지도학습 : 분류
Classification

이 건 명
충북대학교 소프트웨어학과

인공지능 : 튜링 테스트에서 딥러닝까지

학습 내용

- 분류 문제에서의 데이터의 특성을 알아본다
- 학습 모델의 과적합과 부적합에 대해서 알아본다.
- 학습 모델의 성능 평가 방법을 알아본다.
- 불균형 부류 데이터 문제와 해결 방법에 대해서 알아본다.
- 이진 분류기의 성능 평가 방법에 대해서 알아본다.

지도학습

❖ **지도학습**(supervised learning)

- 주어진 (입력, 출력) 에 대한 데이터 이용 : 학습(training) 데이터
  - $\{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$
- 새로운 입력이 있을 때 결과를 결정할 수 있도록 하는 방법 찾아내는 것
  - $y = f(x)$

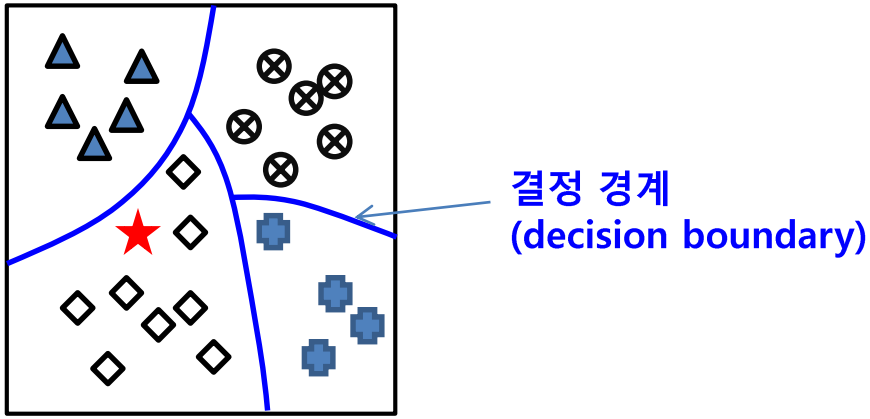
- **분류 (classification)**
  - **출력**이 정해진 **부류**(class, category) 중의 하나로 결정

- **회귀 (regression)**
  - **출력**이 연속인 **영역**(continuous domain)의 값 결정

1. 분류

❖ 분류(classification)

- 데이터들을 정해진 몇 개의 부류(class)로 대응시키는 문제



- **분류 문제의 학습**
  - **학습 데이터**를 잘 **분류**할 수 있는 **함수**를 찾는 것
  - 함수의 형태는 **수학적 함수**일 수도 있고, **규칙**일 수도 있음

- **분류기(classifier)**
  - 학습된 함수를 이용하여 데이터를 분류하는 프로그램

분류

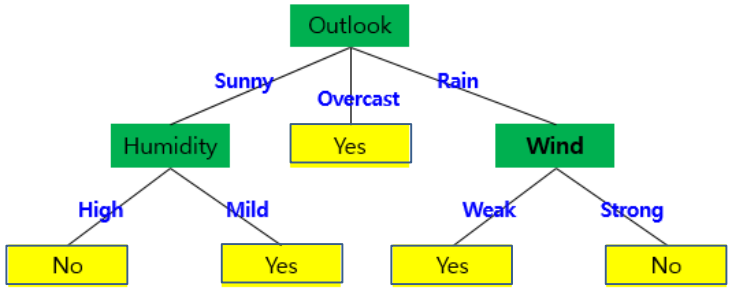
❖ 분류 학습 데이터와 학습 결과의 예

- 범주형 속성

표 4.1 PlayTennis 데이터

Day 날짜	Outlook 조망	Temperature 기온	Humidity 습도	Wind 바람	PlayTennis 테니스 여부
Day1	Sunny	Hot	High	Weak	No
Day2	Sunny	Hot	High	Strong	No
Day3	Overcast	Hot	High	Weak	Yes
Day4	Rain	Mild	High	Weak	Yes
Day5	Rain	Cool	Normal	Weak	Yes
Day6	Rain	Cool	Normal	Strong	No
Day7	Overcast	Cool	Normal	Strong	Yes
Day8	Sunny	Mild	High	Weak	No
Day9	Sunny	Cool	Normal	Weak	Yes
Day10	Rain	Mild	Normal	Weak	Yes
Day11	Sunny	Mild	Normal	Strong	Yes
Day12	Overcast	Mild	High	Strong	Yes
Day13	Overcast	Hot	Normal	Weak	Yes
Day14	Rain	Mild	High	Strong	No

