핸즈온 머신러닝

PART3. 분류성능평가지표

- 분류성능평가지표란? 기계학습에서 모델이나 패턴의 분류 성능 평가에 사용되는 지표이다. 어느 모델이든 간에 발전을 위한 피드백은 현재 모델의 퍼포먼스를 올바르게 평가하는 것에서부터 시작한다.

- 모델의 분류와 정답

모델을 평가하는 요소는 결국, 모델이 내놓은 답과 실제 정답의 관계로써 정의 내릴 수 있다.

		실제 정답	
		True	False
분류 결과	True	True Positive	False Positive
	False	False Negative	True Negative

<Fig1. Confusion matrix>

- True Positive(TP) : 실제 True인 정답을 True라고 예측 (정답)
- False Positive(FP) : 실제 False인 정답을 True라고 예측 (오답)
- False Negative(FN) : 실제 True인 정답을 False라고 예측 (오답)
- True Negative(TN) : 실제 False인 정답을 False라고 예측 (정답)

핸즈온 머신러닝

PART3. 분류성능평가지표

1. Precision, Recall and Accuracy

Precision, Recall, Accuracy는 논문에서도 사용하는 지표들이며 가장 눈에 익는 지표들이다.

1.1 Precision(정밀도)
정밀도란 모델이 True라고 분류한 것 중에서 실제 True인 것의 비율이다

$$(Precision) = \frac{TP}{TP + FP}$$

		실제 정답	
		True	False
분류 결과	True	True Positive	False Positive
	False	False Negative	True Negative

<Fig1. Confusion matrix>

Positive 정답률, **PPV(Positive Predictive Value)**라고도 불린다. 날씨 예측 모델이 맑다로 예측했는데, 실제 날씨가 맑았는지를 살펴보는 지표라고 할 수 있다.

한달 30일 동안 맑은 날이 20일이었는데, 확실한 2일만 맑다고 예측한다면, 당연히 맑다고 한 날 중에 실제 맑은 날(Precision)은 100%가 나오게 됩니다. 하지만 과연, 이러한 모델이 이상적인 모델일까요?

핸즈온 머신러닝

PART3. 분류성능평가지표

1. Precision, Recall and Accuracy

Precision, Recall, Accuracy는 논문에서도 사용하는 지표들이며 가장 눈에 익는 지표들이다.

1.2 Recall(재현율)

$$(Recall) = \frac{TP}{TP + FN}$$

재현율이란 실제 True인 것 중에서 모델이 True라고 예측한 것의 비율이다.
통계학에서는 **sensitivity**으로, 그리고 다른 분야에서는 **hit rate**라는 용어로도 쓰임

Precision이나 Recall은 모두 실제 True인 정답을 모델이 True라고 예측한 경우에 관심이 있으나, 바라보고자 하는 관점만 다르다. Precision은 모델의 입장에서, 그리고 Recall은 실제 정답(data)의 입장에서 정답을 정답이라고 맞춘 경우를 바라보고 있다.

따라서, 우리는 실제 맑은 20일 중에서 예측한 맑은 날의 수도 고려해 보아야합니다. 이 경우에는 Precision만큼 높은 결과가 나오지 않습니다. Precision과 함께 Recall을 함께 고려하면 실제 맑은 날들(즉, 분류의 대상이 되는 정의역, 실제 data)의 입장에서 우리의 모델이 맑다고 예측한 비율을 함께 고려하게 되어 제대로 평가할 수 있습니다. Precision과 Recall은 상호보완적으로 사용할 수 있으며, 두 지표가 모두 높을 수록 좋은 모델입니다.

핸즈온 머신러닝

PART3. 분류성능평가지표

1. Precision, Recall and Accuracy

Precision, Recall, Accuracy는 논문에서도 사용하는 지표들이며 가장 눈에 익는 지표들이다.

