ADsP 데이터 분석 준전문가

3-2 과목
정형 데이터 마이닝

데이터 마이닝의 개요

- ❖ 대용량 데이터에서 의미 있는 데이터 패턴을 파악, 예측하여 의사결정에 활용
- ❖ 가설이나 가정없이 다양한 수리 알고리즘을 이용
- ❖ 데이터 마이닝 도구가 체계화되어 환경에 적합한 제품을 선택하여 활용
- ❖ 알고리즘에 대한 깊은 이해가 없어도 분석에 큰 어려움이 없음
- ❖ 분석 결과의 품질을 위해서 풍부한 경험을 가진 전문가가 하면 좋음

지도학습 vs 비지도학습

지도학습	비지도학습
의사결정나무	
인공신경망	OLAP
일반화 선형 모형	연관성 규칙발견
선형 회귀분석	군집분석
로지스틱 회귀분석	SOM
사례기반 추론	

❖ 데이터 마이닝 추진 단계 : 목적설정 → 데이터 준비 → 가공 → 기법 적용 → 검증

데이터 마이닝을 위한 데이터 분할

❖ 구축용(train data) : 데이터 마이닝 모델을 만드는데 활용(50%)

❖ 검정용(validation data) : 구축된 모형의 과대추정 또는 과소추정을 미세 조정을 하는데 활용(30%)

❖ 시험용(test data) : 모델의 성능을 검증하는데 활용(20%)

❖ 데이터 양이 충분치 않은 경우

↳ 홀드아웃 방법 : 주어진 데이터를 학습용과 시험용 데이터로만 분리하여 사용

↳ k-fold 교차분석 방법 : 주어진 데이터를 k개의 집단으로 구분하여 k-1개의 집단을 학습용으로, 나머지는 검증용으로 설정해 학습. k번 반복 결과 값의 평균을 최종 값으로 사용.

성과분석

		예측	
		Positive	Negative
실제	Positive	True Positive (1)	False Negative (3)
	Nagative	False Positive (2)	True Negative (4)

❖ 정확도(Accuracy) = (1+4)/(1+2+3+4)

↳ 실제와 맞게 예측한 확률

❖ 특이도(Specificity) = 4/(2+4)

↳ 실제로 거짓인 사건을 예측도 거짓으로 한 확률

❖정밀도(Precision) = 1/(1+2)

