Data 분석 특강

117회 정보관리기술사 표기수

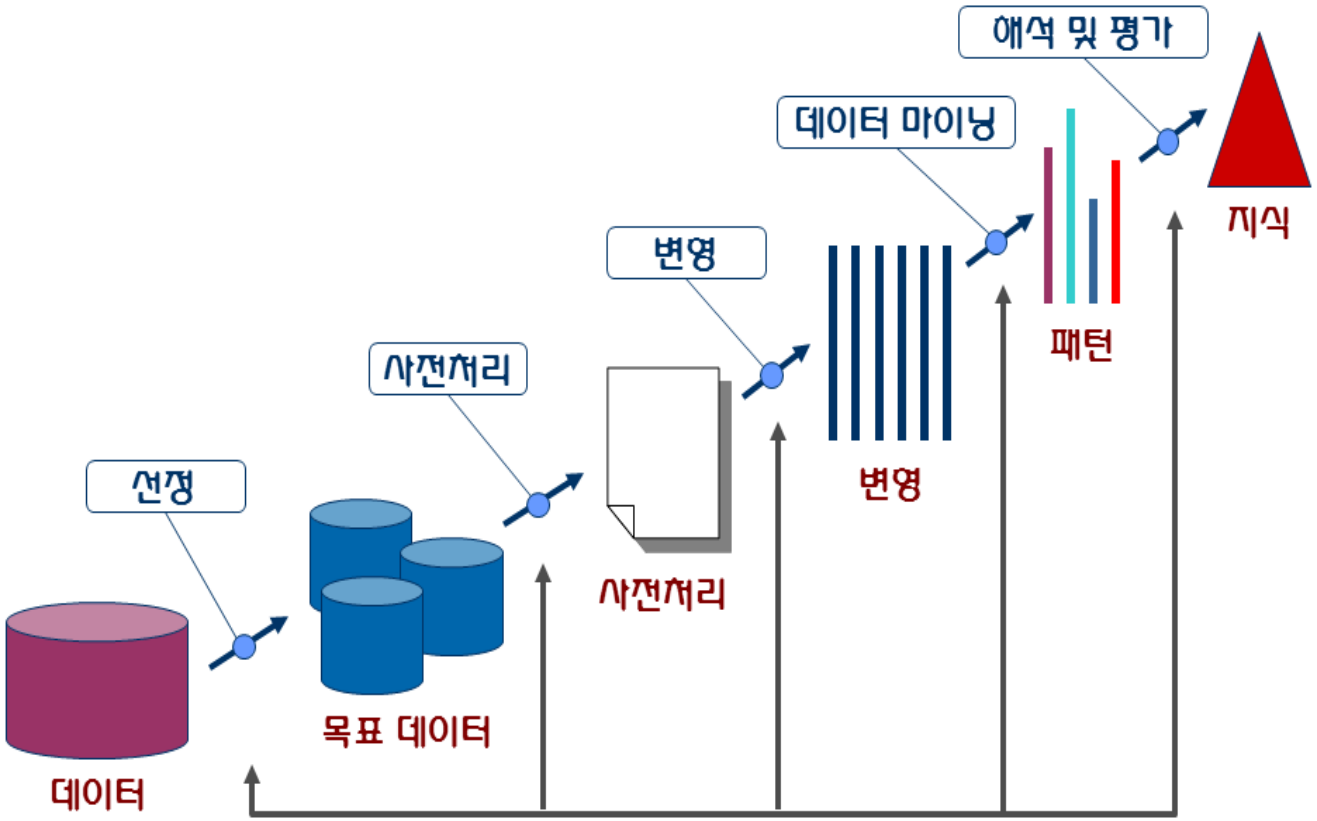
목차

데이터 분석의 이해

I. 데이터 분석의 개요

1. 데이터 분석 단계 (방법론)

✓ **KDD (Knowledge Discovery in Database)**



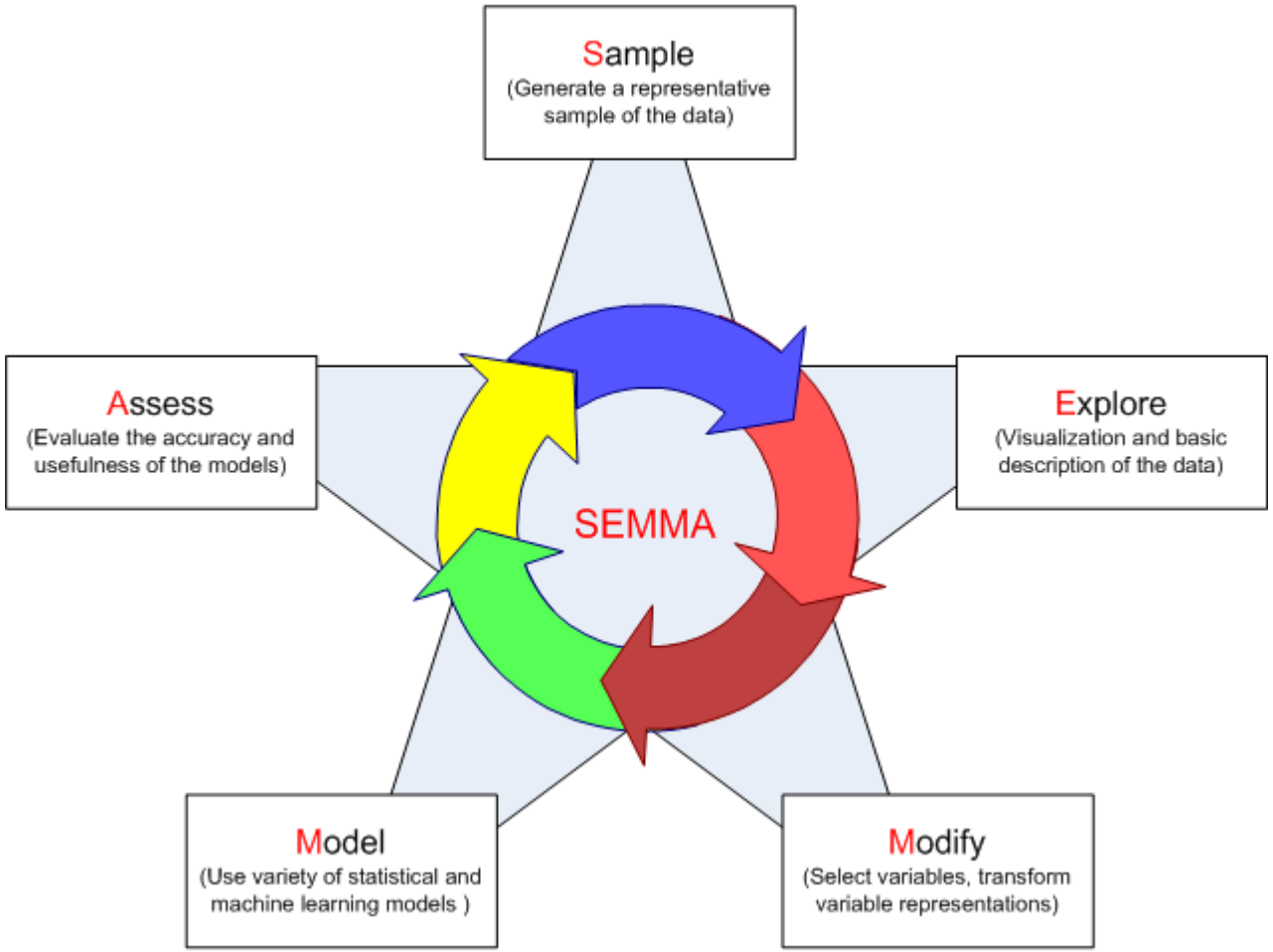
단계	상세
Sampling/Selecting (선택)	방대한 양의 데이터로부터 모집단의 유형과 닮은 작은 양의 데이터 추출
Data cleaning/ Preprocessing (정제, 보완)	확보 데이터의 정확성을 높이기 위해 모호성, 중복 제거, 오류 값 보정. 데이터 양/깊이 늘림
Transformation (변환)	불필요한 레코드, 항목삭제, 파생항목을 만들거나 항목의 값을 세분화 또는 그룹핑 하는 작업
Data Mining (Pattern)	이전 단계에서 선정된 주요한 변수를 사용하여 다양한 모형을 접합해 보는 단계. 데이터 마이닝 기술 적용하여 결과 해석
Interpretation / Evaluation	사용자들에게 보기 편하고 이해하기 쉬운 형태로 제공

데이터 분석의 이해

I. 데이터 분석의 개요

1. 데이터 분석 단계 (방법론)

