2020 D&A

MACHINE LEARNING SESSION

4주차

D&A

Contents

1. 손실함수, 평가지표
2. 파라미터 튜닝

D&A

손실함수(loss function)

학습을 통해 얻은 데이터의 추정치와 실제 데이터의 차이를 평가하는 지표

값이 작을수록 완벽하게 예측했다고 할 수 있음

손실함수에는 여러 종류가 있음(MSE, Cross Entropy)

손실함수 = 목적함수 = 비용함수

손실함수(loss function)

평균 제곱 오차(Mean Squared Error, MSE)

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

(Y_i : 실제 값, $\hat{Y}_i$: 예측값)

가장 많이 사용되는 손실함수

모델의 출력 값과 실제값의 거리 차이를 오차로 사용

실제 오차보다 줄어드는 경우를 막기 위해 제곱하여 평균

MSE 가 작을수록 모델의 정확성이 높음

손실함수(loss function)

교차 엔트로피 오차 (Cross Entropy Error , CEE)

$$H(p,q) = -\sum_{x \in X} p(x) \log q(x)$$

(p(x) : 실제 분포, q(x) : 측정 분포)

분류하는 문제에서 사용

p와 q 두 확률 분포가 비슷할수록 오차는 작아짐

CEE가 작을수록 모델의 정확성이 높음

평가지표

혼동행렬(confusion matrix)

분류 모형의 성능을 보여주기 위한 표

실제 값과 예측 값의 교차표 형태로 제시

다양한 지표를 계산(precision, recall …)

평가지표

혼동행렬(confusion matrix)

		예측값	
		Negative	positive
실제 값	Negative	True Negative (TN)	False Positive (FP)
	positive	False Negative (FN)	True Positive (TP)

* 관심있는 대상을 positive로 둠

평가지표

혼동행렬(confusion matrix) - 정확도 (Accuracy)

		예측값	
		Negative	positive
실제값	negative	True Negative (TN)	False Positive (FP)
	positive	False Negative (FN)	True Positive (TP)

$$\frac{TP + TN}{\text{전체}}$$

Negative/positive 구분에 관심이 없는 경우에 많이 사용

평가지표

혼동행렬(confusion matrix) - 정밀도 (precision)

		예측값	
		Negative	positive
실제값	negative	True Negative (TN)	False Positive (FP)
	positive	False Negative (FN)	True Positive (TP)

$$\frac{TP}{TP + FP}$$

positive 예측이 중요한 경우에 사용

평가지표

혼동행렬(confusion matrix) - 재현도 (recall)

		예측값	
		Negative	positive
실제값	negative	True Negative (TN)	False Positive (FP)
	positive	False Negative (FN)	True Positive (TP)

$$\frac{TP}{TP + FN}$$

positive을 찾아 내는 것이 중요한 경우

민감도(specificity)라고도 함

평가지표

혼동행렬(confusion matrix) -특이도 (specificity)

		예측값	
		Negative	positive
실제값	negative	True Negative (TN)	False Positive (FP)
	positive	False Negative (FN)	True Positive (TP)

$$\frac{TN}{TN + FP}$$

Negatvie 을 찾아 내는 것이 중요한 경우

재현도와 반대되는 경우

평가지표

혼동행렬(confusion matrix) -F1

$$\frac{1}{\frac{\frac{1}{p}+\frac{1}{r}}{2}} = \frac{2pr}{p+r}$$

(p는 정밀도, r은 재현도)

