

BIG DATA ANALYTICS

WEEK-15 | Application-2 Classification

**Yonsei University
Jungwon Seo**

손글씨 분류하기



이진 분류

- binary-classification
- 예측하고자 하는 값이 True(1) 또는 False(0)
- MNIST 데이터셋에서 예를들어 해당하는 이미지가 5인지(True) 아닌지에 (False) 대한 분류기를 만들었다고 가정
- 정확도(accuracy)가 95% 이상
- 가짜 모델을 만들어 모든 예측값이 False라고 했을 때 정확도는?

모델 검증

- 정확도가 과연 올바른 성능 측정 지표일까?
- 불균형 (imbalanced) 데이터셋을 다룰시 문제가 발생함
 - 이상탐지에서 흔히 겪는 문제처럼 데이터셋의 레이블이 한쪽이 월등히 많은 경우 정확도는 성능을 표현할 수 없음
 - 예) 정상 폐 X-ray 9만 9천개, 코로나 감염 폐 X-ray 1000개
 - 모두 정상이라고 예측해도 정확도 99%

오차 행렬

Confusion matrix

- [맞춤] - [예측결과]
 - True-Positive: 예측과 실제값이 모두 True 인것
 - True-Negative: 예측과 실제값이 모두 False 인것
 - False-Positive: 예측과 실제값이 틀리고 예측은 True
 - False-Negative: 예측과 실제값이 틀리고 예측은 False





n=190	예측: True	예측: False		
	실제: True	TP 100	FN 10	110
	실제: False	FP 30	TN 50	80
		130	60	

n=190	예측: True	예측: False	
	실제: True	TP 100	FN 0
실제: False	FP 0	TN 90	90
	100	90	

완벽한 분류기의 경우

오차 행렬

Confusion matrix

		Actual Values	
		1	0
Predicted Values	1	<p>TRUE POSITIVE</p>  <p>1</p>	<p>FALSE POSITIVE</p>  <p>TYPE 1 ERROR</p>
	0	<p>FALSE NEGATIVE</p>  <p>TYPE 2 ERROR</p>	<p>TRUE NEGATIVE</p>  <p>0</p>

정밀도와 재현율

- 정밀도 : 모델이 True라고 예측한 것중 실제 True의 비율

- $Precision = \frac{TP}{TP + FP}$
- 예) 암이라고 예측한 결과중 실제 암인 경우의 비율

- 재현율 : 실제 True인 것 중 모델이 True라고 예측한 비율

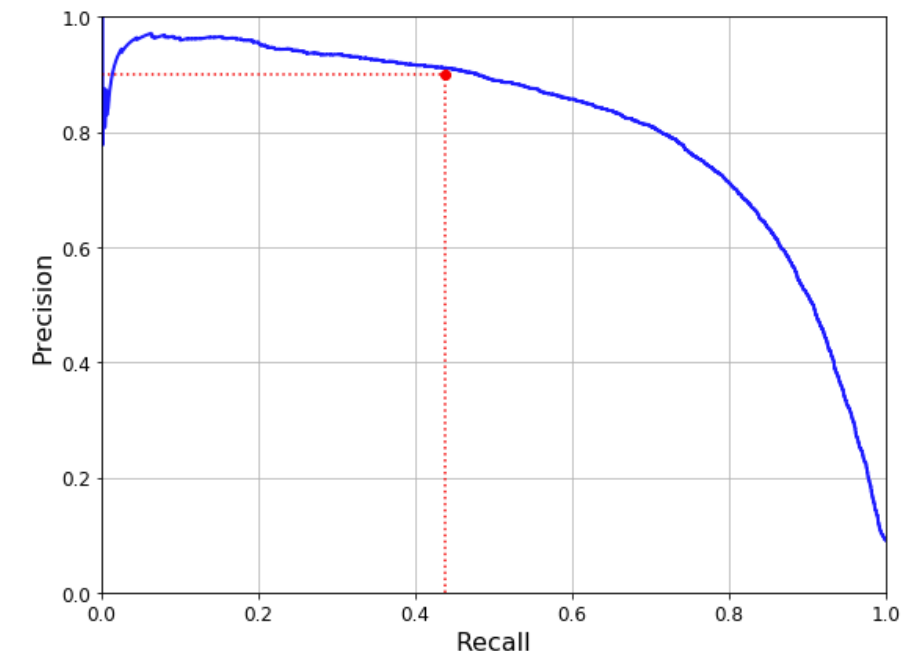
- $Recall = \frac{TP}{TP + FN}$
- 예) 실제 암인 경우 중에 모델이 암이라고 예측한 비율