✓ **SEMMA 방법론**



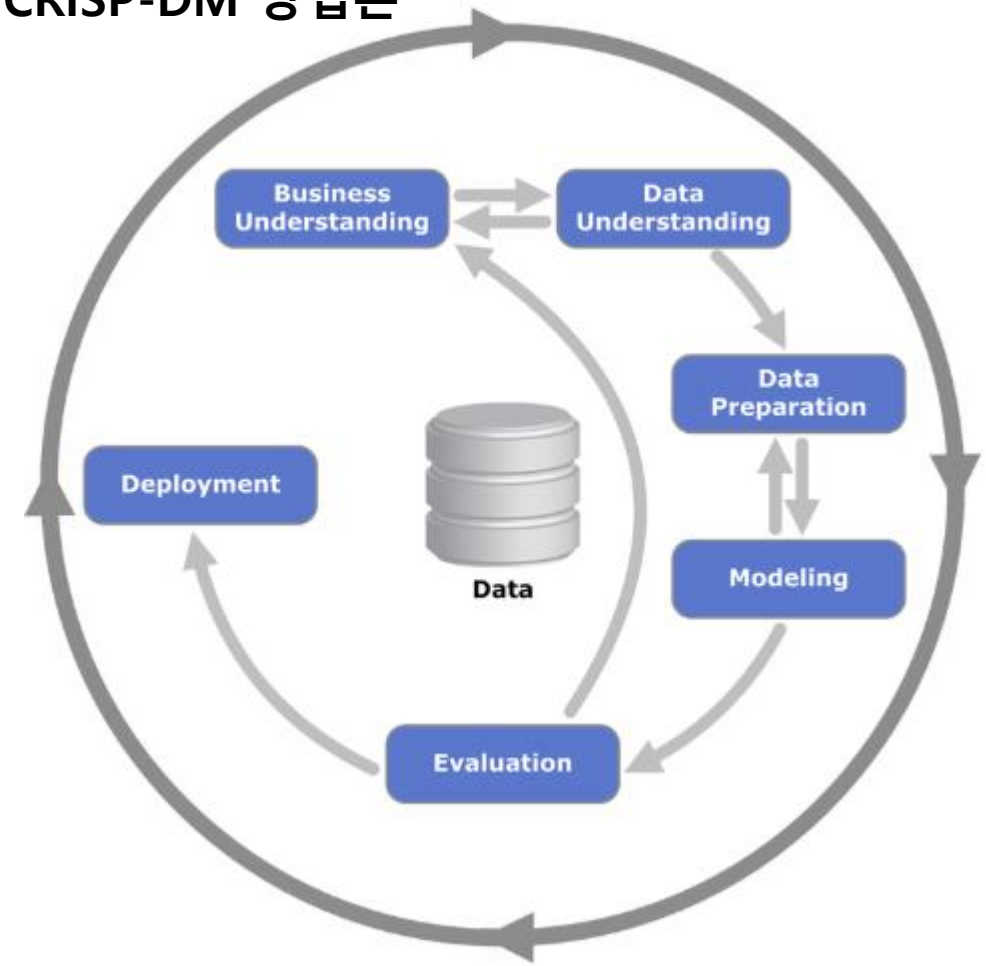
단계	상세
Sample	분석 데이터 생성 비용 절감 및 모델 평가를 위한 데이터 준비
Explore	분석 데이터 탐색 데이터 조감을 통한 데이터 오류 검색 모델의 효율 증대 데이터 현황을 통해 비즈니스 이해, 아이디어를 위해 이상현상, 변화 등을 탐색
Modify	분석 데이터 수정/변환 - 수량화, 표준화, 각종 변환, 그룹화 최적의 모델 구축 위해 변수를 생성, 선택, 변형
Model	모델 구축 데이터의 숨겨진 패턴 발견 하나의 비즈니스 문제 해결을 위해 특수의 모델과 알고리즘 적용 가능
Assess	모델 평가 및 검증 서로 다른 모델을 동시에 비교 추가 분석 수행 여부 결정

데이터 분석의 이해

I. 데이터 분석의 개요

1. 데이터 분석 단계 (방법론)

✓ **CRISP-DM 방법론**



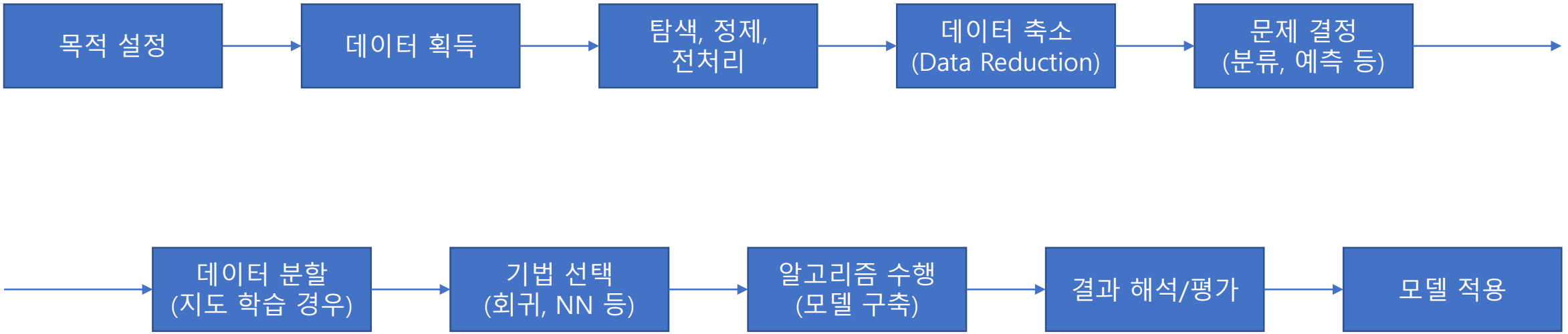
단계	상세
비즈니스 이해	- 비즈니스 목표 결정 - 상황 파악 - Data Mining 목표 설정 - 프로젝트 계획 설정
데이터 이해	- 최초 데이터 수집 - 데이터 기술 / 탐색 / 품질 검증
데이터 준비	- Select / Clean / Construct / Integrated data
모델링	- 모델링 기법 선택 - Test 설계 - 모델 구축
모델 평가	- 결과 평가 - 프로세스 검토 - 다음 단계 결정
고객 전달	- 모니터링 및 유지보수 - 최종 보고서 작성 - 프로젝트 검토

데이터 분석의 이해

I. 데이터 분석의 개요

1. 데이터 분석 단계 (방법론)

✓ **실무적 단계**



데이터 분석의 이해

I. 데이터 분석의 개요

✓ 데이터 획득 : Sampling

- 일부 데이터 샘플링 추출 후, 전체 데이터를 대상으로 "score"를 통해 모델 평가
- Oversampling : 빈도가 현저히 낮은 사건의 경우, 균형 있는 Training set을 통한 모델 구축을 위해 oversampling 수행
  oversampling 수행 후 결과 조정 필요
  예) 유방암 발병 비율

✓ 데이터 전처리 (Pre-processing)

- **Data Type 파악** : Numeric(수치) vs Categorical(범주)

→범주형 변수 처리 : n개의 범주가 있을 경우 n-1개의 변수를 사용하는 Dummy Coding 수행

- **Outlier(이상치) 탐지** : Outlier 탐지 후 도메인 지식을 활용하여 오류 값인지, 실제 극단적인 값인지를 판단
- **Missing Value(결측값) 처리** : Null값 혹은 Error값들에 대한 처리

→ 레코드를 삭제하거나 Mode, Mean, Median, User specific value 등으로 대체

데이터 분석의 이해

I. 데이터 분석의 개요

✓ 데이터 탐색 (Exploration)

- 기술적 통계 (Descriptive Statistics)

: 평균(Mean), 중앙값, 최빈값, 표준편차, 왜도, 첨도, 범위, 최대/최소값, 총합 등

- 데이터 시각화 (Data Visualization)

: 막대그래프(Bar Charts), 선그래프(Line Graphs), 산점도(Scatterplots), Boxplots, 히스토그램

데이터 분석의 이해

I. 데이터 분석의 개요

✓ 데이터 변형 (Transformation)

- **데이터 정규화(표준화) 및 리스케일링(Rescaling)** : 데이터 range 조정 $z = \frac{x-\mu}{\sigma}$, $\frac{x-\mu}{x_{max} - xmin}$

