빅데이터 모델링

분석절차 수립 - 분석모형 선정

분석 모형의 설계는 what과 why를 명확히 하는 것으로 목표 설정의 역할

빅데이터 분석 기법 설계 : 기획단계 - 계획 수립단계 - 데이터 분석 수행

분석 모형 선택 : 분석 목적에 맞는지 확인, 모형의 안정성(입력변수에 따른 예측력), 모델 갱신의 유연성

분석 모형 선정 :
- 분석 모형은 과거 데이터 기반 어떤 수식이나 패턴을 생성하는 것
- 확보된 데이터 값을 변형 및 변환하여 원하는 상태로 데이터 변환
- 선정된 모형의 방법론에 따라서 필요한 자료 파악
- 빅데이터의 특성을 파악하여 올바른 통계분석 기법 선택
- 통계분석 기법의 선택은 변수들 간의 관계, 변수 혹은 자료의 수, 변수에 포함된 자료의 형태 및 집단의 수
- 독립변수 중에서 회귀 모형에 사용할 변수를 축소하거나 선택하거나 제거하여 가장 좋은 회귀 모형 선택
- 변수를 선택하는 방법에는 전진선택법, 후진제거법, 단계별선택법이 있음
- 변수 선택 및 차원 축소 모형 검토 : 데이터 정제 → 변수 간의 상관관계 파악 → 전체 모형 결정
  → 변수 선택 → 최적 모형 결정

분석절차 수립 - 분석모형 정의

- 분석모델링은 분석 목적에 따라 상세 분석 기법을 적용해 모델을 개발하는 과정
- 데이터 저장 기술의 발달로 양질의 데이터 확보 및 데이터 가공 방법 발전
- 다양한 방법으로 분석 대상에 대한 데이터 획득
- 분석데이터는 전 처리 등을 통해서 변수 식별 및 구조화를 통한 모델 설계
- 데이터의 정형, 비정형 등의 형태, 데이터의 타입(연속형, 범주형)도 고려
- 분석목표에 따른 가설이 확실해야 하며 가설을 통해서 데이터의 접근 방법과 분석 방법을 설정
- 분석모델링의 설계 순서 : 데이터 확인 → 데이터에 대한 분석방법 설정 → 분석 진행 → 결과 도출
- 분석 목적에 기반하여 질문을 던지고 가설을 세우고 주요 모수에 대해 검정 가능한 이론을 제시
- 가설검정 절차 : 가설설정 → 의사결정규칙 설정 → 데이터 수집 및 통계량 계산 → 의사결정 → 사후조치
- 통계적 가설은 모집단의 모수에 대해 설정하는 것이며 가설검정은 모수에 대한 서로 경쟁적이며 상호배타적이며 포괄적인 두 개의 가설을 설정하고 어느 가설을 택할지 결정
- 전통적 추정 방법으로 최대우도 추정법이나 최소제곱 추정법을 이용해 표본의 데이터로부터 표본평균을 구하여 추정치로 사용
- 베이지안 추정은 초기 추정치를 설정한 후 더 그럴듯한 추정치로 바꾸어 나가면서 최종의 추정치를 사용
- 베이즈 정리 도입 → 추정문제의 설정 → 사전분포의 설정 → 초기 추정치 설정 → 확률분포 수정 → 사후분포 도출 → 최종 추정치 도출

분석절차 수립 - 분석모형 구축 절차

분석 모형은 하향식, 상향식, 그리고 프로토타이핑 접근 방법을 통해서 구축절차 수립

- 하향식 분석 프로세스
  현황분석을 통해서 인식된 문제점, 전략으로 부터 기회나 문제를 탐색
  → 데이터 가용성 분석, 가설 설정과 샘플 데이터 수집, 가설 검증,
  → 사용자 관점에서의 해결 방안 및 과제 추진 방안 설계, 빅데이터 분석 과제의 타당성 검토
  → 빅데이터 분석 과제의 확정 및 분석계획 수립 과정

- 상향식 분석 프로세스
  하향식과 달리 명확한 문제 해결 절차를 수용하지 않으며 이미 보유하고 있는 DW나 DM 데이터 기반
  → 기술적 통계분석, 군집분석, 시각화 기법, 상관분석, 인과분석 등을 통해 유의미한 패턴 및 관계 도출
  → 패턴과 관계를 비즈니스 관점에서 해석하고 그 의미를 문제나 기회 발견에 사용

- 프로토타이핑 프로세스
  사용자의 개괄적인 요구 사항을 반영한 초기 프로토타입 모델의 개발로부터 시작
  → 사용자의 기본적인 요구만을 반영하여 최대한 짧은 시간 내에 만들어낸 분석 모형 시스템
  → 사용자들은 모델을 사용해 실제로 분석을 수행해 보고 그 프로토타입이 적절한 지 유용한 지 확인

분석환경 구축 - 분석도구 선정

분석 도구 : SPSS, SAS, MINITAB, 오픈소스 툴

오픈소스 툴
- R : 여러 패키지를 통해 데이터를 보기 쉽게 시각화 하는 강점 반면 머신러닝, 딥러닝의 경우
  해당 패키지가 늦게 업데이트되는 점 등이 단점
- Python : 데이터 핸들링, 엔지니어링, 모델링, 머신러닝 등 여러 가지를 할 수 있는 범용성 도구
  대규모로 머신러닝을 하고 배포할 수 있으며 최근 기계학습 및 인공지능을 위한 API를 신속하게
  제공받을 수 있는 반면 시각화 툴이 부족하며 컴퓨터 공학 관련 지식이 있어야 보다 쉽게 접근 할 수 있음

분석환경 구축 - 데이터 분할

데이터 분할
- 의미 있는 분석 결과 도출을 위해서는 수집된 데이터를 분석을 위한 데이터로 전 처리 작업 필요
- 전 처리 과정은 필요 없는 변수나 문제가 있는 이상치 제거
- 데이터가 너무 크거나 모형을 만들거나 검증을 위해서 데이터 분할
- 데이터 분할에는 학습/테스트 데이터 분할과 변수를 줄이는 차원 축소 등이 있음
- 학습 데이터로 모델을 학습시키고 이렇게 학습된 모델을 통해서 예측이나 분석
- 시행을 통해서 얻은 결과값 혹은 예측 값을 실제 결과값과 비교하여 모델의 성능을 평가
- 데이터 분할을 위해서는 학습/테스트 데이터의 크기와 랜덤 샘플링 여부 등을 결정
- 일반적으로 학습 데이터와 테스트 데이터의 크기는 7:3, 8:2로 분리
- 학습 데이터를 다시 분할해서 학습데이터와 학습된 모델을 일차 검증하는 검증 데이터 3가지로 분할

분석기법 - 회귀분석

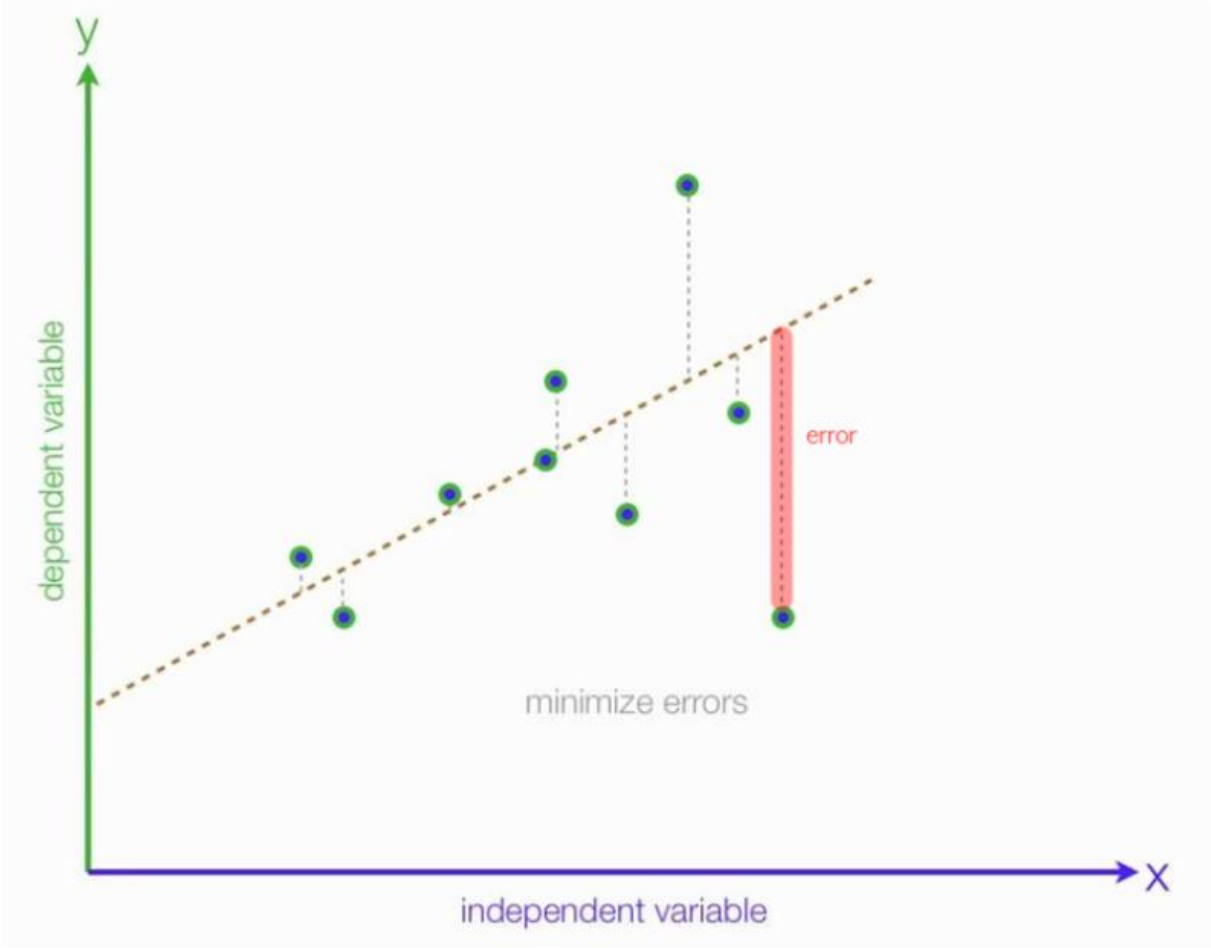
회귀선
주어진 데이터를 대표하는 하나의 직선

회귀식
회귀선을 함수로 표현한 것

단순 선형 회귀분석
y = wx + b인 회귀식에서 x가 1개

잔차
관측값의 y와 예측값의 y간의 차이

최소제곱법
잔차의 제곱의 합이 최소가 되도록 회귀계수를 구함

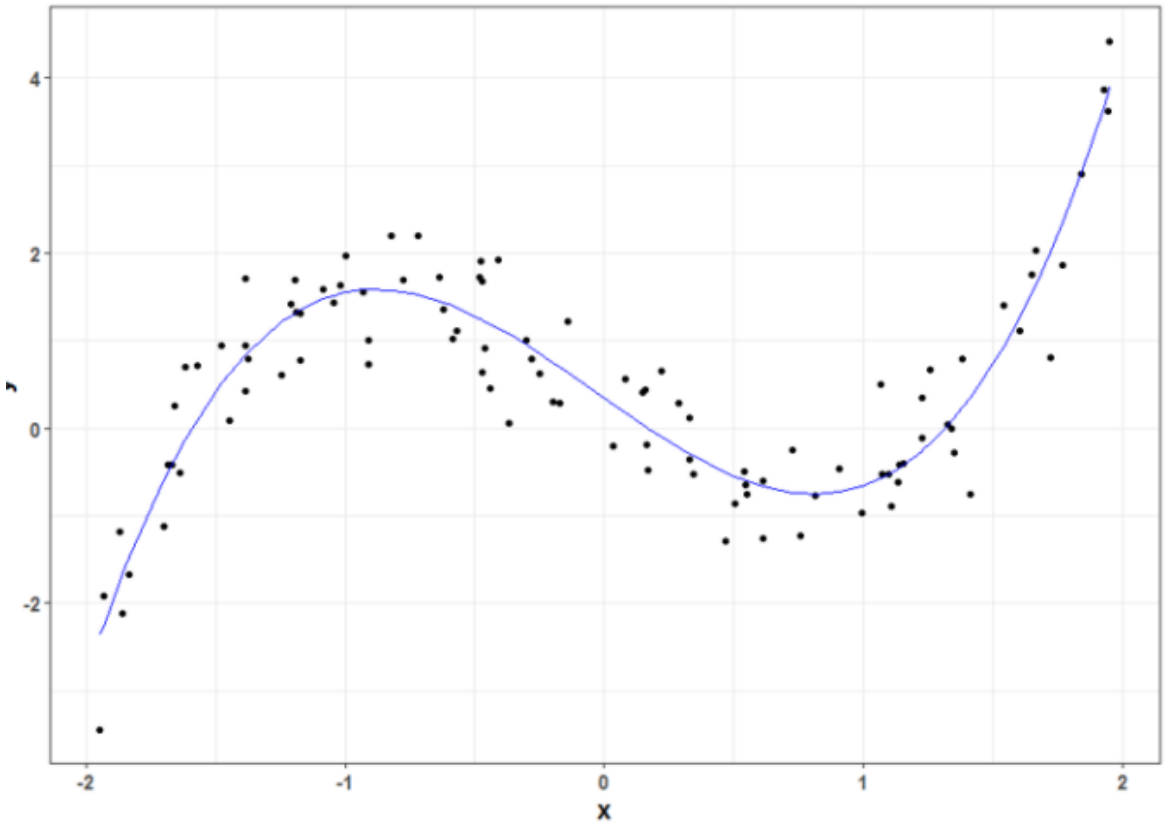


분석기법 - 회귀분석

독립 변수가 다항식으로 구성되는 회귀 모델

만약 종속변수인 y와 독립변수인 x가 선형 관계가 아닌 곡선 형태를 갖는다면 독립변수에 지수승을 붙여서 여러 개의 변수로 만들어 회귀 모델을 구성하는 기법

