

## Unit 3

# 분류기반 ML 모형의 성능 측정 방법

**11주차. 앙상블(Ensemble)**

## Classification based

- » 기계학습에는 다양한 분류 모형이 존재
- » 적합한 분류 모형 선택을 위해 성능을 비교·분석이 필요

## 학습 내용

- Confusion matrix (오차행렬)
- 정확도, 정밀도, 재현율
- Miss Rate, False Negative Rate (FNR)
- False Alarm Rate
- F1 score
- Receiver Operating Characteristic (ROC) Curve

## 학습 목표

- 분류기의 다양한 성과척도를 이해하고 선택할 수 있다.
- 분류 기반 모형의 성능을 비교, 분석할 수 있다.

## Confusion matrix

### ☑ Confusion matrix란?

#### Confusion Matrix

➡ 오차행렬 또는 혼돈행렬

- » Discrete 분류 모형의 성능을 도식화하는 데 사용되는 정방행렬
- » 학습을 통해 Prediction의 성능을 측정하기 위해서 예측값과 실제값을 비교하는 표

## Confusion matrix

## ☑ Confusion matrix

사용자에 따라 Column과 Row의 위치가 바뀔 수 있다! ⚠

		Predictive values	
		Positive (1)	Negative (0)
Actual values	Positive (1)	TP	FN
	Negative (0)	FP	TN

True Positive(TP)

- 실제 True인 정답을 True라고 예측 (정답)

False Negative(FN)

- 실제 True인 정답을 False라고 예측 (오답)

False Positive(FP)

- 실제 False인 정답을 True라고 예측 (오답)

True Negative(TN)

- 실제 False인 정답을 False라고 예측 (정답)

네 가지 Element로 다양한 모형의 **성과 지표를 계산**

## *Confusion matrix*

### Confusion matrix 출력

## Confusion matrix

### ☑ Confusion matrix 출력

Error Rate

$$= \frac{FP + FN}{TP + TN + FP + FN}$$

1-accuracy와 동일

sklearn.metrics

```
from sklearn.metrics import confusion_matrix
df_result = logistic_reg_model.df_result
actual = list(df_result['Actual'])
predicted = list(df_result['Predicted'])
matrix = confusion_matrix(actual, predicted)
print(matrix)
```

## Accuracy

### ☑ Accuracy란?

#### 정확도(Accuracy)

➡ 전체 중에서 모형을 바르게 분류한 비율

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

		Predictive values	
		Positive (1)	Negative (0)
Actual values	Positive (1)	TP	FN
	Negative (0)	FP	TN



## Accuracy

### ☑ Accuracy 출력

정확도(Accuracy)

sklearn.metrics

```
from sklearn.metrics import accuracy_score  
print('accuracy_score:', accuracy_score(actual, predicted))
```

## Accuracy

### ☑ Accuracy의 단점



## Accuracy의 단점

도메인 데이터셋의 **편중**(Bias 발생)

### 예 희귀병

- 발생빈도가 낮아 Accuracy Rate 자체가 낮음
- 정확한 모형 판단이 어려움

Accuracy는 Bias가 없는 데이터인 경우에 **적합**한 척도

## Precision

### ☑ Precision이란?

정밀도(Precision 또는 Specificity)

➡ 모형이 Positive라고 분류한 것 중에서 실제로 Positive인 비율

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

		Predictive values	
		Positive (1)	Negative (0)
Actual values	Positive (1)	TP	FN
	Negative (0)	FP	TN

## Precision

### ☑ Precision 출력

정밀도

sklearn.metrics

```
from sklearn.metrics import precision_score  
print('precision_score:', precision_score(actual, predicted))
```

## Recall Score

### ☑ Recall Score란?

재현율(Sensitivity, Hit rate 또는 TPR)

➔ 실제 데이터가 Positive인 것 중  
모형이 Positive로 분류한 비율

$$\frac{TP}{TP + FN} = TPR = 1 - FNR$$

False Negative Rate

Precision(정밀도)

▪ 모형의 관점

Recall Score(재현율)

▪ 데이터 관점

## Recall Score

### ☑ Recall Score 출력

재현율

sklearn.metrics

```
from sklearn.metrics import recall_score  
print('recall_score:', recall_score(actual, predicted))
```

## Miss Rate, False Negative Rate (FNR)

☑ Miss Rate, False Negative Rate (FNR) 이란?