- 데이터 구분 및 통합 : binning 및 dummy coding을 통해서 데이터 구분 및 통합 수행

예) 봄/여름/가을/겨울을 나타내는 '계절' 변수가 있을 경우, 3개의 변수로 Dummy coding 가능

✓ 데이터 축소 (Reduction)

- **변수 축소** : 상관계수 분석, PCA 등을 활용하여 분석에 사용될 변수를 축소
- **레코드 수 축소** : Random sampling, Stratified sampling(계층적 샘플링), 전문가 지식 기반의 목적성을 가진 샘플링

데이터 분석의 이해

I. 데이터 분석의 개요

✓ 데이터 축소 (Data Reduction)

- **Data Value에 대한 축소**

① Category 축소 : 분류되는 카테고리 개수를 축소시켜서 실제 Data value의 개수를 축소

② Binning : 데이터 Binning 재수행을 통해 그룹핑 되는 데이터 개수를 조정

- **Data Variable에 대한 축소**

① Feature Selection : 상관계수 기반 검증, 통계 검증(Chi-squared, T-test, F-test)

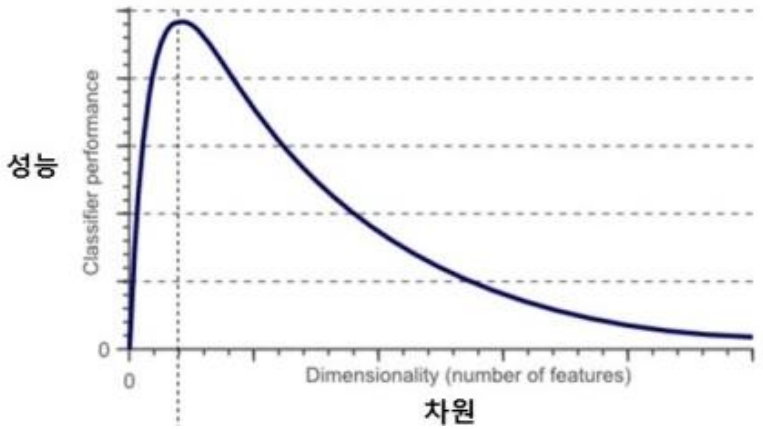
② Feature Extraction : PCA (Principal Components Analysis)

[Feature Selection]

[Feature Extraction]

[참고] 차원의 저주

- 변수를 추가하여 차원이 증가함으로써 복잡성이 기하급수적으로 증가하는 현상
- 체스판의 2차원(8x8)에서 3차원으로, 차원을 50% 증가시키면
  선택지는 512(8x8x8)로 800% 증가함



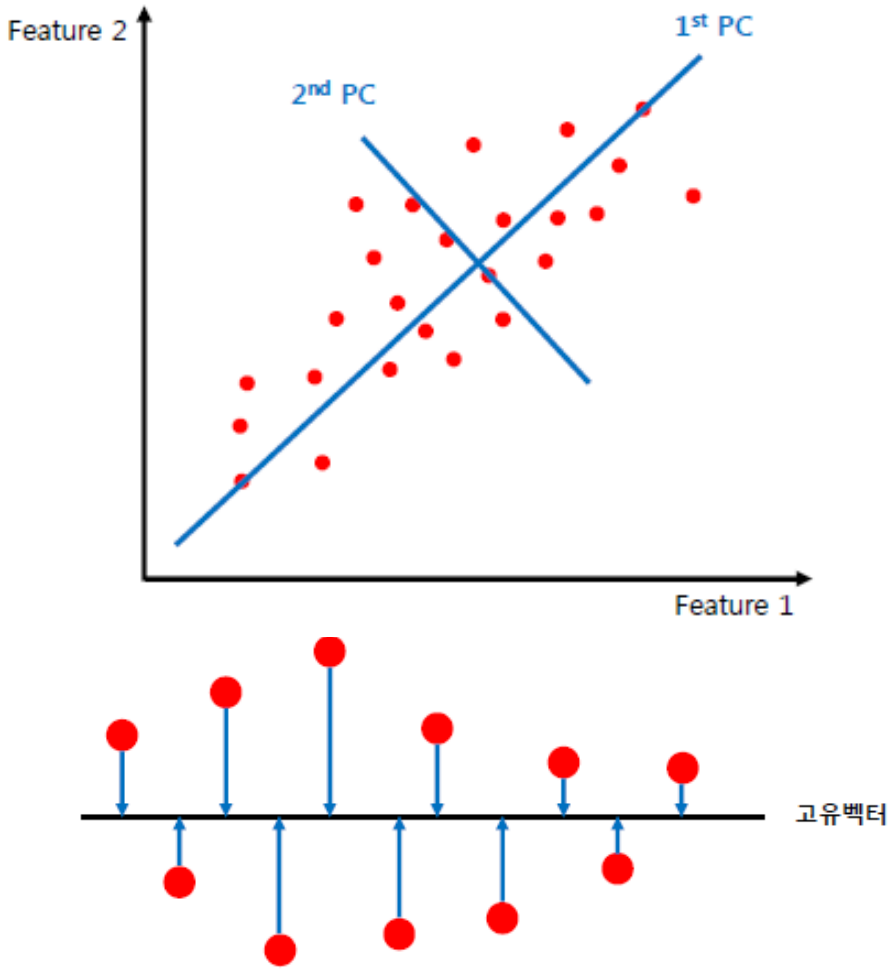
데이터 분석의 이해

I. 데이터 분석의 개요

✓ 데이터 축소 (Data Reduction) - PCA

- PCA 수행 절차

프로세스	설명
1. 공분산 계산	① 입력데이터 X의 평균 μ_x와 공분산 Σ_x를 계산한다. $\mu_x = \frac{1}{N}\sum_{i=1}^{N} x_i \qquad \Sigma_x = \frac{1}{N}(X - M_x)(X - M_x)^T \qquad M_x = \mu_x 1^T$ (1은 모든 원소의 값이 1인 n차원 열벡터)
2. 고유벡터 계산	② 고유치 분석을 통해 공분산 Σ_x의 고유치행렬 Λ과 고유벡터행렬 U을 계산 $\Sigma_x = U\Lambda U^T = [u_1, u_2, \ldots, u_n] \begin{bmatrix} \lambda_1 & 0 & \ldots & 0 \\ 0 & \lambda_2 & \ldots & 0 \\ \vdots & \vdots & \ldots & \vdots \\ 0 & 0 & \ldots & \lambda_n \end{bmatrix} [u_1, u_2, \ldots, u_n]^T$
3. 고유치 선택	③ 고유치 값이 큰 것부터 순서대로 m개의 고유치 $\{\lambda_1, \lambda_2, \ldots, \lambda_m\}$를 선택
4. 변환행렬 생성	④ 선택한 고유치에 대응되는 고유벡터를 열벡터로 가지는 변환행렬 W 생성 $W = [u_1, u_2, \ldots, u_m]$
5. 선형변환	⑤ W에 의한 선형변환에 의해 특징데이터 Y를 얻는다. $Y = W^T X$



- 고유벡터 : 데이터의 분포를 나타내는 선
- 고유값 : 고유벡터에 대한 데이터의 분산

데이터 분석의 이해

I. 데이터 분석의 개요

✓ 데이터 축소 (Data Reduction) - PCA

- PCA 예시 : 시리얼의 각 구성 성분이 체중에 미치는 영향

Variable	1	2	3	4	5	6
calories	0.07624155	-0.01066097	0.61074823	-0.61706442	0.45754826	0.12601775
protein	-0.00146212	0.00873588	0.00050506	0.0019389	0.05533375	0.10379469
fat	-0.00013779	0.00271266	0.01596125	-0.02595884	-0.01839438	-0.12500292
sodium	0.98165619	0.12513085	-0.14073193	-0.00293341	0.01588042	0.02245871
fiber	-0.00479783	0.03077993	-0.01684542	0.02145976	0.00872434	0.271184
carbo	0.01486445	-0.01731863	0.01272501	0.02175146	0.35580006	-0.56089228
sugars	0.00398314	-0.00013545	0.09870714	-0.11555841	-0.29906386	0.62323487
potass	-0.119053	0.98861349	0.03619435	-0.042696	-0.04644227	-0.05091622
vitamins	0.10149482	0.01598651	0.7074821	0.69835609	-0.02556211	0.01341988
shelf	-0.00093911	0.00443601	0.01267395	0.00574066	-0.00823057	-0.05412053
weight	0.0005016	0.00098829	0.00369807	-0.0026621	0.00318591	0.00817035
cups	0.00047302	-0.00160279	0.00060208	0.00095916	0.00280366	-0.01087413
rating	-0.07615706	0.07254035	-0.30776858	0.33866307	0.75365263	0.41805118
Variance%	55.52834702	37.25226212	3.84177661	2.75336623	0.55865192	0.0334504
Cum%	55.52834702	92.78060913	96.62238312	99.37575531	99.93440247	99.96785736