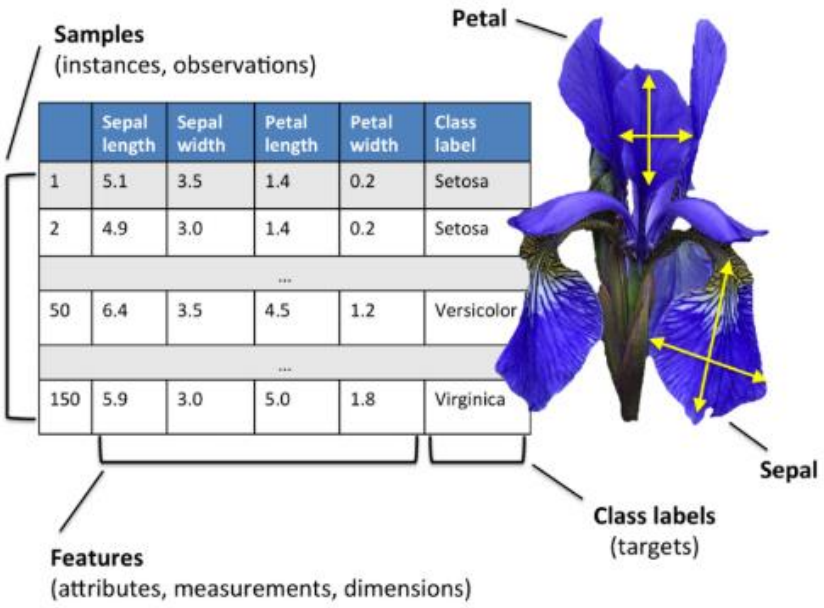
(출처: Machine Learning, Tom Mitchell, 1995)