- 한가지만 높다고 좋은 모델이 아니고, 서로 상호보완적

- 예1) 실제 10개의 암 데이터 중에 확실한 2개만 암이라고 예측
 - 정밀도 100% -> 스팸 분류
- 예2) 모든 경우를 다 암이라고 예측
 - 재현율 100% -> 사기 결제

- 정밀도/재현율 트레이드오프

- 모델이 전반적으로 좋아지지 않는한 (TP TN만 높음) FP가 낮아지면, FN이 높아지고 그 반대도 마찬가지
- 그러므로 정밀도와 재현율을 둘다 높이는 힘듦



정확도와 조화평균

- 정확도 : 재현율/정밀도와 달리 False를 False라고 답한 경우 (TN)도 고려

- $Accuracy = \frac{TP + TN}{TP + FP + FN + TN}$
- Imbalance데이터에 유의미한 수치를 보이기 어려움

- 조화평균: 정밀도와 재현율을 동시에 고려하는 지표

- $F_1 = \frac{2}{\frac{1}{Prec} + \frac{1}{Recall}} = 2 * \frac{Prec * Recall}{Prec + Recall} = \frac{TP}{TP + \frac{FN + FP}{2}}$
- 현실적으로 레이블이 균형잡힌 데이터의 확보가 어렵기때문에 F1스코어를 많이 평가지표로 사용함

연습문제

- 왼쪽 표의 Accuracy, Precision, Recall, F1-score는?
- 오른쪽 표의 Accuracy, Precision, Recall, F1-score는?