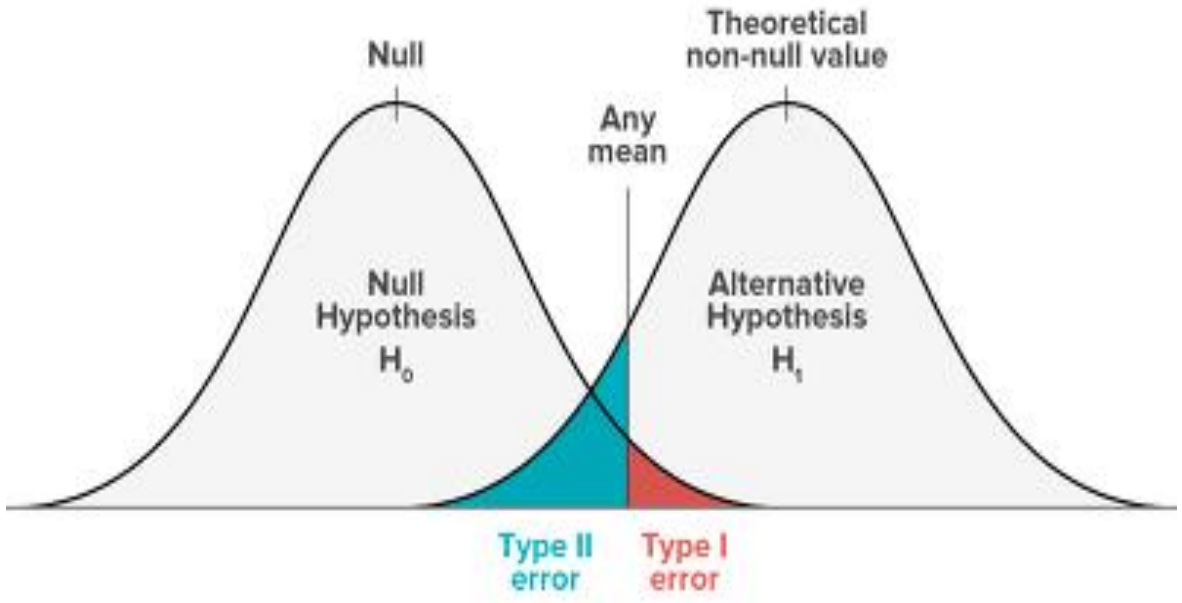
1.3 Precision-Recall Trade-off

Any mean을 좌우로 조정하게되면 Type 1, 2 error의 크기가 변한다. 하지만 둘다 커지거나 둘다 작아지는 경우가 없는 trade-off관계이다.

$$(Type\ 1\ error) = P(\ accept\ H_0\ |\ H_0\ is\ not\ true\ )$$
$$(Type\ 2\ error) = P(\ reject\ H_0\ |\ H_0\ is\ true\ )$$

		H_0	
		True	False
Test result	Accept		Type 1 error
	Reject	Type 2 error	

<Fig2. Type 1, 2 error>



핸즈온 머신러닝

PART3. 분류성능평가지표

1. Precision, Recall and Accuracy

Precision, Recall, Accuracy는 논문에서도 사용하는 지표들이며 가장 눈에 익는 지표들이다.

1.3 Precision-Recall Trade-off

Any mean을 좌우로 조정하게되면 Type 1, 2 error의 크기가 변한다. 하지만 둘다 커지거나 둘다 작아지는 경우가 없는 trade-off관계이다.

		실제 정답	
		True	False
분류 결과	True	True Positive	False Positive
	False	False Negative	True Negative

<Fig1. Confusion matrix>

		H_0	
		True	False
Test result	Accept		Type 1 error
	Reject	Type 2 error	

<Fig2. Type 1, 2 error>

다시 본론으로 돌아와서, Precision과 Recall은 TP를 분자로써 같이하고 분모에는 TP에 Type 1, 2 error에 해당하는 FN, FP를 더하여 계산한다.

이때, FN, FP는 각각 Type 1, 2 error에 있으므로 Precision과 Recall 또한 trade-off 관계에 있다고 할 수 있다.