$$\hat{y} = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \cdots$$

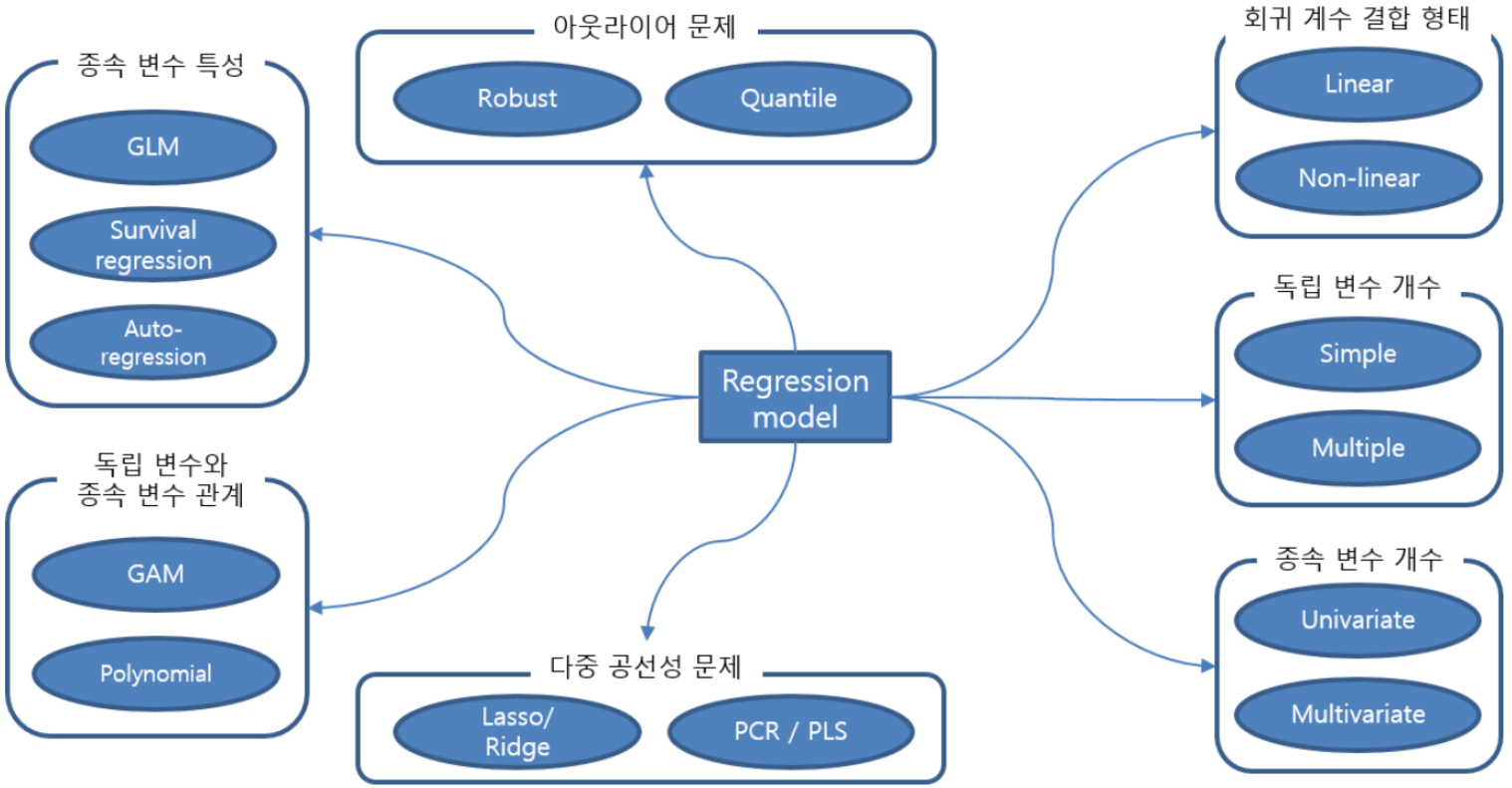


분석기법 - 회귀분석

어떤 자료에 대해서 그 값에 영향을 주는 조건을 고려하여 구한 평균
회귀모델 : $y=h(x1,x2,x3,\ldots,xk;\beta1,\beta2,\beta3,\ldots,\beta k)+\epsilon$
h : 평균을 구하는 함수, x가 주어지면 x의 영향력 β을 고려하여 해당 조건에서의 평균값을 계산
e : 오차항을 의미하여 다양한 현실적인 한계로 발생하는 불확실성을 포함, 잡음(noise)라고도 함
잡음은 이론적으로 평균이 0이고 분산이 일정한 정규 분포를 띄는 성질을 가지고 있음

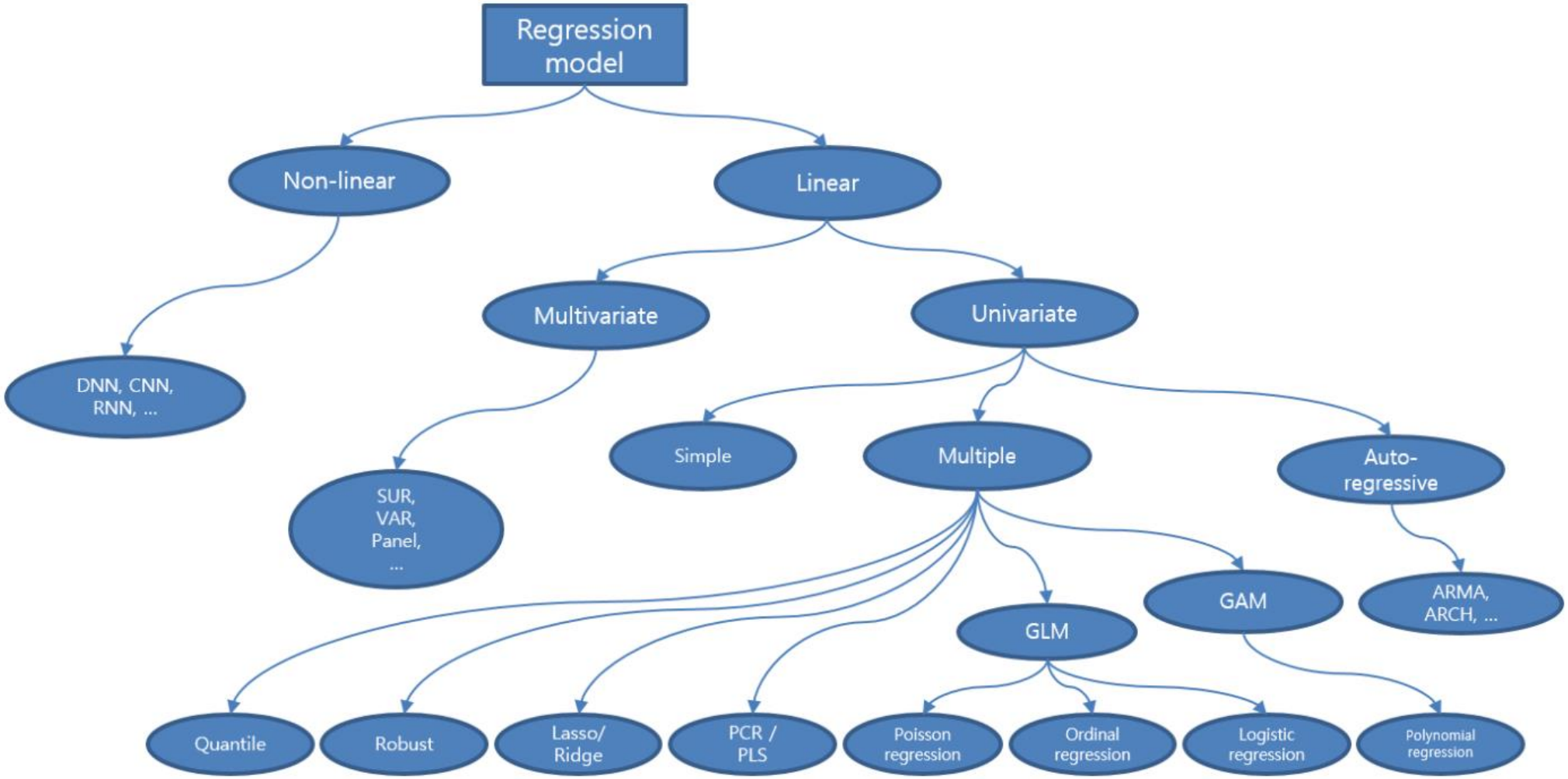
모델 검정 :
회귀모델의 예측치와 실측치의 차이인 잔차가 우리가 가정한 오차항(e)의 조건을 충족하는지 확인

underfitting :
중요한 조건 미반영
overfitting:
단순한 잡음을 반영



분석기법 - 회귀분석

일반화 수준에 따라 계층적으로 표현



분석기법 - 회귀분석

선형회귀의 정확한 정의는 종속변수의 평균이 독립변수와 회귀계수(Regressin Coefficient)들의 선형 결합(Linear Combination)으로 된 회귀모형을 말하며, 회귀계수를 선형 결합으로 표현할 수 있는 모형을 의미. 주의할 점은 "선형의 의미가 독립변수와 종속변수가 꼭 직선의 그래프 형태(1차식)를 나타내는 것이 아니다."라는 것임

일반선형회귀
(Linear Regression)

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p$$

[일반선형회귀의 가정]

1. 독립변수와 종속변수의 선형성
2. 오차항의 독립성
3. 오차항의 등분산성
4. 오차항의 정규성

가정이 깨진다면?

일반화 선형회귀
(Generalized Linear Regression)

$$f(x) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p$$

[일반화 선형회귀 예시]

- 로지스틱 회귀분석

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p$$

- Cox의 비례위험 모형

$$\log\left(\frac{h(t)}{h_0(t)}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p$$

분석기법 - 회귀분석

일반선형모델 vs 일반화선형모델

선형(Linear) 모형 : 회귀계수를 독립변수의 선형결합으로 나타낸 모형
일반화선형회귀 : 종속변수를 변환하여, 회귀계수를 독립변수의 선형결합으로 나타낸 모형

	일반선형모델 General linear model	일반화선형모델 Generalized linear model
모델구하는 수학적 방법	Least squares Best linear unbiased prediction	Maximum likelihood Bayesian
이 부류에 속하는 통계방법들	ANOVA ANCOVA MANOVA MANCOVA Linear regression Mixed model	Linear regression Logistic regression Poisson regression Gamma regression

분석기법 - 로지스틱 리그레션

로지스틱 회귀분석은 반응변수가 1 또는 0인 이진형 변수에서 쓰이는 회귀분석 방법. 종속변수에 로짓변환을 실시하기 때문에 로지스틱 회귀분석이라고 불리며 우선 계수가 Log Odds ratio가 되기 때문에 해석이 매우 편리하고, case-control과 같이 반응 변수에 따라 샘플링된 데이터에 대해서 편의(bias)가 없는 타당한 계수 추정치를 계산할 수 있다.

$$ln(\frac{p}{1-p}) = Z \quad ,(Z = B_0 + B_1X_1...)$$

$$e^{ln(\frac{p}{1-p})} = e^z$$

$$\frac{p}{1-p} = e^z$$

$$p = (1-p)e^z$$

$$p = e^z - pe^z$$

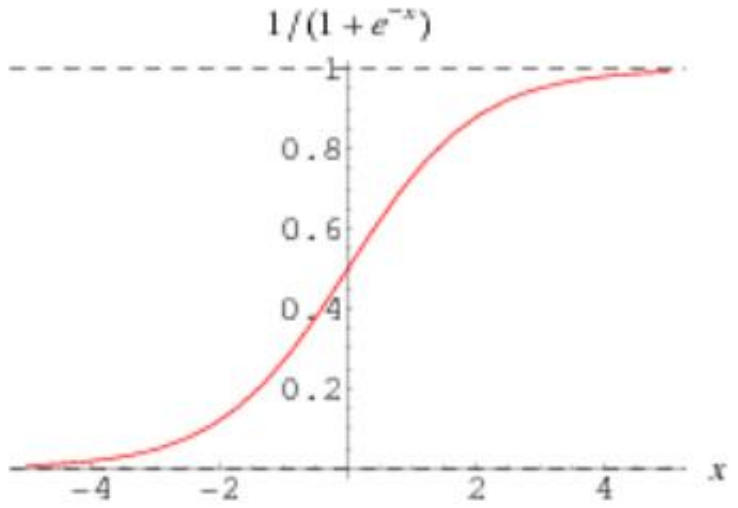
$$p + pe^z = e^z$$

$$p(1+e^z) = e^z$$

$$p = \frac{e^z}{1+e^z}$$

$$= \frac{1}{1+e^{-z}}$$ (신경망의 Sigmoid 함수)

$$P(y=1) = \frac{1}{(1+e^{-(\beta_0+\beta X)})}$$



분석기법 - 로지스틱 리그레션

case-control 연구는 인구집단에서 case와 control을 샘플링해서 이를 대상으로 질병의 원인이 되거나 연관성이 있는 요인을 원인을 찾는 연구입니다. 예를 들어 전체 인구집단 중에 유병률이 0.001 밖에 되지 않는 질병이 있다면, 전체 인구 집단을 대상으로 연구를 수행하면 질병에 걸린 사람(case)를 연구에 포함하기가 매우힘들어지겠죠. 따라서 case-control study에서는 질병이 있는 사람들을 더 많이 샘플링을 해서 case와 control의 비율을 최대 1:1 까지 맞춥니다. 그렇다면 이 연구용 집단은 전체 모집단과는 특성이 다르게 됩니다. 따라서 많은 수치가 잘못된 계산값을 내놓게 됩니다. 예를 들어 상대위험도(Relative Risk)의 경우 참값과 다른 값이 나오게 됩니다.

이상적으로 case-control 연구는 case, control 여부에 따라서만 샘플링이 결정되어야 합니다. 로지스틱 회귀분석이 좋은 한 가지 이유는 이 가정이 맞은 경우, 편의가 없는 계수 추정치를 추정하게 됩니다. 편의가 없다는 말은 연구 집단에서 구한 계수의 기댓값이 전체 인구집단에서 구할 수 있는 계수의 "참값" 이라는 것입니다.