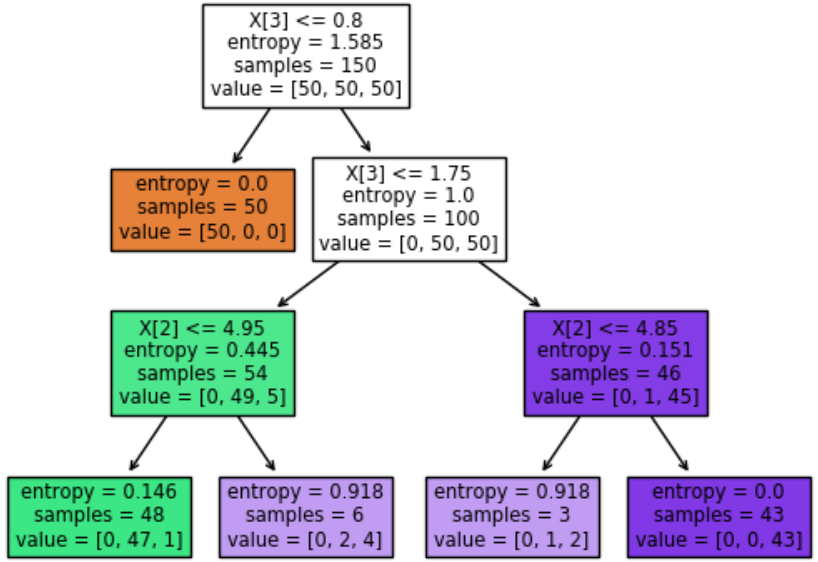
분류

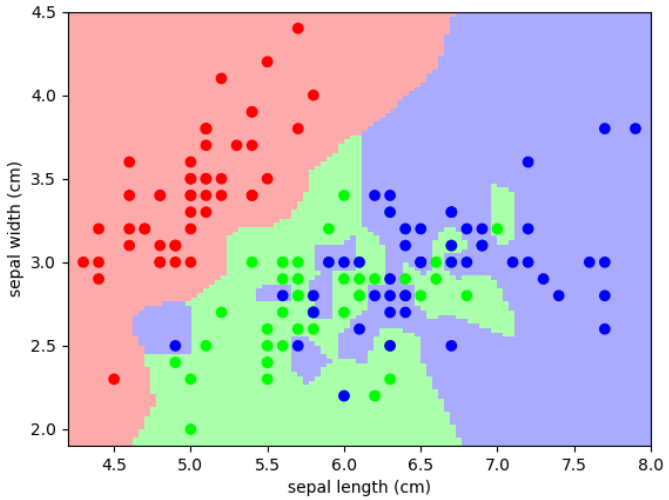
❖ 분류 학습 데이터와 학습 결과의 예

- 수치형 속성



	Sepal length	Sepal width	Petal length	Petal width	Class label
1	5.1	3.5	1.4	0.2	Setosa
2	4.9	3.0	1.4	0.2	Setosa
...					
50	6.4	3.5	4.5	1.2	Versicolor
...					
150	5.9	3.0	5.0	1.8	Virginica





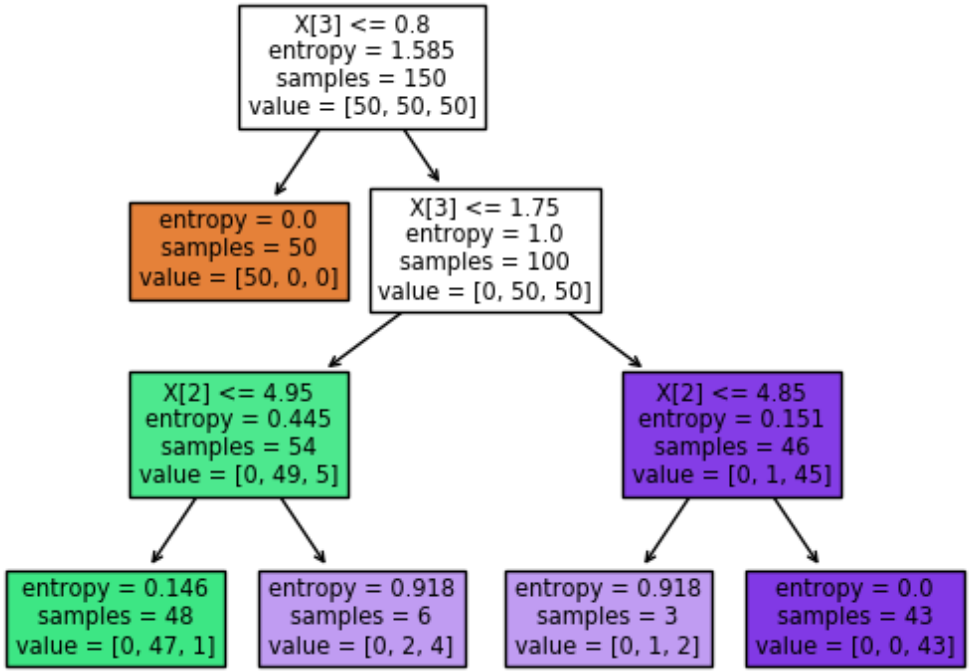
https://medium.com/@jebaseelanravi96/machine-learning-iris-classification-33aa18a4a983
https://pythonmachinelearning.pro/supervised-learning-using-decision-trees-to-classify-data/

분류

```
import matplotlib.pyplot as plt
from sklearn.datasets import load_iris
from sklearn.tree import DecisionTreeClassifier, plot_tree

iris = load_iris()
decision_tree = DecisionTreeClassifier(criterion="entropy", random_state=0, max_depth=3)
decision_tree = decision_tree.fit(iris.data, iris.target)

plt.figure()
plot_tree(decision_tree, filled=True)
plt.show()
```