핸즈온 머신러닝

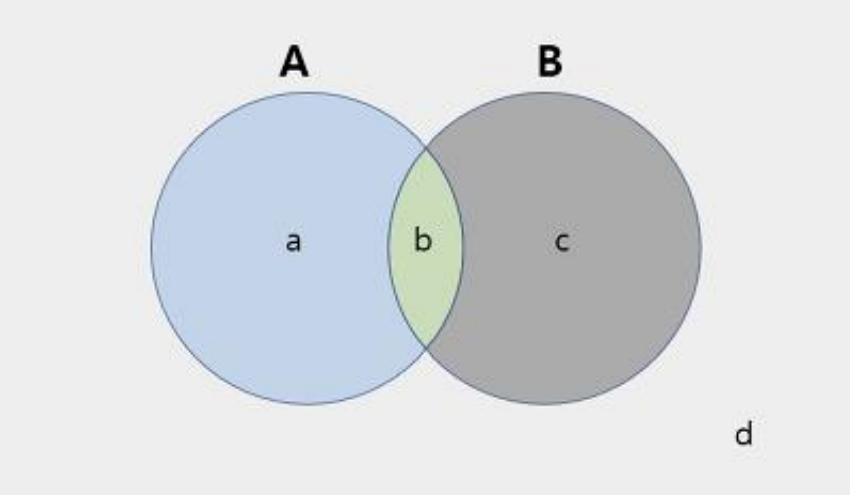
PART3. 분류성능평가지표

1. Precision, Recall and Accuracy

Precision, Recall, Accuracy는 논문에서도 사용하는 지표들이며 가장 눈에 익는 지표들이다.

1.3 Precision-Recall Trade-off

A는 실제 날씨가 맑은 날입니다. 그리고 B는 모델에서 날씨가 맑은 날이라고 예측한 것이다.
이때 b의 영역은 TP로 실제 맑은 날씨를 모델이 맑다고 제대로 예측한 영역이다.



<Fig3. classification with venn-diagram>

$$(Precision) = \frac{b}{b+c}$$

$$(Recall) = \frac{b}{a+b}$$

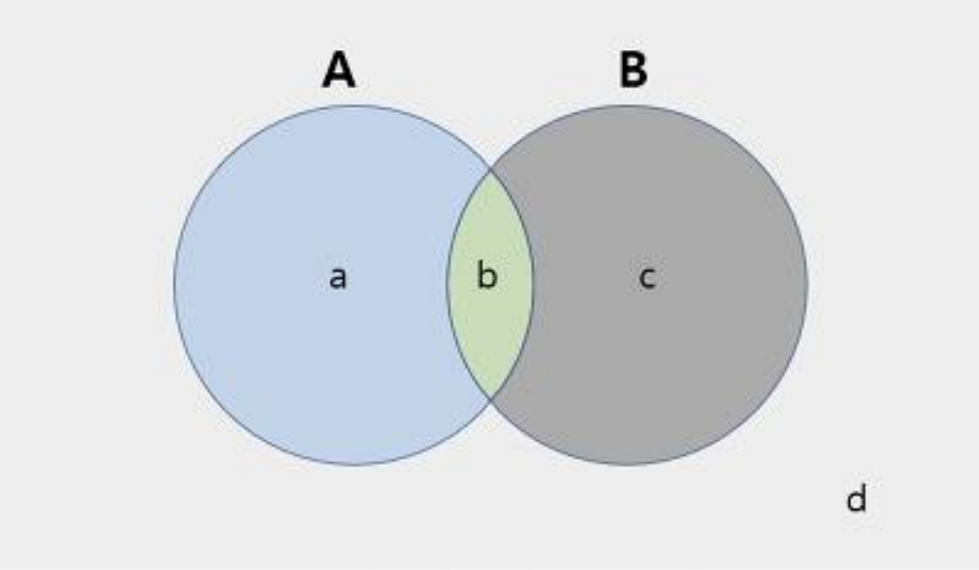
핸즈온 머신러닝

PART3. 분류성능평가지표

1. Precision, Recall and Accuracy

1.3 Precision-Recall Trade-off

모델의 입장에서 모두 맑은 날이라고만 예측하는 경우를 생각해보면 TN(d)의 영역이 줄어들게 되고 그에 따라 FN(a)의 영역 또한 줄게 된다. 그러므로 Recall은 분모의 일부인 FN(a)영역이 줄기 때문에 Recall은 100%가 된다.
즉, 여기서 A⊂B인 관계를 형성한다. 하지만, 주의할 것은 단순히 a의 영역만 줄어드는 것이 아니라 d의 영역과 a의 영역이 모두 c로 흡수된다는 것이다. Precision의 경우에는 기존보다 FP(c)의 영역이 커져 Precision은 줄게 된다.