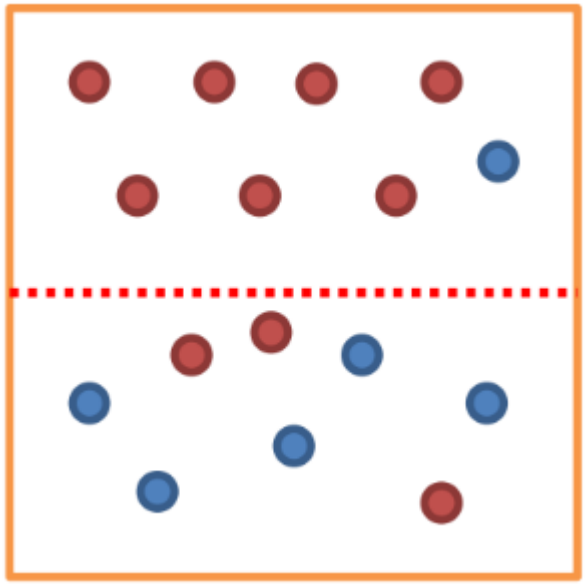
샘플링 결과에 따라 계수의 추정량에는 영향이 없고 오직 절편만 바뀐다.
https://3months.tistory.com/327

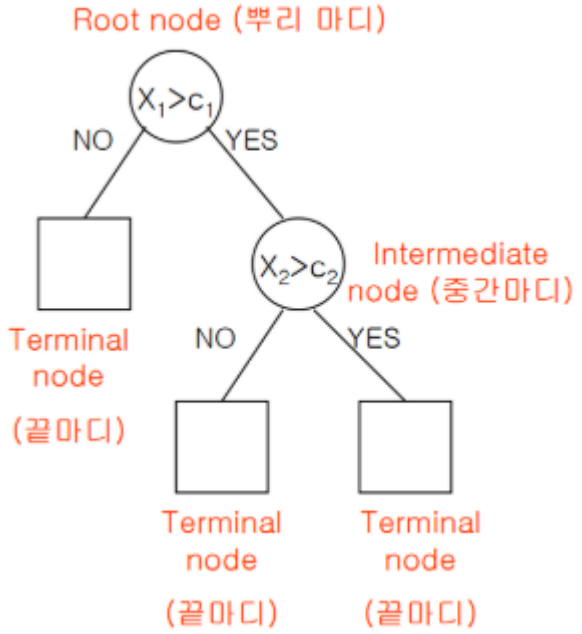
분석기법 - 의사결정나무

Decision Tree : 의사결정 나무라는 의미. 트리 구조를 사용, 각 분기점(node)에는 분석 대상의 속성들이 위치
- 각 분기점마다 목표 값을 잘 분류할 수 있는 속성을 찾아서 배치
- 해당 속성이 갖는 값을 이용하여 새로운 가지(branch)를 만듬
- 데이터 분류 시 최대한 많은 데이터 세트가 해당 분류에 속할 수 있도록 결정 노드의 규칙이 정해져야 함
- 결정 노드는 정보 균일도가 높은 데이터 세트를 먼저 선택할 수 있도록 규칙 조건을 만듬
- 정보의 균일도라는 룰을 기반으로 알고리즘이 쉽고 직관적이며 어떻게 규칙 노드와 리프 노드가 만들어지는지 알 수 있고 시각화로 표현할 수 있는 장점
- 균일도가 다양하게 존재할 수록 트리의 깊이가 커지고 복잡해 지며 과적합으로 정확도가 떨어진다는 단점
- 모든 데이터 상황을 만족하는 완벽한규칙은 없다고 인정하고 더 나은 성능을 보장하기 위하여 성능 튜닝을 통하여 트리의 크기를 사전에 제한하는 것이 요구됨

정보 이득 지수
1- 엔트로피(혼잡도) 지수
정보 이득이 높은 속성을 기준으로 분할

지니 계수
다양성이 낮을 수록 균일도가 높다는 의미로 1로 갈수록 균일도가 높으며 지니 계수가 높은 속성을 기준으로 분할





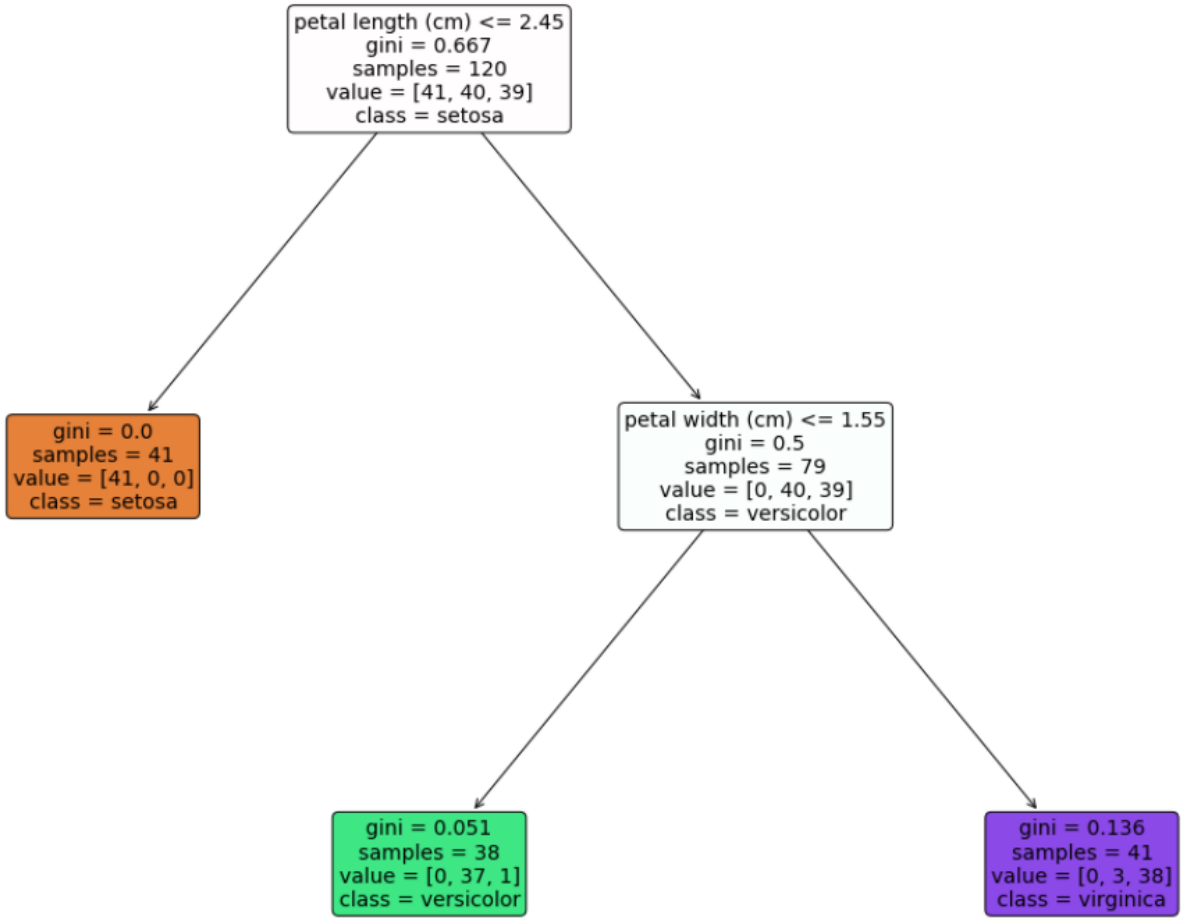
분석기법 - 의사결정나무

분꽃 종류 분류 :
SETOSA, VERSICOLOR, VIRGINICA

분류 속성 :
Petal length(꽃잎 길이),
Petal width(꽃잎 넓이)

규칙 조건 : petal length <= 2.45
gini : value=[] 데이터 분포에서의 지니 계수
samples : 현 규칙에 해당하는 데이터 건수
value = [] 클래스 값 기반의 데이터 건수
class=setosa는 가장 많은 하위 노드

각 노드의 색깔은 붓꽃 데이터의 레이블 값을 의미. 색깔이 짙어 질수록 지니 계수가 낮고 해당 레이블에 속하는 샘플 데이터가 많음



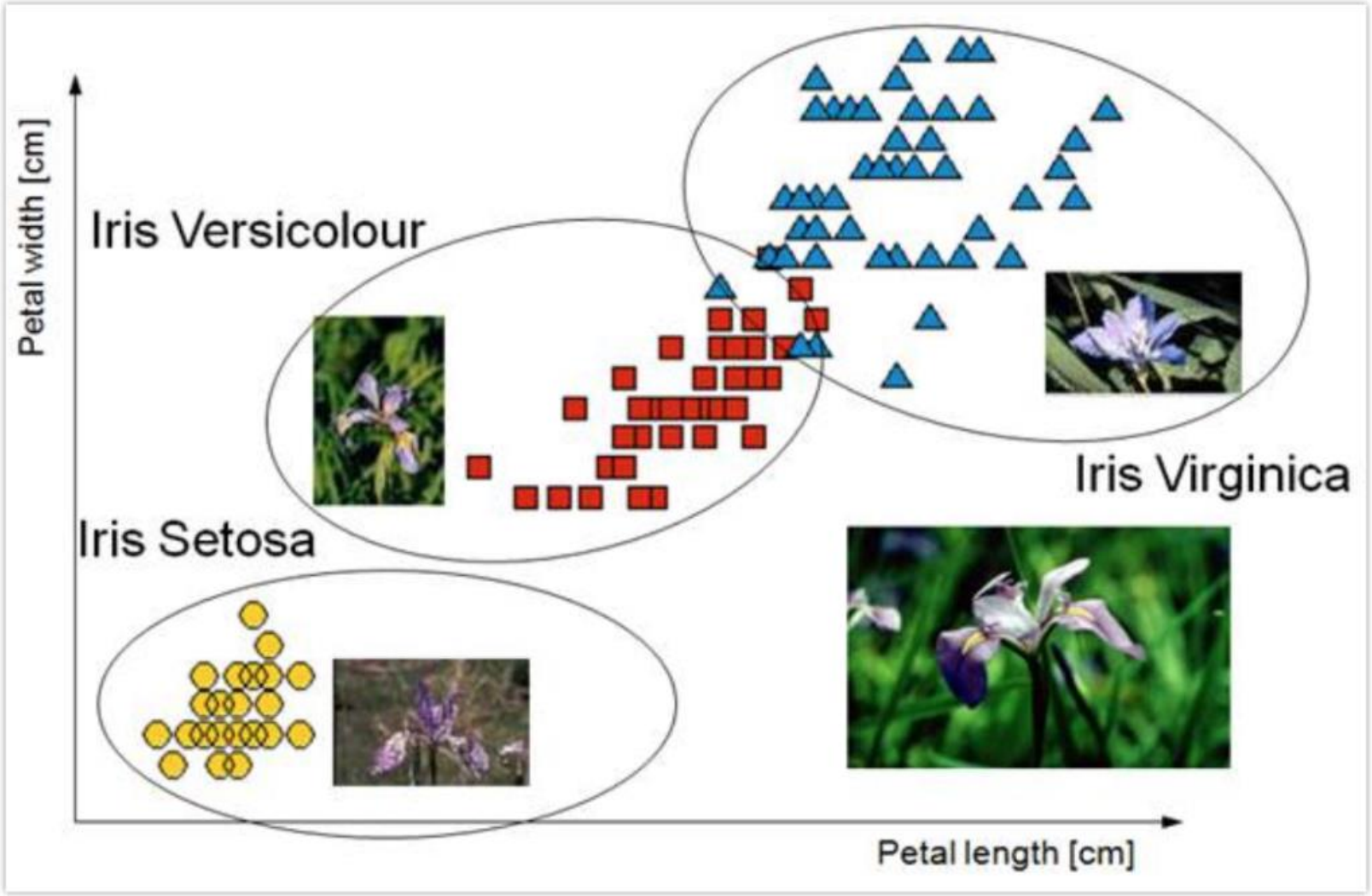
분석기법 - 의사결정나무

DT 파라미터

- max_depth : 트리의 최대 깊이
- max_features : 최적의 분할을 위해 고려할 최대 피처 개수
- max_leaf_nodes : 말단 노드의 최대 개수
- min_samples_split : 노드를 분할하기 위한 최소한의 샘플 데이터. 디폴트 2. 작게 설정할 수록 분할되는 노드 증가, 과적합 가능성 증가
- min_samples_leaf : 말단 노드가 되기 위한 최소한의 샘플 데이터 수

분석기법 - 의사결정나무

분류 - 꽃받침(Sepal)과 꽃잎(Petal)의 길이와 폭을 가지고 세 개의 종을 분류



분석기법 - 의사결정나무

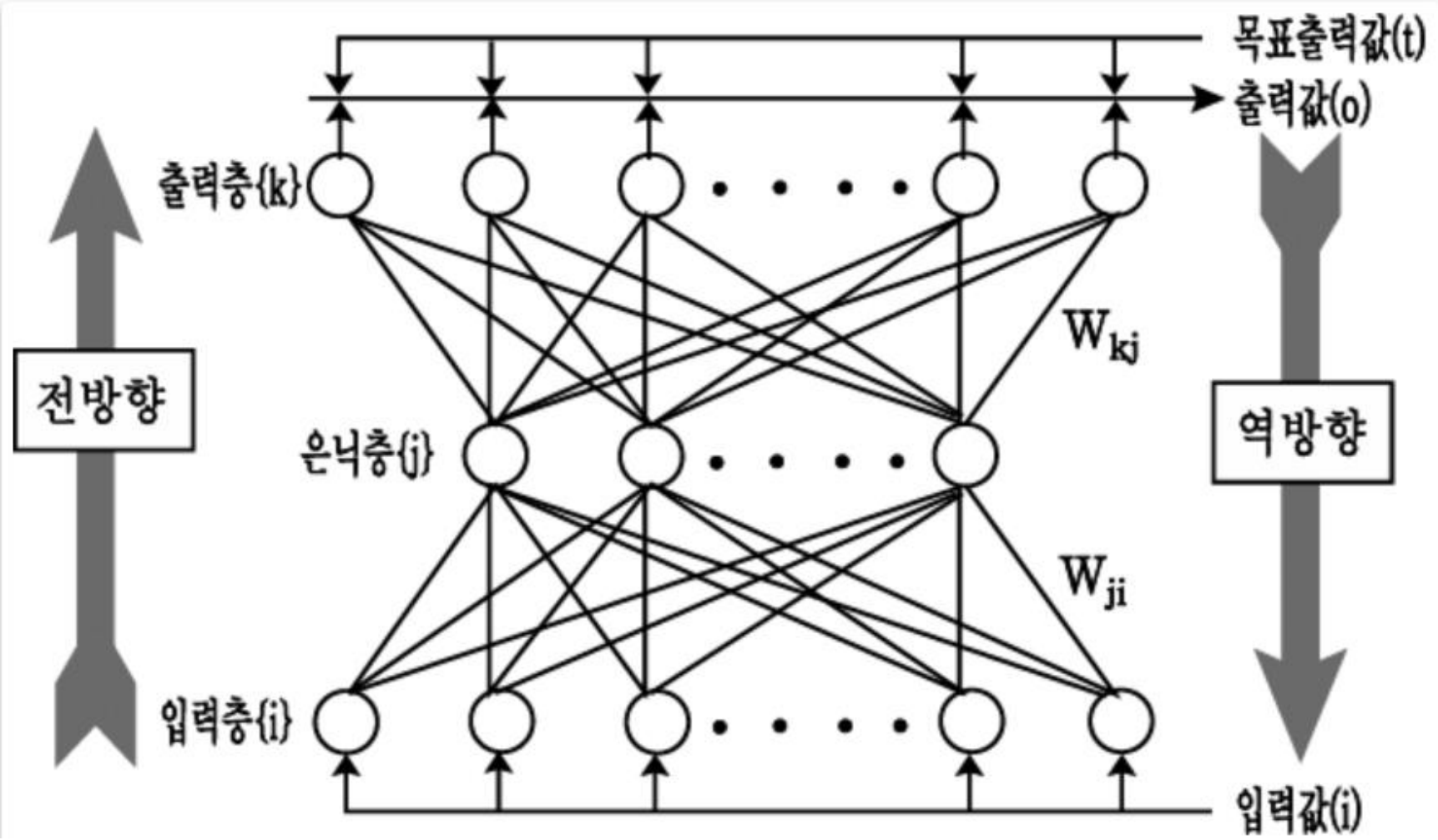
의사결정 나무는 간단하게 말해서 if~else와 같이 특정 조건을 기준으로 O/X로 나누어 분류/회귀를 진행하는 tree구조의 분류/회귀 데이터마이닝 기법이다.

이해도가 매우 높고 직관적이라는 장점이 있다. 그렇기에 많이 사용되며, 의사결정나무도 많은 머신러닝 기법과 동일하게 종속변수의 형태에 따라 분류와 회귀 문제로 나뉜다.
종속변수가 범주형일 경우 Decision Tree Classification으로 분류를 진행하고, 종속변수가 연속형일 경우 Decision Tree Regression으로 회귀를 진행한다.

