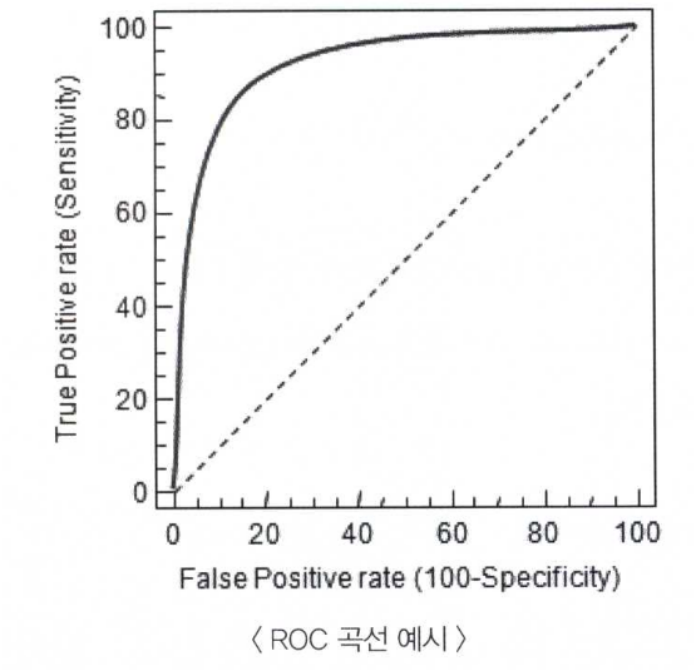
ROC 곡선과 AUC

```
# ROC 곡선
```

- Receiver Operating Characteristic Curve의 약자
- ROC 곡선은 원래 군사 영역에서 유래된 개념으로 나중에 의학 영역에서 발전하였음
- 수신자 조작 특성 곡선이라는 명칭도 의학 영역에서 유래된 것

FPR = FP / (FP + TN) = 1 - TNR = 1- 특이성



〈 ROC 곡선 예시 〉

의미

- ROC 곡선의 가로축은 거짓 양성 비율(False Positive Rate, FPR)을 나타내고, 세로축은 실제 양성 비율(True Positive Rate, TPR)을 나타냄
- FPR(특이성)= FP/N , TPR(민감도)= TP/P
  - 특이성은 음성 클래스를 정확하게 예측하는 능력
  - 특이성이 높을수록 모델은 음성 클래스를 더 잘 예측
  - 민감도는 양성 클래스를 정확하게 예측하는 능력
  - 민감도가 높을수록 모델은 양성 클래스를 더 잘 예측
- 위 식에서 P는 실제 양성 샘플 수, N은 실제 음성 샘플 수를 의미

- TP는 P의 양성 샘플 중에서 분류기가 양성 샘플로 예측한 샘플의 개수를 나타내고, FP는 N개의 음성 샘플 중에서 분류기가 양성 샘플로 예측한 샘플의 개수를 나타냄

예시 1

- 10명의 암 의심 환자가 있는데 여기서 3명만 실제 암에 걸렸다고 가정(P=3). 그 외 7명은 암에 걸리지 않았음(N=7)
- 병원에서 10명의 환자에 대한 진단을 해서 3명의 암환자가 있다고 결론을 내렸습니다. 하지만 여기서 실제 암환자는 2명뿐(TP=2) 그렇다면 실제 양성 비율 TPR = TP / P = 2/3으로 계산할 수 있습니다.
- 불행하게도 7명의 암에 걸리지 않은 환자들 중 한 명이 오진을 받았습니다. (FP = 1) 그렇다면 FPR = FP / N = 1 / 7
- 이 분류기의 분류 결과는 ROC 곡선상의 점 (1/7, 2/3)이 됩니다.

예시 2

표 2.1 이진분류 모델의 출력 결과 샘플

샘플 인덱스	실제 레이블	모델 출력확률	샘플 인덱스	실제 레이블	모델 출력확률
1	p	0.9	11	p	0.4
2	p	0.8	12	n	0.39
3	n	0.7	13	p	0.38
4	p	0.6	14	n	0.37
5	p	0.55	15	n	0.36
6	p	0.54	16	n	0.35
7	n	0.53	17	p	0.34
8	n	0.52	18	n	0.33
9	p	0.51	19	p	0.30
10	n	0.505	20	n	0.1

- 테스트 세트에 20개의 샘플이 있고, 오른쪽 표와 같이 결과를 출력
- 샘플은 예측확률이 높은 순서대로 정렬
- 모델은 양성, 음성의 값으로 출력하기 전에 임곗값을 정해 주어야 함
- 예측확률이 임곗값보다 높다면 양성으로 판별되고,
- 임곗값보다 작다면 음성으로 분류

- 예를 들어 임곗값이 0.9라면 첫 번째 샘플만 양성으로 예측되고, 나머지는 모두 음성으로 예측
- 임곗값은 동적으로 조절할 수 있는데, 높은 점수부터 시작해서 낮은 점수로 이동시키고, 각 임곗값은 모두 하나의 FPR과 TPR에 대응합니다.
- ROC 그림에서 각 절단점에 대응하는 위치를 그리고 모든 점을 연결하면 최종적으로 ROC 곡선을 얻을 수 있습니다.

