3.1~3.3

Chapter 3. 분류

Contents

- Loss func, GD, MSGD
- 오차 행렬 – 정밀도, 재현률, F1 score
- 결정 함수, 임곗값
- ROC 곡선, PR곡선

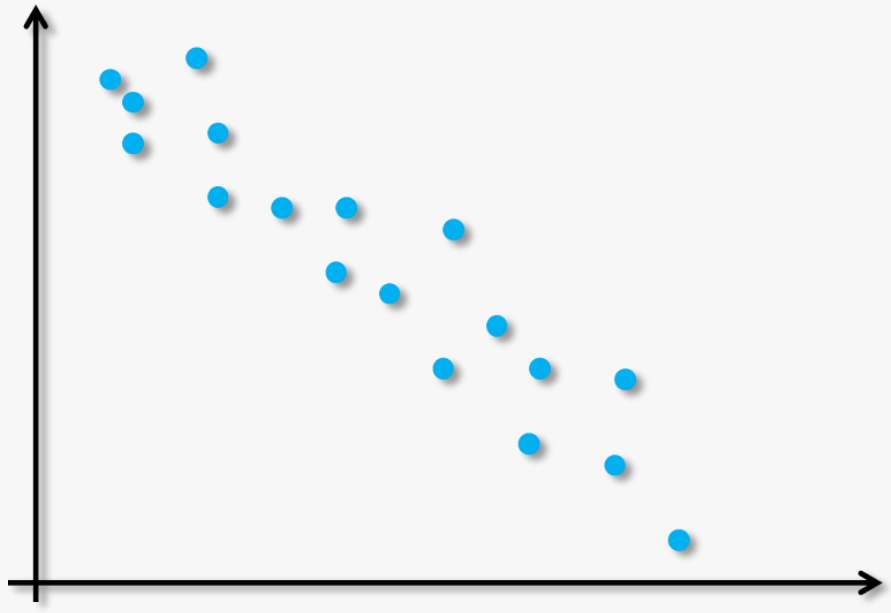
손실 함수 (Loss Func.)

- 신경망이 학습할 수 있도록 해주는 지표
- 머신러닝 모델의 출력값과 사용자가 원하는 출력값의 차이 (=오차)
- 손실 함수 값이 최소화되도록 하는 가중치, 편향을 찾는 것이 바로 학습
- ex.평균 제곱 오차(MSE), 교차 엔트로피 오차

경사 하강법 (Gradient Descent)

- 학습률과 손실함수의 순간기울기(gradient)를 이용하여 가중치(weight)를 업데이트하는 방법.

경사 하강법 (Gradient Descent)