분석기법 - 인공신경망

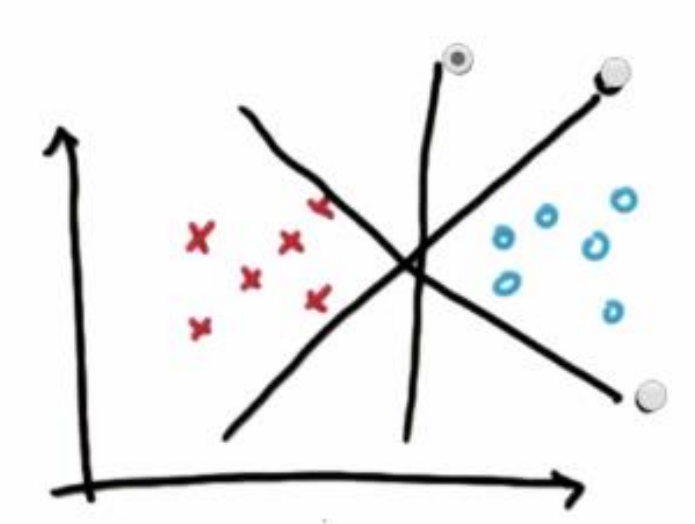
- 인공신경망은 생물학적 뇌의 작동 원리를 그래로 모방한 방법으로 데이터 안의 독특한 패턴이나 구조를 인지하는 데 필요한 모델을 구축하는 기법
- 간단한 계산능력을 가진 처리 단위, 뉴런 또는 노드들이 서로 복잡하게 연결된 컴퓨터 시스템으로서 외부에서 주어진 입력에 대하여 반응할 수 있으며 이러한 특징은 인공신경망을 구성하고 있는 다수의 뉴런끼리의 상호연결성에 기인
- 뉴런은 생체 내의 신경세포와 비슷한 것으로 가중치화 되어 상호 연결되어 있으며 가장 일반적인 인공신경망 모형은 다계층 퍼셉트론 모형으로서 입력층에서 은닉층, 은닉층에서 출력층으로 각 뉴런이 서로 연결되어 있음
- 인간의 신경학적 뉴런과 비슷한 노드와 층으로 구성되며 노드는 신경망 모형에서 가장 기본적인 요소를 의미함. 노드는 정보를 입력물로 받아들여 작동하는 인간의 뇌와 비슷하며 학습패러다임에 근거한 인공신경망은 입력 데이터를 기초로 가중치를 통해서 의사결정을 하게 함
- 다양한 뉴런이 서로 연결된 구조를 이용하여 의사결정이 이루어지는 구조를 이용
- 뇌의 신경시스템을 응용하여 예측을 최대화하기 위한 조직화을 찾기 위해 반복적으로 학습
- 복잡하고 비선형적이며 관계성을 갖는 다변량을 분석할 수 있음
- 인공신경망 기법은 회귀분석과 같은 선형 기법과 비교하여 비선형 기법으로서의 예측력이 뛰어나며 자료에 대한 통계적 분석 없이 결정을 수행할 수 있음

분석기법 - 인공신경망

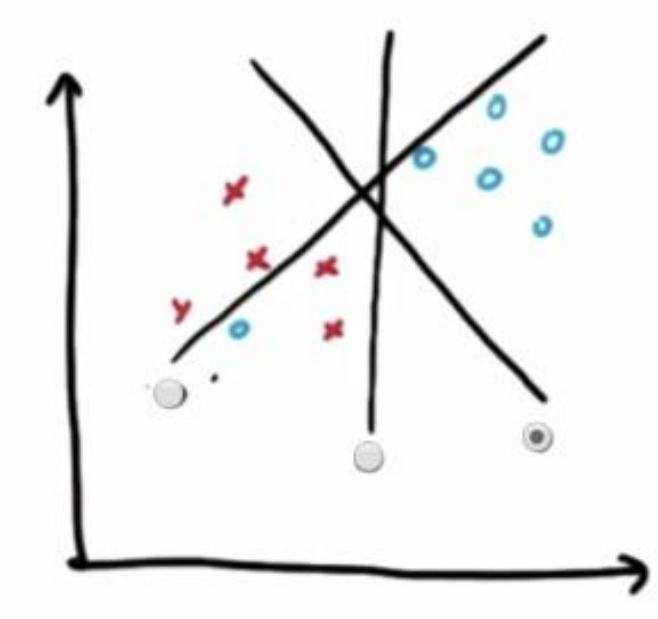


분석기법 - 서포트벡터머신

- Margin이란 선과 가장 가까운 양 옆 데이터와의 거리
  - 선과 가장 가까운 포인트를 서포트 벡터(Support vector)
  - 데이터를 정확히 분류하는 범위를 먼저 찾고, 그 범위 안에서 Margin을 최대화하는 구분선을 선택
  - 로버스트하다는 것은 아웃라이어(outlier)의 영향을 받지 않는다는 의미
  - 어느 정도 outlier를 무시하고 최적의 구분선



출처: Udacity

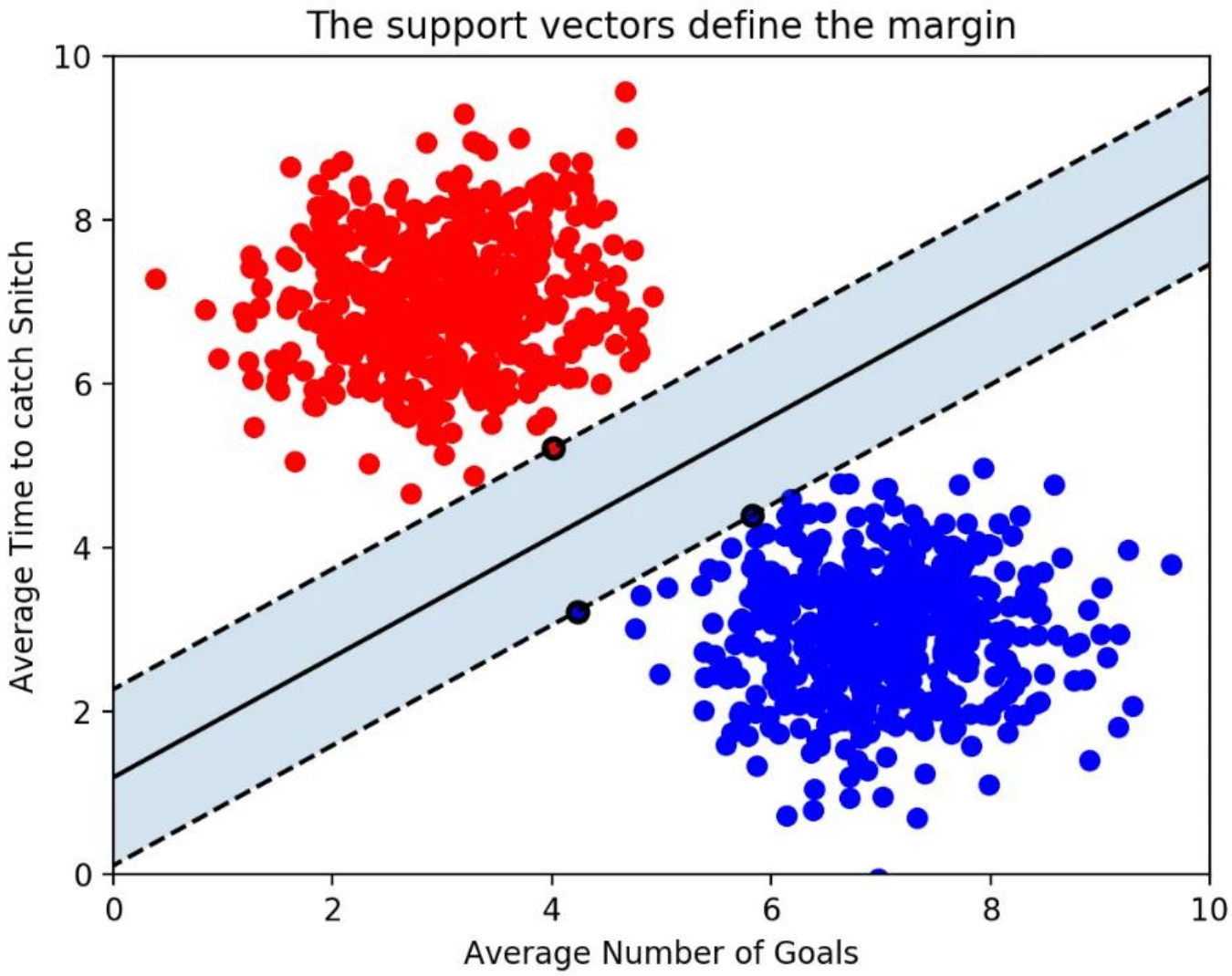


분석기법 - 서포트벡터머신

가운데 실선이 하나 그어져있는데, 이게 바로 '결정 경계'가 되겠다. 그리고 그 실선으로부터 검은 테두리가 있는 빨간점 1개, 파란점 2개까지 영역을 두고 점선을 그어놓았다. 점선으로부터 결정 경계까지의 거리가 바로 '마진(margin)'이다.

최적의 결정 경계는 마진을 최대화 한다.

x축과 y축 2개의 속성을 가진 데이터로 결정 경계를 그었는데, 총 3개의 데이터 포인트(서포트 벡터)가 필요. 즉, **n개의 속성을 가진 데이터에는 최소 n+1개의 서포트 벡터가 존재**

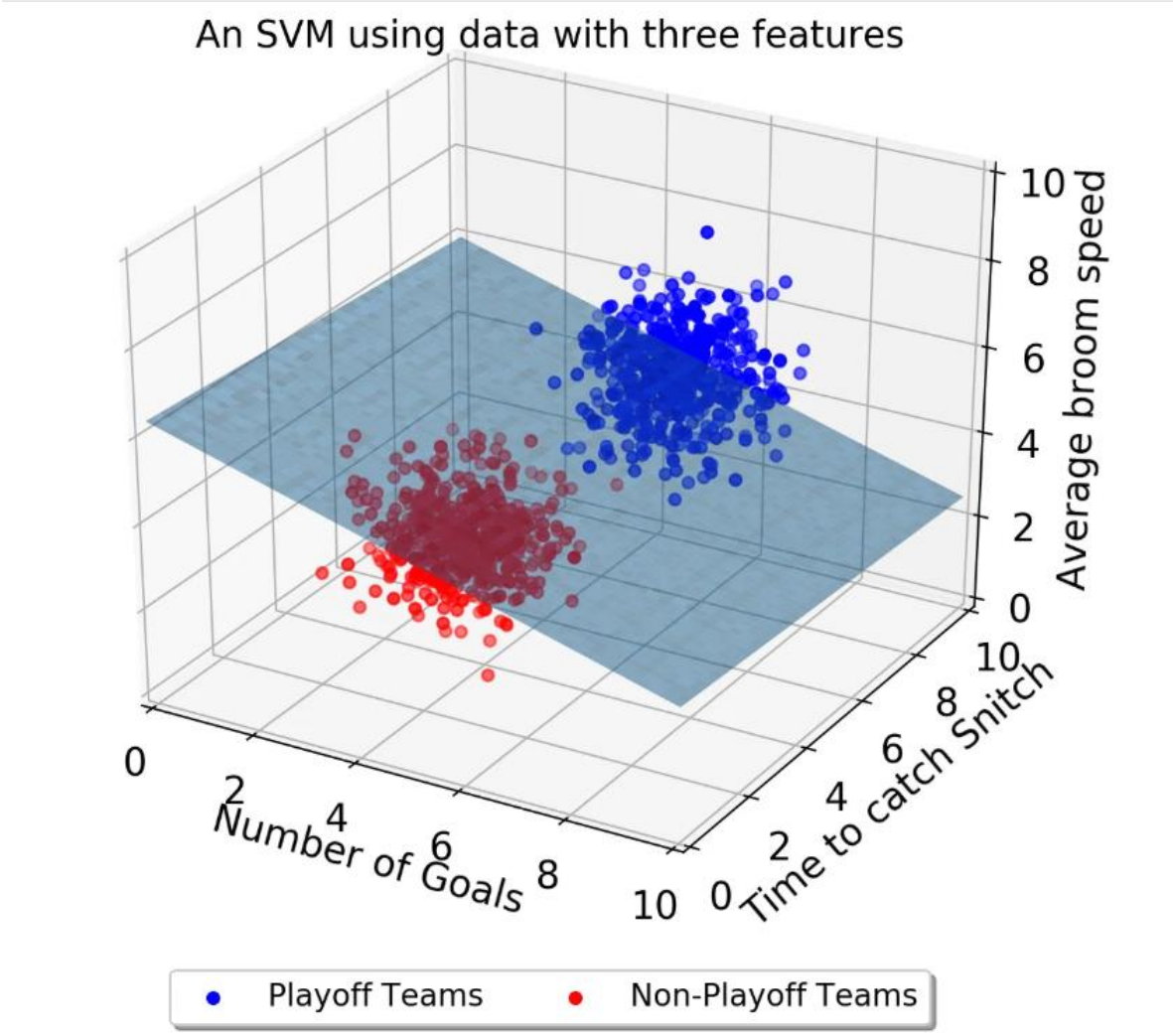


분석기법 - 서포트벡터머신

결정 경계는 '선'이 아닌 '평면'

시각적으로 인지할 수 있는 범위는 딱 3 차원까지다.

차원, 즉 속성의 개수가 늘어날수록 당연히 복잡해지며 **결정 경계**도 단순한 평면이 아닌 고차원이 된다. 이를 **"초평면(hyperplane)"**이라고 부른다.