데이터 분석의 이해

I. 데이터 분석의 개요

✓ 모델 성능 평가

- 분류 또는 예측 모델의 유용성을 판단하고 상이한 모델들을 서로 비교 평가하는 활동

예측모델 성능평가 : MAPE, RMSE, 향상차트(Lift Chart)

분류모델 성능평가 : 정오행렬(분류행렬), 정확도(Accuracy), 오분류율(Error Rate), 민감도(Sensitivity), 특이도(Specificity), 향상차트

데이터 분석의 이해

I. 데이터 분석의 개요

✓ 모델 성능 평가

1) 예측(Prediction) 모델 성능평가

평가 척도	산술식	설명
절대평균오차/편차 (MAE : Mean Absolute Error/deviation)	MAE = $\frac{1}{n}\sum_{i=1}^{n}\lvert e_i \rvert$	평균 절대오차의 크기
평균백분율 오차 (MPE : Mean Percentage Error)	MPE = 100 x $\frac{1}{n}\sum_{i=1}^{n}\frac{e_i}{y_i}$	오차의 방향 고려하여 예측이 실제값에서 얼마나 벗어나는지에 대한 퍼센트
절대평균백분율 오차 (MAPE : Mean Absolute Percentage Error)	MAPE = 100 x $\frac{1}{n}\sum_{i=1}^{n}\left\lvert\frac{e_i}{y_i}\right\rvert$	예측이 실제 값에서 평균적으로 벗어나는 정도를 백분율 점수로 표현
평균제곱오차의 제곱근 (RMSE : Root Mean Squared Error)	RMSE = $\sqrt{\frac{1}{n}\sum_{i=1}^{n}e_i^2}$	오차들의 제곱의 평균에 제곱근을 계산하여 성능 평가

- 이상치나 잡음에 대해 민감하게 반응하므로, 이에 대한 처리가 필요

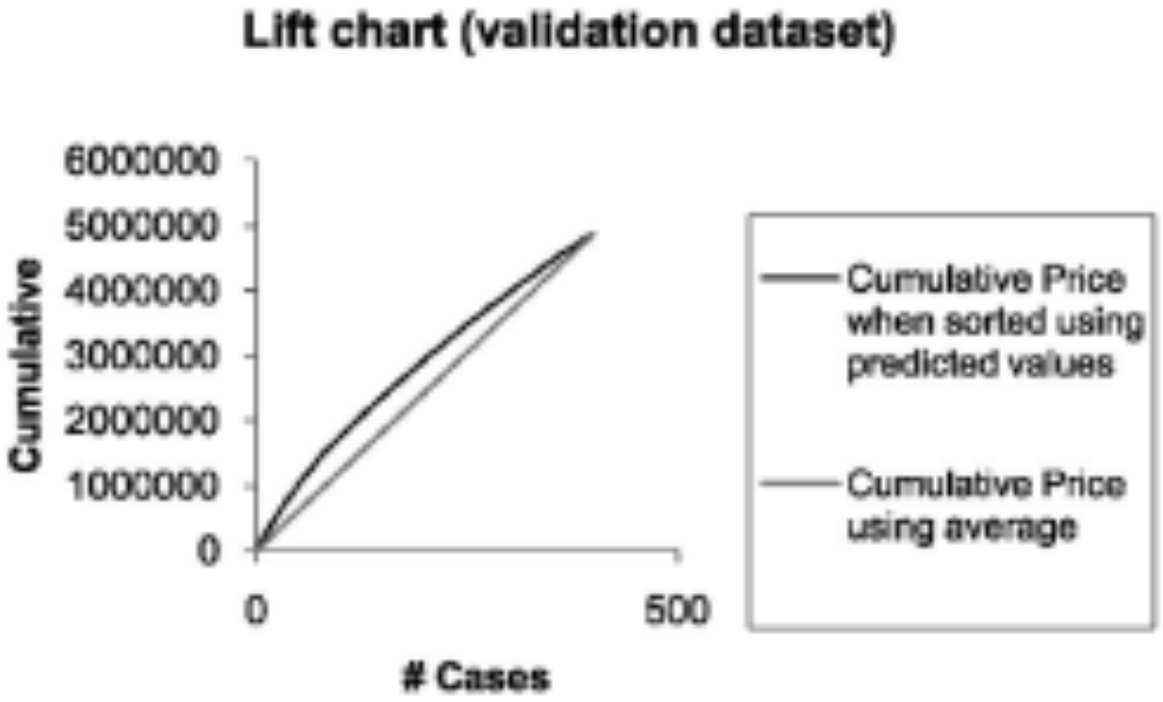
데이터 분석의 이해

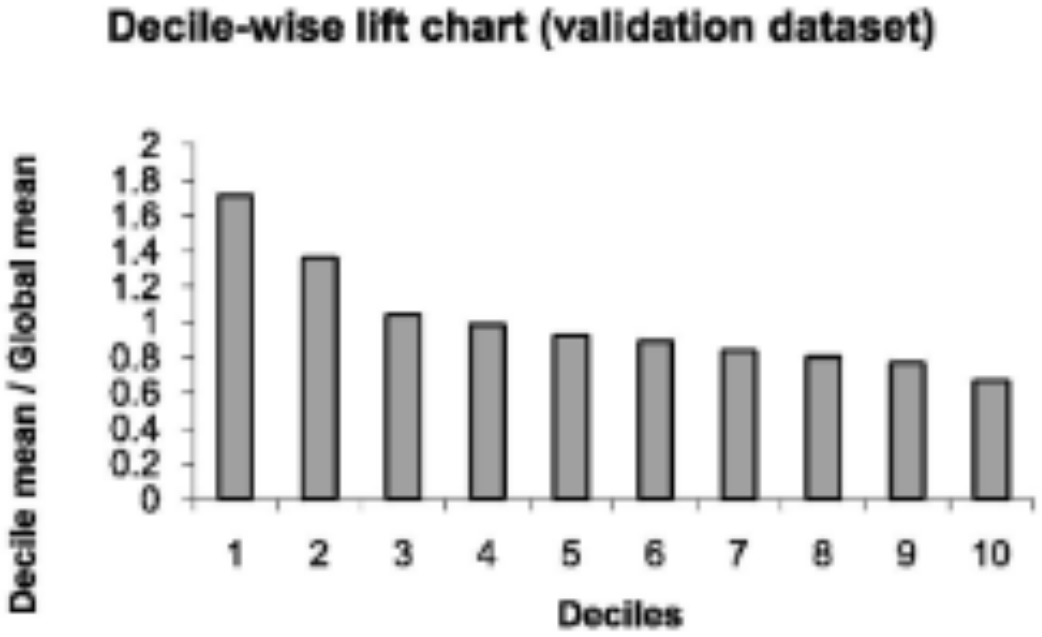
I. 데이터 분석의 개요

✓ 모델 성능 평가

1) 예측(Prediction) 모델 성능평가

- 향상차트(Lift Chart)





데이터 분석의 이해

I. 데이터 분석의 개요

✓ 모델 성능 평가

2) 분류(Classification) 모델 성능평가

: 정오행렬(분류행렬)을 통한 성능 평가

구분		실제 클래스	
		C_1	C_2
예측 클래스	C_1	True 올바르게 분류된 경우 ($n_{1,1}$)	False C2를 C1으로 잘못 분류된 경우 ($n_{2,1}$)
	C_2	False C1을 C2로 잘못 분류된 경우 ($n_{1,2}$)	True 올바르게 분류된 경우 ($n_{2,2}$)

- 오차율(Error Rate) = $\frac{n_{2,1}+n_{1,2}}{n}$
- 정확도(Accuracy) = 1 – Error Rate = $\frac{n_{1,1}+n_{2,2}}{n}$
- 민감도(Sensitivity/Recall) = $\frac{n_{1,1}}{n_{1,1}+n_{1,2}}$: 중요한 클래스(C_1)을 감지
- 특이도(Specificity) = $\frac{n_{2,2}}{n_{2,1}+n_{2,2}}$: C_2를 올바르게 제거

데이터 분석의 이해

I. 데이터 분석의 개요

✓ 모델 성능 평가

2) 분류(Classification) 모델 성능평가

: 정오행렬(분류행렬)을 통한 성능 평가

구분		실제 클래스	
		C_1	C_2
예측 클래스	C_1	$n_{1,1}$ = 2,689	$n_{2,1}$ = 85
	C_2	$n_{1,2}$ = 25	$n_{2,2}$ = 201