사람은 직관적으로 회귀선을 그을 수 있으나,

기계의 경우는 불가능

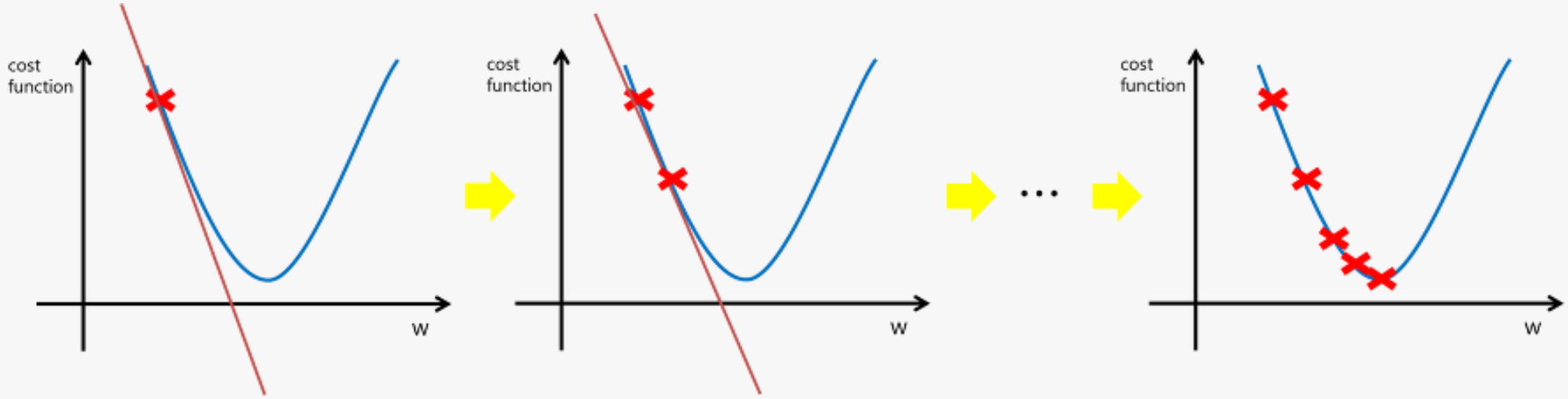
➔ 임의로 기울기와 절편값을 주고,

MSE를 구하게 한다

➔ 기계는 MSE를 최소화 시킬 수 있는

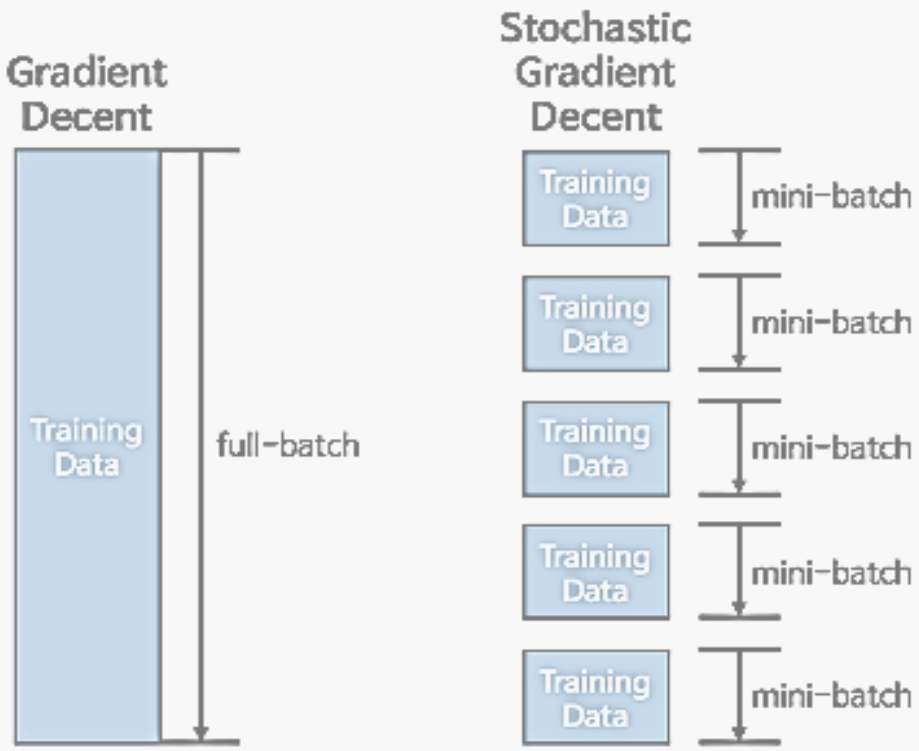
기울기와 절편값을 찾는다

경사 하강법 (Gradient Descent)



- 편미분을 통한 w의 값 갱신으로 최적값을 찾아간다. (기울기가 0이 될 때 까지)
- 이 때 점차 접선의 경사가 감소하여 경사 하강법이라고 한다.

미니 배치 경사 하강법 (Mini Batch Gradient Descent)



- GD를 전체 데이터(batch)가 아닌 일부 데이터(mini batch)의 모음으로 사용
- 각각의 batch는 gd보다 부정확하나 여러번 반복하면 정답으로 수렴
- 훈련시 무작위성을 사용하기 때문에 이름에 “확률적”이 붙음.

