증에 사용하는 방법이다.

● 방법 : 테스트 데이터를 제외한 데이터를 무작위로 중복되지 않는 K개의 데이터로 분할 -> (K-1)개의 데이터를 학습 데이터로 사용하고, 나머지 1개 데이터를 검증 데이터 로 사용 -> 검증 데이터를 바꾸며 K번 반복해서 분할된 데이터가 한 번씩 검증 데이터로 사용된다.

● LOOCV보다 연산량이 작고, 중간 정도의 편향(Bias)과 분산(Variance)을 가진다.

02 다음 설명에서 빈칸에 알맞은 명칭은?

(㉠)는(은) 모집단을 조사하여 얻을 수 있는 통계적인 특성 수치를 의미하고, 모집단 분포의 특성을 규정짓는 척도가 된다.

① 표본　② 표준편차

③ 분산　④ 모수

● 모수(Population Parameter)란 모집단을 조사하여 얻을 수 있는 통계적인 특성 수치를 의미하고, 모집단 분포의 특성을 규정짓는 척도가 된다. 예를 들면, 모평균, 모분산, 모표준편차 등이 있다.

● 표본(Sample)이란 모집단에 대한 분석을 위해 표집되는 부분 집합이다.

4. 분석 모형 평가 및 개선 – 분석 모형 평가

03 다음과 같은 특징을 갖는 검정 방법은?

모집단의 분산이나 표준편차를 알지 못할 때, 표본으로부터 추정된 분산이나 표준편차를 이용하여 두 모집단의 평균의 차이를 알아보는 검정 방법이다.

① F-검정 ② Z-검정

③ T-검정 ④ 카이제곱 검정

1. Z-검정

● 정규분포를 가정하고, 추출된 표본이 동일 모집단에 속하는지 가설을 검증하기 위해 사용된다.

● 분산 또는 표준편차를 알고 있는 경우 사용된다.

● 수식 : $Z_0 = \left|\frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}}\right|$

2. T-검정

● 모집단의 분산이나 표준편차를 알지 못할 때, 표본으로부터 추정된 분산이나 표준편차를 이용하여 두 모집단 의 평균의 차이를 알아보는 검정 방법이다.

3. 분산 분석(ANOVA)

● 두 개 이상의 집단 평균 차이를 비교할 때 사용하는 가설

04 다음 중 학습된 데이터가 충분하지 않아 학습 데이터의 구조 및 패턴을 정확히 확인하지 못하는 경우를 뜻하는 명칭은?

① 과소표집 ② 과대표집

③ 과소적합 ④ 과대적합

일반화 오류

● 분석 모형을 만들 때 주어진 데이터의 특성이 지나치게 반영되어 발생하는 오류를 의미하고, 이를 **과대적합(Over-Fitting)**되었다고 한다.

학습 오류

● 분석 모형을 만들 때 주어진 데이터의 특성을 지나치게 덜 반영되어 발생하는 오류를 의미하고, 이를 **과소적합 (Under-Fitting)**되었다고 한다.

4. 분석 모형 평가 및 개선 – 분석 모형 평가

05 다음 중 연구자가 연구를 통해 실제로 알고 싶은 전체 집단을 의미하는 용어는?

① 표본 ② 모수

③ 모집단 ④ 표본집단

06 다음 중 K-fold 교차 검증에 대한 설명으로 옳지 않은 것은?

① 데이터를 K개의 fold로 나눈다.

② 데이터를 K-1개는 검증 데이터로, K개는 학습 데이터에 사용한다.

③ 검증 데이터를 바꾸며 K번 반복하므로 분할된 데이터가 한 번씩 한 번씩 검증 데이터로 사용된다.

④ K-fold 교차 검증은 교차 검증 방법 중 하나로 교차 검증은 데이터를 훈련 데이터와 평가 데이터로 나누어 여러 차례 검증하는 방법이다.

K-fold 교차 검증에서는 데이터를 K개의 fold로 나누고, K-1개는 학습 데이터로 하고 나머지 하나를 검증 데이터로 사용한다.

4. 분석 모형 평가 및 개선 – 분석 모형 평가

07 다음 중 교차 검증 방법에 속하지 않는 것은?

① K-fold 교차 검증

② K-means clustering

③ 부트스트랩

④ LOOCV

교차 검증 방법에는 K-fold 교차 검증, 홀드 아웃(Hold out) 교차 검증, LOOCV, LpOCV, 부트스트랩(Bootstrap)이 있다.
K-means Clustering(K-평균 군집)은 그룹을 할당해서 군집화하는 비지도 학습 알고리즘이다.

08 다음에서 설명하는 교차 검증 기법은?

데이터를 무작위로 7:3 또는 8:2 비율로 학습 데이터와 검증 데이터로 나누는 방법이다.
가장 보편적으로 랜덤 추출을 통해 데이터를 분할하는 방법으로 학습 데이터와 검증 데이터가 60~80%이고, 테스트 데이터가 20~40%이다.

① Hold-out 교차 검증

② LOOCV

③ LpOCV

④ K-fold 교차 검증

4. 분석 모형 평가 및 개선 – 분석 모형 평가

05 적합도 검정

● 적합도 검정(Goodness of Fit Test)은 가정된 확률이 정해져 있을 때와 가정된 확률이 정해져 있지 않을 때 데이터가 가정된 확률에 적합하게 따르고 있는가를 검정하는 것이다.

● 가정된 확률이 정해져 있는 경우에는 카이제곱 검정을 이용하여 검정을 수행하고, 가정된 확률이 정해져 있지 않은 경우에는 정규성 검정(Normality Test)을 사용하여 검정한다.

1) 가정된 확률 검정

① 카이제곱 검정(Chi-Squared Test)

▶ 카이제곱 검정 방법은 어떤 그룹이 서로 독립인지 아닌지 확인하는 방법으로 카이제곱 검정 유형으로 독립성 검정, 적합성 검정, 동질성 검정이 있다.

▶ R언어에서 chisq.test() 함수를 사용하여 나온 결과의 $p-value$가 0.05보다 큰 경우 관측된 데이터가 가정된 확률을 따른다고 할 수 있고, 이 경우 귀무가설(Ho)을 채택한다.