Miss Rate 또는 FNR(False Negative Rate)

➡ 실제 데이터가 Positive인 것 중  
모형이 Negative로 잘못 분류한 비율

$$\frac{FN}{TP + FN} = 1 - TPR$$

완전한 분류기 = 0

## False Alarm Rate

### ☑ False Alarm Rate이란?

#### False Alarm Rate

➔ 실제 데이터가 Negative인 것 중  
모형이 Positive로 잘못 분류한 비율

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

		Predictive values	
		Positive (1)	Negative (0)
Actual values	Positive (1)	TP	FN
	Negative (0)	FP	TN



## F1 Score

### ☑ F1 Score란?

F1 Score

➡ Precision과 Recall의 조화평균

? 조화평균을 사용하는 이유

↓  
극단적인 값에 대해서 **패널티**를 주기 위해

## F1 Score

☑ F1 Score란?

F1 Score

$$\frac{2 \times \textit{Precision} \times \textit{Recall}}{\textit{Precision} + \textit{Recall}}$$

		Predictive values	
		Positive (1)	Negative (0)
Actual values	Positive (1)	TP	FN
	Negative (0)	FP	TN

## F1 Score

### ☑ F1 Score 출력

F1 Score

sklearn.metrics

```
from sklearn.metrics import f1_score  
print('f1_score:', f1_score(actual, predicted))
```

## 분류기 성과 척도 계산

## ☑ 예시

		Predictive values	
		Positive (1)	Negative (0)
Actual values	Positive (1)	TP= 55	FN= 5
	Negative (0)	FP= 10	TN= 30

Accuracy

$$= \frac{TP + TN}{TP + TN + FP + FN} = \frac{55 + 30}{55 + 30 + 10 + 5} = \frac{85}{100} = 0.85$$

Error rate

$$= 1 - \text{accuracy} = 0.15$$

Recall

$$= \frac{TP}{TP + FN} = \frac{55}{55 + 10} = \frac{55}{65} = 0.85$$

## 분류기 성과 척도 계산

## ☑ 예시

		Predictive values	
		Positive (1)	Negative (0)
Actual values	Positive (1)	TP	FN
	Negative (0)	FP	TN

Precision

$$= \frac{TP}{TP + FP} = \frac{55}{55 + 5} = \frac{55}{60} = \mathbf{0.92}$$

F1 Score

$$= \frac{2 \times Precision \times Recall}{Precision + Recall} = \frac{2 \times 0.92 \times 0.85}{0.92 + 0.85} = \mathbf{0.88}$$

False Alarm

$$= \frac{FP}{TN + FP} = \frac{5}{30 + 5} = \frac{5}{35} = \mathbf{0.14}$$

## ROC Curve

### ☑ ROC Curve란?

ROC Curve(Receiver Operating Characteristic Curve)

- » 오차행렬과 분류기 성능 척도는 임계값에 따라 상이
- » 임계값 =  $\theta$ , 이를 조정할 경우 재현율과 정밀도가 서로 Trade off