n=20	예측: True	예측: False	
	실제: True	TP 5	FN 10
실제: False	FP 0	TN 5	5
	5	15	

n=20	예측: True	예측: False	
	실제: True	TP 5	FN 0
실제: False	FP 10	TN 5	15
	15	5	

ROC 곡선

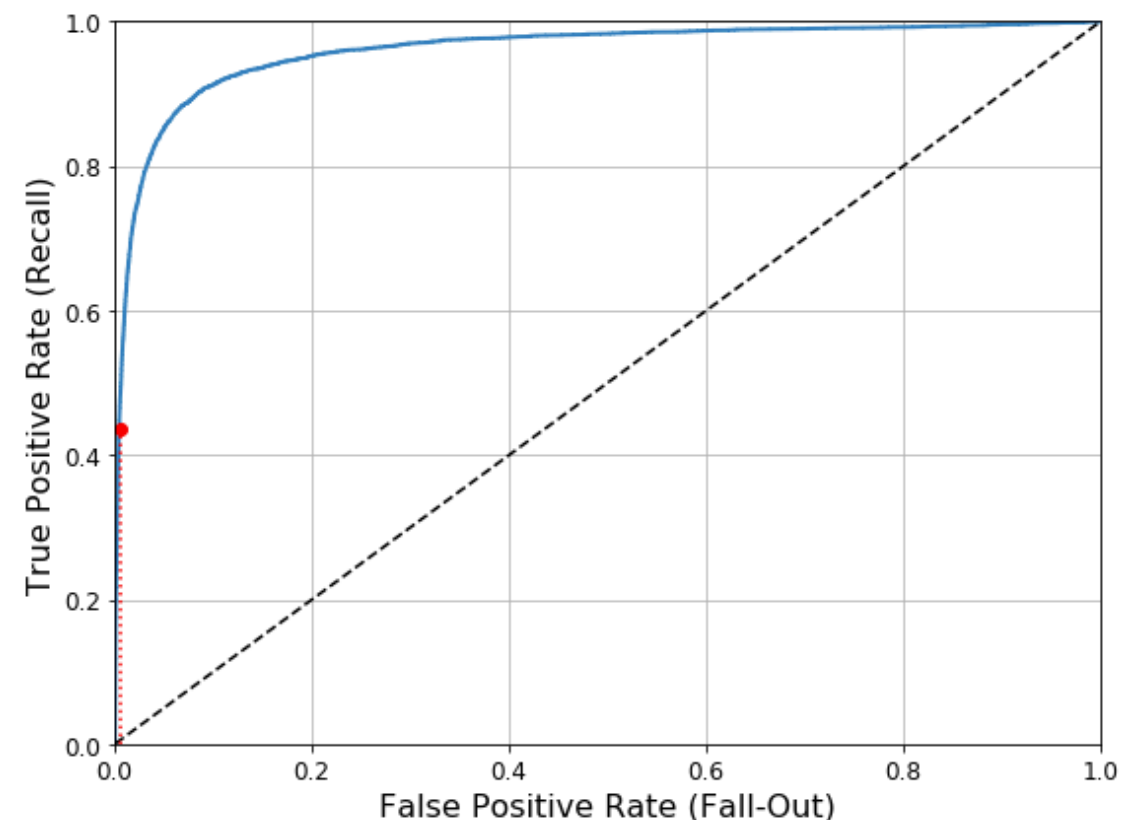
- 거짓양성비율(FPR, Specificity)과 진짜양성비율(TPR, Recall)에 대한 곡선

- $$FPR = \frac{FP}{FP + TN}, TPR = \frac{TP}{TP + FN}$$

- 재현율이 높을수록 거짓 양성 비율이 높아짐

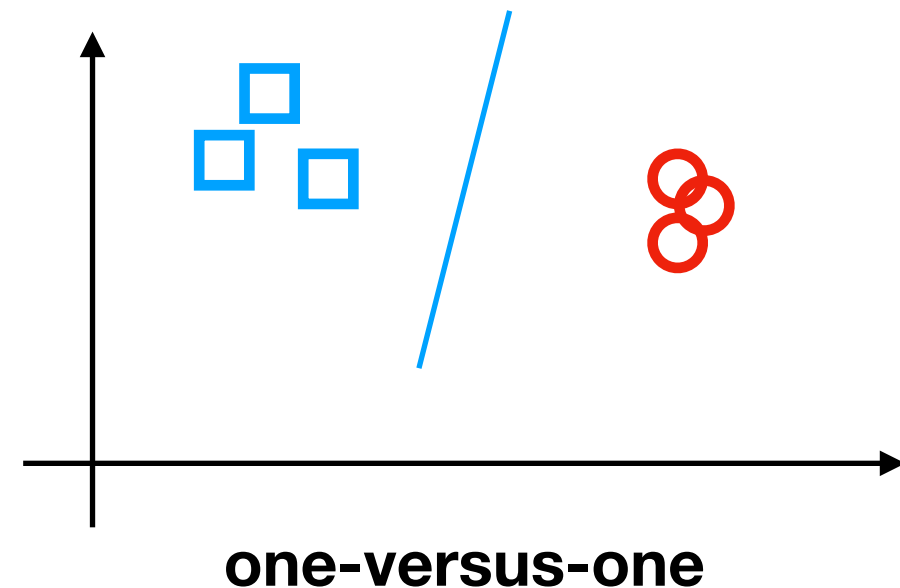
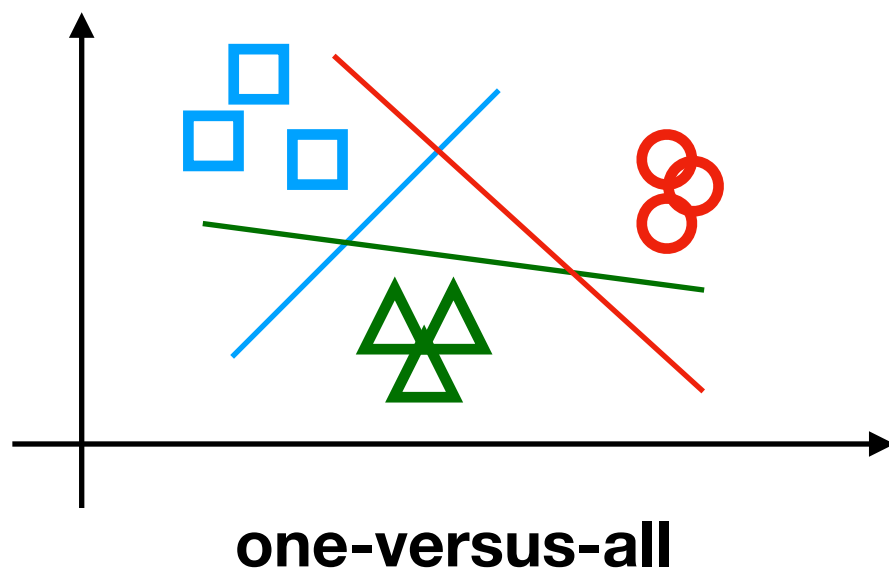
- 좋은 모델일 수록 곡선이 왼쪽 상단 모서리에 가까워짐