4. 분석 모형 평가 및 개선 – 분석 모형 평가

2) 정규성 검정

① 샤피로-윌크 검정(Shapiro-Wilk Test)

- ▶ 샤피로-윌크 검정은 데이터가 정규분포를 따르는지 확인하는 검정방법이다.
- ▶ R언어에서 sharpiro.test() 함수를 사용하여 검정하며, $p-value$가 0.05보다 작은 경우 귀무가설(Ho)을 기각하고, 대립가설(H_1)을 채택한다.
- ▶ 다만 언어의 sharpiro.test() 함수를 사용하는 경우 데이터의 수는 3개에서 5,000개 이하로만 사용 가능하다.

② 콜모고로프-스미르노프 적합성 검정(Kolmogorov-Smirnov Goodness of Fit Test, K-S검정)

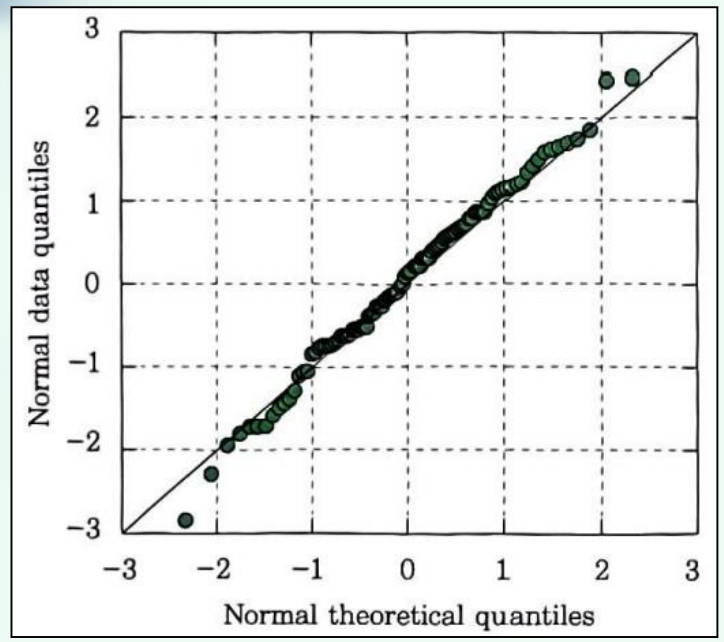
- ▶ 콜모고로프-스미르노프 적합성 검정은 데이터의 누적 분포 함수와 비교하고자 하는 분포의 누적 분포 함수 간의 최대 거리를 통계량으로 사용하는 가설 검정 방법이다.
- ▶ R언어에서 ks.test() 함수를 사용하여 검정하며, $p-value$가 0.05보다 작은 경우 귀무가설(Ho)을 기각하고, 대립가설(H_1)을 채택한다.

4. 분석 모형 평가 및 개선 – 분석 모형 평가

2) 정규성 검정

③ Q-Q plot

- ▶ Q-Q plot은 그래프를 통해 정규성 가정을 시각적으로 검정하는 방법이다.
- ▶ Q-Q plot에서 대각선 참조선을 따라서 데이터가 분포할 경우 정규성 가정을 만족한다고 할 수 있다.
- ▶ Q-Q plot 해석은 주관적일 수 있기 때문에 보조 수단으로 사용하는 것이 좋다.



4. 분석 모형 평가 및 개선 – 분석 모형 평가

개념 체크

01 다음 중 그래프를 통해 정규성 가정을 시각적으로 검정하는 방법은?

① 산점도 ② Q-Q plot

③ 히스토그램 ④ Box-plot

Q-Q plot

● 그래프를 통해 정규성 가정을 시각적으로 검정하는 방법

● 대각선 참조선을 따라서 데이터가 분포할 경우 정규성 가정을 만족한다고 할 수 있다.

● 해석이 주관적일 수 있기에, 보조 수단으로 사용하는 것이 좋다.

산점도(Scatter Plot)

● 가로축과 세로축의 좌표 평면상에서 각각의 관측값들을 표시하는 시각화 방법이다. 2개의 연속형 변수 간의 관계를 보기 위해서 사용한다.

히스토그램(Histogram)

● 히스토그램은 자료의 분포가 직사각형 형태이다.

● 가로축은 수치형 데이터이다.

02 다음과 같은 특징을 갖는 정규성 검정 방법은?

데이터의 누적 분포 함수와 비교하고자 하는 분포의 누적 분포 함수 간의 최대 거리를 통계량으로 사용하는 가설 검정 방법이다.

R언어에서 ks.test() 함수를 사용하여 검정하며, $p-value$가 0.05보다 작은 경우 귀무가설(Ho)을 기각하고, 대립가설(H_1)을 채택한다.

① 콜모고로프-스미르노프 적합성 검정

② 샤피로-윌크 검정

③ Q-Q plot

④ 카이제곱 검정

샤피로-윌크 검정

● 샤피로-윌크 검정은 데이터가 정규분포를 따르는지 확인하는 검정방법이다.

● R언어에서 sharpiro.test()함수를 사용하여 검정하며, p-value가 0.05보다 작은 경우 귀무가설을 기각하고, 대립가설을 채택한다.

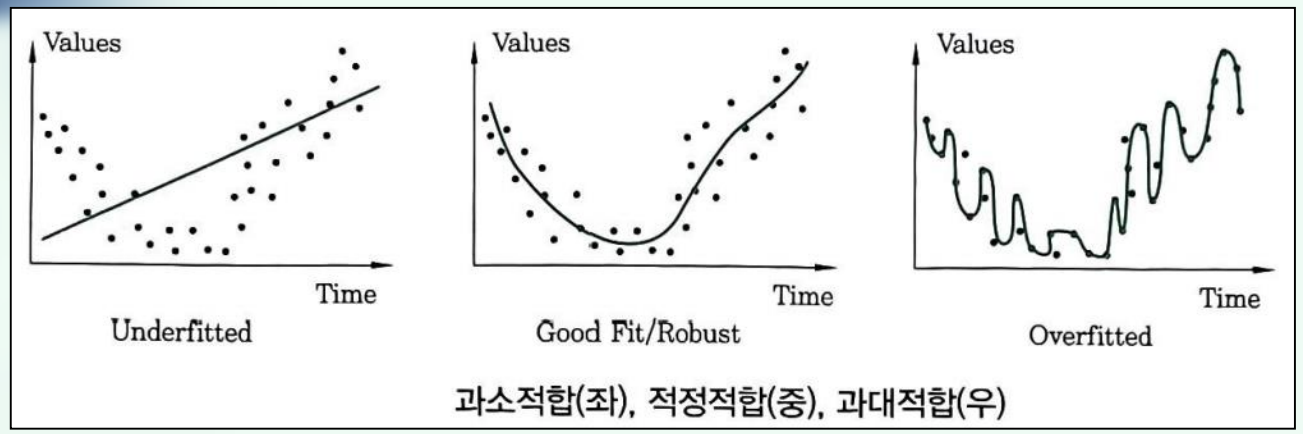
● 다만 언어의 sharpiro.test()함수를 사용하는 경우 데이터의