분류

❖ **분류기 학습 알고리즘**

- **결정트리(decision tree) 알고리즘**
- **K-근접이웃 (K-nearest neighbor, KNN) 알고리즘**
- **다층 퍼셉트론 신경망**
- **딥러닝(deep learning) 신경망**
- **서포트 벡터 머신(Support Vector Machine, SVM)**
- **에이다부스트(AdaBoost)**
- **랜덤 포리스트(random forest)**
- **확률 그래프 모델** (probabilistic graphical model)

분류

❖ 이상적인 분류기

- **학습에 사용되지 않은 데이터**에 대해서 분류를 잘 하는 것
- **일반화**(generalization) **능력**이 좋은 것

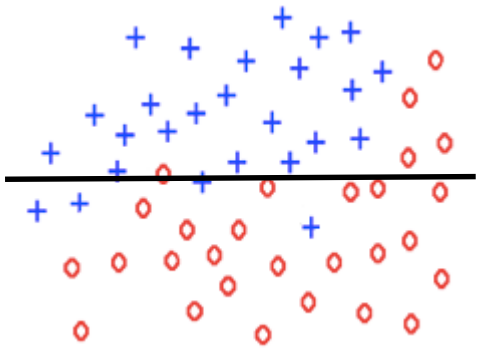
❖ 데이터의 구분

- **학습 데이터**(training data)
  - 분류기(classifier)를 학습하는데 사용하는 데이터 집합
  - 학습 데이터가 많을 수록 유리
- **테스트 데이터**(test data)
  - 학습된 모델의 성능을 평가하는데 사용하는 데이터 집합
  - 학습에 사용되지 않은 데이터
- **검증 데이터**(validation data)
  - 학습 과정에서 학습을 중단할 시점을 결정하기 위해 사용하는 데이터 집합

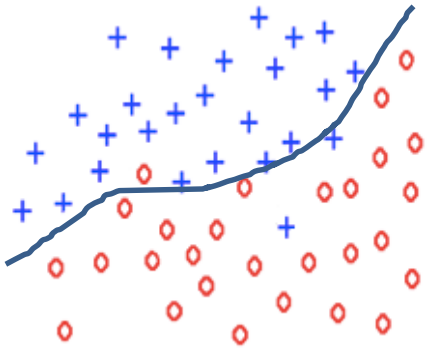
분류

❖ **과적합**(overfitting)과 **부적합**(underfitting)

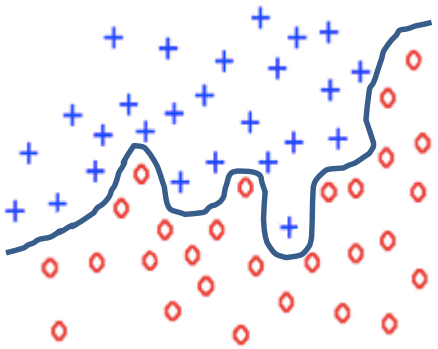
- **과적합**
  - **학습 데이터**에 대해서 **지나치게 잘 학습**된 상태
  - 데이터는 **오류**나 **잡음**을 포함할 개연성이 크기 때문에, 학습 데이터에 대해 매우 높은 성능을 보이더라도 **학습되지 않은 데이터**에 대해 좋지 않은 **성능**을 보일 수 있음
- **부적합**
  - 학습 데이터를 충분히 학습하지 않은 상태



부적합(underfitting)



정적합(good fitting)