- 즉 곡선 아래 면적(AUC)이 1에 수렴함



다중 분류

- 로지스틱 회귀나 서포트 벡터머신 분류기는 이진분류만 가능
- 다음과 같은 방법으로 다중 분류기로 변환
 - One-versus-all: 1과 나머지, 2와 나머지 ...
 - One-versus-one: 0과 1구별, 0과 2구별... 9와 10 구별
 - 클래스 N개일때 분류기는 $\frac{N * (N - 1)}{2}$ 개가 필요
 - scikit-learn 라이브러리에서는 위의 두가지 방법중 알아서 선택후 분류

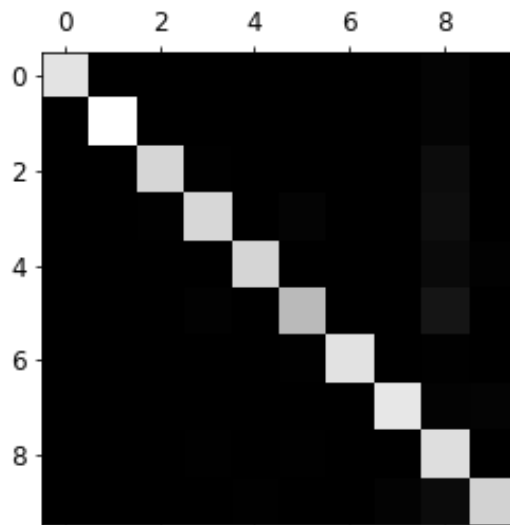


에러분석

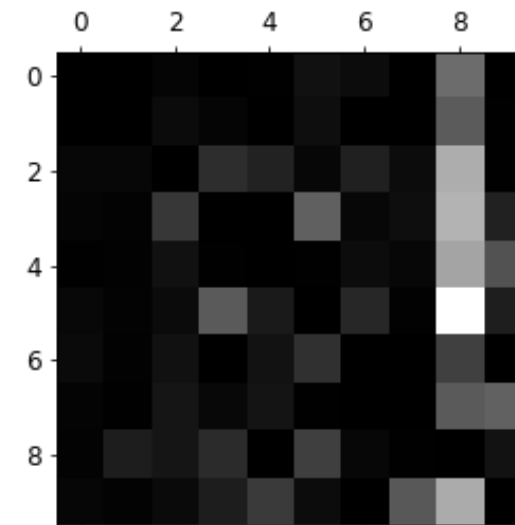
- 멀티 클래스 데이터에 대한 confusion matrix 출력시 주대각선 외의 값들은 모두 오류 분류