4. 분석 모형 평가 및 개선 – 분석 모형 개선

01 과대적합 방지

1) 과대적합(Over-Fitting)과 과소적합(Under-Fitting)

- 과대적합(Over-Fitting)이란 학습 모델을 지나치게 복잡하게 학습하여 학습 데이터셋에서는 모델 성능이 높지만, 새로운 데이터가 주어진 경우 정확도가 낮아지는 경우를 의미한다.
- 과소적합(Under-Fitting)이란 학습된 데이터가 충분하지 않아 학습 데이터의 구조 및 패턴을 정확히 확인하지 못하는 경우를 의미한다.



과소적합(좌), 적정적합(중), 과대적합(우)

4. 분석 모형 평가 및 개선 – 분석 모형 개선

2) 과대적합 방지 방법

● 과대적합 방지 방법으로는 데이터 증강, 모델의 복잡도 감소, 가중치 규제 적용, 드롭아웃이 있다.

① 데이터 증강(Data Augmentation)

▶ 데이터의 개수가 적을 경우 지나치게 세세한 학습이 진행될 수 있기 때문에 과적합을 유발할 수 있어 데이터를 증강시켜 데이터 분석을 위한 충분한 데이터셋을 확보해야 한다.

▶ 데이터의 양이 적을 경우 데이터 변형, 데이터 표집 등의 방법을 활용하여 데이터의 수를 늘릴 수 있다.

② 모델의 복잡도 감소

▶ 모델의 복잡도가 높은 경우 데이터 과대적합의 위험이 있다.

▶ 이 경우 모델의 복잡도와 관련되는 인공신경망 은닉층 수 감소, 매개변수의 수 조절 등의 방법으로 모델의 복잡도를 감소시킬 수 있다.

4. 분석 모형 평가 및 개선 – 분석 모형 개선

2) 과대적합 방지 방법

③ 가중치 규제 적용

▶ 가중치 규제(Weight Regularization)란 가중치의 값을 제한하여 모델의 복잡도를 간단하게 만드는 것을 의미한다.

▶ 가중치 규제의 종류에는 라쏘(L1 노름 규제), 릿지(L2 노름 규제), 엘라스틱 넷이 있다.

가중치 규제의 종류

구분	설명
L1 노름 규제 (라쏘, Lasso Regression)	• 기존 비용 함수에 모든 가중치(w)들의 절댓값 합계를 추가하여 값이 최소가 되도록 하는 방법 • 불필요한 회귀 계수를 급격하게 감소시켜 0으로 만들고 제거하는 방법 $\frac{1}{N}\sum_{i=1}^{N}(y_i-\hat{y}_i)^2+\lambda\sum_{j=1}^{M}\lvert w_j\rvert$ ($\frac{1}{N}\sum_{i=1}^{N}(y_i-\hat{y}_i)^2$: 기존 비용 함수, $\lambda\sum_{j=1}^{M}\lvert w_j\rvert$: 절댓값 합계, λ :규제의 강도를 정하는 초매개 변수, y : 실젯값, $\hat{y}$: 예측값, w : 가중치)
L2 노름 규제 (릿지, Ridge Regression)	• 기존 비용 함수에 모든 가중치(w)들의 제곱합을 추가하는 방법 • 회귀 계수의 크기를 감소시키는 방법 $\frac{1}{N}\sum_{i=1}^{N}(y_i-\hat{y}_i)^2+\frac{\lambda}{2}\sum_{j=1}^{M}\lvert w_j\rvert^2$ ($\frac{1}{N}\sum_{i=1}^{N}(y_i-\hat{y}_i)^2$: 기존 비용 함수, $\frac{\lambda}{2}\sum_{j=1}^{M}\lvert w_j\rvert^2$: 제곱합)

엘라스틱 넷 (Elastic Net)	기존 비용 함수에 L1 노름 규제와 L2 노름 규제를 결합한 방법 $\frac{1}{N}\sum_{i=1}^{N}(y_i-\hat{y}_i)^2+\alpha\sum_{j=1}^{M}\lvert w_j\rvert^1+\beta\sum_{j=1}^{M}\lvert w_j\rvert^2$ ($\frac{1}{N}\sum_{i=1}^{N}(y_i-\hat{y}_i)^2$: 기존 비용 함수, $\alpha\sum_{j=1}^{M}\lvert w_j\rvert^1$: L1 규제, $\beta\sum_{j=1}^{M}\lvert w_j\rvert^2$: L2 규제)

노름(norm) : 가중치 벡터의 길이 또는 크기를 측정하는 방법
비용 함수(Cost Function) : 실제값과 가장 오차가 작은 가설 함수를 도출하기 위해 사용되는 함수로 예측값에서 실제값의 차의 제곱의 평균과 같음

4. 분석 모형 평가 및 개선 – 분석 모형 개선

2) 과대적합 방지 방법

④ 드롭아웃(Drop Out)

- 드롭아웃은 학습 과정에서 신경망 일부를 사용하지 않는 방법이다.
- 드롭아웃은 서로 연결된 연결망에서 0~1 사이의 확률(Drop Out Rate)로 뉴런을 제거하는 방법이다.
- 제거되는 신경망의 종류와 개수는 랜덤하게 드롭아웃 확률(Drop Out Rate)에 의해 결정된다.
- 드롭아웃은 신경망 학습 시에만 사용하고, 예측 시에는 사용하지 않는다.
- 드롭아웃의 유형에는 초기 드롭아웃, 공간적 드롭아웃, 시간적 드롭아웃이 있다.