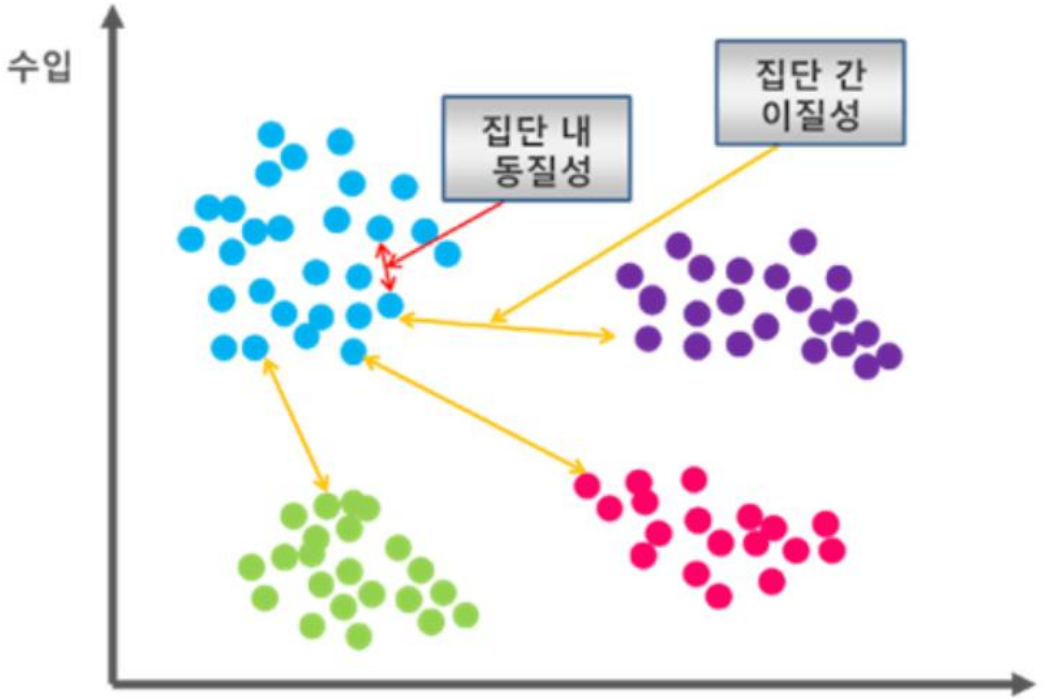


분석기법 - 연관성 분석

- 연관성 분석은 상품 혹은 서비스 간의 관계를 살펴보고 이로부터 유용한 규칙을 찾아내고자 할 때 이용될 수 있는 기법
- 동시 구매될 가능성이 큰 상품들을 찾아내는 기법으로 장바구니 분석과 관련된 문제에 많이 적용
- 측정의 기본은 얼마나 자주 구매되었는가 하는 빈도를 기본으로 연관정도를 정량화하기 위해서는 지지도, 신뢰도, 향상도를 계산하여 기준으로 함
- 연관성 규칙의 기본적인 개념은 장바구니 품목들을 식별하는 것에서부터 시작됨
- 사건들은 동시자발적으로 발생하며 이러한 사건들은 상호 영향을 주면서 결과가 나타나게 되는데 이와 같이 사건 또는 품목 간에 일어나는 연관성을 규명하려는 것이 연관성 규칙

분석기법 - 군집분석

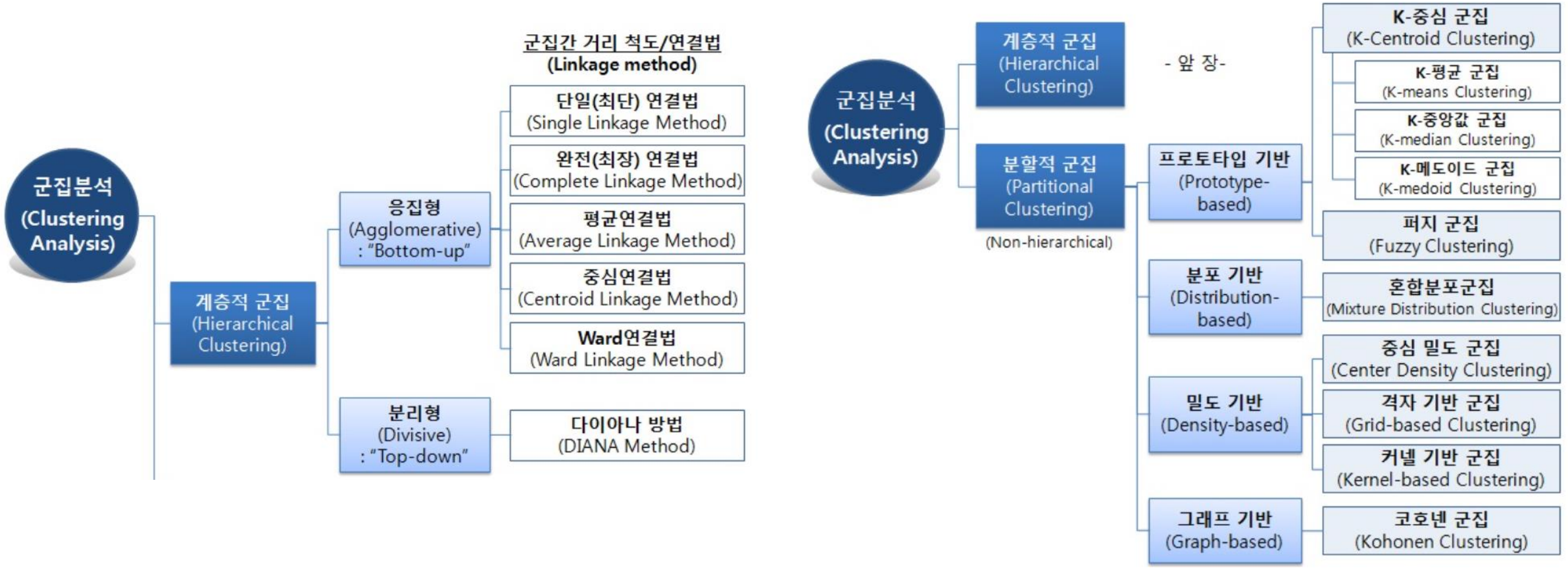
- 군집분석은 종속변수에 대한 독립변수의 영향과 같이 사전에 정의된 특수한 목적이 없음
- 데이터 자체에 의존해서 데이터 구조와 자료를 탐색하고 요약하는 기법
- 대용량 데이터의 경우 전체에 대한 의미 있는 정보를 얻어낼 수 있으며 전체를 유사한 군집으로 구분한다면 전체에 대한 의미 있는 정보를 얻을 수 있음
- 동일한 군집 내 개체들은 유사한 성격을 가짐, 즉 서로 다른 군집은 이질적인 성격을 갖도록 군집 형성
- 군집 간의 유사도를 평가하기 위하여 유클리드 거리, 마할라노비스 거리, 해밍 거리 등의 측정 함수 사용
- 대상을 어떻게 분석할 지에 따라 계층적 군집분석과 비계층적 군집분석으로 구분



분석기법 - 군집분석

군집분석의 분류

- 계층적 군집분석은 개별 대상 간의 거리에 의하여 가장 가까이 있는 대상들로부터 시작하여 결합해 감으로써 나무모양의 계층구조를 형성해 가는 방법
- 비계층적 군집분석은 군집의 수가 한 개씩 감소하는 것이 아니라 사전에 정해진 군집의 숫자에 따라 대상들이 할당됨. 많은 데이터를 빠르고 쉽게 분류할 수 있어야 하나 군집 형성을 위한 초기 값에 따라 군집 결과가 달라질 수 있음



분석기법 - 군집분석

계층적 군집 분석의 특징

- 거리 계산
  관측치들의 거리 계산은 유클리드 거리 계산식을 이용
  거리 기반으로 데이터를 축소(군집)
  기존 관측치들 간의 모든 거리를 계산
  군집화 방식

  단일기준결합방식 : 각 군집에서 중심으로부터 거리가 가까운 것 1개씩 비교하여 가장 가까운 것끼리 군집화
  완전기준결합방식 : 각 군집에서 중심으로부터 가장 먼 대상끼리 비교하여 가장 가까운 것끼리 군집화
  평균기준결합방식 : 각 군집의 모든 대상의 쌍 집합에 대한 거리를 평균 계산하여 가장 가까운 것끼리 군집화
  ward법 : 단순한 거리 기준이 아닌 구성 가능한 군집들을 구성하는 대상들의 측정치의 분산을 기준으로 사용

- 데이터의 축소
  예측 목적이 아닌 데이터 축소 목적의 분석 기법 -> 전체적인 데이터의 구조 파악
  전체 집단을 파악하기 어려우므로, 군집으로 나누어 집단을 파악
  (A : ~를 좋아하는 집단, B : ~를 싫어하는 집단

분석기법 - 군집분석

비계층적 군집 분석의 특징

- 계층적 군집분석보다 속도 빠름
- 군집의 수를 알고 있는 경우 이용
- k는 미리 정하는 군집 수
- 확인적 군집분석
- 계층적 군집화의 결과에 의거한 군집 수 결정
- 변수보다 관측대상 군집화에 많이 이용
- 군집의 중심(Cluster Center)은 사용자가 정함

> k-평균(k-Menas) 군집분석 알고리즘

① k값을 초기값으로, k개의 centroid 선정 (랜덤)

② 각 데이터 포인터를 가장 가까운 centroid에 할당

③ centroid에 할당된 모든 데이터 포인트의 중심 위치 계산 (centroid 재조정)

④ 재조정된 centroid와 가장 가까운 데이터 포인트 할당

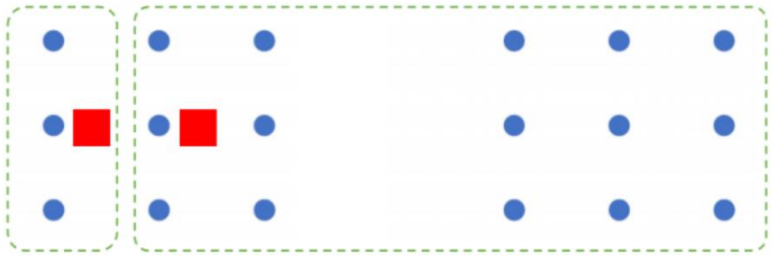
⑤ centroid 재조정이 발생되지 않을 때까지 ③, ④단계 반복

분석기법 - 군집분석

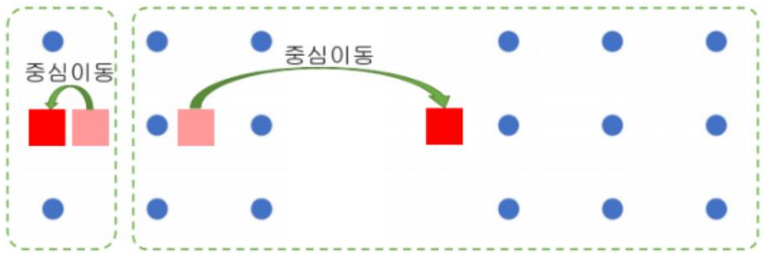
k-평균 알고리즘의 예시

1) 프로세스 1: 군집수 k를 2로 정함. -> 랜덤으로 2개의 시드를 초기화 함. (이 값은 random seed에 따라 달라짐)

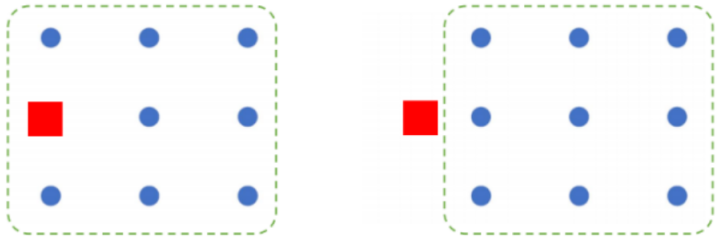
2) 프로세스 2: 모든 개체들과 두 시드와의 거리를 계산하여 가장 가까운 시드에 배정함으로써 군집을 생성.



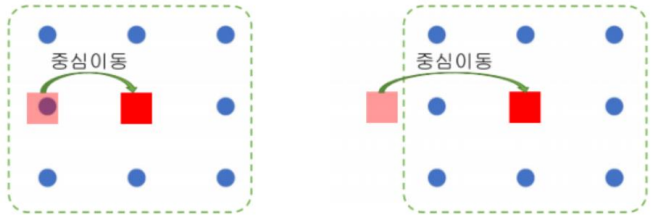
3) 프로세스 3: 묶여진 군집 내에서 중심을 이동.



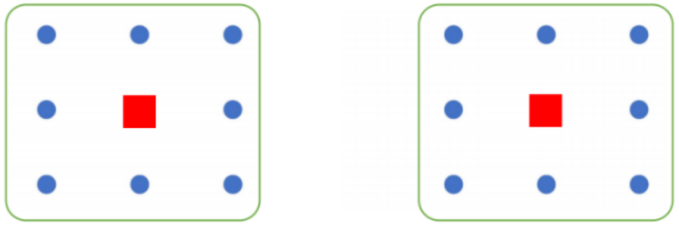
4) 프로세스 2': 이동한 중심을 기준으로 가장 가까운 시드에 배정하여 군집을 생성.



5) 프로세스 3': 묶여진 군집 내에서 중심이 이동됨.



6) 이동한 중심을 기준으로 가장 가까운 시드에 배정하여 군집을 생성. 만약 군집이 변함이 없다면 프로세스 종료.



고급분석기법 - 범주형 자료분석

- 범주형 자료분석은 변수들이 이산형 변수일 때 주로 사용하는 분석
  두 제품간의 선호도가 성별에 따라 연관이 있는지 여부를 판단하는 경우 각 집단 간의 비율 차이가 있는지 확인

- 변수의 속성이 일련의 범주 또는 불연속적 척도로 구성
  명목적 변수 : 성별, 종교, 지역, 정당 선호도
  서열적 변수 : 사회계층, 태도
- 일반적으로 빈도를 세서 표를 작성하며 두 변수의 범주가 교차되어 있다면 분할표로 요약
- 분할표를 기반으로 범주형 변수의 독립성, 동질성 검정 등의 카이제곱 검정을 수행

비율의 비교 : 두 개의 비율 비교 척도를 다루는 방법
- 상대 위험도 : 한 변수의 범주 별 다른 변수의 비율을 상대적으로 비교할 때 쓰이며 첫 범주에 속할 확률 추정량과 두 번째 범주에 속할 확률 추정량의 비로 정의. 상대 위험도가 1에 가까울수록 연관성이 없다는 의미
  ex) 대졸자의 합격 확률이 고졸의 1.6배
- 오즈비 : 오즈(성공확률/실패확률)의 각 범주 별 비율로 정의되며 상대위험도 보다 유연한 특성을 보임
  ex) 심장질환이 있을 오즈는 알콜 중독자 집단이 비중독자 집단의 약 4.26배

고급분석기법 - 다변량 분석

일반적으로 다변량분석이란 연구자의 연구대상으로부터 측정된 두 개 이상의 변수들의 관계를 동시에 분석할 수 있는 모든 통계적인 기법을 의미
각 개인 혹은 대상물에 대한 다수의 측정치를 동시에 분석하는 모든 통계적 방법이며 일변량분석과 이변량분석의 확장 형태로 볼 수 있음

다변량 분석 기법의 종류

- 다중회귀분석 : 하나의 계량적 종속변수와 하나 이상의 계량적 독립변수 간에 관련성이 있다고 가정되는 연구문제에 적합한 분석 기법으로 다수의 독립변수의 변화에 따른 종속변수의 변화를 예측

- 다변량 분산분석 : 두개 이상의 범주형 종속변수와 다수의 계량적 독립변수 간의 관련성을 동시에 알아볼 때 이용되는 통계적 방법으로 일변량 분산분석의 확장된 형태. 다수의 관광행동집단의 다수의 관광만족도 차원

- 다변량 공분산분석 : 두 개 이상의 계량적 종속변수에 대한 각 집단 반응치의 분산에 대한 가설 검증에 유용
  직원의 학력을 통제한 상태에서 이론시험 성적과 실무 성적이 두 가지의 교육방법(강의/학습참여)에 따른 차이

- 정준상관분석 : 다수의 계량적 종속변수와 다수의 계량적 독립변수 간의 상관관계를 알아보고자 할 때 사용
  다수의 외식동기 항목과 다수의 레스토랑 선택속성 변수들 간의 관계분석을 통해 고객의 외식동기가 레스토랑 선택에 미치는 영향을 분석