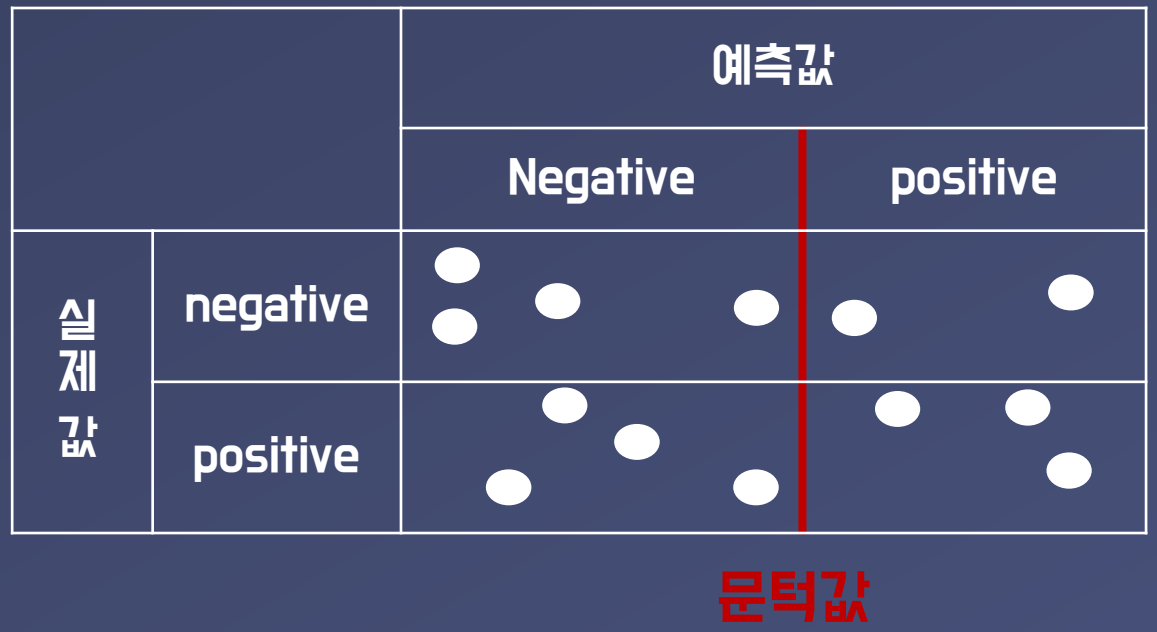
정밀도와 재현도의 조화평균

조화평균 : 역수의 평균을 구해 다시 역수

평가지표

문턱값(threshold)

Negative와 positive를 구분하는 기준선

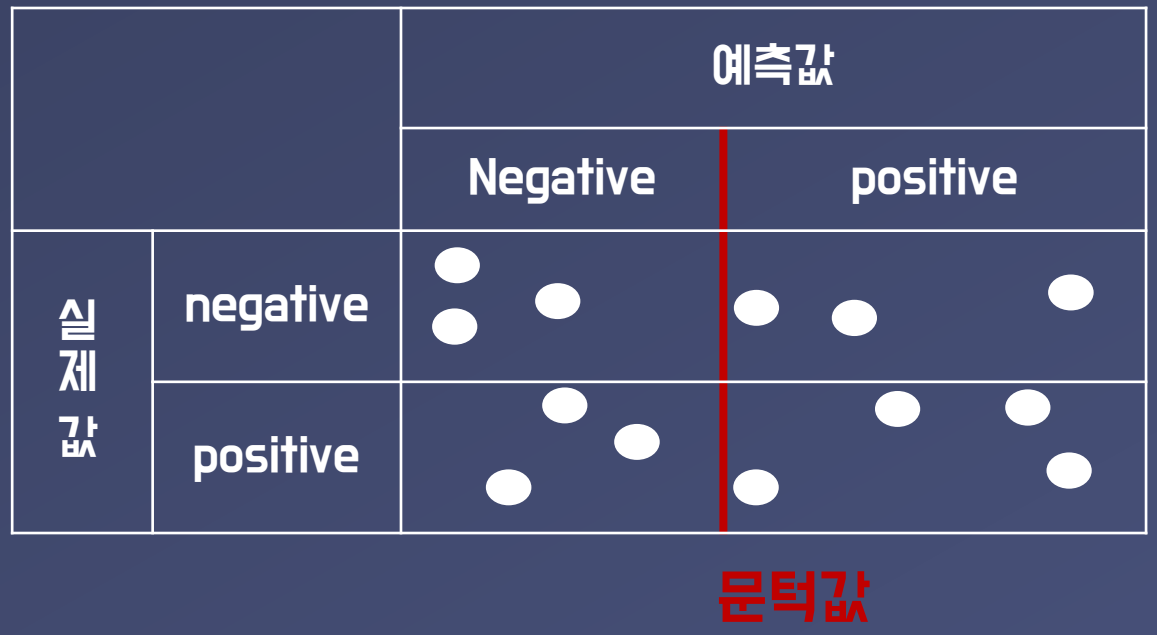


정밀도 : 3/5
재현도 : 3/7
특이도 : 4/6

평가지표

문턱값(threshold)