종류	설명
초기 드롭아웃	•학습과정에서 노드들을 p 의 확률(보통의 경우 0.5)로 학습 횟수마다 임의로 생략하고 남은 노드들과의 연결선들만을 이용하여 학습 및 추론 하는 방법 •심층신경망(DNN)에서 사용된다.
공간적 드롭아웃	•Feature Map 내의 노드 전체에 대해 드롭아웃의 적용 여부를 결정하는 방법 •합성곱신경망(CNN)에서 사용된다.
시간적 드롭아웃	•노드들을 생략하지 않고 노드들의 연결선 일부를 생략하는 방법 •순환신경망(RNN)에서 사용된다.

4. 분석 모형 평가 및 개선 – 분석 모형 개선

개념 체크

01 다음 중 과대적합 방지 방법 중 하나로 학습과정에서 신경망 일부를 사용하지 않는 방법을 의미하는 용어는?

① 가지치기　　② 드롭아웃

③ 데이터 분할　　④ 데이터 삭제

드롭아웃(Drop Out)

● 드롭아웃은 학습 과정에서 신경망 일부를 사용하지 않는 방법이다.

● 드롭아웃은 서로 연결된 연결망에서 0~1 사이의 확률로 뉴런을 제거하는 방법이다.

● 제거되는 신경망의 종류와 개수는 랜덤하게 드롭아웃 확률(Drop Out Rate)에 의해 결정된다.

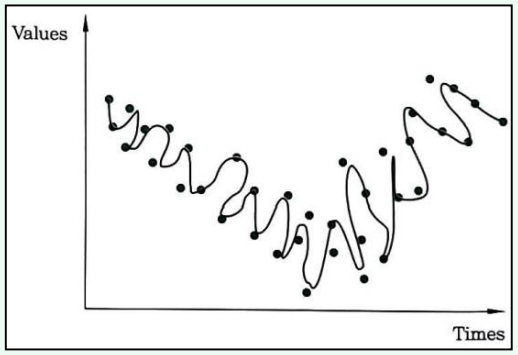
● 드롭아웃은 신경망 학습 시에만 적용하고, 예측 시에는 사용하지 않는다.

● 드롭아웃의 유형에는 초기 드롭아웃, 공간적 드롭아웃, 시간적 드롭아웃이 있다.

가지치기(Prunning)

● 의사결정나무에서 가지를 생성하는 과정을 가지

02 다음 그래프와 같은 데이터 형태에 해당하는 것은?



① 과소적합　　② 과대적합

③ 과대표집　　④ 과소표집

● **과대적합**(Over-Fitting)이란 학습 모델을 지나치게 복잡하게 학습하여 학습 데이터셋에서는 모델 성능이 높지만, 새로운 데이터가 주어진 경우 정확도 되려 낮아지는 경우를 의미한다.

● **과소적합**(Under-Fitting)이란 학습된 데이터가 충분하지 않아서 학습 데이터의 구조 및 패턴을 정확히 확인하지 못하는 경우를 의미한다.

● **과대표집**(Over-Sampling)이란 소수 클래스의 데이터를 복제 또는 생성하여 데이터 비율을 맞추는 방법이다.

과적합 가능성이 존재하고, 알고리즘 성능은 높지만, 검증

4. 분석 모형 평가 및 개선 – 분석 모형 개선

03 다음 수식의 가중치 규제 기법은?

$$\frac{1}{N}\sum_{i=1}^{N}(y_i-\hat{y}_i)^2+\frac{\lambda}{2}\sum_{j=1}^{M}|w_j|^2$$

① 엘라스틱 넷 ② 라쏘

③ 릿지 ④ 혼합 방법

L1 노름 규제(라쏘, Lasso regression)

● 기존 비용 함수에 모든 가중치(ω)들의 합계를 추가하여 값이 최소가 되도록 하는 방법

● 불필요한 회귀 계수를 급격하게 감소시켜 0으로 만들고 제거하는 방법

L2 노름 규제(릿지, Ridge regression)

● 기존 비용 함수에 모든 가중치(ω)들의 제곱합을 추가 하는 방법

● 회귀 계수의 크기를 감소시키는 방법

엘라스틱 넷(Elastic Net)

● 기존 비용 함수에 L1 노름 규제(라쏘)와 L2 노름 규제(릿지)를 결합한 방법

4. 분석 모형 평가 및 개선 – 분석 모형 개선

02 매개변수 최적화

1) 매개변수 최적화(Parameter Optimization)의 정의

- 매개변수(parameter)는 함수를 호출할 때 인수로 전달된 값을 함수 내부에서 사용할 수 있게 해주는 변수를 말한다.
- 분석 모형의 결괏값과 실젯값 차이를 손실함수라고 하고, 손실함수를 최소화하는 매개변수(가중치, 편향)를 찾아가는 과정을 매개변수 최적화라고 한다.

2) 매개변수 최적화 기법

① 경사하강법(GD: Gradient Descent)

- ▶ 경사하강법이란 예측값과 실젯값의 차이인 손실함수의 크기를 최소화 시키는 매개변수(parameter)를 찾는 방법이다.
- ▶ 경사하강법에는 배치 경사하강법, 확률적 경사하강법, 미니 배치 경사하강법이 있다.