고급분석기법 - 다변량 분석

- 요인분석 : 많은 수의 변수들 간의 상호 관련성을 분석하고 이들 변수들을 어떤 공통요인들로 설명하고자 할 때 이용하는 기법. 관광객이 여행사 선택 변수(속성)들이 많을 때 이들 변수 모두를 개별적으로 분석하기 보다는 좀 더 이해하기 쉬운 몇 개의 요인으로 축소하거나 요약할 때 사용

- 군집분석 : 집단에 관한 사전 정보가 전혀 없는 각 표본에 대하여 그 분류체계를 찾을 때 사용하며 판별분석과 달리 집단이 사전에 정의되지 않음. 주제공원 운영자가 고객들로부터 각종 레저활동에 대한 관심도, 다양한 실/내외 시설에 대한 선호도 등을 조사하여 각종 주제시설의 세분시장을 발견하려는 경우

- 다중 판별분석 : 각 표본을 여러 개의 범주를 가진 종속변수에 기초한 여러 개의 집단으로 분류 시 적합. 주목적은 집단 간의 차이를 판별하며 어떤 사례가 여러 개의 계량적 독립변수에 기초하여 특정 집단에 속할 가능성을 예측하는 데 있음. 새 패키지 상품을 구매할 고객과 구매하지 않을 고객을 예측하는 것과 새 상품을 평가하는 어떤 척도가 구매자와 비구매자를 가장 잘 판별해 줄 수 있는가? 새 상품을 살 것이라는 반응이 가격척도 점수가 높은 것과 항상 관련이 있고 반면 새 상품을 사지 않을 것이라는 반응이 가격척도 점수가 낮은 것과 관련이 있다면 가격은 구매자와 비구매자를 판별하는 데 좋은 척도라는 결론 도출

- 다차원 척도법 : 두 표본의 유사성을 다차원 공간상의 거리로 나타낼 때 사용. 특정 관광지를 대상으로 관광객의 지각에 대한 유사성 연구. 응답자들이 경쟁관광지와 비교하여 자기 지역 관광상품에 대한 이미지를 어떻게 지각하는지를 알 수 있으며 이를 통해 자기 지역의 차별화 방안을 구체화할 수 있음

고급분석기법 - 시계열 분석

목적
- 시간에 따라 관측되는 자료를 시계열 자료하고 함
- 가장 큰 목적은 현재까지 수집된 시계열자료를 분석하여 미래를 예측하는 것
- 변동은 우연변동과 계통변동으로 구분할 수 있는데 계통변동에 의한 특성이 유지되고 우연변동이 작을수록 통계적 모형을 통하여 더 정확하게 미래를 예측

평활법
- 예측 시 과거의 모든 자료를 동일하게 취급하여 계산한 예측값 보다는 최근의 자료를 더 비중 있게 취급.
- 이동평균법은 최근 일정 시점 자료들의 평균값을 이용한 예측, 일정 시점의 크기에 따라 그 결과가 달라짐
- 지수평활법은 가중치를 현시점에서 과거로 갈수록 지수적으로 작게 주는 방법. 단순지수평활법, 이중지수평활법, 계절지수평활법

자기회귀 이동평균과정
- 자기회귀과정(AR) : 현재의 관측 값이 과거의 관측 값에 영향을 받는 경우를 자기회귀(AR)과정이라고 하며 과거의 p개 관측 값에 영향을 받는 경우 p차 자기회귀과정(AR(p) 과정) 이라고 함
- 이동평균과정(MA) : 현재의 관측 값이 현재 및 과거의 오차항들의 가중평균으로 결정되는 과정
- 자기회귀 이동평균과정(ARMA) : 자기회귀과정이나 이동평균과정으로만 시계열자료를 적합시킬 경우 p나 q의 값이 너무 커질 수가 있으며 자기회귀과정과 이동평균과정을 동시에 포함하는 확률과정을 사용할 수 있는데 이를 자기회귀 이동평균과정이라 함

고급분석기법 - 시계열 분석

정상성(stationarity)을 나타내는 시계열은 시계열의 특징이 해당 시계열이 관측된 시간에 무관합니다. 따라서, 추세나 계절성이 있는 시계열은 정상성을 나타내는 시계열이 아닙니다 — 추세와 계절성은 서로 다른 시간에 시계열의 값에 영향을 줄 것이기 때문입니다. 반면에, 백색잡음(white noise) 시계열은 정상성을 나타내는 시계열입니다 — 언제 관찰하는지에 상관이 없고, 시간에 따라 어떤 시점에서 보더라도 똑같이 보일 것이기 때문입니다.

몇 가지 경우는 헷갈릴 수 있습니다 — 주기성 행동을 가지고 있는 (하지만 추세나 계절성은 없는) 시계열은 정상성을 나타내는 시계열입니다. 왜냐하면 주기가 고정된 길이를 갖고 있지 않기 때문에, 시계열을 관측하기 전에 주기의 고점이나 저점이 어디일지 확실하게 알 수 없습니다.

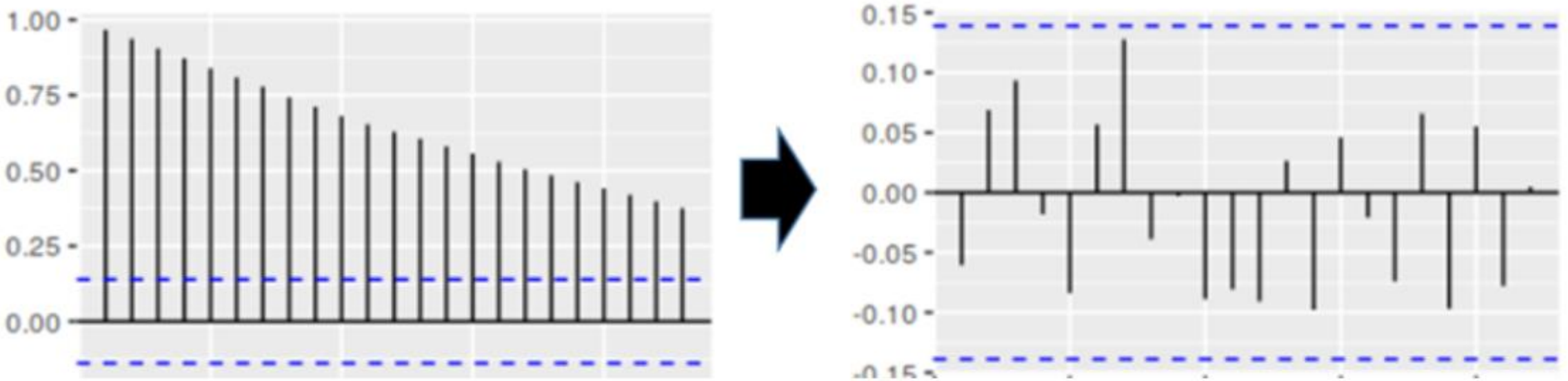
일반적으로는, 정상성을 나타내는 시계열은 장기적으로 볼 때 예측할 수 있는 패턴을 나타내지 않을 것입니다. (어떤 주기적인 행동이 있을 수 있더라도) 시간 그래프는 시계열이 일정한 분산을 갖고 대략적으로 평평하게 될 것을 나타낼 것입니다.

고급분석기법 - 시계열 분석

ARIMA(Autoregressvie integrated MovingAverage)

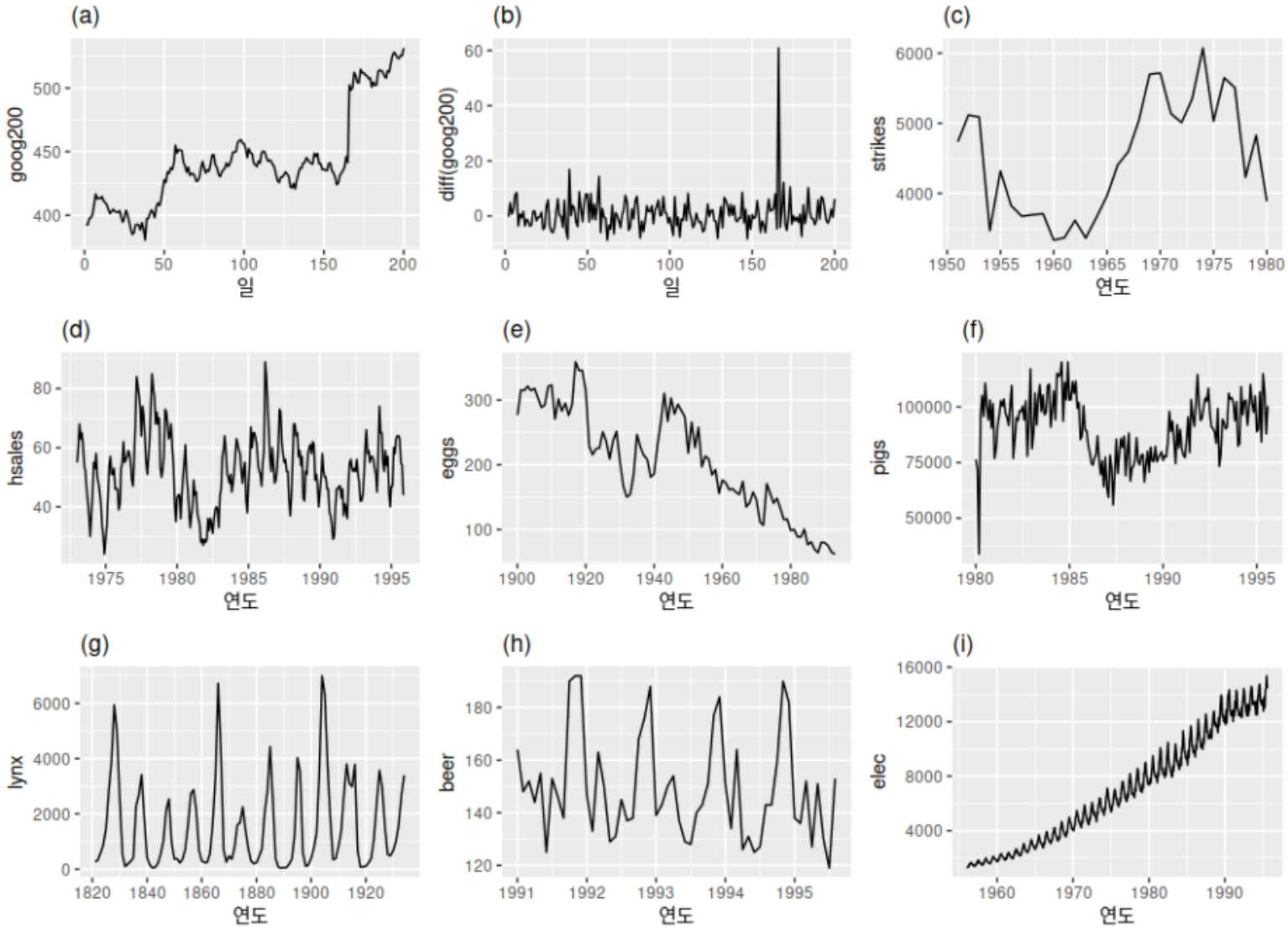
AR(자기상관) : 이전의 값이 이후의 값에 영향을 미치고 있는 상황
MA(이동평균) : 랜덤 변수의 평균값이 지속적으로 증가하거나 감소하는 추세

차분은 비정상성을 정상성으로 만들기 위해, 관측값들의 차이를 계산하는 것.
(아래 그림처럼 비정상성에서 정상성으로)



고급분석기법 - 시계열 분석

정상성을 나타내는 시계열?
분명하게 계절성이 보이는 (d), (h), (i)는 후보가 되지 못합니다. 추세가 있고 수준이 변하는 (a), (c), (e), (f), (i)도 후보가 되지 못합니다. 분산이 증가하는 (i)도 후보가 되지 못합니다. 그러면 (b)와 (g)만 정상성을 나타내는 시계열 후보로 남았습니다.
언뜻 보면 시계열 (g)에서 나타나는 뚜렷한 주기(cycle) 때문에 정상성을 나타내는 시계열이 아닌 것처럼 보일 수 있지만 이러한 주기는 불규칙적(aperiodic)입니다 — 먹이를 구하기 힘들만큼 살쾡이 개체수가 너무 많이 늘어나 번식을 멈춰서, 개체수가 작은 숫자로 줄어들고, 그 다음 먹이를 구할 수 있게 되어 개체수가 다시 늘어나는 식이기 때문입니다.
장기적으로 볼 때, 이러한 주기의 시작이나 끝은 예측할 수 없습니다.
따라서 이 시계열은 정상성을 나타내는 시계열



고급분석기법 - 베이지안 기법

- 확률을 '지식의 상태를 측정 ' 하는 것이라고 해석하는 확률론

- 사건 A와 B가 있을 때, '사건 B가 일어난 것을 전제로 한 사건 A의 조건부 확률'을 구하고 싶다고 하자. 그런데 지금 알고 있는 것은 사건 A가 일어난 것을 전제로 한 사건 B의 조건부 확률, A의 확률, B의 확률 뿐이다. 그럴 때, 원래 구하고자 했던 '사건 B가 일어난 것을 전제로 한 사건 A의 조건부 확률'은 다음과 같이 구할 수가 있다.
- 새로운 정보를 토대로 어떤 사건이 발생했다는 주장에 대한 신뢰도를 갱신해 나가는 방법
- 베이즈 정리는 사전확률과 사후확률 간의 관계에 대해 설명하는 정리, P(A|B) = P(B|A)P(A)/P(B)
- 사례 : 만약 어떤 사람이 질병에 걸렸다고 검진받았을 때 이 사람이 정말로 질병에 걸렸을 확률
  질병 D의 발병률 0.1%, 질병이 실제로 있을 때 질병이 있다고 검진할 확률(민감도) 99%,
  질병이 없을 때 질병이 없다고 검진할 확률(특이도) 98%
  D : True, 실제로 병이 있다.
  P : Positive, 병이 있다고 진단받았다.
  P(P) = 병일 때 Positive일 확률 + 병이 아닐 때 Positive일 확률 = P(P|D)*P(D) + P(P|NO D)*P(NO D)
  P(P|D) = 0.99, P(NO P|NO D) = 0.98, P(D) = 0.001, P(P|NO D)=0.02, P(NO D) = 0.999

  P(D|P) = P(P|D) * P(D) / P(P) = 0.99 * 0.001 / (0.99 * 0.001 + 0.02 * 0.999)= 0.047

$$P(A|B) = \frac{P(B \cap A)}{P(B)} = \frac{P(A)P(B|A)}{P(B)} = \frac{P(A)P(B|A)}{P(A)P(B|A) + P(A^c)P(B|A^c)}$$