<Fig3. classification with venn-diagram>

$$(Precision) = \frac{b}{b+c}$$

$$(Recall) = \frac{b}{a+b}$$

핸즈온 머신러닝

PART3. 분류성능평가지표

1. Precision, Recall and Accuracy

1.4 Accuracy(정확도)

위 두 지표는 모두 True를 True라고 옳게 예측한 경우에 대해서만 다루었다.
하지만, False를 False라고 예측한 경우도 옳은 경우이다. 이때, 해당 경우를 고려하는 지표가 바로 **정확도(Accuracy)이다.**

$$(Accuracy) = \frac{TP + TN}{TP + FN + FP + TN}$$

정확도는 가장 직관적으로 모델의 성능을 나타낼 수 있는 평가 지표이다.
하지만, 여기서 domain의 편중(bias)을 고려해야 한다. 만약 우리가 예측하고자 하는 한달 동안이 특정 기후에 부합하여 비오는 날이 흔치 않다고 가정하면, 이 경우에는 해당 data의 domain이 불균형하게되므로 맑은 것을 예측하는 성능은 높지만, 비가 오는 것을 예측하는 성능은 매우 낮을 수 밖에 없다.
따라서 이를 보완할 지표가 필요하다.

핸즈온 머신러닝

PART3. 분류성능평가지표

1. Precision, Recall and Accuracy

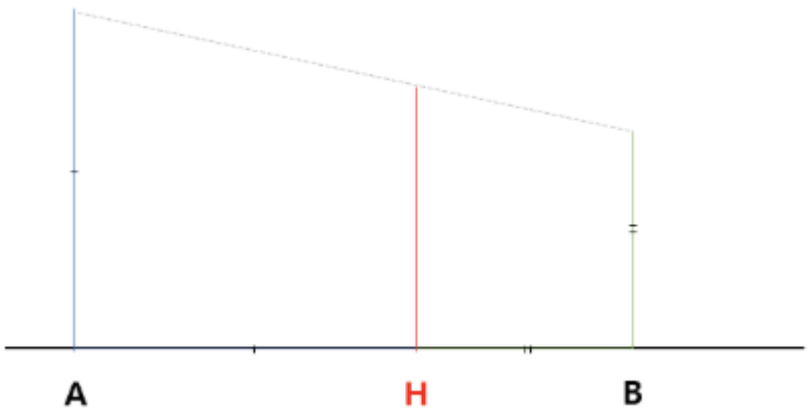
$$(F1\text{-}score) = 2 \times \frac{1}{\frac{1}{Precision} + \frac{1}{Recall}} = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

1.5 F1 score

F1 score는 Precision과 Recall의 조화평균이다.
데이터 label이 불균형 구조일 때, 모델의 성능을 정확하게 평가할 수 있으며, 성능을 하나의 숫자로 표현할 수 있다.

- **조화평균의 기하학적 접근**

서로 다른 길이의 A, B와 이 두 길이의 합만큼 떨어진 변(AB)으로 이루어진 사다리꼴을 생각하면
이 AB에서 각 변의 길이가 만나는 지점으로부터 맞은 편의 사다리꼴의 변으로 내린 선분이 바로 조화평균을 나타낸다.



기하학적으로 봤을 때, 단순 평균이라기보다는 작은 길이 쪽으로 치우치게 된,
그러면서 작은 쪽과 큰 쪽의 사이의 값을 가진 평균이 도출된다.
이렇게 조화평균을 이용하면 산술평균을 이용하는 것보다, 큰 비중이 끼치는 bias가 줄어든다고 볼 수 있다.

핸즈온 머신러닝

PART3. 분류성능평가지표

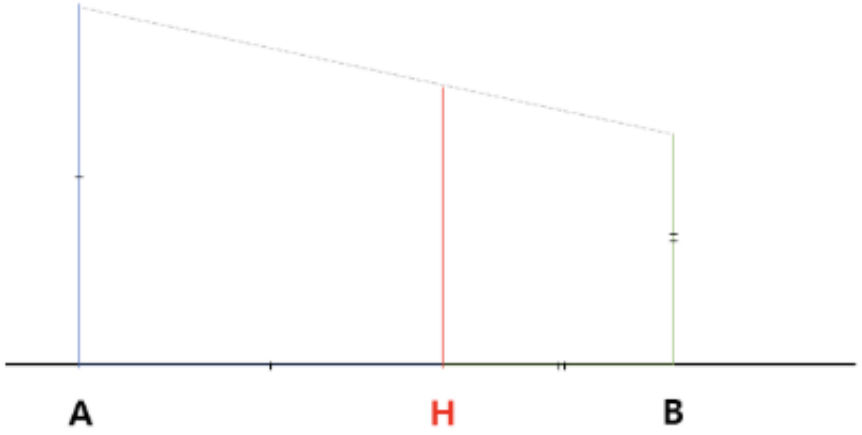
$$(F1\text{-}score) = 2\times\frac{1}{\frac{1}{Precision}+\frac{1}{Recall}} = 2\times\frac{Precision \times Recall}{Precision + Recall}$$