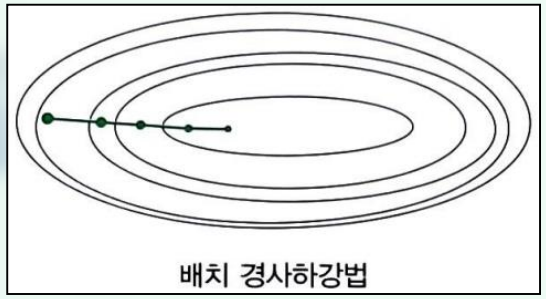
4. 분석 모형 평가 및 개선 – 분석 모형 개선

2) 매개변수 최적화 기법

① 경사하강법(GD: Gradient Descent)

▶ 배치 경사하강법(BGD: Batch Gradient Descent)

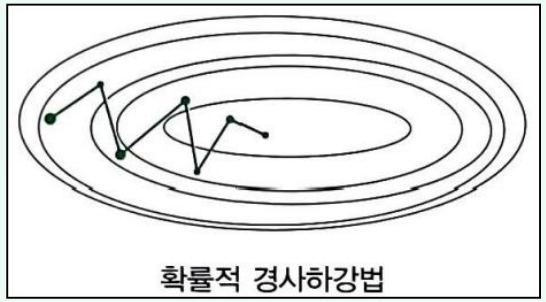
- 전체 학습 데이터를 하나의 배치(batch, 데이터 소분단위)로 묶어 학습시키는 방법



배치 경사하강법

▶ 확률적 경사하강법(SGD: Stochastic Gradient Descent)

- 전체 데이터 중 단 하나의 데이터를 사용하여 경사하강법을 1회(batch size=1) 진행하는 방법



확률적 경사하강법

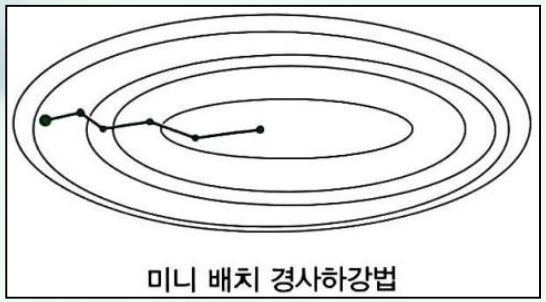
4. 분석 모형 평가 및 개선 – 분석 모형 개선

2) 매개변수 최적화 기법

① 경사하강법(GD: Gradient Descent)

▶ 미니 배치 경사하강법(Mini-Batch Gradient Descent)

- SGD와 BGD의 절충안으로 전체 데이터를 사용자가 정한 크기의 batch size개씩 나누고, 나뉜 배치로 학습시키는 방법



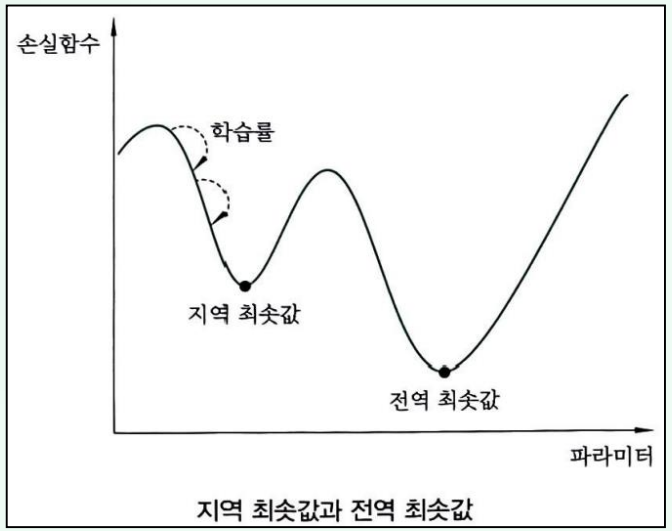
미니 배치 경사하강법

4. 분석 모형 평가 및 개선 – 분석 모형 개선

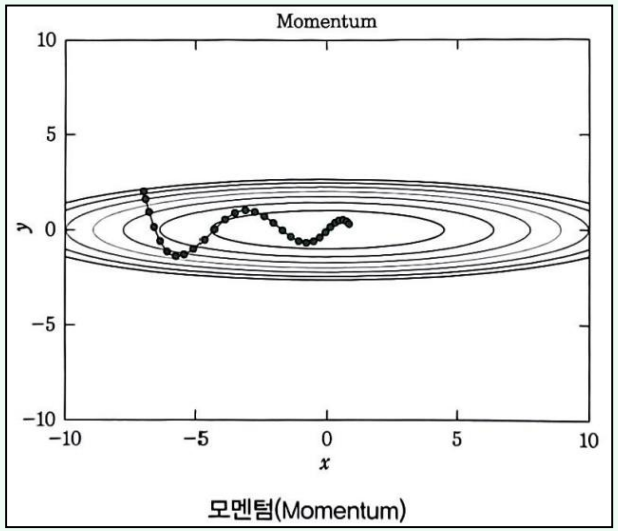
2) 매개변수 최적화 기법

② 모멘텀(Momentum)

- ▶ 모멘텀은 확률적 경사하강법의 매개변수 변경 방향에 가속도를 부여하는 방식으로 공이 구르는 듯한 모습을 보인다.
- ▶ 모멘텀에서 x의 한 방향으로 일정하게 가속하고, y축 방향의 속도는 일정하지 않다.
- ▶ 모멘텀은 확률적 경사하강법에 비해 효율적인 학습을 할 수 있고, 확률적 경사하강법이 갖는 지역 최솟값(Local Minimum)을 해결할 수 있다.



지역 최솟값과 전역 최솟값



모멘텀(Momentum)

지역 최솟값(Local Minimun) : 경사하강법 손실함수 곡선에서 작은 언덕 부위로 함수의 일부 구간의 최솟값
전역 최솟값(Global Minimum) : 경사하강법 손실함수 곡선에서 갖는 큰 언덕 부위로 전체 구간의 최솟값

4. 분석 모형 평가 및 개선 – 분석 모형 개선

2) 매개변수 최적화 기법

② 모멘텀(Momentum)

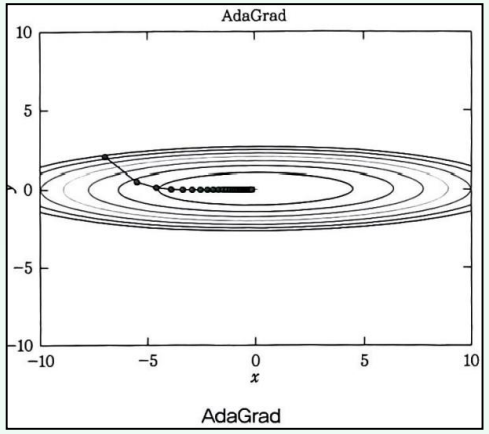
▶ 모멘텀은 경사가 가파른 곳에서는 빠른 속도의 관성을 이기지 못하고 최소지점을 지나쳐서 불필요하게 재연산을 하게 되는 오버슈팅(Over Shooting) 문제가 발생할 수 있다.