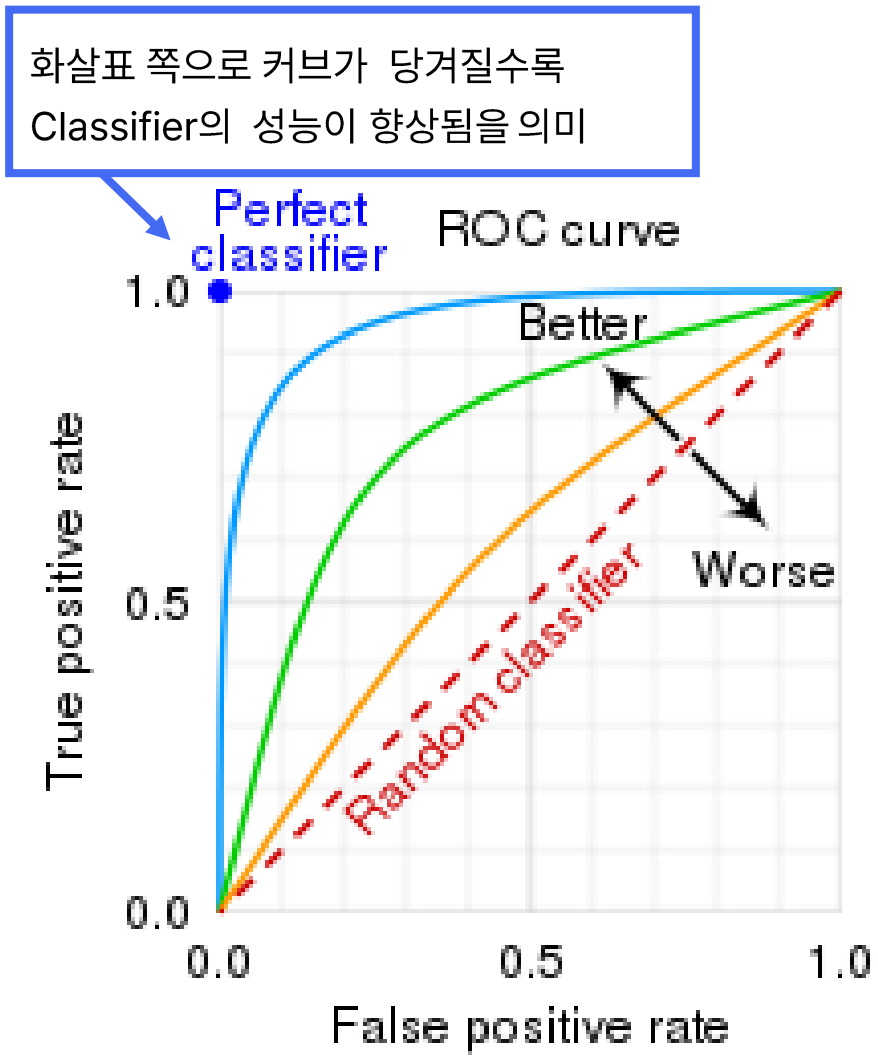
↳ 예측을 참이라고 했는데 실제로도 참일 확률

❖ 재현율(Recall) = 민감도(Sensitivity)= 1/(1+3)

↳ 실제로 참인 경우 중에 예측을 참으로 한 확률

❖ $F1\ Score = 2 * \frac{Precision * Recall}{Precision + Recall}$

ROC Curve



❖ 가로축(FPR) : 1-특이도(Specificity)

❖ 세로축(TPR) : 민감도(Sensitivity)

❖ 그래프의 면적(Area Under Curve)이 클수록 모형의 성능이 좋다고 평가함

분류분석

분류분석

❖ 지도학습(교사학습)에 해당, 데이터가 어떤 그룹에 속하는지 예측하는데 사용되는 기법

❖ 로지스틱 회귀분석, 의사결정나무, 인공신경망, SVM 등 지도학습의 대부분이 분류분석에 속함

로지스틱 회귀분석

❖ 반응변수가 범주형인 경우에 적용되는 회귀분석 모형

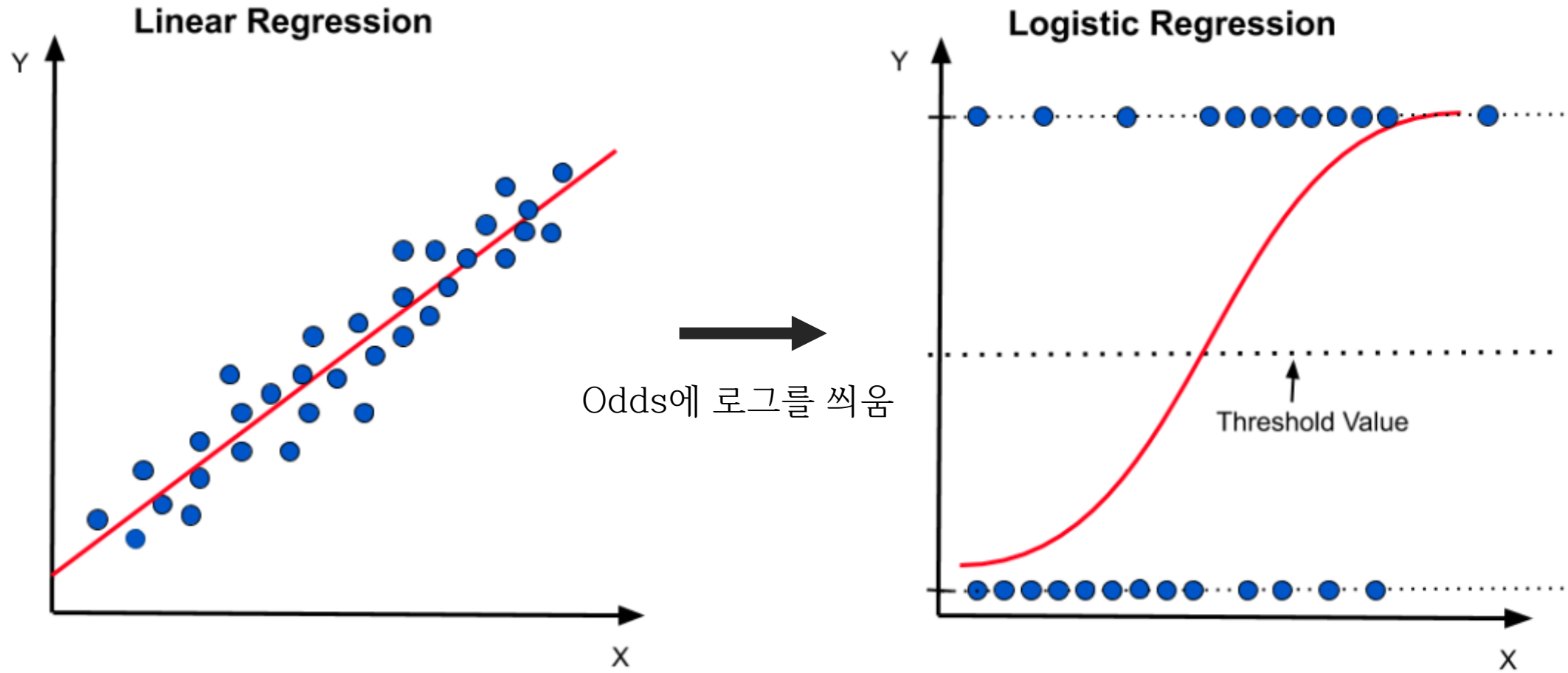
❖ exp(β)는 나머지 변수가 주어질 때 x1이 한 단위 증가할 때마다 성공의 오즈가 몇 배 증가하는지를 나타내는 값