		예측값									
		0	1	2	3	4	5	6	7	8	9
실제값	0	5577	0	22	5	8	43	36	6	225	1
	1	0	6400	37	24	4	44	4	7	212	10
	2	27	27	5220	92	73	27	67	36	378	11
	3	22	17	117	5227	2	203	27	40	403	73
	4	12	14	41	9	5182	12	34	27	347	164
	5	27	15	30	168	53	4444	75	14	535	60
	6	30	15	42	3	44	97	5552	3	131	1
	7	21	10	51	30	49	12	3	5684	195	210
	8	17	63	48	86	3	126	25	10	5429	44
	9	25	18	30	64	118	36	1	179	371	5107

- 히트맵으로 표현시, 어떤 클래스가 오류가 많은지 쉽게 확인 가능



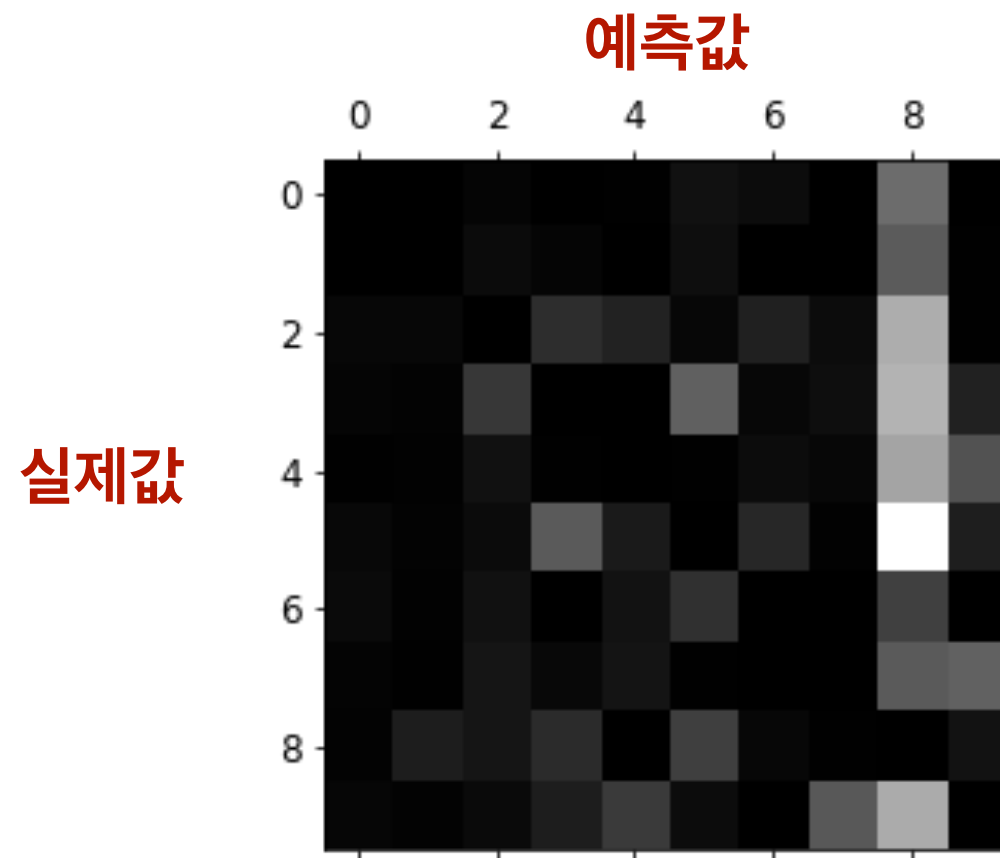
Heatmap



대각행렬을 0으로 변경 뒤 heatmap

에러분석

- 행 기준(실제)으로 보면, 8자체는 잘예측이 되었지만, 열 기준(예측)으로 보면 8이 아닌 것들을 8로 예측함
 - 8처럼 보이지만 8은 아닌 데이터를 더 확보 한뒤 훈련할 필요성이 있음



참고문헌

- Géron, Aurélien. Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow: Concepts, tools, and techniques to build intelligent systems. O'Reilly Media, 2019.

E.O.D