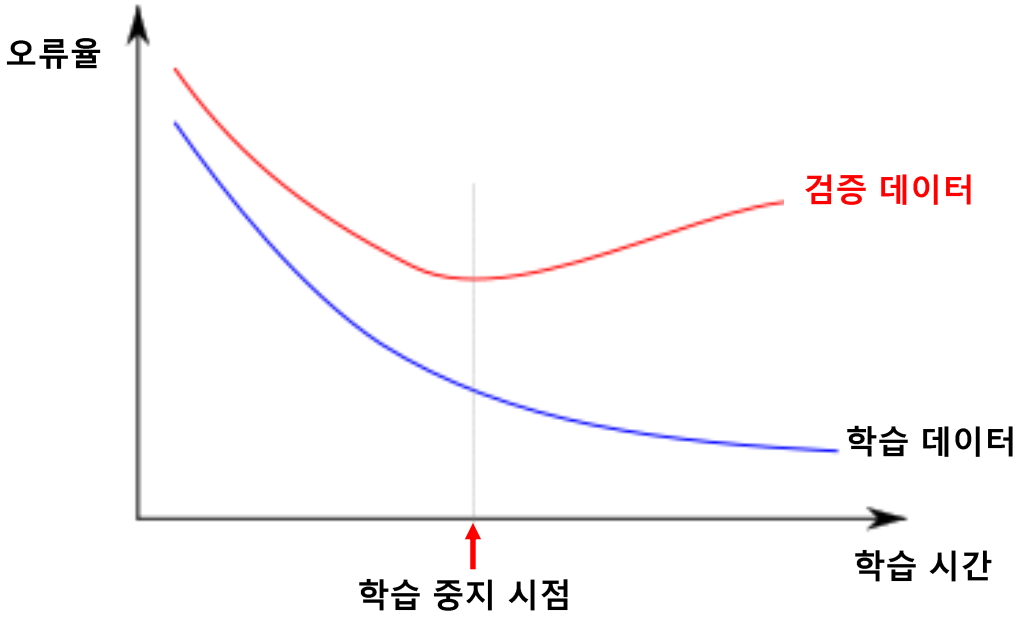


과적합(overfitting)

분류

❖ 과적합 회피 방법

- **학습데이터**에 대한 **성능**
  - **학습**을 **진행**할 수록 오류 **개선** 경향
  - 지나치게 학습이 진행되면 과적합 발생
- 학습과정에서 별도의 **검증 데이터**(validation data)에 대한 성능 평가
  - 검증 데이터에 대한 오류가 감소하다가 증가하는 시점에 학습 중단



분류

❖ 분류기의 성능 평가

- **정확도** (accuracy)
  - 얼마나 정확하게 분류하는가
  - **정확도** = (**옳게 분류한 데이터 개수**)/(**전체 데이터 개수**)
  - **테스트 데이터**에 대한 정확도를 분류기의 정확도로 사용
- 정확도가 높은 분류기를 학습하기 위해서는 **많은 학습데이터**를 사용하는 것이 유리
- **학습데이터**와 **테스트 데이터**는 **겹치게 않도록** 해야 함

분류

❖ 데이터 부족한 경우 성능평가

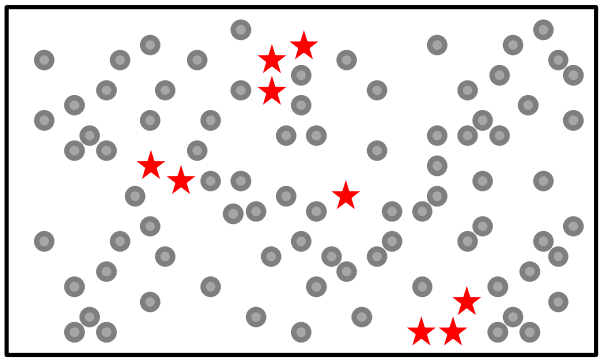
- 별도로 테스트 데이터를 확보하면 비효율적
- 가능하면 많은 데이터를 학습에 사용하면서, 성능 평가하는 방법 필요

- **K-겹 교차검증(k-fold cross-validation)** 사용
  - 전체 데이터를 **k 등분**
  - **각 등분**을 한번씩 **테스트 데이터**로 사용하여, 성능 평가를 하고 **평균값** 선택

분류

❖ 불균형 부류 데이터(imbalanced class data) 문제

- **특정 부류**에 속하는 **학습 데이터**의 **개수**가 다른 부류에 비하여 **지나치게 많은 경우**
- **정확도**에 의한 성능 평가는 **무의미**할 수 있음
  - 예. A 부류의 데이터가 전체의 99%인 경우, 분류기의 출력을 항상 A 부류로 하더라도 정확도는 99%가 됨.