↳ 오즈 = p/(1-p) = 확률/(1-확률) : 성공할 확률과 실패할 확률의 비율

↳ 회귀계수(β) > 0 이면 S자, β < 0 이면 역 S자(반대) 모양이 됨

❖ 계수 추정법 : 최대우도측정법(MLE) 사용.

로지스틱 회귀분석



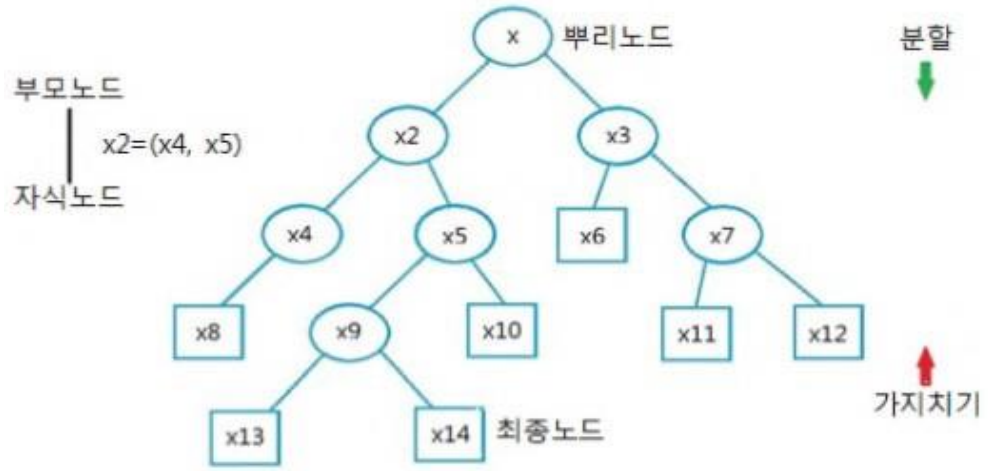
https://images.app.goo.gl/hPYpP6TNjP44npuV6

의사결정나무

❖ 연속적으로 발생하는 의사결정 규칙을 나무 모양으로 시각화.

❖ 해석이 간편하고, 연산이 빠르다.

❖ 성장 → 가지치기 → 타당성평가 → 해석 및 예측



❖ 단점 : 새로운 자료에 대한 과대적합이 발생할 수 있다.

↳ 과대적합(OverFitting): 학습이 과해 기존 자료에 대한 성능이 매우 뛰어남. 단, 새로운 데이터에 대한 대응이 미흡함.