③ 네스테로프 모멘텀(Nesterov momentum)

▶ 모멘텀의 오버슈팅(Over Shooting)문제를 개선한 방법으로 모멘텀 방향과 현재 위치에서의 기울기를 반영하여 계산량을 줄이고 정확도를 향상시키는 방법이다.

④ AdaGrad(Adaptive gradient)

▶ AdaGrad는 매개변수 값을 업데이트하면서 각 변수마다 학습률을 다르게 적용하는 방법이다.



오버 슈팅 : 경사가 가파른 경우 빠른 속도로 내려오다가 최소 지점을 만나면 기울기는 순간적으로 작아지지만 속도는 여전히 커서 최소 지점을 지나는 현상이다.

4. 분석 모형 평가 및 개선 – 분석 모형 개선

2) 매개변수 최적화 기법

⑤ RMSProp

▶ AdaGrad에서 최적값에 도달하기 전에 학습률이 0에 가까워지는 상황을 방지하기 위해 개선된 방법이다.

⑥ Adam

▶ 모멘텀(Momentum)과 AdaGrad가 합쳐진 방법이다.

3) 초매개변수 최적화(Hyperparameter Optimization)

● 초매개변수 최적화는 최적값이 존재하는 범위를 조금씩 줄여가면서, 최종적으로 최적값을 찾아내는 방법이다.

● 초매개변수 최적화 방법으로는 매뉴얼 탐색, 그리드 탐색, 랜덤 탐색, 베이지안 최적화가 있다.

매뉴얼 탐색	사용자가 선택한 조합에서 최적의 조합을 찾는 방법
그리드 탐색	초매개변수의 경우의 수 중에서 최적의 조합을 찾는 방법
랜덤 탐색	초매개변수의 최소, 최댓값을 정하고 정해진 범위 내에서 무작위의 값을 정해진 횟수만큼 추출하여 최적의 조합을 찾는 방법
베이지안 최적화	단순히 무작위로 반복해서 추출하는 것이 아니라 기존에 추출되어 평가된 결과를 바탕으로 향후 탐색할 범위를 조율하여 효율적으로 최적화하는 방법

4. 분석 모형 평가 및 개선 – 분석 모형 개선

개념 체크

01 다음 설명에서 빈칸에 알맞은 명칭은?

모멘텀은 경사가 가파른 곳에서는 빠른 속도의 관성을 이기지 못하고 최소 지점을 지나쳐서 불필요하게 재연산을 하게 되는 (㉠) 문제가 발생할 수 있다.

① 다중공선성

② 기울기 소실

③ 오버슈팅

④ 손실함수

모멘텀(Momentum)

● 모멘텀은 확률적 경사하강법의 매개변수 변경 방향에 가속도를 부여하는 방식으로 공이 구르는 듯한 모습을 보인다.

● 모멘텀에서 x의 한 방향으로 일정하게 가속하고, y축 방향의 속도는 일정하지 않다.

● 모멘텀은 확률적 경사하강법에 비해 효율적인 학습을 할 수 있고, 확률적 경사하강법이 갖는 지역 최솟값(Local Minimum)을 해결할 수 있다.

02 다음 중 손실함수를 최소화하는 매개변수를 찾아가는 과정을 의미하는 용어는?

① 매개변수 최소화

② 매개변수 최적화

③ 매개변수 최대화

④ 매개변수 다중화

손실함수를 최소화하는 매개변수를 찾아가는 과정은 '매개변수 최적화(Parameter Optimization)'이다.

4. 분석 모형 평가 및 개선 – 분석 모형 개선

03 다음 중 매개변수 값을 업데이트하면서 각 변수마다 학습률을 다르게 적용하는 매개변수 최적화 기법은?

① Momentum

② Adam

③ AdaGrad

④ RMSProp

AdaGrad(Adaptive gradient)

● AdaGrad는 매개변수 값을 업데이트 하면서 각 변수마다 학습률을 다르게 적용하는 방법이다.

Adam

● 모멘텀과 AdaGrad가 합쳐진 방법이다.

RMSProp

● AdaGrad에서 최적값에 도달하기 전에 학습률이 0에 가까워지는 상황을 방지하기 위해서 개선된 방법이다.

04 다음 전체 데이터 중 단 하나의 데이터를 사용하여 경사하강법을 1회 진행하는 방법은?

① 모멘텀

② 미니 배치 경사하강법

③ 배치 경사하강법

④ 확률적 경사하강법

경사하강법(GD; Gradient Descent)

● 경사하강법이란 예측값과 실젯값의 차이인 손실함수의 크기를 최소화시키는 매개변수(parameter)를 찾는 방법이다.

● 배치 경사하강법(BGD; Batch Gradient Descent)

- 전체 학습 데이터를 하나의 배치(Batch, 데이터 소분단위)로 묶어 학습하는 방법

● 확률적 경사하강법(SGD; Stochastic Gradient Descent)

- 전체 데이터 중 단 하나의 데이터를 사용하여 경사하강법 을 1회(batch size=1) 진행하는 방법은

● 미니 배치 경사하강법(Mini-Batch Gradient Descent)

- SGD와 BGD의 절충안으로 전체 데이터를 사용자가 정한 크기의 batch size개씩 나누고, 나뉘 배치로 학습시키는 방법

4. 분석 모형 평가 및 개선 – 분석 모형 개선

03 분석 모형 융합

- 분석 모형 융합은 여러 분석 모형을 결합한 것을 의미하고, 이는 앙상블(Ensemble) 모형으로 설명할 수 있다.
- 앙상블은 여러 종류의 분석 모형을 결합하여 보다 좋은 분석 모형을 만드는 것을 의미한다.
- 앙상블 방법에는 보팅(Voting), 배깅(Bagging), 스태킹(Stacking), 부스팅(Boosting)이 있다.