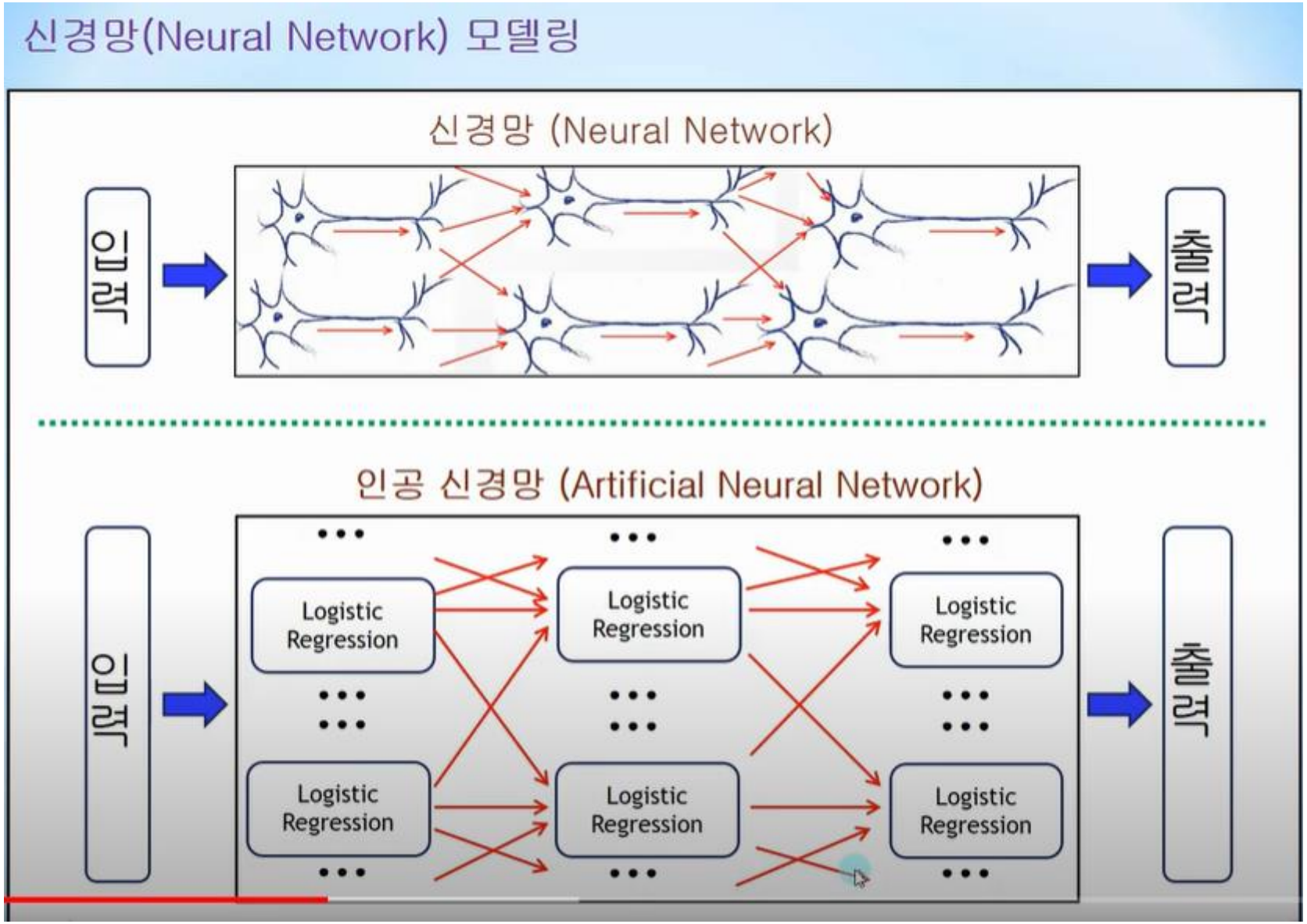
고급분석기법 - 딥러닝 분석

- 심층 학습 또는 딥러닝은 여러 비선형 변환 기법의 조합을 통해 높은 수준의 추상화(다량의 데이터나 복잡한 자료들 속에서 핵심적인 내용 또는 기능을 요약하는 작업)를 시도하는 기계 학습 알고리즘의 집합으로 정의되며 큰 틀에서 사람의 사고방식을 컴퓨터에게 가르치는 기계학습의 한 분야라고 할 수 있음
- 현재 영상처리 패턴인식 등 다양한 분야에 사용되고 있는 딥러닝 모델에는 영상, 이미지와 같은 데이터를 처리하는 Convolutional Neural Network과 음성인식 및 음악, 시퀀스가 있는 문자열 데이터를 처리하는 Recurrent Neural Network 등이 대표적으로 사용되고 있으며 각 분야별 특화된 기법으로 진화하며 인공지능의 발전 주도
- 딥러닝의 딥이라는 말은 신경방의 층이 깊고 각 층마다 고려되는 변수가 많다는 의미. 2~3개의 층으로 구성되어 있는 천층망과 그 이상의 층으로 구성되어 있는 심층망으로 구분. 딥러닝의 깊이를 나타내는 층의 개수는 입력층과 출력층 사이에 은닉층 개수를 추가하는 형태이며 이러한 이유로 심층신경망(DNN)이라 함
- CNN : 사람의 시각 인지 과정을 모방하여 컴퓨터 비전 분야에서 특화된 방법으로 사용. 컨볼루션을 사용하는 역전파 기반의 인공신공망의 유형으로서 다양한 형태의 컨볼루션을 사용하면 3차원 데이터의 공간적 정보를 유지한 채 원하는 특성을 추출하는 데 탁월한 성능을 보여줌. 인공신경망 알고리즘의 앞부분에 컨볼루션 기법을 추가한 알고리즘으로 입력하는 데이터를 컨볼루션으로 특징 추출 등의 전처리를 한 후 추출된 특징을 기반으로 신경망을 이용하여 분류해 내는 방식. 페이스북의 딥페이스는 두개의 인물사진을 비교해서 동일인인지 여부를 판단하는 프로그램인데 그 정확도가 97.25%에 이름
- RNN : 음성이나 언어 등 연속적으로 되풀이되는 입력 데이터를 사용하는 모델로서 음성인식, 자연어처리 등 다양한 분야에서 사용. 기본적인 아이디어는 순차적인 정보를 처리한다는데 있으며 대표적으로 언어 데이터에 쓰이는데 그 이유는 단어 다음에 단어가 올 것이 확실한 데이터이므로 음성인식, 단어의 의미 파악, 대화 등에 사용됨. 이 이외에도 음악, 문자열, 동영상 등 순차적인 정보가 담긴 데이터라면 적용 가능. 일반 신경망에서 추가된 순환가중치는 과거의 데이터에 대한 정보를 기억할 수 있는 기능이 있으며 새로운 데이터를 처리할 때 과거의 기억을 사용할 수 있음

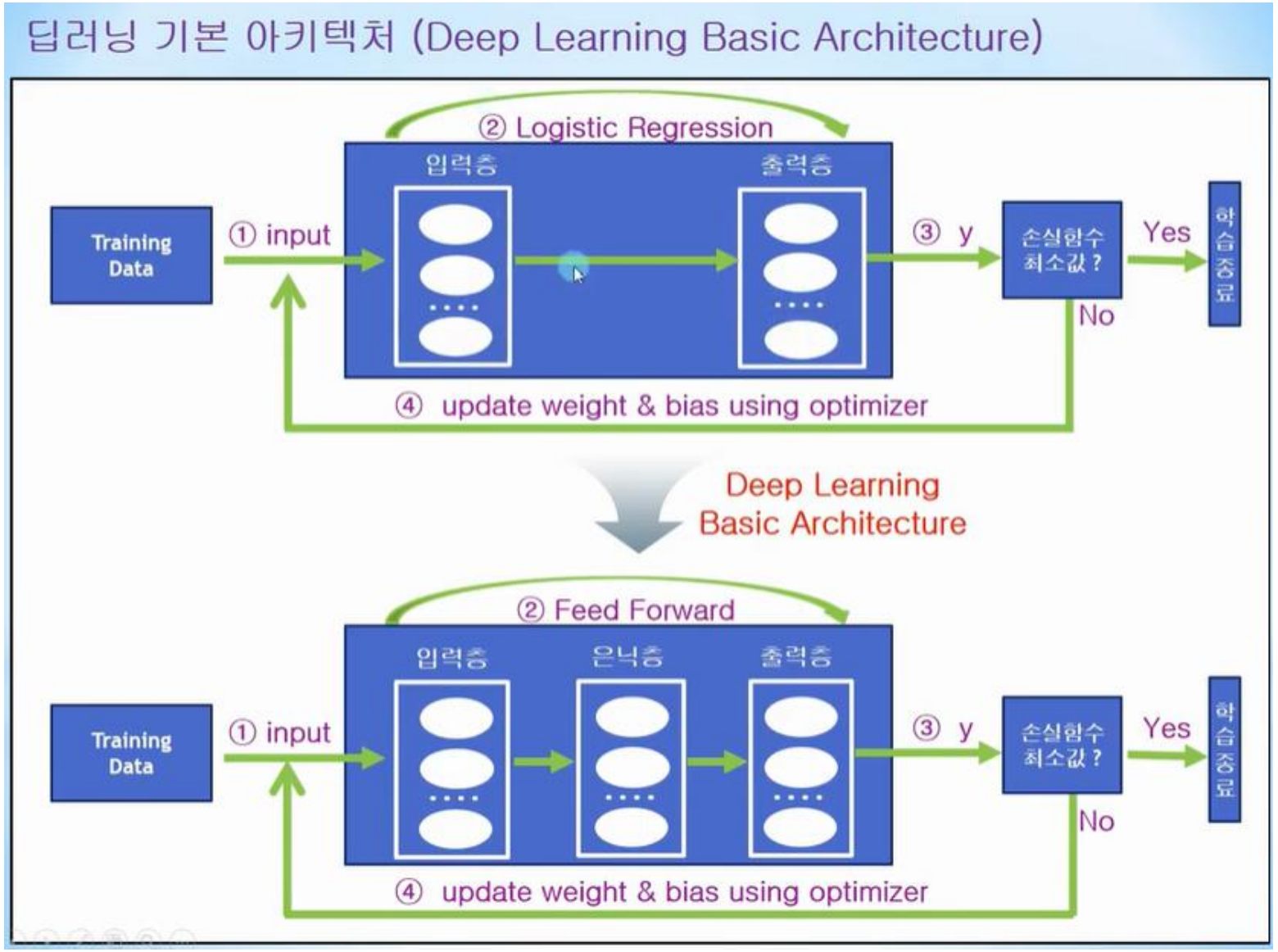
고급분석기법 - 딥러닝 분석

- 심층 신뢰망(DBN; Deep Belief Network) : 입력층과 은닉층으로 구성되어 있는 볼츠만 머신(RBM)을 층층히 쌓아 올린 형태로 연결한 신경망이며 RMB을 여러 층으로 쌓아 올려 연결하기 때문에 심층신뢰망이라고 하는데 이를 통해 사전 학습을 하여 깊은 신경망 구조에서 나타나는 경사감소소멸 현상을 해결. 여기서 볼츠만 머신이란 여러 가지 형태의 레이블된 데이터 또는 레이블되지 않은 데이터을 확률적인 방법으로 판별하는 생성모델을 의미. RBM 기반으로 사전학습을 하게 되면 순차적으로 초기화된 가중치를 얻게 되고 그 다음은 레이블이 있는 학습 데이터를 가지고 지도학습을 진행하여 가중치의 튜닝 과정을 거치는데 이때 역전파 기법이 적용되며 최종 가중치를 계산하는 것이 심층신뢰망의 목표임

고급분석기법 - 딥러닝 분석

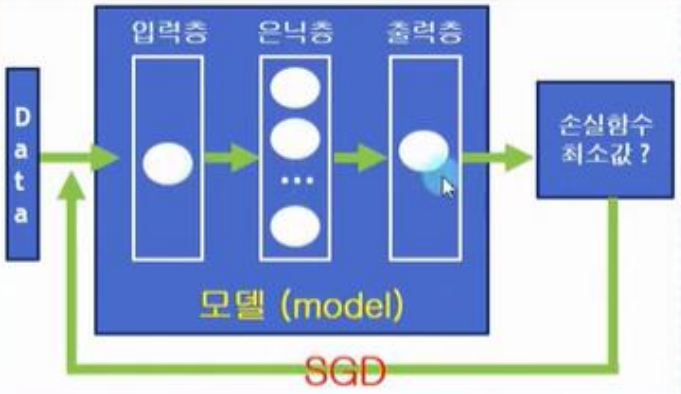


고급분석기법 - 딥러닝 분석



고급분석기법 - 딥러닝 분석

딥러닝 기본 코드 (TensorFlow 2.x)



Training Data

공부시간 (x)	Fail/Pass (t)
2	0
4	0
6	0
8	0
10	0
12	0
14	1
16	1
18	1
20	1

Logistic Regression

```
model = tf.keras.models.Sequential()
model.add(tf.keras.layers.Dense(1, input_shape=(1,),
                                activation='sigmoid'))
model.compile(tf.keras.optimizers.SGD(learning_rate=0.1),
              loss='binary_crossentropy', metrics=['accuracy'])
model.fit(x_data, t_data, epochs=500)
```

(출력층 노드 개수)

기본 코드 비교

Deep Learning

```
model = tf.keras.models.Sequential()
model.add(tf.keras.layers.Dense(8, input_shape=(1,),
                                activation='sigmoid'))
model.add(tf.keras.layers.Dense(1, activation='sigmoid'))
model.compile(tf.keras.optimizers.SGD(learning_rate=0.1),
              loss='binary_crossentropy', metrics=['accuracy'])
model.fit(x_data, t_data, epochs=500)
```

(은닉층 노드 개수, 출력층 노드 개수)

고급분석기법 - 비정형 데이터 분석

- 정형화되지 않은 데이터로서 구체적으로 미리 정의된 데이터 모델을 가지고 있지 않은 데이터를 의미
- 대표적인 비정형 데이터의 예로는 아주 많은 양의 데이터를 가지고 있으면서 구조와 형태가 다르고 정형화되지 않은 문서, 영상, 음성 등을 들 수 있음
- 빅데이터 환경에서 거의 80% 이상이 비정형 데이터이므로 빅데이터에서의 데이터 마이닝은 비정형 데이터 마이닝에 초점이 맞추어져 있음
- 빅데이터에서 데이터 마이닝은 통계 기반의 데이터 분석 도구를 사용하거나 OLAP 분석을 통해 데이터를 다양한 각도의 관점으로 조명하여 의미 있는 것으로 해석하는 것에 덧붙여 데이터 사이의 숨겨진 관계화 패턴, 경향 등을 추출함. 이것은 비정형 데이터를 일단 정련 과정을 통해 정형 데이터로 만들고 난 다음에 일반적인 데이터 마이닝 작업인 분류, 군집화, 회귀분석 요약, 이상감지 등에 적용하여 의미 있는 정보를 발굴해 낸다는 것임
- 비정형 데이터 마이닝 과정을 살펴보면 탐색, 이해, 분석의 과정으로 진행함.
  탐색 과정 : 질의, 집합연산, 재귀 및 팽창 등의 작업 수행
  이해 과정 : 통계, 분배, 특징 선택, 군집화, 분류 편집, 시각화 등의 작업 수행
  분석 과정 : 경향, 상관관계, 분류 등의 작업 수행
- 정제된 데이터베이스를 기반으로 일정한 기준이 적용된 상식적인 범위에서 부분적인 데이터를 다루는 정형 데이터 마이닝의 한계를 뛰어넘는 대표적인 비정형 데이터 마이닝 기법으로 텍스트 마이닝, 웹마이닝, 오피니언 마이닝, 소셜 네트워크 분석 등이 있음
- 텍스트 마이닝은 인간의 언어로 이루어진 비정형 텍스트 데이터들을 자연어 처리 방식을 이용하여 대규모 문서에서 정보를 찾아 추출하거나 연계성을 파악하거나 분류 또는 군집화, 요약 등 빅데이터에 숨겨진 의미를 발견하는 기법