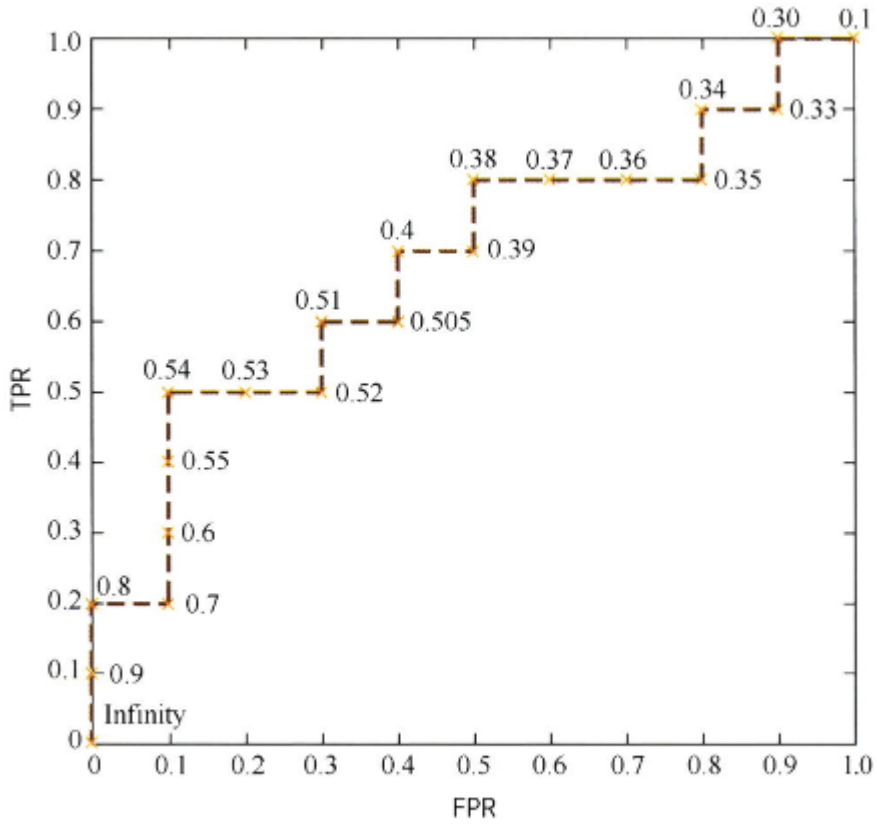


그림 2.2 ROC 곡선

- 임곗값 0.9일 때 모델은 1번 샘플을 양성 샘플로 예측
- 예제에서 양성 샘플의 수는 10개(P=10)
- TP는 1이 되고, TPR = TP / P = 1/ 10
- 이때 잘못 예측한 양성 샘플이 없기 때문에 FP는 0이 됨
- FPR = 0 / 10 = 0
- 곡선 (0, 0.1)
- 임곗값이 0.7일 때 TP는 2이 되고, 2 / 10
- FP는 1이 되고, 1/ 10 곡선 (0.1 , 0.2)
  Threshold -> 0.52
  1 ~ 8 -> 양성으로 모델이… 판별 5 / 10 = 0.5
  3 / 10 = 0.3 -> (0.3, 0.5)

AUC

•AUC(Aear Under Curve)는 ROC 곡선 아래의 면적
•ROC 곡선에 기반해 모델 성능을 정량화하여 나타낼 수 있음
•AUC값을 계산하기 위해서는 ROC 곡선의 x축을 따라 적분만 하면 됨
•일반적으로 0.5 ~ 1 사이에 있고, AUC가 클수록 분류기의 성능이 더 좋다는 것을 나타냄

ROC 곡선의 유래

- 레이더 병사들의 보고 정성확을 연구하기 위해 관리자는 모든 레이더병의 보고 특징을 종합
- 잘못 보고하거나 누락시킨 보고에 대한 각각의 확률을 2차원 좌표계에 그림
- y축은 민감성(실제 양성율), 즉 모든 적군의 기습 사건 중 각 레이더병이 정확하게 예측한 확률
- x축은 1-특이성(거짓 양성율), 모든 적군이 아닌 신호 중에서 레이더병이 잘못 보고한 확률
- 각 레이더병의 보고 기준이 다르기 때문에 얻은 민감성과 특이성의 조합도 다름
- 레이더병의 보고 성능에 대한 종합한 후, 관리자는 그들이 하나의 곡선상에 놓여 있다는 것을 발견
- 이 곡선이 바로 ROC곡선