Confusion Matrix

- 지도 학습의 성능 시각화 목적 (비지도 학습 : Matching Matrix)

	Actually Positive (1)	Actually Negative (0)
Predicted Positive (1)	True Positives (TPs)	False Positives (FPs)
Predicted Negative (0)	False Negatives (FNs)	True Negatives (TNs)

정밀도와 재현률

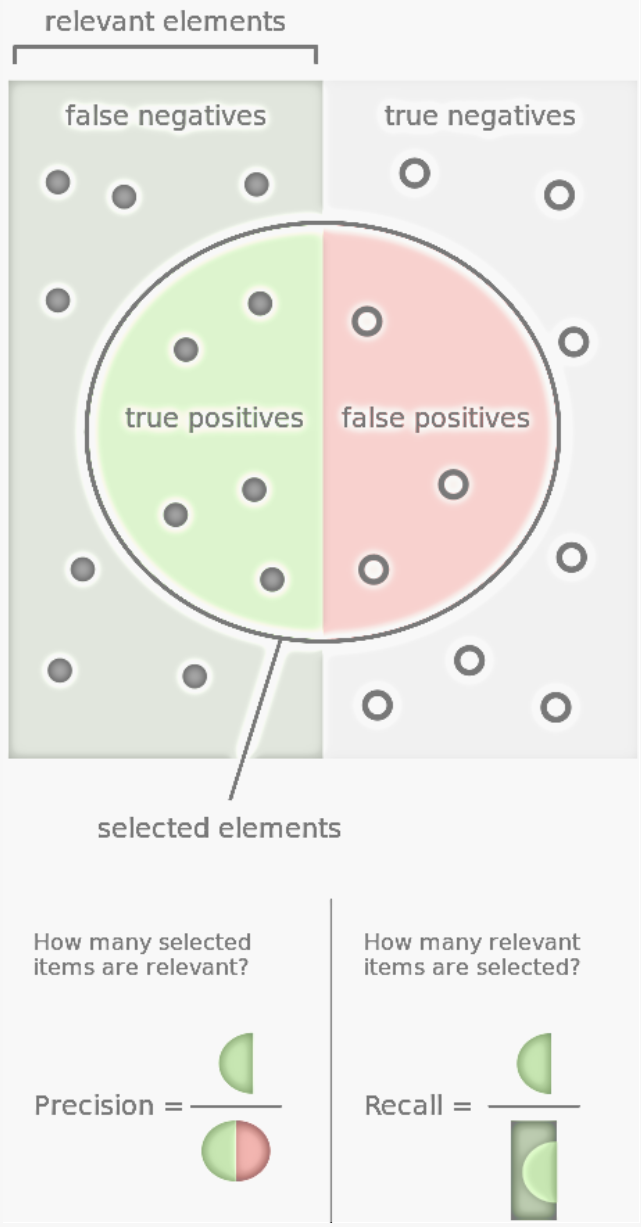
- 정밀도 : 날씨가 맑다고 예측 시, 진짜 맑을 비율
- 재현률 : 맑은 날, 맑을 것이라고 예측한 비율