- 오차율(Error Rate) = $\frac{85+25}{3,000}$ = 3.67%
- 민감도(Sensitivity/Recall) = $\frac{2,689}{2,689+25}$ = 99.1 %
- 정확도(Accuracy) = 1 – 3.67% = $\frac{2,689+201}{3,000}$ = 96.33%
- 특이도(Specificity) = $\frac{201}{85+201}$ = 70.3%

데이터 분석의 이해

I. 데이터 분석의 개요

✓ 모델 성능 평가

2) 분류(Classification) 모델 성능평가

: 결과가 경향 값일 때 컷오프값 설정 필요

Actual Class	"Owner" 확률	Actual Class	"Owner" 확률
Owner	0.99	Owner	0.50
Owner	0.98	Nonowner	0.47
Owner	0.97	Nonowner	0.33
Owner	0.96	Owner	0.21
Owner	0.95	Nonowner	0.19
Owner	0.86	Nonowner	0.18
Owner	0.85	Nonowner	0.17
Nonowner	0.76	Nonowner	0.15
Owner	0.72	Nonowner	0.14
Owner	0.69	Nonowner	0.06
Owner	0.68	Nonowner	0.04
Nonowner	0.62	Nonowner	0.02

```
## cutoff = 0.5
> confusionMatrix(ifelse(owner.df$Probability>0.5,
# note: "reference" = "actual"
Confusion Matrix and Statistics

          Reference
Prediction nonowner owner
  nonowner        10     1
  owner            2    11

               Accuracy : 0.875
```

```
## cutoff = 0.25
> confusionMatrix(ifelse(owner.df$Probability>0.25,
Confusion Matrix and Statistics

          Reference
Prediction nonowner owner
  nonowner         8     1
  owner            4    11

               Accuracy : 0.791667
```

데이터 분석의 이해

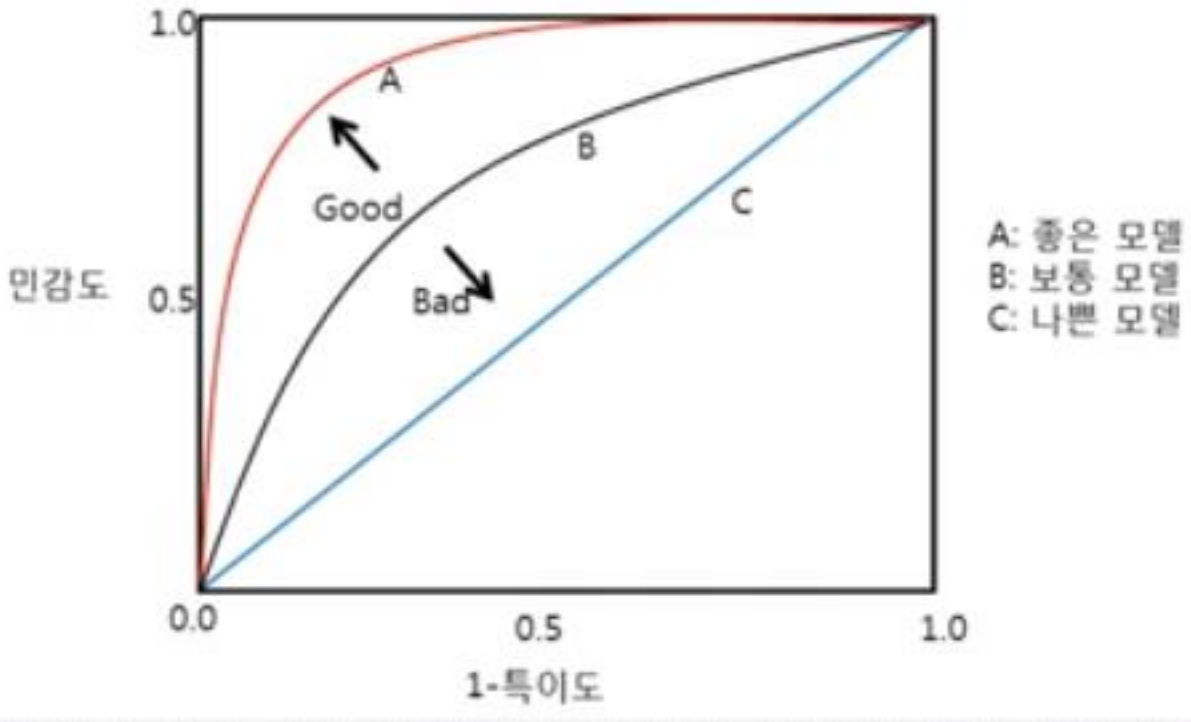
I. 데이터 분석의 개요

✓ 모델 성능 평가

2) 분류(Classification) 모델 성능평가

: 민감도와 특이도를 반영한 ROC(Receiver Operating Characteristic)로 평가 가능

→ AUC (Area Under the Curve) : 곡선 아래 영역, AUC가 1에 가까워질수록 좋은 성능을 나타냄



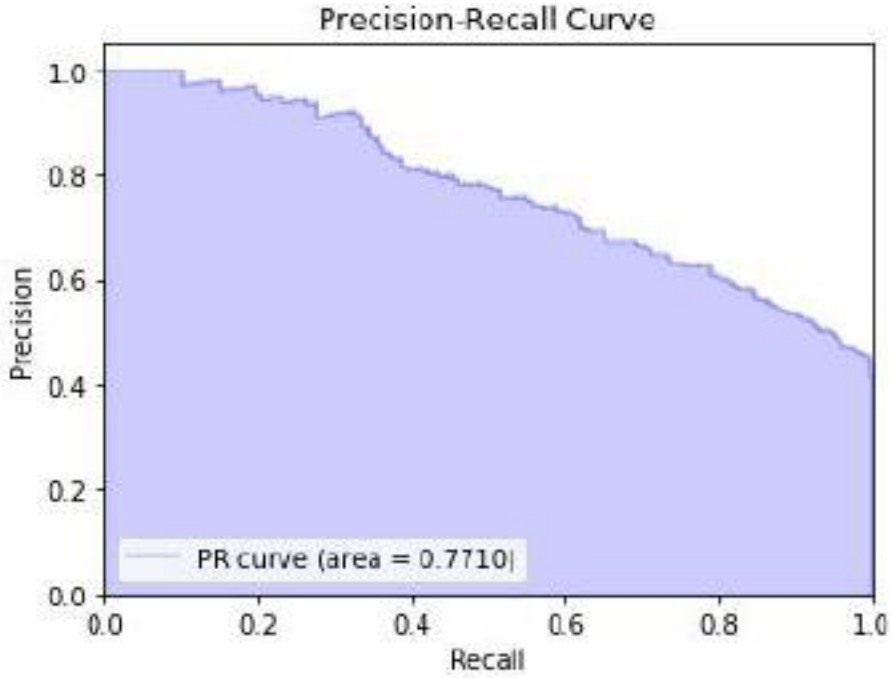
데이터 분석의 이해

I. 데이터 분석의 개요

✓ 모델 성능 평가

2) 분류(Classification) 모델 성능평가

		실제 정답 (Gold Standard에 의해 결정)	
		True	False
실험 결과	Positive	True Positive	False Positive (en:Type I error)
	Negative	False Negative (en:Type II error)	True Negative



정밀도 (Precision)	$\frac{T_p}{T_p + F_p}$	- Positive라 예측한 사례 중 TruePositive 비율
재현율 (Recall)	$\frac{T_p}{T_p + T_n}$	- True라 예측한 사례 중 TruePositive 비율
F1 Score	$2 * \frac{Precision * Recall}{Precision + Recall}$	- Precision과 Recall의 트레이드오프를 잘 통합하여 정확성을 한번에 나타내는 지표

예측 및 분류 방법

I. 회귀분석

1. 다중 선형 회귀분석

✓ 회귀분석의 개념

- 독립변수들과 종속변수 간에 존재하는 관련성을 분석하기 위하여, 관측된 자료에서 이들 간의 함수적 관계를 통계적으로 추정하는 방법
- 인과관계를 검정하는 분석방법
- 하나 또는 여러 개의 독립변수가 종속변수에 영향을 미칠 때 변수들 간의 관계를 설명하고 예측하는 분석 방법