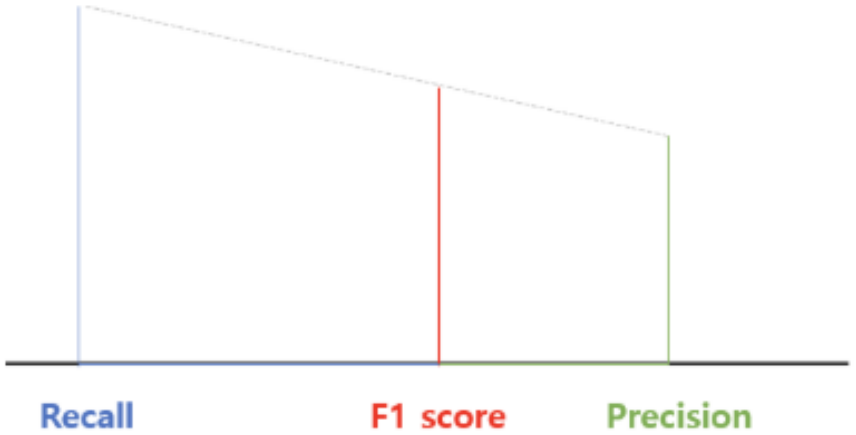
1. Precision, Recall and Accuracy

1.5 F1 score

F1 score는 Precision과 Recall의 조화평균이다.
데이터 label이 불균형 구조일 때, 모델의 성능을 정확하게 평가할 수 있으며, 성능을 하나의 숫자로 표현할 수 있다.

- **조화평균의 기하학적 접근**
서로 다른 길이의 A, B와 이 두 길이의 합만큼 떨어진 변(AB)으로 이루어진 사다리꼴을 생각하면
이 AB에서 각 변의 길이가 만나는 지점으로부터 맞은 편의 사다리꼴의 변으로 내린 선분이 바로 조화평균을 나타낸다.





핸즈온 머신러닝

PART3. 분류성능평가지표

1. Precision, Recall and Accuracy

$$(F1\text{-}score) = 2 \times \frac{1}{\frac{1}{Precision} + \frac{1}{Recall}} = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

1.5 F1 score

정밀도와 재현율이 비슷한 분류기에서는 F1 점수가 높다. 하지만 항상 바람직한 것은 아니다. 상황에 따라 정밀도가 중요할 수도 있고, 재현율이 중요할 수도 있다.

예를 들어 어린아이에게 안전한 동영상을 걸러내는 분류기를 훈련시킨다 가정하면 재현율이 높으나
정말 나쁜 동영상이 몇 개 노출되는 것보다 좋은 동영상이 많이 제외되더라도 (낮은 재현율)
안전한 것들만 노출시키는 (높은 정밀도) 분류기를 선호할 것이다.

다른 예로 감시 카메라를 통해 좀도둑을 잡아내는 분류기를 훈련시킨다고 가정하면 분류기의 재현율이 99%라면
정확도가 30%만 되더라도 괜찮을지 모른다. 아마도 경비원이 잘못된 호출을 종종 받게 되겠지만 거의 모든 좀도둑을 잡을 것이다.

⇒ 그러나 이 둘을 모두 얻을 수는 없다. 정밀도를 올리면 재현율이 줄고 그 반대도 마찬가지이다.
⇒ 이를 정밀도/재현율 트레이드오프라고 한다.