Negative와 positive를 구분하는 기준선



정밀도 : 4/7 -> 높아짐
재현도 : 4/7 -> 높아짐
특이도 : 6/10 -> 낮아짐

평가지표

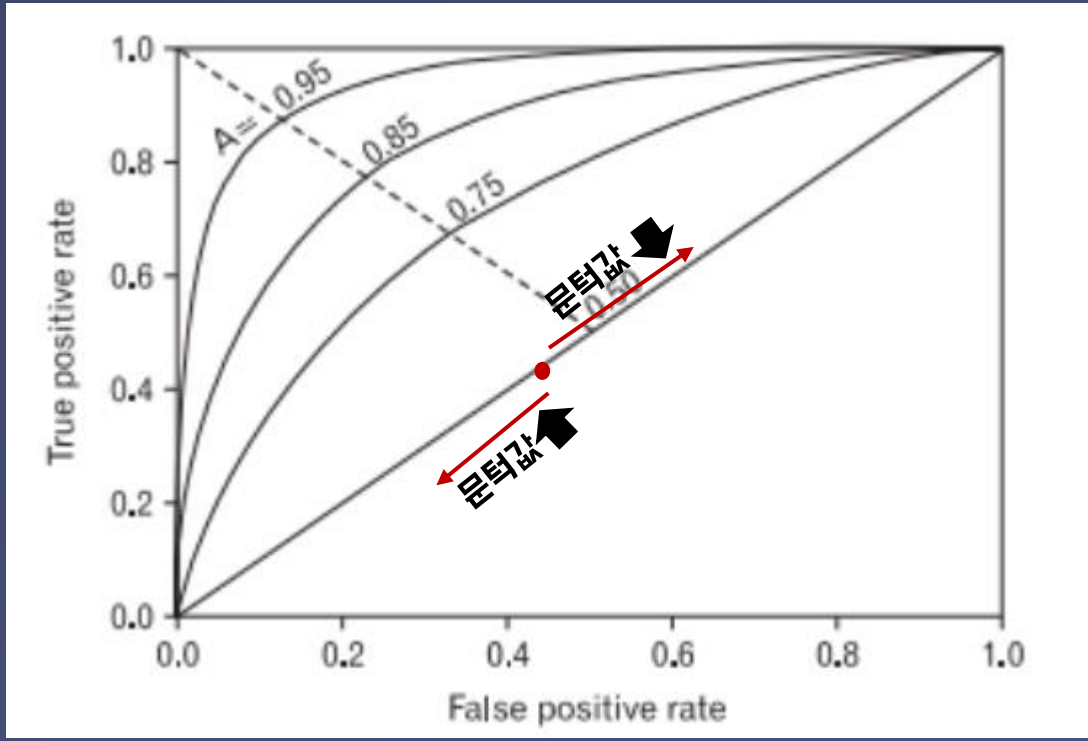
ROC 곡선 (Receiver operating Characteristic)

가로축은 1 - 특이도, 세로축은 재현도

문턱값(threshold)을 변화시키면서 특이도와 재현도의 변화를 곡선으로 표시

평가지표

ROC 곡선 (Receiver operating Characteristic)



평가지표

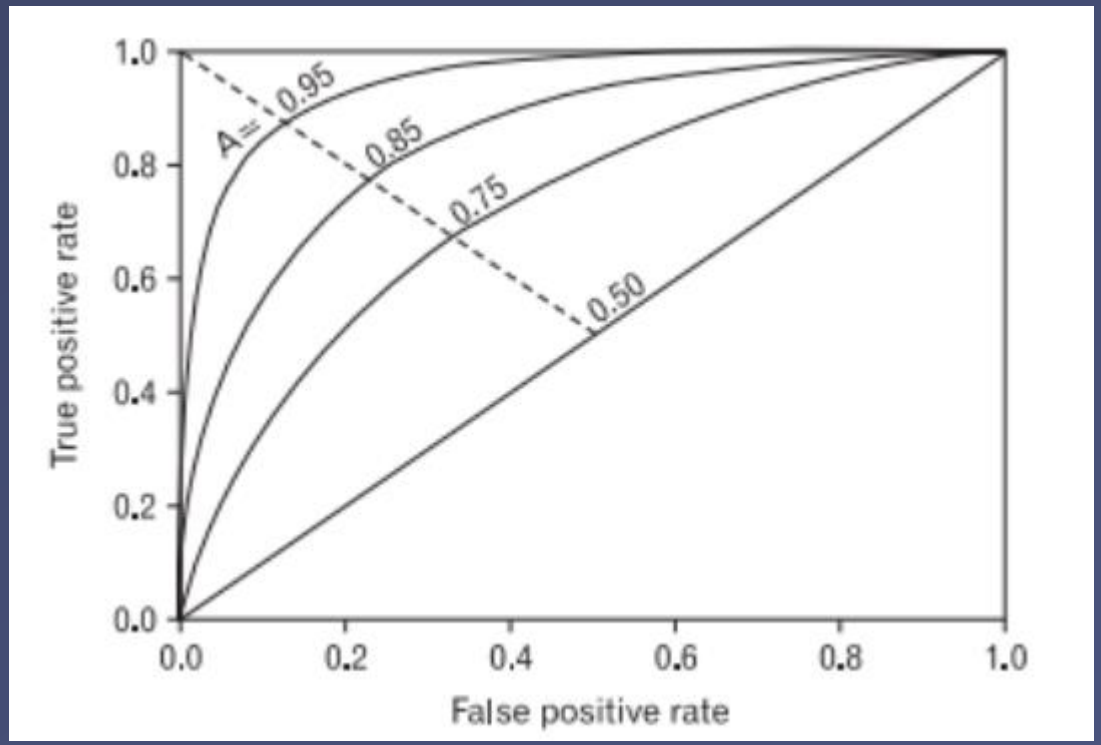
ROC- AUC(Area Under the Curve)