↳ 과소적합(UnderFitting): 모형이 단순하거나 데이터 부족으로 성능이 매우 떨어짐

의사결정나무 불순도 측정

❖ 카이제곱 통계량

$$X^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$$

k: 범주의 수, O: 실제 도수, E: 기대 도수

❖ 지니지수

$$Gini = 1 - \sum_{i=1}^{c} (p_i)^2$$

$$Gini = 1 - (\frac{3}{5})^2 - (\frac{2}{5})^2 = 0.48$$

❖ 엔트로피 지수

$$Entropy(A) = -\sum_{k=1}^{m} p_k \log_2(p_k)$$

$$Entropy = -(0.5 * \log_2 0.5) * 2 = 1$$

의사결정나무 알고리즘

❖ CART : 의사결정나무의 기본형으로 가장 많이 활용되는 알고리즘

↳ 불순도 : 출력변수가 범주형일 경우 지니지수, 연속형일 경우 이진분리 사용

❖ C4.5와 C5.0 : 범주형 입력변수에 대해서는 범주의 수만큼 가지분리가 일어남

↳ 불순도 : 엔트로피지수 사용

❖ CHAID : 가지치기 없이 적당한 크기에서 나무 성장을 중지

↳ 불순도 : 범주형일 경우 카이제곱 통계량 사용, 연속형인 경우 F통계량 사용

앙상블 분석

❖ 앙상블 분석 : 여러 개의 예측모형들을 만든 후 조합

❖ 배깅 : 여러 개의 부스트랩 자료(랜덤 샘플링)를 생성한 후 각 자료에 예측모형을 만든 후 결합

↳ 가지치기를 하지 않고 최대로 성장한 의사결정나무 활용

❖ 부스팅 : 예측력이 약한 모형을 결합하여 예측력이 강한 모형을 만드는 방법

↳ 각 자료에 동일한 가중치를 주는 것이 아니라 자료마다 다른 가중치가 부여됨

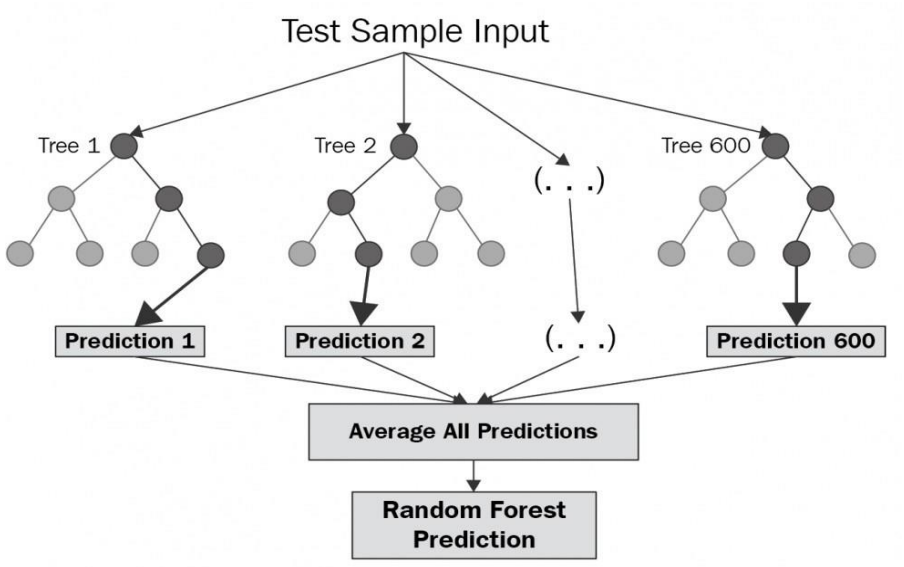
↳ 훈련 오차를 쉽고 빠르게 줄일 수 있음

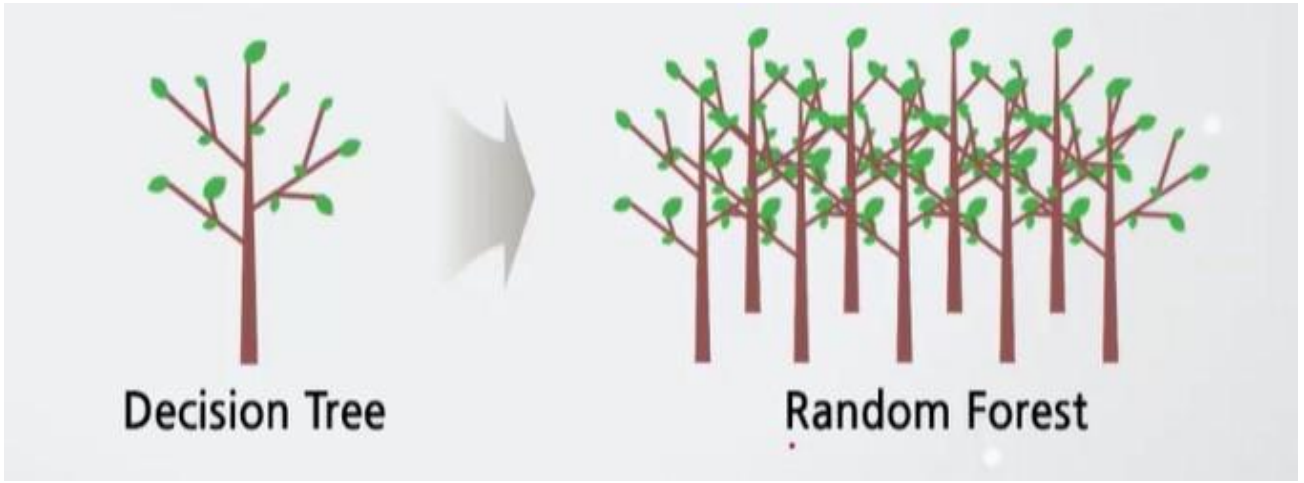
앙상블 분석

❖ 랜덤포레스트 : 배깅 + 랜덤성

여러 개의 결정 트리들을 임의적으로 학습하는 방식의 앙상블 방법

여러가지 학습기들을 생성한 후 이를 선형 결합하여 최종 학습기를 만드는 방법



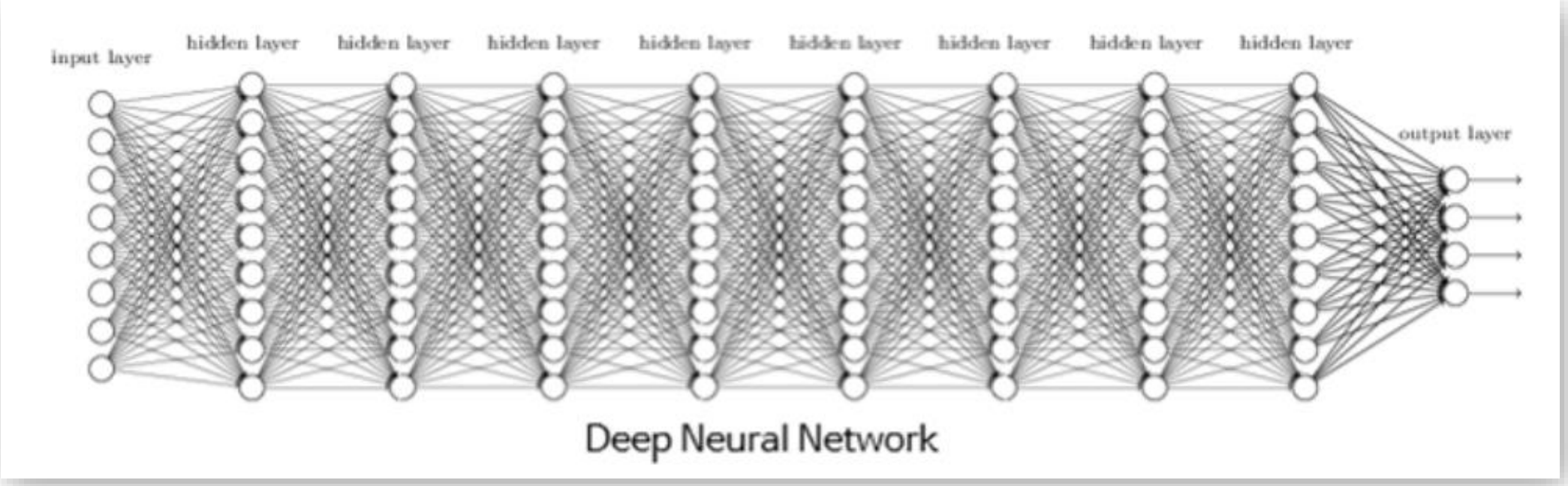


인공신경망 분석

❖ 인간의 뇌를 기반으로 한 추론 모델, 뉴런과 뉴런사이가 가중치가 있는 링크들로 연결되어 있음.