핸즈온 머신러닝

PART3. 분류성능평가지표

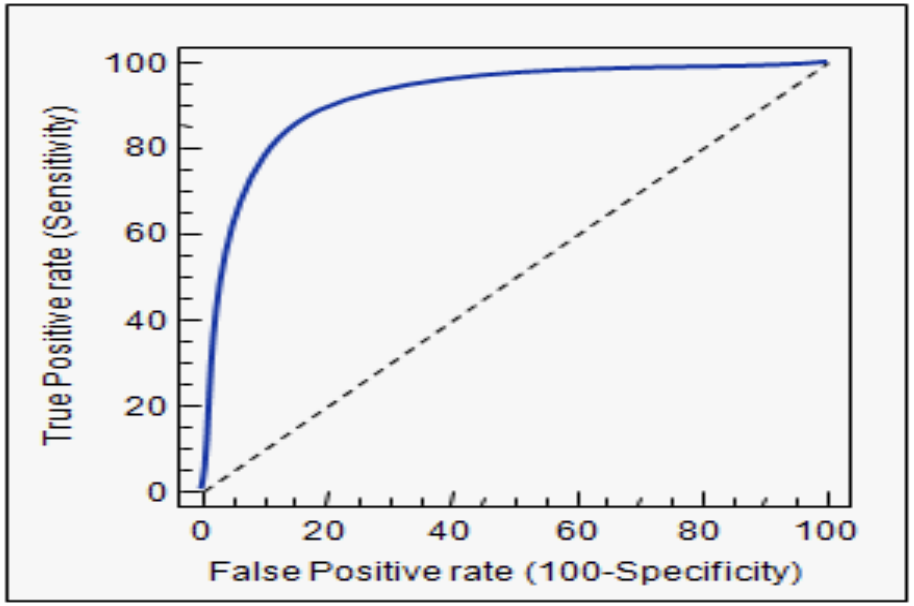
$$Fall\text{-}out(FPR) = \frac{FP}{TN + FP}$$

2. Fall-out

Fall-out은 **FPR(False Positive Rate)**으로도 불리며, 실제 False인 data 중에서 모델이 True라고 예측한 비율이다.
즉, 모델이 실제 false data인데 True라고 잘못 예측(분류)한 것으로 다음과 같이 표현할 수 있습니다.

2.1 ROC(Receiver Operating Characteristic) curve

여러 임계값들을 기준으로 Recall-Fallout의 변화를 시각화한 것이다.
Fallout은 실제 False인 data 중에서 모델이 True로 분류한,
Recall은 실제 True인 data 중에서 모델이 True로 분류한 비율을 나타낸
지표로써, 이 두 지표를 각각 x, y의 축으로 놓고 그려지는 그래프를 해석.



curve가 왼쪽 위 모서리에 가까울수록 모델의 성능이 좋다고 평가한다.
즉, Recall이 크고 Fall-out이 작은 모형이 좋은 모형인 것이다.
y=x 그래프보다 상단에 위치해야 어느정도 성능이 있다고 말할 수 있다.

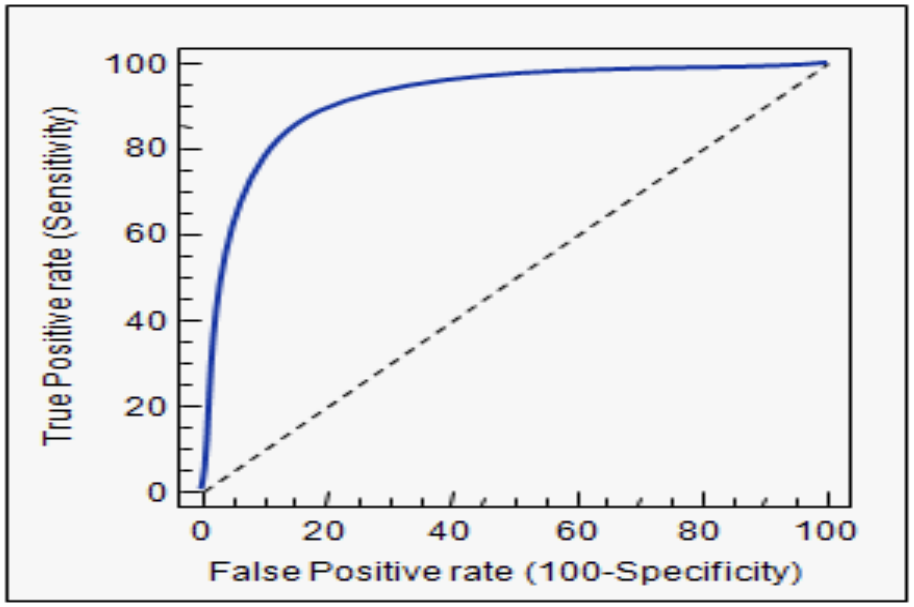
2.2 AUC(Area Under Curve)