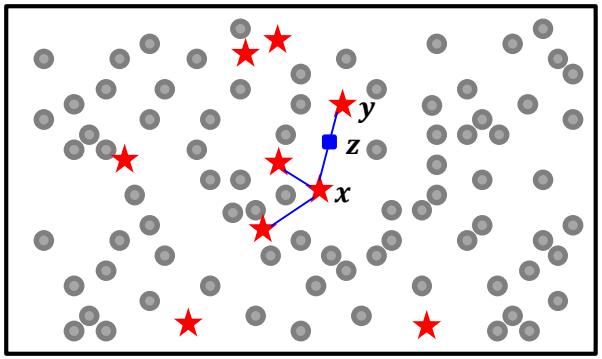
- 대응방안
  - **가중치**를 고려한 **정확도 척도** 사용
  - 많은 학습데이터를 갖는 부류에서 **재표본추출**(re-sampling)
  - 적은 학습데이터를 갖는 부류에 대해서 인공적인 **데이터 생성**

분류

❖ 불균형 부류 데이터 문제 – cont.

- **SMOTE**(Synthetic Minority Over-sampling Technique) 알고리즘
  - 빈도가 낮은 부류의 학습 데이터를 인공적으로 만들어 내는 방법

1. 임의로 닞은 빈도 부류의 학습 데이터 $\boldsymbol{x}$ 선택
2. $\boldsymbol{x}$의 k-근접이웃(k-nearest neighbor, KNN)인 같은 부류의 데이터 선택
3. k-근접이웃 중에 무작위로 하나 y를 선택
4. x와 y를 연결하는 직선 상의 무작위 위치에 새로운 데이터 생성



분류

❖ 이진 분류기의 성능 평가

- **이진 분류기**(binary classifier)
  - 두 개의 부류만을 갖는 데이터에 대한 분류기

표 4.2 이진 분류기의 혼동행렬

		예 측	
		양성	음성
실제	양성	진양성(True Positive) TP	위음성(False Negative) FN (type 2 error)
	음성	위양성(False Positive) FP (type 1 error)	진음성(True Negative) TN

		예측	
		P	N
실제	P	40	10
	N	15	35

- 민감도(sensitivity)/재현율(recall)/진양성율(true positive rate)

$$\text{민감도} = \frac{TP}{TP+FN}$$

$$\text{민감도} = \frac{40}{40+10} = \frac{40}{50}$$

- 특이도(specificity)/진음성율(true negative rate)

$$\text{특이도} = \frac{TN}{FP+TN}$$

분류

❖ 이진 분류기의 성능 평가 – cont.

표 4.2 이진 분류기의 혼동행렬

		예 측	
		양성	음성
실제	양성	진양성(True Positive) *TP*	위음성(False Negative) *FN*
	음성	위양성(False Positive) *FP*	진음성(True Negative) *TN*

- 정밀도(precision)

$$\text{정밀도} = \frac{TP}{TP+FP}$$

- 음성 예측도

$$\text{음성 예측도} = \frac{TN}{TN+FN}$$

- 위양성율

$$\text{위양성율} = \frac{FP}{FP+TN} = 1 - \text{특이도}$$

- 위발견율

$$\text{위발견율} = \frac{FP}{TP+FP} = 1 - \text{정밀도}$$

- 정확도

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

- F1 측도

$$F1 = 2\frac{(\text{정밀도}) \cdot (\text{재현율})}{(\text{정밀도}) + (\text{재현율})}$$

		예측	
		P	N
실제	P	40	10
	N	15	35

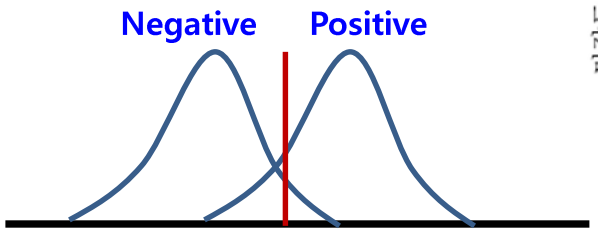
$$\text{정밀도} = \frac{40}{40+15} = \frac{40}{55}$$