↳ 뉴런 : 기본적인 정보처리 단위

↳ 역전파 알고리즘을 활용하여 비선형성을 극복한 새로운 모형 등장



인공신경망 분석

❖활성화 함수 : 출력을 결정하는 함수

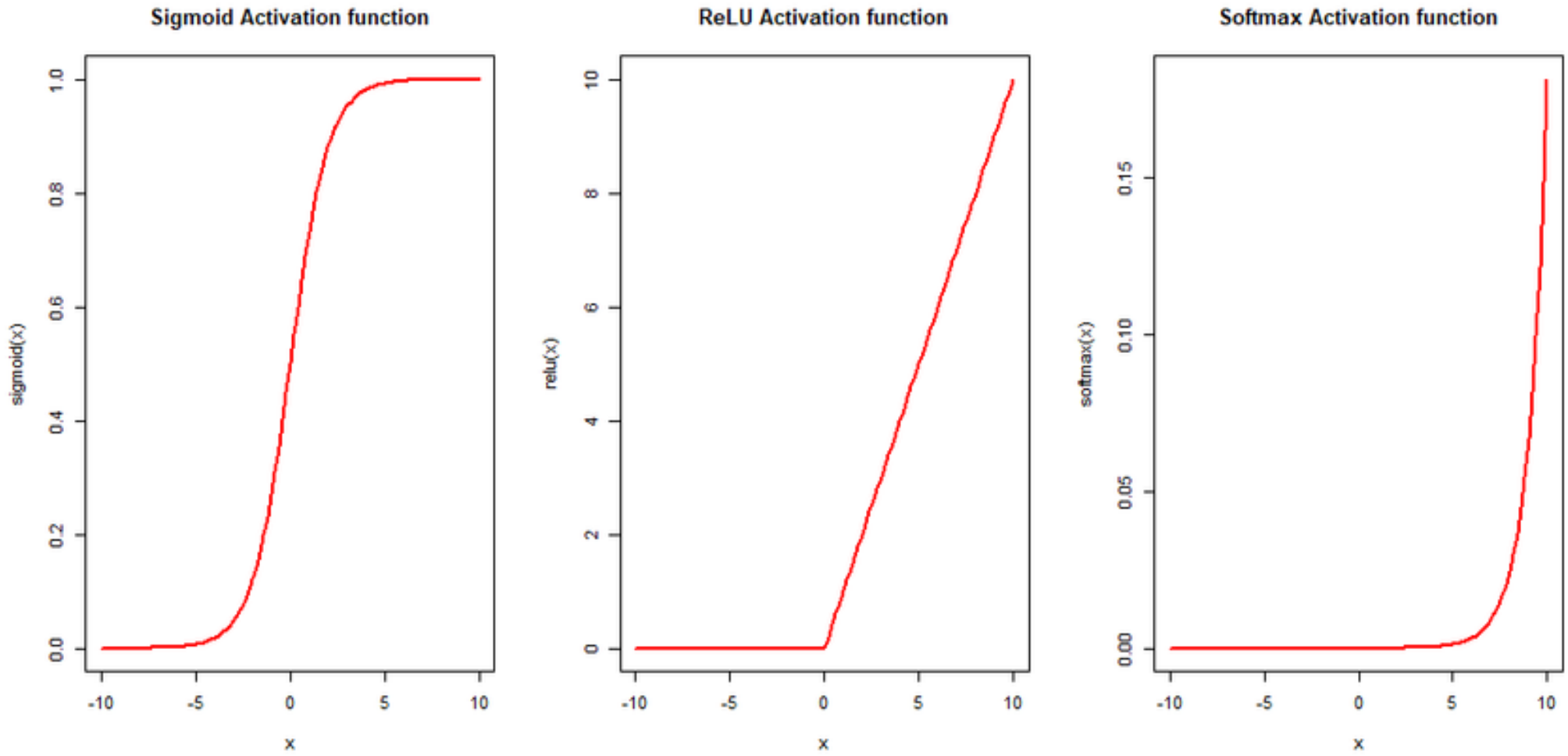
↳ 계단 함수 : 0 또는 1과 같은 정수의 출력 값

↳ 시그모이드(sigmoid) 함수 : 0~1사이의 확률 값으로 출력(단일 출력)