구분	유형	설명
독립변수 수	단순회귀분석	- 하나의 종속변수와 하나의 독립변수 사이의 관계
	다중회귀분석	- 하나의 종속변수와 다수의 독립변수 사이의 관계
예측/분류 모델	선형회귀분석	- 종속변수 y와 하나 이상의 독립변수 x와의 선형 상관 관계
	로지스틱 회귀분석	- 독립 변수의 선형 결합을 이용하여 사건의 발생 가능성을 예측

✓ 회귀분석 (Regression) 의 목적

① 예측 : 독립변수의 변화에 따라 종속변수의 값이 어떻게 변할지를 예측
예) 콜레스테롤이 높으면 중성지방도 높음

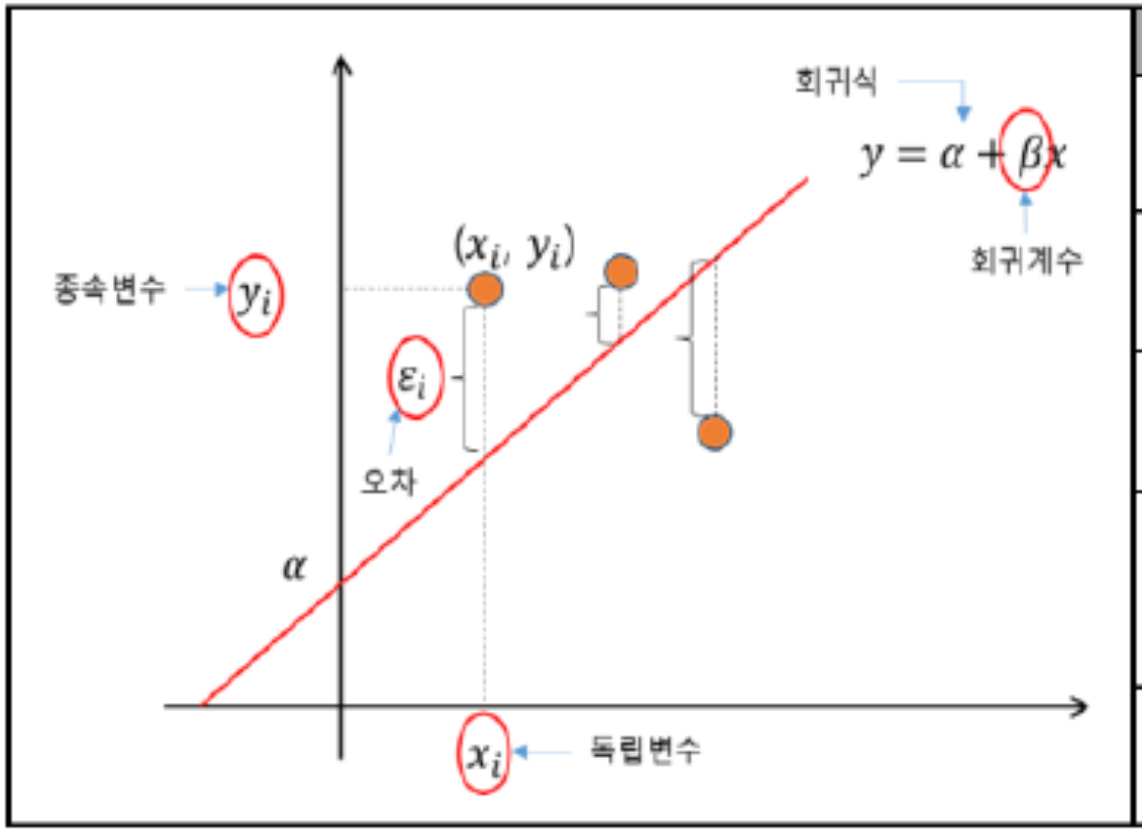
② 설명 : 두 변수 사이의 영향 관계를 설명
예) 온라인 게임의 몰입(즐거움)에 영향을 주는 요인

예측 및 분류 방법

I. 회귀분석

1. 다중 선형 회귀분석

✓ 다중 선형 회귀분석의 구성요소



요소	설명
독립변수	-입력값이나 원인이 되는 변수
종속변수	-독립변수에 의해 결과물이나 효과
회귀계수	-단위 시간에 따라 변하는 양 (기울기)
최소자승법	-산포도에 위치한 각 점에서 회귀선에 수직으로 이르는 값의 제곱의 합이 최소가 되는 선을 최적의 회귀선으로 추정
회귀방정식	-회귀선의 수학적 함수로 표현한 식

예측 및 분류 방법

I. 회귀분석

1. 다중 선형 회귀분석

✓ 다중 선형 회귀분석의 절차

순서	절차	설명
(1)	산점도 작성	-변수간의 관계가 직선관계인지 곡선관계인지를 파악 -상관계수를 통해 양의 상관관계인지 음의 상관관계인지를 확인
(2)	단순선형회귀모형 선정	-독립변수 X와 종속변수 Y의 관계가 1차 직선의 상관관계인경우 단순 선형 회귀모형을 선정한다. ($y_i = \alpha + \beta_{x_i} + \varepsilon_i$ $\varepsilon_i \sim N(0, \sigma^2)$, i = 1,2, ..., n) -단순선형회구 모형은 선형성, 정규성, 등분산성, 독립성을 가정함
(3)	회귀식의 추정	-오차항 ε_i 의 평균을 0으로 가정하고, 추정된 회귀식은 $\hat{y}_i = \alpha + \beta x_i$ -최소자승법을 통해 오차의 제곱합으로 편차를 최소화하는 α, β 값을 추정 $-\beta = \frac{\sum(x-\bar{x})(y-\bar{y})}{\sum(x-\bar{x})^2}$ $-\alpha = \bar{y} - \beta x_i$
(4)	회귀식의 정도 측정	-추정된 회귀식이 정말 원래의 관측값들을 잘 대표하는지 점검 -결정계수 $r^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$ (SST : 총제곱합, SSE : 잔차제곱합, SSR : 회귀제곱합) 값이 1에 가까울수록 추정된 회귀식 주위에 자료가 밀집되어 있으므로 자료를 잘 대표하고 있다고 판단
(5)	회귀식을 통한 모집단의 추정	-도출된 회귀식을 통해 모집단에서 추정하고자 하는 종속변수를 도출하여 추정함

[참고] 최소제곱법

- 적합된 회기식에 의한 예측치와 관찰치의 차이인 잔차들의 제곱의 합이 최소가 되도록 회귀계수를 추정하는 방법(미분)

$$\sum_{i=1}^{n}(e_i)^2 = \sum_{i=1}^{n}(y_i - b_0 + b_1 x_i)^2$$

$$\frac{\partial D}{\partial b_0} = -2\sum_{i=1}^{n}(y_i - b_0 + b_1 x_i) = 0$$

$$\frac{\partial D}{\partial b_1} = -2\sum_{i=1}^{n} x_i(y_i - b_0 + b_1 x_i) = 0$$

예측 및 분류 방법

I. 회귀분석

1. 다중 선형 회귀분석

✓ 선형 회귀분석의 가정

구분	가정	설명
모형에 대한 가정	선형성	- 예측하고자 하는 종속변수 y 와 독립변수 x 간에 선형성을 만족한다고 가정
오차에 대한 가정	독립성	- 독립변수 x 간에 상관관계가 없음을 가정
	정규성	- 잔차는 정규분포를 따른다고 가정
	등분산성	- 잔차가 특정한 패턴 없이 고르게 분포한다고 가정

✓ 다중 공선성 문제 (Multicollinearity)

독립성 위배 → 다중공선성 발생

- 독립변수간 종속성 발생
- R^2값은 높지만 p-value 값이 커서 유의하지 않는 경우 발생

해결방안

- 상관관계 높은 변수 일부 제거
- 변수 변형 및 상관관계 이유 해결
- PCA(Principle Component Analysis) 이용

예측 및 분류 방법

I. 회귀분석

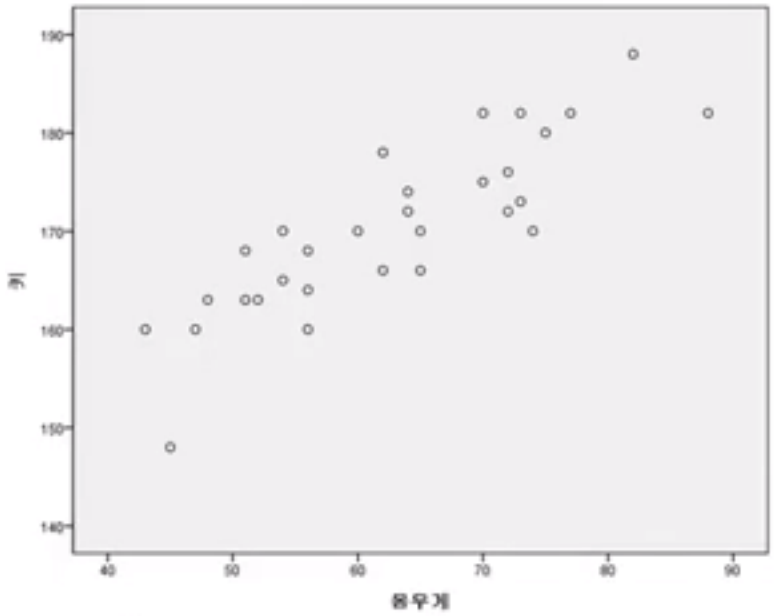
2. 로지스틱 회귀분석

✓ **개념**

- 분석 대상들이 여러 집단으로 나누어진 경우, 독립 변수의 선형 결합을 이용하여 개별 관측치가 어느 집단에 속하는지 확률을 계산하는 분류 기법
- 회귀분석에서 종속변수가 범주형인 자료로 확장