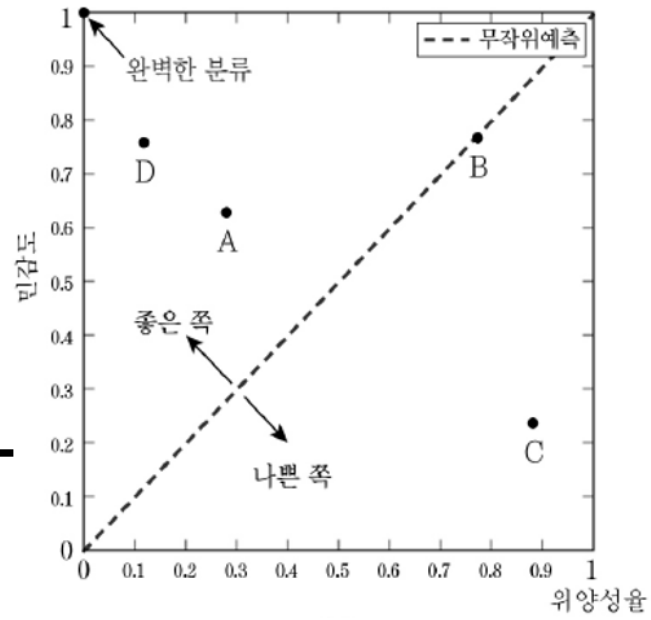
$$\text{정확도} = \frac{40+35}{40+15+10+35} = \frac{75}{100}$$