보팅(Voting)	여러 개의 분석 모형 결과를 조합하는 방법으로 직접투표(Hard Voting)와 간접투표(Soft Voting)가 있다.
배깅(Bagging)	부트스트랩(Bootstrap) 샘플링으로 추출한 여러 개의 표본에 각각 모형을 병렬적으로 학습하고, 추출된 결과를 집계 (aggregation)하는 기법이다.
스태킹(Stacking)	여러 분석 모형의 예측값을 최종 모델의 학습 데이터로 사용하는 예측 방법이다.
부스팅(Boosting)	예측력이 약한 모형들을 결합하여 예측력이 강한 모형을 만드는 알고리즘으로 분류가 잘못된 데이터에 가중치를 적용하여 표본을 추출하는 기법이다.

4. 분석 모형 평가 및 개선 – 분석 모형 개선

04 최종 모형 선정

● 분석 모형 개발 단계에서 구성한 여러 개의 분석 모델을 대상으로 실제 업무에 적용할 수 있는 최종 모형을 선정한다.

● 최종 모형 선정 절차는 최종 모형 평가 기준 선정, 최종 모형 분석 결과 검토, 알고리즘별 결과 비교 순 이다.

① 최종 모형 평가 기준 선정

▶ 분석 모형 개발 후 분석 알고리즘 수행 결과를 검토하여 최종 모형을 선정한다.

정확도(Accuracy), 정밀도(Precision), 재현율(Recall) 등의 평가 지표를 활용한다.

② 최종 모형 분석 결과 검토

▶ 최종 모형 선정 시 업무관계자(데이터 분석가, 데이터 처리자, 고객 등)의 리뷰를 종합하여 최적의 분석 모형을 선정한다.

③ 알고리즘별 결과 비교

▶ 분석 알고리즘에 따라 매개변수를 변경하여 결과를 비교하고, 이를 바탕으로 최종 모형을 선정한다.

4. 분석 모형 평가 및 개선 – 예상 문제

예상 문제

1. 다음 중 분석 모형 설정에 대한 설명으로 옳은 것은?

① 이상적인 분석 모형은 높은 편향과 높은 분산으로 설정 되어야 한다.

② 이상적인 분석 모형은 높은 편향과 낮은 분산으로 설정 되어야 한다.

③ 이상적인 분석 모형은 낮은 편향과 낮은 분산으로 설정 되어야 한다.

④ 이상적인 분석 모형은 낮은 편향과 높은 분산으로 설정 되어야 한다.

2. 다음 중 회귀 모형 평가 지표에 대한 설명으로 틀린 것은?

① 회귀 모형 평가 지표에는 평균절대오차(MAE), 평균제곱오차(MSE), 평균제곱근오차(RMSE), 평균절대백분율오차(MAPE)가 있다.

② 평균제곱오차(MSE)의 수식은 $\frac{1}{n}\sum_{i=1}^{n}|(y_i - \hat{y}_i)|^2$ 와 같다.

③ 평균절대오차(MAE)는 모델의 실젯값과 예측값 차이에 절댓값을 취하여 평균한 값이다.

④ 평균제곱근오차(RMSE)는 평균제곱오차(MSE)에 제곱근을

3. 다음 중 분석 모형 평가 방법에 대한 설명으로 틀린 것은?

① 종속변수가 범주형인 경우 RMSE를 활용하여 분석 모형을 평가한다.

② 회귀 모형을 평가하기 위해서는 회귀 모형 평가 지표 혹은 결정계수를 사용한다.

③ 분류 모형을 평가하기 위해서는 분류 모형 평가 지표 혹은 혼동행렬을 사용한다.

④ 결정계수는 0에서 1 사이의 범위를 갖고, 수치가 1에 가까울수록 모델의 설명력이 높다고 할 수 있다.

종속변수가 범주형인 경우 혼동행렬을 활용하여 분석 모형 을 평가한다. RMSE(평균제곱근오차)는 종속변수가 연속형일 때 사용한다.

4. 다음 중 결정계수에 대한 설명으로 틀린 것은?

① 결정계수는 선형회귀 모형의 성능 검증 지표로 많이 사용되고, 회귀 모형의 예측값이 실젯값과 얼마나 유사한지를 나타내는 지표이다.

② 결정계수의 수식은 $R^2 = \left(1 - \frac{SSE}{SST}\right)$ 와 같다.

4. 분석 모형 평가 및 개선 – 예상 문제

다음 혼동행렬을 보고 물음에 답하시오.(5~6)

예측값 \ 실젯값	Positive	Negative
Positive	60	30
Negative	40	70

5. 주어진 혼동행렬에서 재현율은 얼마인가?

① 2/3 ② 3/5 ③ 3/10 ④ 4/11

참 긍정률(TP Rate, 재현율(Recall), 민감도(Sensitivity)

● 실제 긍정 범주 중 긍정의 비율

● 수식 : $\frac{TP}{TP+FN} = \frac{60}{60+30} = \frac{60}{90} = \frac{2}{3}$ 가 된다.

6. 주어진 혼동행렬에서 정밀도는 얼마인가?

① 1/3 ② 3/7 ③ 3/5 ④ 3/11

정밀도(Precision)

● 예측 긍정 범주 중 긍정의 비율

● 수식 : $\frac{TP}{TP+FP} = \frac{60}{60+40} = \frac{60}{100} = \frac{3}{5}$ 이 된다.

7. 다음 중 교차 검증 방법에 속하지 않는 것은?

① K-fold 교차 검증

② K-means clustering

③ 부트스트랩

④ LOOCV

교차 검증 방법에는 K-fold 교차 검증, 홀드아웃(Hold-Out) 교차 검증, LOOCV, LpOCV, 부트스트랩(Bootstrap)이 있다.

K-means Clustering은 그룹을 할당해서 군집화하는 비지도 학습 알고리즘이다.

8. 다음에서 설명하는 교차 검증 기법은?

데이터를 무작위로 7:3 또는 8:2 비율로 학습 데이터와 검증 데이터로 나누는 방법이다.

가장 보편적으로 랜덤 추출을 통해 데이터를 분할하는 방법으로 학습 데이터와 검증 데이터가 60~80%이고, 테스트 데이터가 20~40%이다.

① Hold-out 교차 검증

② LOOCV

③ LpOCV

4. 분석 모형 평가 및 개선 – 예상 문제

9. 다음 중 모집단에 대한 유의성 검정 방법이 아닌 것은?