ROC 곡선 아래의 면적

0 ~ 1 사이의 범위

무작위로 예측할 경우 0.5

1에 가까울 수록 성능이 높음



D&A

하이퍼 파라미터

파라미터 vs 하이퍼 파라미터

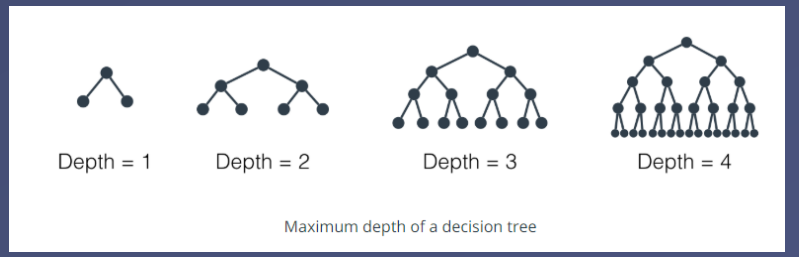
모델 내부적으로 결정

가중치라고도 함

Ex) 선형회귀의 계수

사용자에 의해 결정

Ex) max_depth, min_samples_leaf



D&A

하이퍼 파라미터

Parameter(weight) 는 학습 과정에서 조정

Hyper-parameter는 사용자가 고정한 값으로 학습

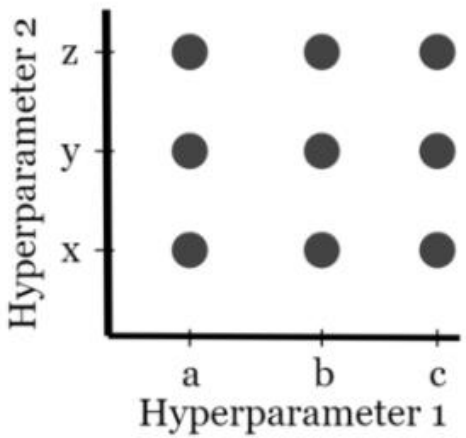
데이터마다 최적의 하이퍼 파라미터가 다르며, 성능을 높일 수 있는
방법이기 때문에 하이퍼 파라미터 튜닝이 필요함

Gridsearch 와 Randomsearch가 대표적

하이퍼 파라미터-Grid Search

하이퍼 파라미터로 적용해볼 값들을 미리 정해 , 모든 조합을 시행

간단하지만, 하이퍼파라미터의 개수가 많을수록 시간이 오래걸림



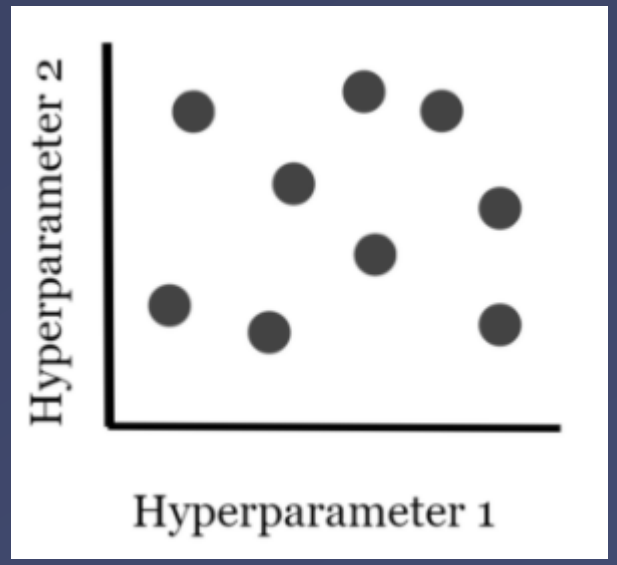

```
'criterion': ['gini', 'entropy'],
'n_estimators': [50, 70, 90, 110],
'max_depth': [3, 5, 7,10],
'max_features': [0.8,0.85,0.9]}
```