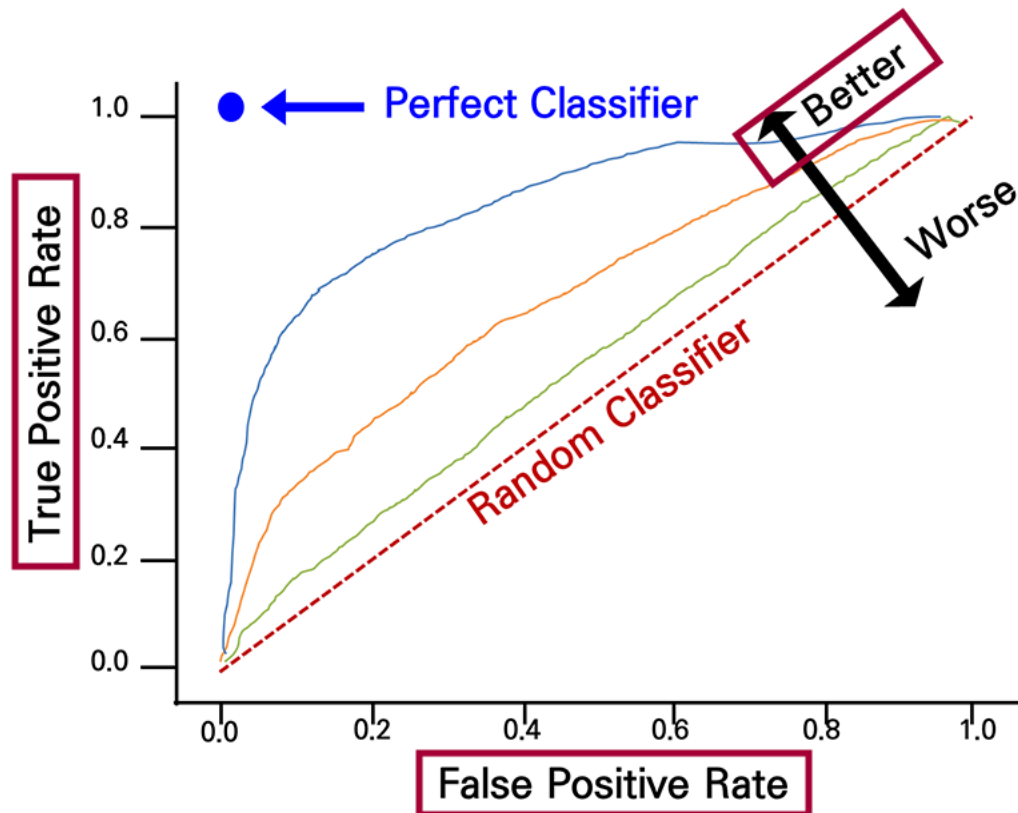


임계값에 독립적인 성과평가 척도가 있을까

ROC Curve

## ROC Curve

### ✓ ROC Curve

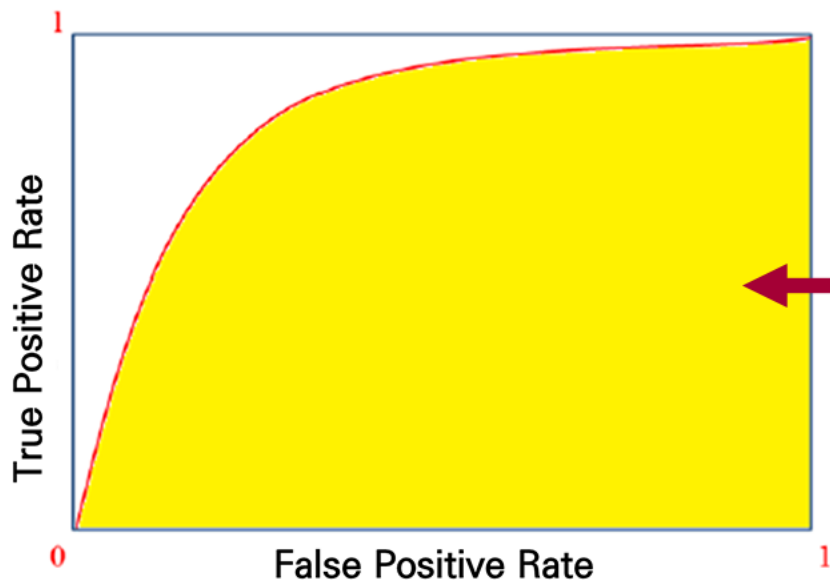


ROC Curve가 세타와 무관하게 독립적인 성과 척도로 많이 활용

## ROC Curve

### ✓ ROC Curve

ROC AUC(Area Under the curve)



면적을 계산하면  
숫자를 통해 쉽게 비교 가능

ROC보다 AUC가 더 쉽게 모형 성과를 비교 분석 할 수 있음



## 분류기 성과 척도

### ☑ 분류기 성과 척도의 선택

“ 절대적 성과 척도는 없다. ”

다양한 성과 척도 고려



분류기의 다양한 특성 파악



목적에 맞는 분류기 선택

## 분류기 성과 척도

### ☑ 분류기 성과 척도의 선택

## ☀ 주의사항



오차행렬 원소들의 의미를 혼동해서는 안 된다!



연구연구자에 따라서 행과 열을 바꾸어서 제시하는 경우도 있으므로 계산할 때 주의한다!

- » Confusion matrix (오차행렬)
- » 정확도, 정밀도, 재현율
- » Miss Rate, False Negative Rate (FNR)
- » False Alarm Rate
- » F1 score
- » Receiver Operating Characteristic (ROC) Curve