- 30일 중 확실히 맑은 2일을 제외한 나머지는 예측 보류

  ➔ 그 이틀은 정밀도가 높지만, 예측을 보류한

  다른 맑은 날들을 고려하면 모델이 유용하지 않음

- Precision과 Recall 모두 고려해야 함



F1 score

- 정밀도와 재현률의 조화평균
- 정확도, 정밀도보다 모델의 성능을 정확히 평가 가능

  ➔ 산술평균에 비하여 상대적으로 큰 값이 끼치는 Bias가 줄어듦

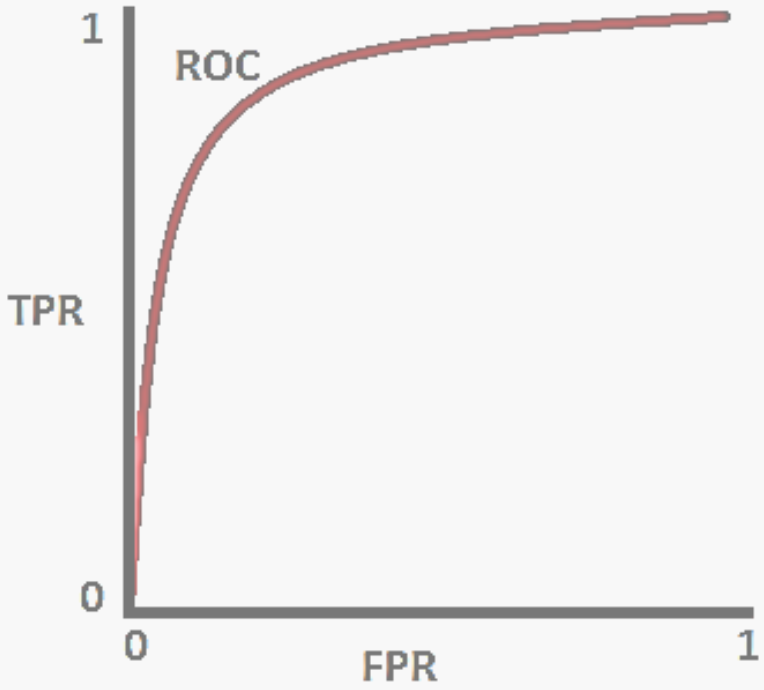
  ➔ 정밀도와 재현률이 모두 높아야 좋은 모델이기 때문에 F1을 성능의 척도로 사용

ROC곡선

- Receiver operating characteristics, ROC
- 머신러닝 모델 평가에 사용
- 검사, 모델링의 임곗값 설정에도 활용
- AUC 면적이 넓을수록 모델의 성능이 좋다고 평가
- AUC(Area Under the Curve)