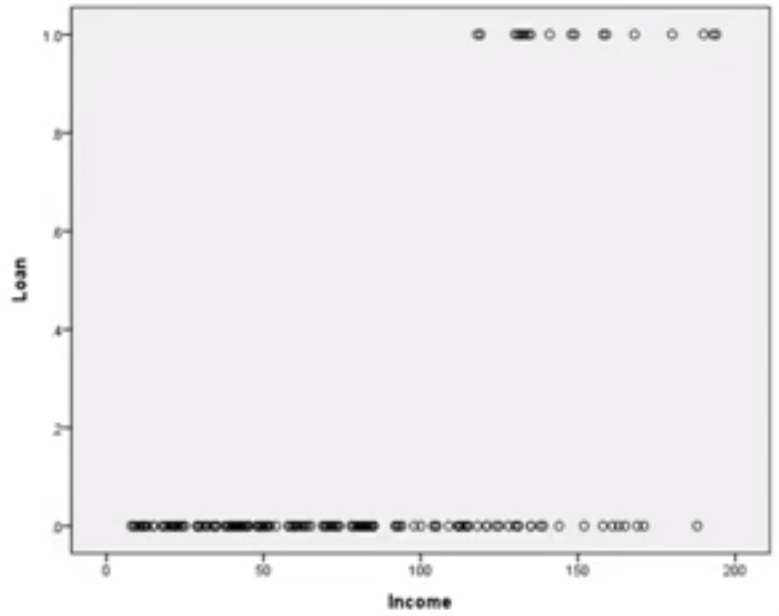
$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \varepsilon_i$$

수치형 변수 + 수치형 변수



$$logit = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \varepsilon_i$$

수치형 변수 + 범주형 변수

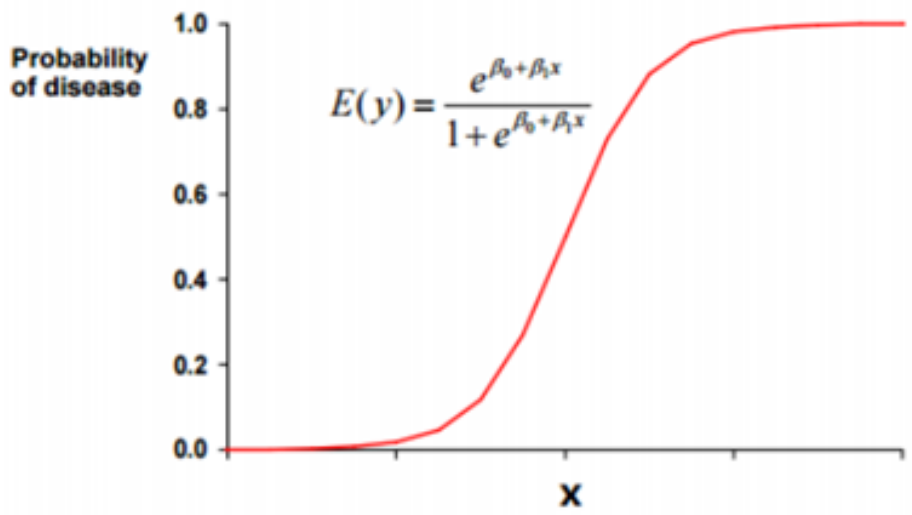


예측 및 분류 방법

I. 회귀분석

2. 로지스틱 회귀분석

✓ 개념도



- 로지스틱 회귀분석에서는, 종속변수가 발생할 확률 p에 대한 승산비(Odds)에 로그를 취한 Logit을 종속변수로 사용

- 승산비(Odds) = $\frac{P}{1-p} = e^{ax+b}$

$$\text{logit} = \log\frac{P}{1-p} = ax + b$$

$$p = \frac{1}{1 + e^{-(ax+b)}}$$

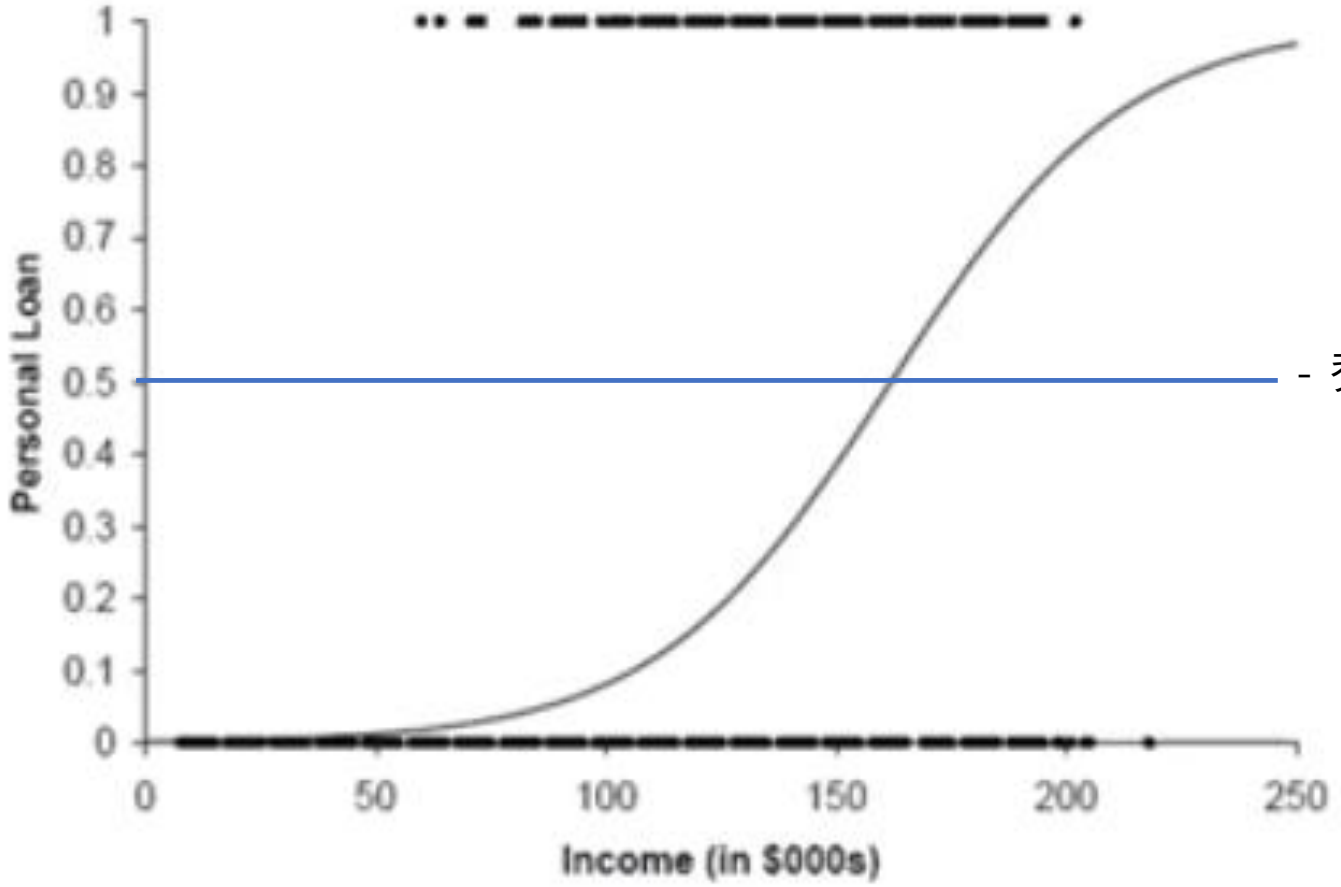
개념	설명	관련 식
승산비	Odds rate (OR) 어떤 사건이 일어날 확률과 일어나지 않을 확률의 비	$\text{odds} = \frac{p(y=1\|x)}{1-p(y=1\|x)}$
Log	log 는 (-∞, ∞) 의 값 가능하여 회귀 모형 성립이 가능	Log (odds) = Log(p/1-p)
Logit	Log 연산을 통한 Logit 획득 (0~1)	$\text{logit}(p) = \log\frac{p}{1-p}$