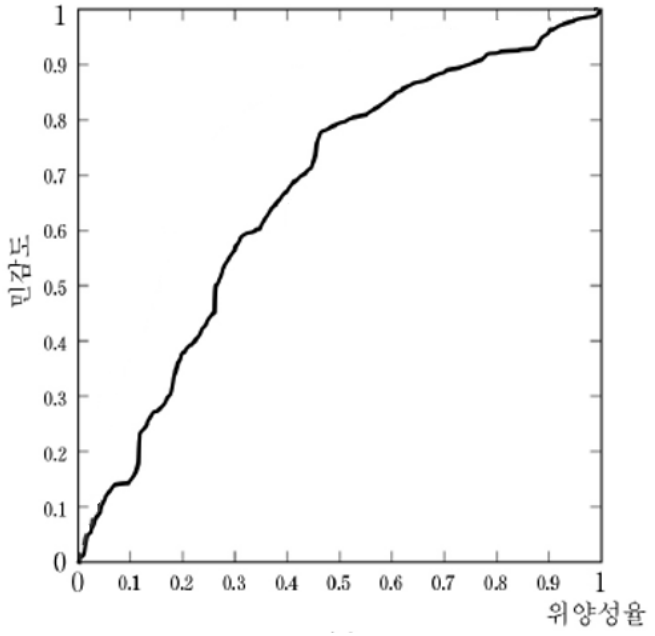
분류

❖ **이진 분류기의 성능 평가 – cont.**

- **ROC 곡선**
  - 부류 판정 임계값에 따른 (위양성율, 민감도) 그래프



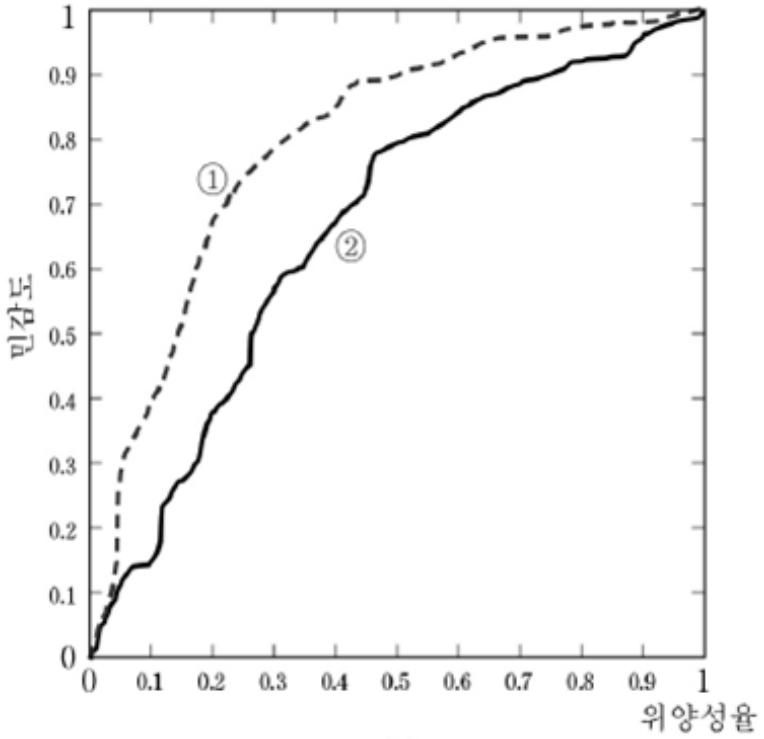




분류

❖ 이진 분류기의 성능 평가 – cont.

- **AUC**(Area Under the Curve)
  - ROC 곡선에서 곡선 아래 부분의 면적
  - 클수록 바람직



Quiz

❖ **분류에 관련한 다음 설명 중에서 옳지 않는 것을 선택하시오.**

① 분류는 데이터를 정해진 몇 개의 부류로 대응시키는 것이다.

② 분류에 사용될 수 있는 기법으로 결정트리 알고리즘, 서포트 벡터 머신, 다층 퍼셉트론, 에이다 부스트, 확률 그래프 모델 등이 있다.

③ 분류기는 학습된 결정경계에 따라 입력 데이터에 대응하는 부류를 출력한다.

④ 학습에서 학습 데이터에 대해서 잘 맞는 성질을 일반화 능력이라고 한다.

❖ **분류기에 관련한 다음 설명 중에서 옳지 않는 것을 선택하시오.**

① 학습 데이터에 대해서는 성능이 높이지만 테스트 데이터에 대해서는 성능이 크게 떨어지는 상황을 과적합이라고 한다.

② 학습 과정에서 과적합을 피하기 위해 별도로 성능 평가를 위해 사용하는 데이터를 검증 데이터라고 한다.

③ 분류기의 정확도는 (전체 데이터 개수)/(바르게 분류한 데이터 개수)로 계산한다.

④ 검증 데이터에 대해 오류율은 학습이 진행되어 감에 따라 감소하다가 증가하는 경향을 보인다.

Quiz

❖ **불균형 데이터의 문제에 대한 설명으로 옳지 않은 것을 선택하시오.**

① 학습 데이터에서 부류별 데이터 비율에 차이가 클 때 정확도만을 고려하여 학습을 하는 것은 바람직하지 않다.

② 불균형 데이터이 경우에는 정확도를 계산할 때 부류별 데이터의 비율을 고려할 수 있다.

③ 불균형 데이터의 문제를 피하기 위해 빈도가 큰 부류의 데이터를 표본추출하여 전체 부류의 비율이 균등해지도록 학습 데이터를 구성할 수 있다.

④ SMOTE 알고리즘은 빈도가 큰 부류의 데이터에 대해서 적용된다.

❖ **이진 분류기에 대한 설명으로 옳지 않은 것을 선택하시오.**

① ROC 곡선은 임계값에 따른 민감도와 위양성율의 위치를 나타낸 것이다.

② AUC의 값이 작을 수록 우수한 이진 분류기이다.

③ 위발견율은 '1 – 정밀도'의 값과 같다.

④ F1 측도는 정밀도와 민감도를 결합한 측도이다.