ROC곡선

1. TPR과 FPR
2. ROC 상의 Point가 의미하는 것
3. Curve가 의미하는 것



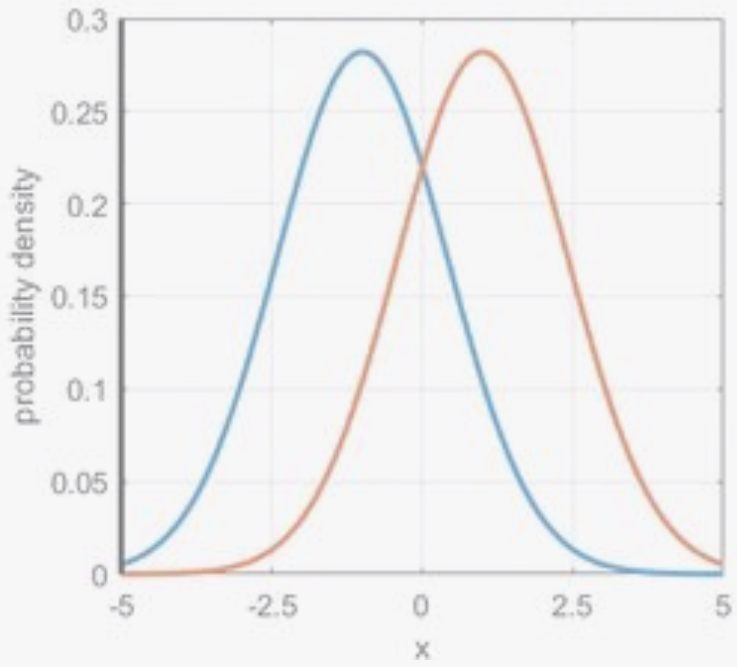
1) TPR과 FPR

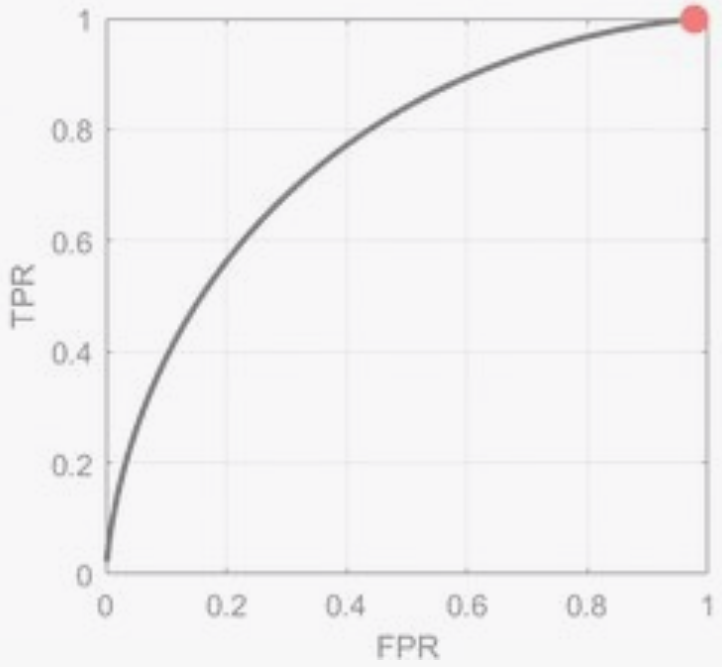
- TPR=민감도 (1을 1로 예측)
- FPR=1-특이도 (0을 1로 예측)
- TPR과 FPR은 임곗값에 따라 변하며, 비례한 경향을 띔
  - 모든 예측을 1로 하면 TPR과 FPR 모두 높음 (=threshold가 낮은 상태)
  - 모든 예측을 0으로 하면 TPR과 FPR 모두 높음 (=threshold가 높은 상태)

  ➔ 정확한 예측을 위한 threshold 설정이 필요

2) ROC 상의 Point가 의미하는 것

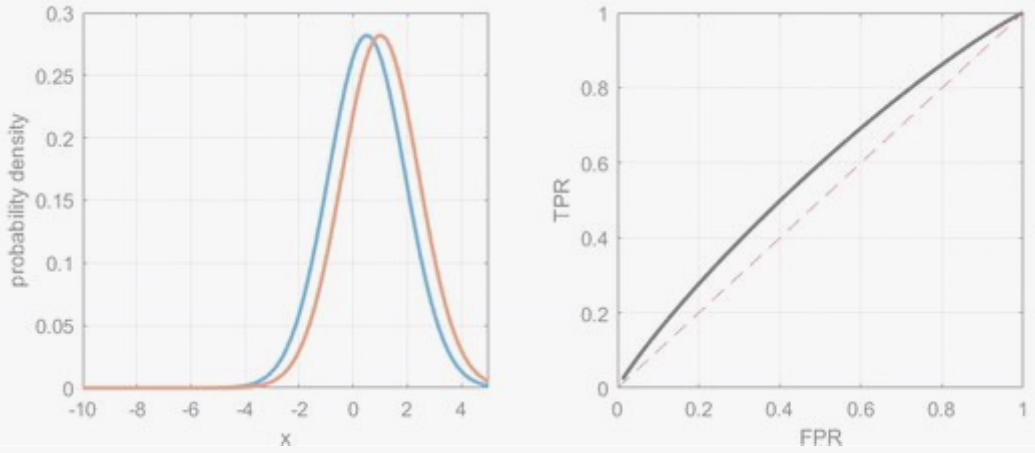
- 현 위의 점 : threshold에 따른 FPR과 TPR을 표현한 것.
- 현재 분류 모델을 그대로 이용하며, 임곗값을 바꾸었을 때 FPR과 TPR의 비율을 표현





3) Curve가 의미하는 것

- 모델의 분류 정확도가 높을수록 ROC Curve는 좌상단에 가까워짐
- 분류 정확도가 높으면 정확도가 낮을 때 보다
  임곗값에 영향을 크게 받지 않는다