① Z-검정 ② T-검정

③ Q-검정 ④ F-검정

모집단에 대한 유의성 검정 방법에는 Z-검정, T-검정, 분산 분석(ANOVA), 카이제곱 검정, F-검정이 있다.

Z-검정

● 정규분포를 가정하고, 추출된 표본이 동일 모집단에 속하는지 가설을 검증하기 위해 사용된다.

● 분산 또는 표준편차를 알고 있는 경우 사용된다.

T-검정

● 모집단의 분산이나 표준편차를 알지 못할 때, 표본으로부터 추정된 분산이나 표준편차를 이용하여 두 모집단의 평균의 차이를 알아보는 검증 방법이다.

F-검정

● 두 모집단의 분산의 차이가 있는지를 검증하는 방법으로 F-값이 클수록 두 집단 간의 분산 차이가 존재하는 것을 의미한다.

10. 다음에서 설명하는 유의성 검정 방법은?

이것은 데이터가 정규분포를 따르는지 확인하는 검정 방법이다.

R언어에서 관련 함수를 사용하여 검정할 수 있고, $p-value$가 0.05보다 작은 경우 귀무가설(Ho)을 기각하고, 대립가설(H_1)을 채택한다.

다만 R언어의 관련 함수를 사용하는 경우 데이터의 수는 3개에서 5,000개 이하로만 사용 가능하다.

① 카이제곱 검정

② Q-Q plot

③ 샤피로-윌크 검정

④ 콜모고로프-스미르노프 적합성 검정

11. 다음에서 설명하는 정규성 검정 방법은?

이것은 데이터가 정규분포를 따르는지 확인하는 검정 방법이다.

R언어에서 관련 함수를 사용하여 검정할 수 있고, $p-value$가 0.05보다 작은 경우 귀무가설(Ho)을 기각하고, 대립가설(H_1)을 채택한다.

다만 R언어의 관련 함수를 사용하는 경우 데이터의 수는 3개에서 5,000개 이하로만 사용 가능하다.

① 카이제곱 검정

② Q-Q plot

③ 샤피로-윌크 검정

④ 콜모고로프-스미르노프 적합성 검정

Q-Q plot

● Q-Q plot은 그래프를 통해서 정규성 가정을 시각적으로 검정하는 방법이다.

● Q-Q plot에서 대각선 참조선을 따라서 데이터가 분포할 경우 정규성을 만족한다고 할 수 있다.

● Q-Q plot은 해석이 주관적일 수 있기에 보조 수단으로

4. 분석 모형 평가 및 개선 – 예상 문제

13. 다음 중 과대적합 방지 기법이 아닌 것은?

① 모델의 복잡도 감소

② 데이터 감소

③ 가중치 규제 적용

④ 드롭아웃

14. 다음의 수식이 의미하는 기법은?

$$\frac{1}{N}\sum_{i=1}^{N}(y_i-\hat{y}_i)^2+\frac{\lambda}{2}\sum_{j=1}^{M}|w_j|^2$$

① Drop Out ② Elastic Net

③ Lasso ④ Ridge

L1 노름 규제(라쏘, Lasso regression)

● 기존 비용 함수에 모든 가중치(ω)들의 합계를 추가하여 값이 최소가 되도록 하는 방법

● 불필요한 회귀 계수를 급격하게 감소시켜 0으로 만들고 제거하는 방법

L2 노름 규제(릿지, Ridge regression)

● 기존 비용 함수에 모든 가중치(ω)들의 제곱합을 추가 하는

15. 다음 중 과대적합 방지를 위해 인공신경망의 일부를 사용하지 않는 기법은?

① Ensemble ② L1 norm

③ Drop Out ④ L2 norm

드롭아웃(Drop Out)

● 드롭아웃은 학습 과정에서 신경망 일부를 사용하지 않는 방법이다.

● 드롭아웃은 서로 연결된 연결망에서 0~1 사이의 확률로 뉴런을 제거하는 방법이다.

● 드롭아웃은 신경망 학습 시에만 사용하고, 예측 시에는 사용하지 않는다.

● 드롭아웃의 유형에는 초기, 공간적, 시간적 드롭아웃이 있다.

16. 다음 중 매개변수 최적화 기법인 경사하강법에 속하지 않는 것은?

① 맥스 배치 경사하강법

② 확률적 경사하강법

③ 배치 경사하강법

4. 분석 모형 평가 및 개선 – 예상 문제

17. 다음 중 분석 모형 융합인 앙상블 모델이 아닌 것은?

① Voting ② ANN

③ Bagging ④ Boosting

앙상블 모델에는 보팅, 배깅, 스태킹, 부스팅이 있다.

ANN은 인공신경망을 의미한다.

18. 다음 빈칸에 들어갈 알맞은 용어는?

(㉠)는 함수를 호출할 때 인수로 전달된 값을 함수 내부에서 사용할 수 있게 해주는 변수를 말한다.
분석 모형의 결괏값과 실젯값 차이를 (㉡)라고 하고 (㉡)를 최소화하는 (㉠)를 찾아가는 과정을 (㉠) 최적화라고 한다.

① ㉠ : Hyper Parameter
㉡ : Activation Function

② ㉠ : Hyper Parameter
㉡ : Loss Function

③ ㉠ : Parameter
㉡ : Loss Function

④ ㉠ : Parameter

19. 다음에서 설명하는 경사하강법은?

이것은 확률적 경사하강법의 매개변수 변경 방향에 가속도를 부여하는 방식으로 공이 구르는 듯한 모습을 보인다.
이것에서 x의 한 방향으로 일정하게 가속하고, y축 방향의 속도는 일정하지 않다.
이것은 확률적 경사하강법에 비해 효율적인 학습을 할 수 있고, 확률적 경사하강법이 갖는 지역 최솟값(Local Minimum)을 해결할 수 있다.

① Adam ② AdaGrad

③ RMSProp ④ Momentum

Adam

● 모멘텀과 AdaGrad가 합쳐진 방법이다.

AdaGrad(Adaptive Gradient)

● 매개변수 값을 업데이트 하면서 각 변수마다 학습률을 다르게 적용하는 방법이다.

RMSProp

● AdaGrad에서 최적값에 도달하기 전에 학습률이 0에

감사합니다.