ROC curve는 그래프이기 때문에 명확한 수치로써 비교하기가 어렵다. 따라서 그래프 아래의 면적값을 이용한다.
최대값은 1이며 좋은 모델(즉, Fall-out에 비해 Recall 값이 클수록) 1에 가까운 값이 나온다

핸즈온 머신러닝

PART3. 분류성능평가지표

2.1 ROC(Receiver Operating Characteristic) curve



거짓 양성 비율(FPR)에 대한 진짜 양성 비율(TPR:재현율의 다른 이름) 의
양성으로 잘못 분류된 음성 샘플의 비율이 FPR이다.
이는 1에서 음성으롯 정확하게 분류한 음성 샘플의 비율인
진짜 음성 비율(TNR)을 뺀 값이다. TNR을 특이도라고 한다.
그러므로 ROC 곡선은 민감도(재현율)에 대한 1-특이도 그래프이다.

재현율이 높을수록 분류기가 만드는 거짓 양성이 늘어난다. 점선은 완전한 랜덤 분류기의 ROC 곡선을 뜻한다.
좋은 분류기는 이 점선에서 멀리 떨어져 있어야 한다.

곡선 아래의 면적(AUC)을 측정하면 분류기들을 비교할 수 있다.

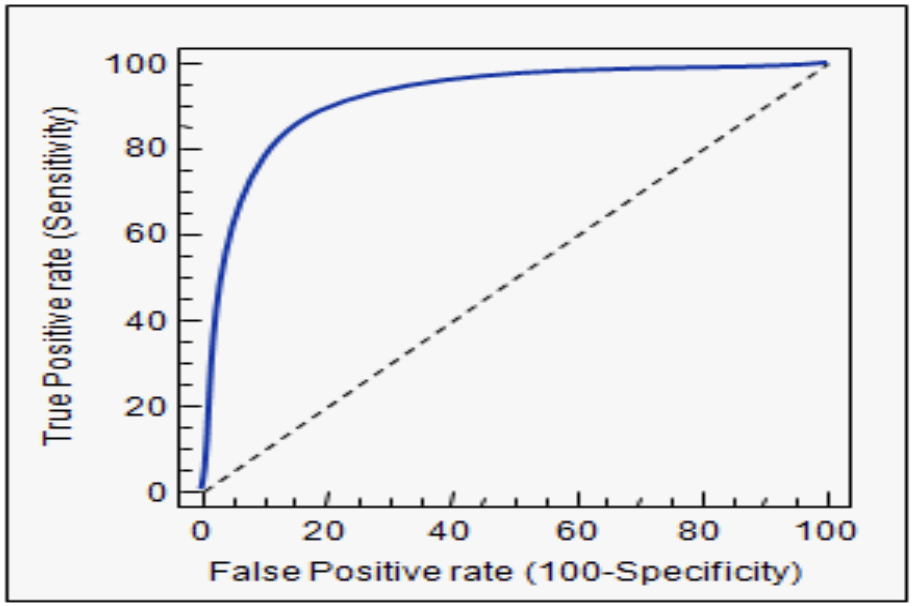
완벽한 분류기는 ROC의 AUC가 1이고,

완전한 랜덤 분류기(훈련 데이터의 클래스 비율에 따라 무작위로 예측하는 것. 실제 클래스가 비슷한 비율의 예측 클래스로 나뉘어 FPR과 TPR 값이 비슷해진다. 결국 ROC 곡선이 직선과 가깝게 되어 AUC 면적이 0.5가 된다) 는 0.5이다.

핸즈온 머신러닝

PART3. 분류성능평가지표

2.1 ROC(Receiver Operating Characteristic) curve



ROC 곡선이 정밀도/재현율 (PR 곡선)과 비슷해서 어떤 것을 사용해야 할지 궁금할 수 있다. 일반적인 법칙은 양성 클래스가 드물거나 거짓 음성보다 거짓 양성이 더 중요할 때 PR 곡선을 사용하고 그렇지 않으면 ROC 곡선을 사용한다.

예를 들어 조금 전의 ROC 곡선(그리고 ROC의 AUC 점수)을 보면 매우 좋은 분류기라고 생각할 수 있다. 하지만 이는 음성(5 아님)에 비해 양성(5임)이 크게 적기 때문이다.

이와는 다르게 PR 곡선은 분류기의 성능 개선 여지가 얼마나 되는지 잘 보여준다. (오른쪽 위 모서리에 가까워질 수 있는지)