↳ 소프트맥스(softmax) 함수 : 다중 클래스 분류 모델을 만들 때 사용, 모든 입력 값에 대한 확률 값을 출력

↳ ReLU 함수 : 입력 값이 0 이하이면 0으로 출력, 그 외 다른 숫자는 그대로 출력

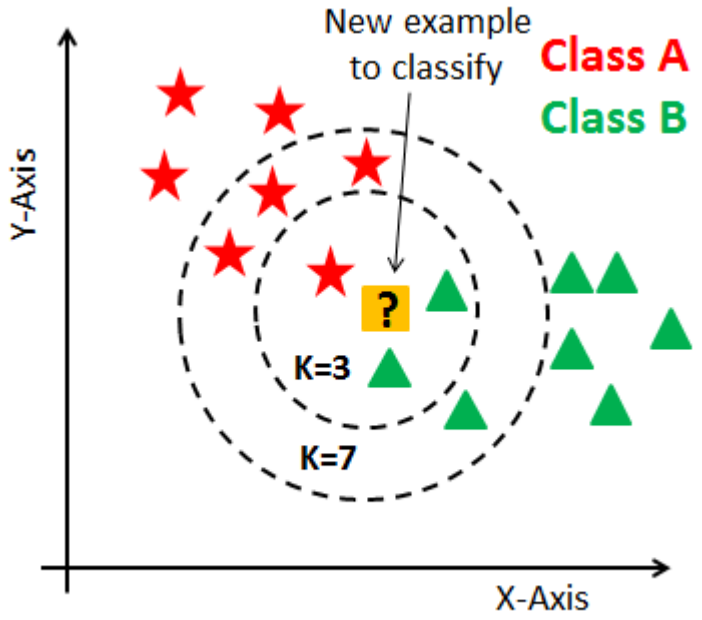
인공신경망 분석



https://images.app.goo.gl/eUy4pT8B9nTVZoodA

K-Nearest-Neighbor

❖ 새로운 입력이 생기면, 주변에 있는 K개 데이터의 클래스를 통해 새로운 입력에 대한 클래스를 결정.



출처 : https://images.app.goo.gl/RJPYFEK59zkwxU5H8

❖ K가 너무 작으면 Overfitting

❖ K가 너무 크면 Underfitting

↳ 최적의 K값을 찾는 것이 중요!

❖ 장점 : 사용이 매우 간단하며, 데이터 처리가 쉬움

❖ 단점 : 수치형 데이터가 아닐경우 유사도를 정의하기 어려움

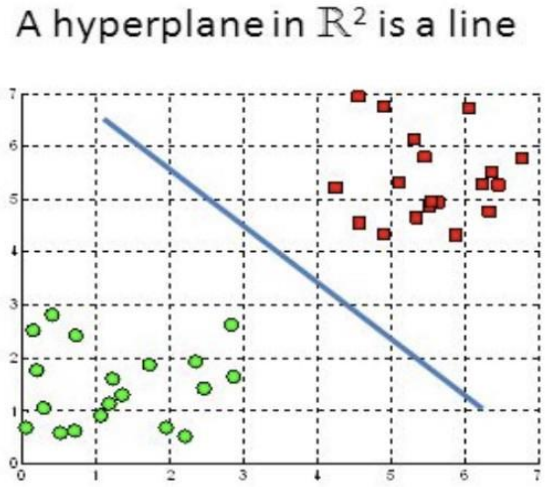
이상치가 존재하는 데이터일 경우 분류 성능에 악영향을 크게 미침

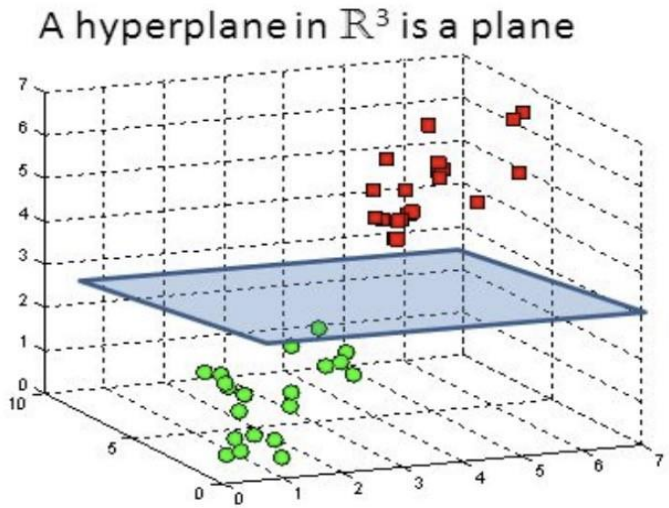
Support Vector Machine

❖ 이진 선형 분류 모델

❖ 데이터가 표현된 공간에서 분류를 위한 경계(선)을 정의

↳ 데이터와 경계 사이의 여백을 최대화 하는 선을 정의해야 한다.