고급분석기법 - 비정형 데이터 분석

- 웹 콘텐츠 마이닝은 웹 페이지에서 유용한 데이터, 정보, 지식을 마이닝하고 추출하고 통합하는 것을 의미
- 웹 Usage 마이닝은 웹상에서 사용자가 찾고자 했던 것을 기록하고 있는 웹 서버 로그에서 유용한 정보를 추출하는 과정을 의미
- 오피니언 마이닝은 어떤 사안이나 인물, 이슈, 이벤트 등과 같은 원천 데이터에서 의견이나 평가, 태도, 감정 등과 같은 주관적인 정보를 식별하고 추출하는 것으로 오피니언 분석, 평판분석, 정서분석이라고도 함. 상품이나 서비스에 대한 시장 규모를 예측하거나 소비자의 반응 및 입소문을 분석하는 데 활용
- 소셜 데이터 마이닝은 소셜 미디어에 올라오는 정보를 이용해 마케팅 전략, 사회의 이슈 및 트랜드, 여론 변화 흐름을 분석해 새로운 정보를 제공. SNA 분석은 명성, 응집력, 범위, 중계, 구조적 등위성의 5가지 속성을 지님
  명성 : 네트워크에서 누가 권력을 가지고 있는지 혹은 누가 책임을 지고 있는지를 의미
  응집력 : 행위자들 간 강한 사회화 관계(직접적 연결)의 존재를 나타냄
  범위 : 행위자의 네트워크 규모를 나타냄
  중개 : 다른 네트워크와 연결해 주는 것을 의미
  구조적 등위성 : 한 네트워크의 구조적 지위와 그 위치가 주는 역할이 동일한 사람들 간의 관계를 의미

고급분석기법 - 앙상블 분석

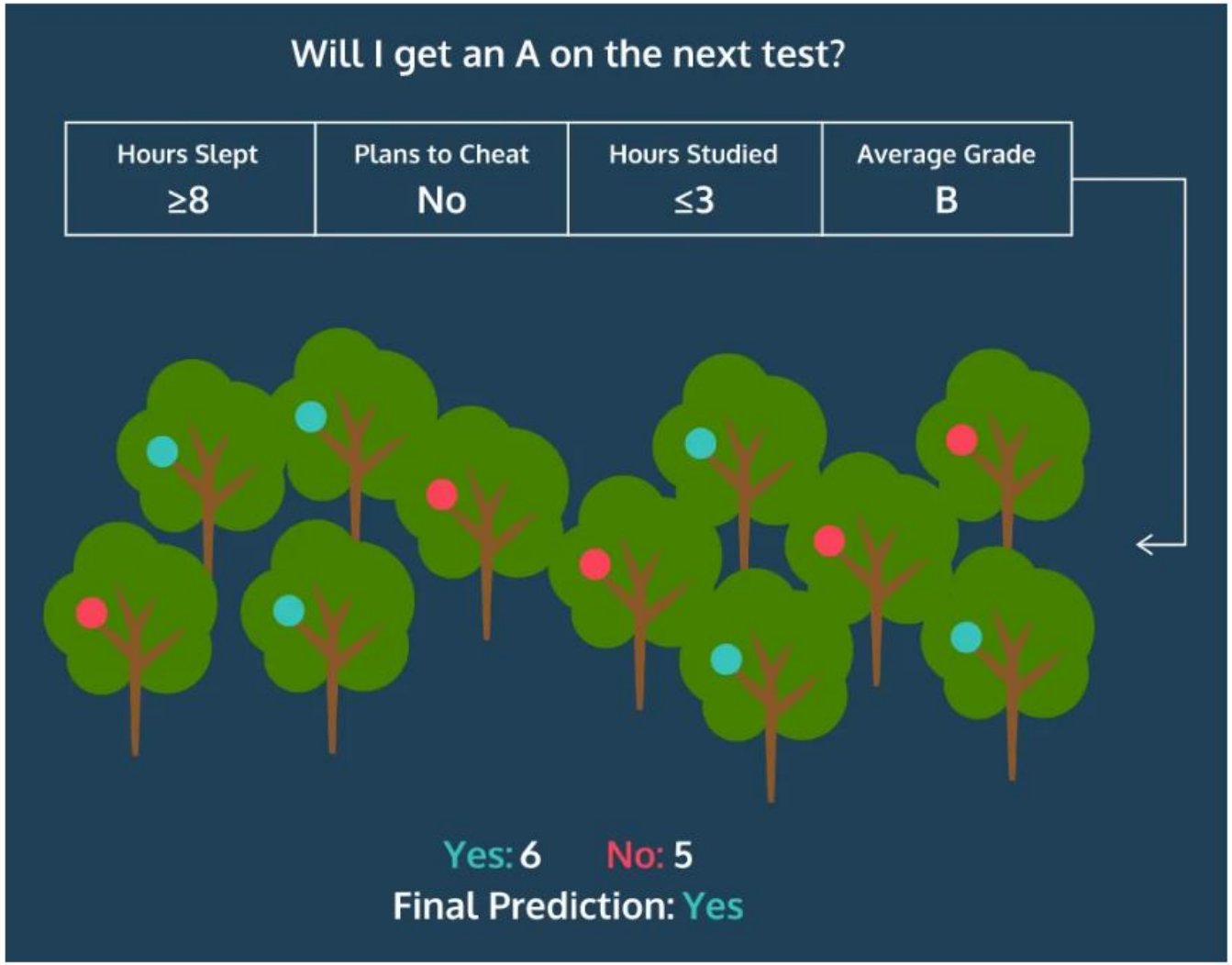
- 앙상블 기법은 여러 가지 모형이 독립적으로 결과를 뽑아낸 후 각 모형들의 결과값을 집계하여 예측 값을 정함
  Response 변수가 Categorical 변수일 때 : 투표 형식으로 각 모형에서 가장 많이 뽑힌 값이 예측 값이 됨
  Response 변수가 Numerical 변수일 때 : 평균 혹은 중위수를 계산함으로써 중심 경향을 나타내는 값이 예측 값
- 목표변수의 형태에 따라 분류분석 혹은 회귀분석에도 모두 사용 가능. 각각 분류 앙상블, 회귀 앙상블이라 함
- 학습에서 나타나는 오류는 bias로 인한 underfitting, 높은 분산으로 인한 overfitting인데 앙상블 모형은 여러 모형의 평균을 취함으로써 어느 쪽에도 치우치지 않는 결과를 얻을 수 있으며 여러 모형의 의견을 취합함으로써 분산을 감소시킬 수 있음
- 앙상블 학습의 유형에는 보팅, 배깅, 부스팅으로 구분
  보팅과 배깅은 여러 개의 분류기가 투표를 통해 최종 예측 결과를 결정하는 방식. 다른점은 보팅의 경우 서로 다른 알고리즘을 가진 분류기를 결합하는 것이고 배깅은 각각의 분류기가 모두 같은 유형의 알고리즘. 배깅은 부트스트랩과 보팅 과정을 거쳐 모델 선정. 부트스트랩은 주어진 자료에서 동일한 크기의 표본을 랜덤 복원추출로 뽑은 자료이며 다수의 샘플데이터를 생성. 보팅은 여러 개의 모형으로부터 산출된 결과를 다수결에 의해서 최종 결과를 선정하는 과정. 대표적이 배깅 방식이 랜덤 포레스트
  부스팅은 예측력이 약한 분류 모형을 결합하여 강한 예측 모형을 만드는 과정으로 가중치 반영 및 표분추출에 의한 분류기 생성 방식이 있음. 분류가 잘못된 데이터에 더 큰 가중치를 주어 표본을 추출하는 기법.
  그래디언트 부스트, XGBoost, LightGBM 등이 있음

고급분석기법 - 앙상블 분석

(**"무작위 숲"**이라는 이름처럼) **랜덤 포레스트는 훈련을 통해 구성해놓은 다수의 나무들로부터 분류 결과를 취합해서 결론을 얻는**, 일종의 **인기 투표**

학습 데이터 세트에 총 1000개의 행이 있다고 해보자. 그러면 임의로 100개씩 행을 선택해서 의사결정 트리를 만드는 게 **배깅(bagging)**이다. 학습 데이터의 일부를 기반으로 생성했다는 것이 중요. 이 때 **중복을 허용**

1000개의 행이 있는 가방(bag)에서 임의로 100개 뽑아 첫 번째 트리를 만들고 그 100개의 행은 가방에 도로 집어 넣는 방식



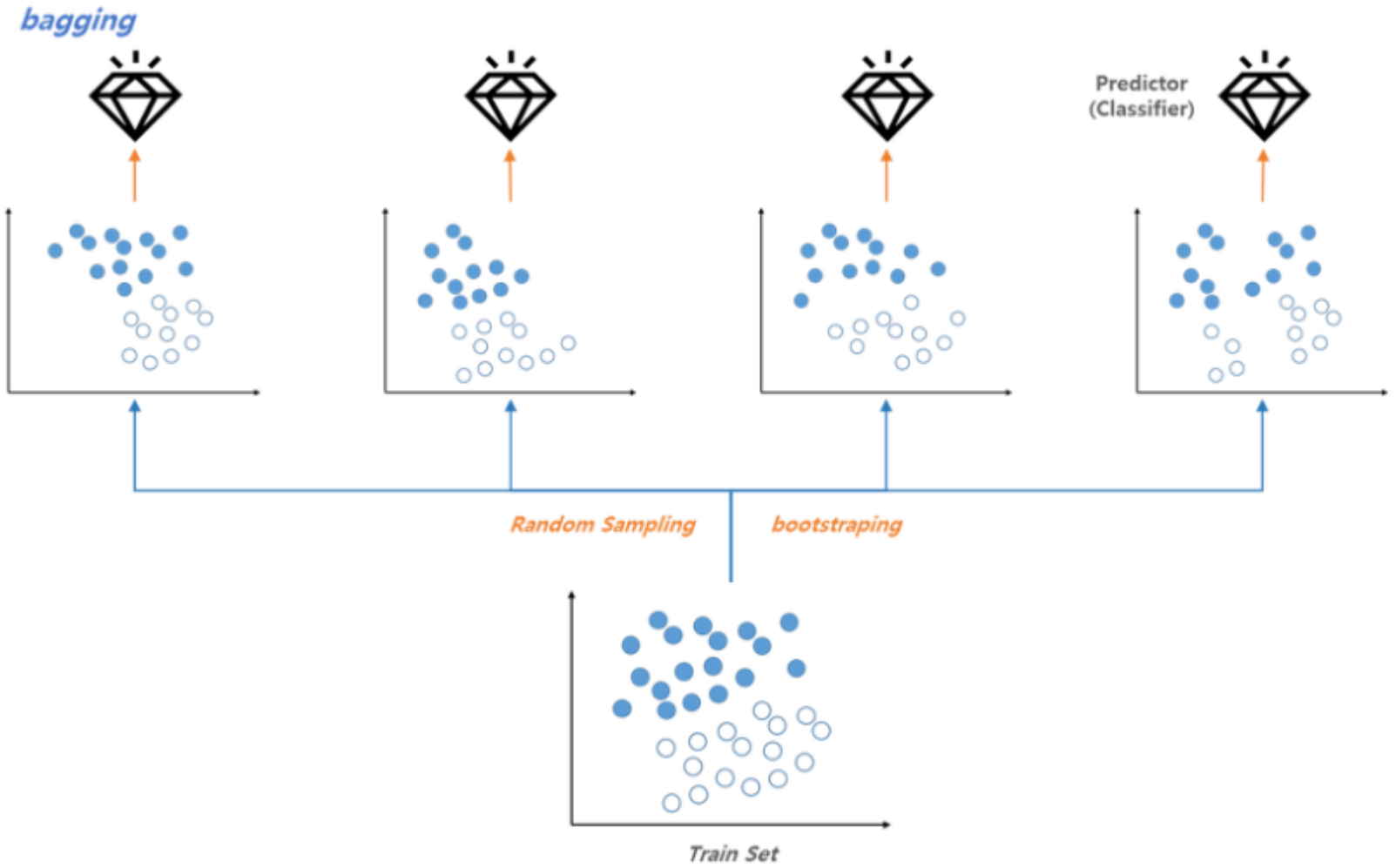
고급분석기법 - 앙상블 분석

하나의 알고리즘 사용
학습 데이터셋을 랜덤하게 추출하여 모델을 학습

투표방식으로 최빈값으로 결정

배깅 :
학습 데이터에서 랜덤하게 추출 시 중복 허용

몇개씩 속성을 뽑는 게 좋을까. 속성이 25개면 5개, 즉 **전체 속성 개수의 제곱근만큼 선택**하는 게 가장 좋다고, 경험적으로 인식



고급분석기법 - 비모수 통계

통계적 검정에서 모집단의 모수에 대한 검정은 모수적 검정과 비모수적 검정으로 구분됨

1. 모수적 방법
   - 검정하고자 하는 모집단의 분포에 대한 가정을 하고, 그 가정하에서 검정통계량과 검정통계량의 분포를 유도해 검정을 실시하는 방법

   (모수적 검정 및 모수적 검정 방법)
   - 가정된 분포의 모수에 대해 가설을 설정
   - 관측된 자료를 이용해 구한 표본평균, 표본분산 등을 이용해 검정을 실시

2. 비모수적 방법 : 순위(rank), 부호(sign) 이라는 단어 나오면 비모수 떠올리자
   - 자료가 추출된 모집단의 분포에 대한 아무 제약을 가하지 않고 검정을 실시하는 방법
   - 관측된 자료가 특정분포를 따른다고 가정할 수 없는 경우에 이용
   - 관측된 자료의 수가 많지 않거나(30개 미만) 자료가 개체간의 서열관계를 나타내는 경우에 이용

   (비모수적 검정 및 비모수적 검정 방법)
   - 가정된 분포가 없으므로 가설은 분포의 형태에 대해 설정
   - (예) '분포의 형태가 동일하다' 또는 '분포의 형태가 동일하지 않다'
   - 관측값의 절대적인 크기에 의존하지 않는 관측값들의 순위(rank)나 두 관측값 차이의 부호(sign) 등을 이용해 검정

   (비모수적 검정의 예)
   - 스피어만의 순위상관계수, 부호검정(sign test), 윌콕슨의 순위합검정(rank sum test), 윌콕슨의 부호순위합검정(Wilcoxon signed rank test), 만-위트니의 U 검정, 런검정(run test)

고급분석기법 - 비모수 통계