2 * 4 * 4 * 3 = 96

96 번을 반복하여
최적의 하이퍼 파라미터를 찾음

D&A

하이퍼 파라미터-Random Search



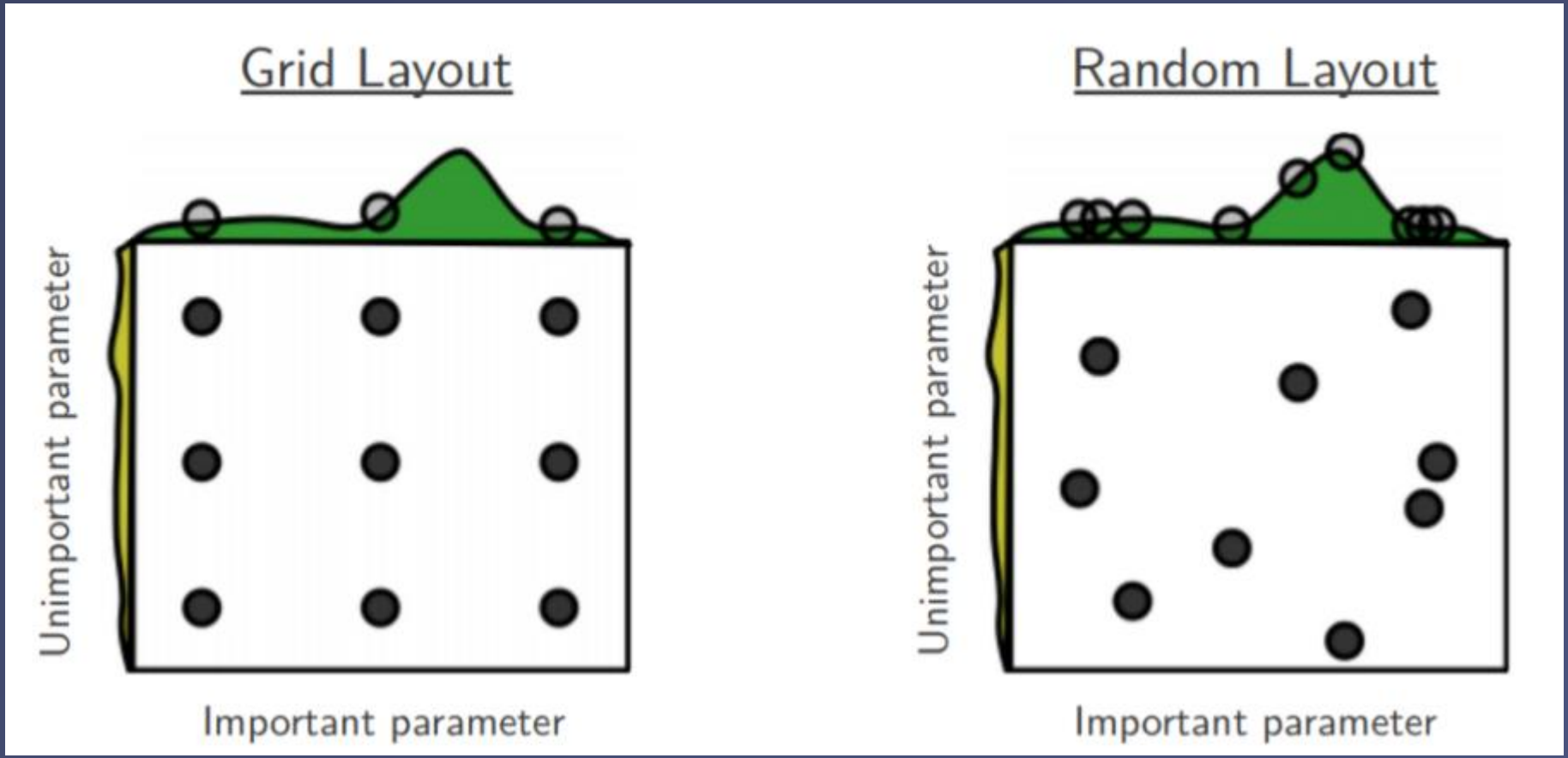
하이퍼 파라미터로 적용해볼 값들을 미리 정해 , 범위 내 무작위 값 추출

시간대비 성능이 뛰어남

Grid search 보다 훨씬 다양한 조합들을 시험

n_iter 값만큼 반복 무작위 추출

하이퍼 파라미터 튜닝



과제

저번 세션 단일 모델에 파라미터 튜닝 해보기
(다음주 수요일 9시전까지 3번 제출)

10월 7일 세션 진행

D&A