❖ 장점 : 데이터의 이해도가 떨어져도 사용가능, 정확도 높음

❖ 단점 : 파라미터의 결정과 모형을 구축하는데 시간이 오래 소요됨

군집분석

군집분석

❖ 군집의 개수나 구조에 대한 고려는 하지 않고, 데이터 사이의 거리를 기준으로 군집화

↳ 각 개체의 유사성을 측정하여 분류하고 서로 다른 군집과 상이성을 규명하는 분석방법

❖ 연속형 변수 거리 측정 방법

↳ 유클리드 거리 : 점과 점 사이의 직선 거리(피타고라스 정리)

↳ 맨하탄 거리 : 한점에서 다른 점까지 가는 최단 거리(빌딩, 숲 등 주변환경 고려)

↳ 표준화거리, 마할라노비스 거리, 캔버라 거리, 민코우스키 거리

❖ 범주형 변수 거리 측정 방법

↳ 코사인 거리(1-코사인 유사도), 코사인 유사도(벡터 내적)

↳ 자카드 거리(집합으로 표현), 자카드 계수, 자카드 유사도

계층적 군집분석

❖ N개의 군집으로 시작해 점차 군집의 개수를 줄여 나가는 기법

❖ 최단연결법: 최단거리를 이용해 군집형성, 고립된 군집을 찾는데 중점

❖ 최장연결법: 최장거리를 이용해 군집형성, 내부 응집성에 중점을 둠

❖ 평균연결법: 계산량이 많지만 모든 데이터를 포함하는 하나의 군집 형성

❖ 와드연결법: 군집내의 편차들의 제곱합에 기초하여 군집 형성, 군집간 정보 손실 최소화

비계층적 군집분석

❖ K-평균 군집분석

❖ 원하는 군집의 개수와 초기값(seed)들을 정해 seed를 중심으로 군집을 형성

❖ 한번 군집이 형성되어도 개체들은 다른 군집으로 이동할 수 있음

❖ 군집이 안정적이지만, 최적이라는 보장이 없음.

❖ 다양한 형태의 데이터에 적용 가능.

❖ 혼합 분포 군집 : EM 알고리즘

↳ E-step : 잠재변수의 기대치 계산

↳ M-step : 잠재변수의 기대치 이용하여 파라미터 수정

자기조직화지도(SOM)

❖ 고차원의 데이터를 이해하기 쉬운 저차원의 뉴런으로 정렬

❖ 입력변수의 위치 관계를 그대로 보존(뉴런수 = 입력 변수의 개수)

❖ 연결강도는 입력패턴과 가장 유사한 경쟁층 뉴런이 승자가 된다.

❖ BMU: 입력 층의 표본 벡터에 가장 가까운 프로토타입 벡터

❖ 단 하나의 전방 패스를 사용함으로써, 속도가 빠름

↳실시간 정보처리 가능

연관분석

❖ 사건들 간의 규칙을 발견하기 위해 적용

❖ A사건이 일어나면 B사건도 일어난다.

❖ 지지도 : $\frac{P(A \cap B)}{\text{전체}}$, A와 B가 동시에 일어나는 경우

❖ 신뢰도 : $\frac{P(A \cap B)}{P(A)}$, 연관 있는 정도

❖ 향상도 : $\frac{P(A \cap B)}{P(A)*P(B)}$, 향상도가 1이면, 관련 없음, 1보다 크면 성능이 우수하다는 것을 의미

연관분석

❖ 순차패턴분석 : 연관분석 + 시간정보, 동시에 진행될 가능성 큰 상품 찾기

❖ Apriori 알고리즘

↳ 최소 지지도보다 큰 항목에 대해서면 연관규칙 계산

❖ FP(Frequent Pattern)-Growth 알고리즘

↳ FP-Tree를 만든 후, 분할정복 방식 적용

↳ Apriori보다 DB를 적게 스캔하고, 빠르게 집합을 추출할 수 있음

문제풀이

3-2과목 문제풀이 정답

1	④	11	②	21	②	31	②	41	④	51	①	61	③	71	④
2	②	12	①	22	①	32	③	42	④	52	④	62	③	72	②
3	③	13	③	23	③	33	③	43	③	53	②	63	③	73	④
4	②	14	①	24	④	34	③	44	④	54	④	64	④	74	④
5	③	15	①	25	③	35	③	45	④	55	①	65	②	75	②
6	④	16	②	26	②	36	②	46	①	56	④	66	②		
7	①	17	④	27	①	37	③	47	④	57	④	67	②		
8	④	18	①	28	②	38	③	48	④	58	④	68	①		
9	④	19	④	29	④	39	③	49	④	59	①	69	②		
10	②	20	①	30	②	40	④	50	②	60	③	70	②		