PR곡선

- 데이터 Lable 분포 불균등이 심할 경우 ROC의 대안으로 주로 사용
- ROC Curve는 불량품 혹은 이상치 검출의 경우 필연적으로
  정상 lable에 데이터가 치우쳐짐

PR곡선

- 성능 평가 방법 :

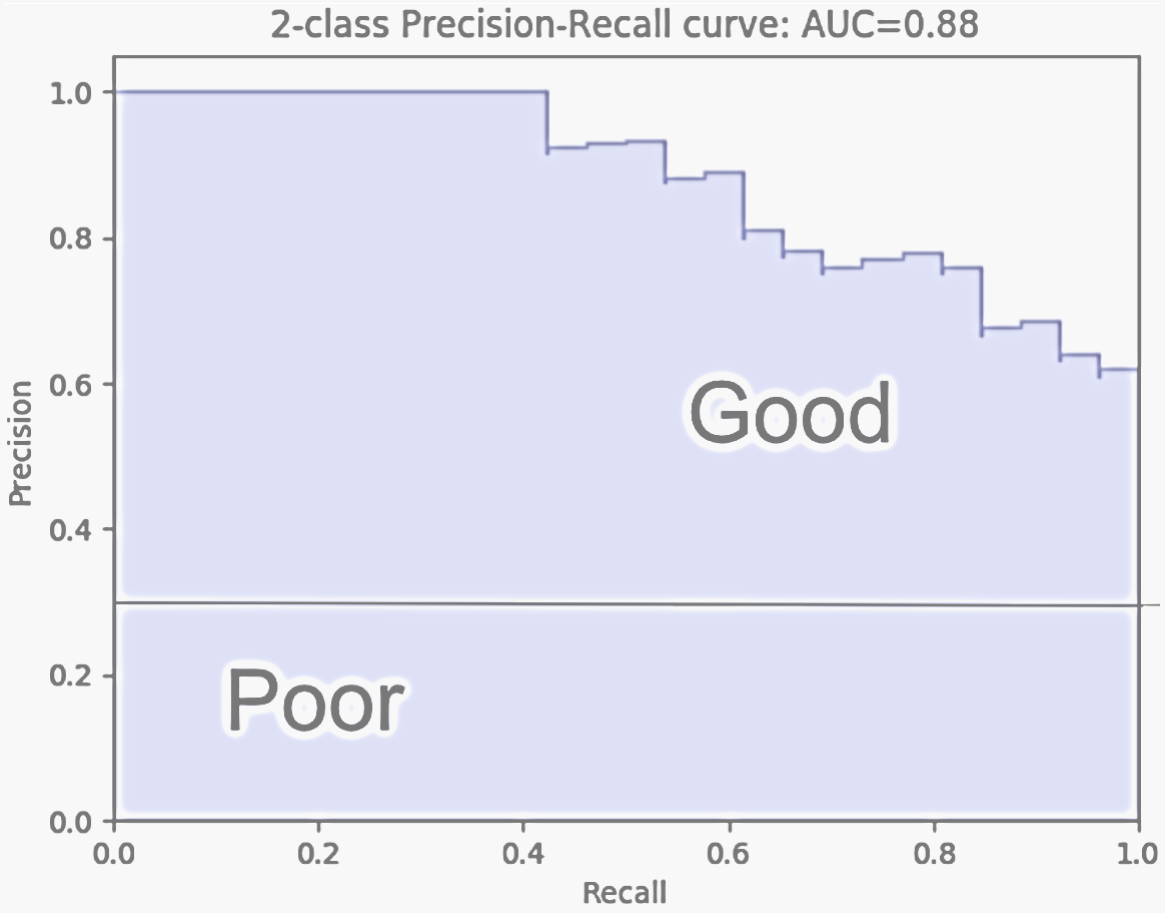
  Base Line을 기준으로

  모델의 성능을 판단

- Baseline= P/(P+N)

  (P : Positive lable 수

  N: 전체 데이터 수)



^0^