예측 및 분류 방법

I. 회귀분석

2. 로지스틱 회귀분석

✓ 개념도



- 컷오프 기준으로 p=0 또는 1로 분류

예측 및 분류 방법

II. KNN (K-Nearest Neighbors, k-최근접이웃 알고리즘)

✓ 개념

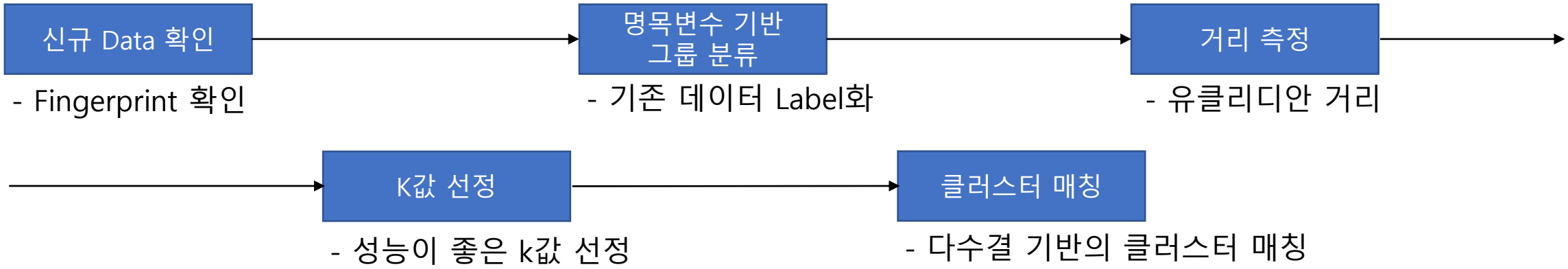
- 새로운 fingerprint를 기존 클러스터 내의 모든 데이터와 거리를 측정하여 가장 많은 속성을 가진 클러스터에 할당하는 분류 알고리즘

✓ 특징

- Data-driven : 별도의 모델 구축(Model-driven)을 통한 분류가 아닌, 데이터 자체에 대한 분류 기법
- 유클리디안 거리 (Euclidean distance)

$$\sqrt{(x_1 - u_1)^2 + (x_2 - u_2)^2 + \cdots + (x_p - u_p)^2}$$

✓ 동작 원리



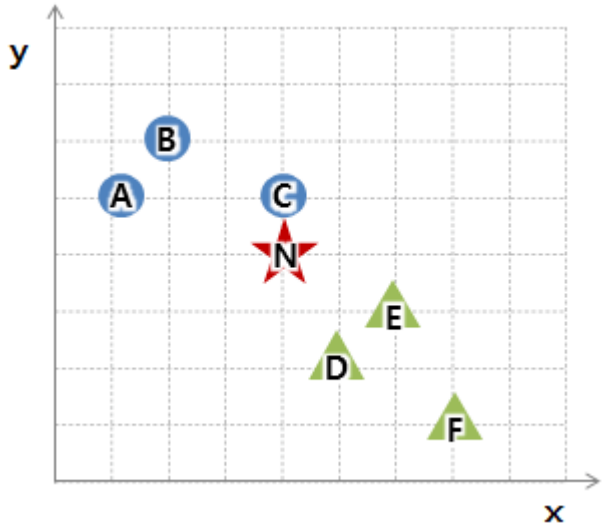
- Fingerprint 확인
- 기존 데이터 Label화
- 유클리디안 거리
- 성능이 좋은 k값 선정
- 다수결 기반의 클러스터 매칭

예측 및 분류 방법

II. KNN (K-Nearest Neighbors, k-최근접이웃 알고리즘)

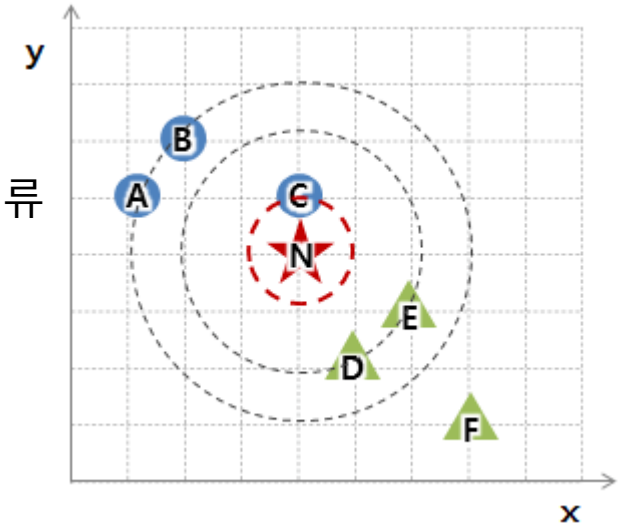
✓ 개념도

데이터	x좌표	y좌표	그룹
A	1	5	●
B	2	6	●
C	4	5	●
D	5	2	▲
E	6	3	▲
F	7	1	▲
N	4	4	?



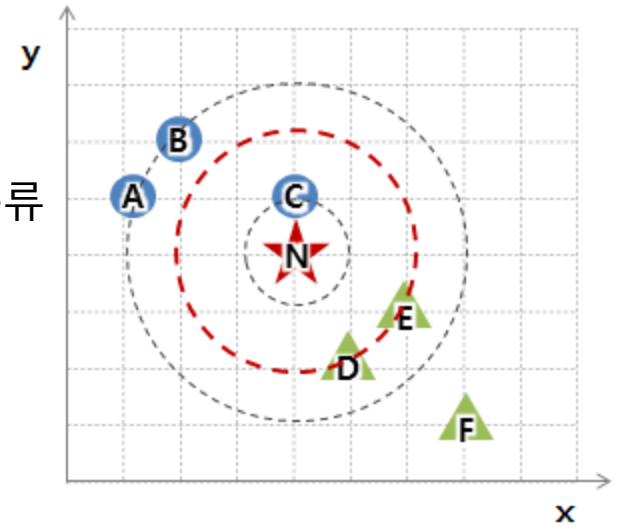
K = 1

→ ●로 분류



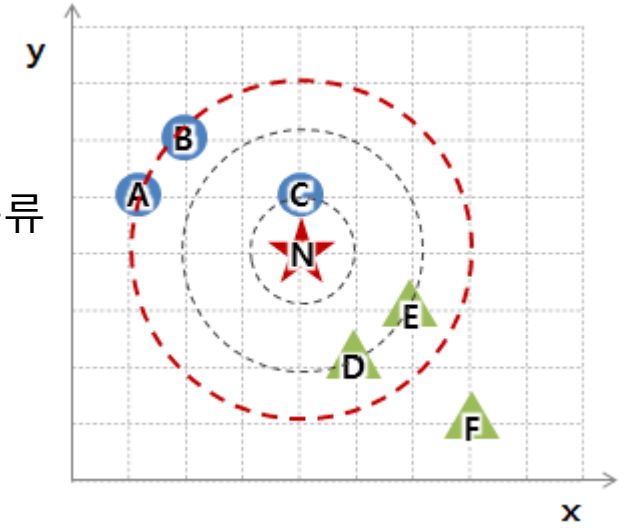
K = 3

→ ▲로 분류



K = 5

→ ●로 분류



예측 및 분류 방법

II. KNN (K-Nearest Neighbors, k-최근접이웃 알고리즘)

✓ K값 선정

- K는 새로운 데이터를 분류하기 위해 사용되는 최근접 데이터의 개수를 의미
- Validation set에서의 Error율을 기반으로 최적화된 k값을 선정

✓ KNN의 장점

- Simple : 유클리디안 거리를 계산하여 다수결에 의한 결정이므로, 설명하기 쉽고 로직이 단순함
- 가정 불필요 : Data-driven 기반이므로 별도의 통계적 가정이 불필요함 (예: 데이터가 정규분포를 따른다고 가정)

✓ KNN의 단점

- 고비용 : 모든 데이터에 대한 거리를 계산해야 하므로, 시간과 비용 소모 많음
- 매번 수행시마다 비용 소모

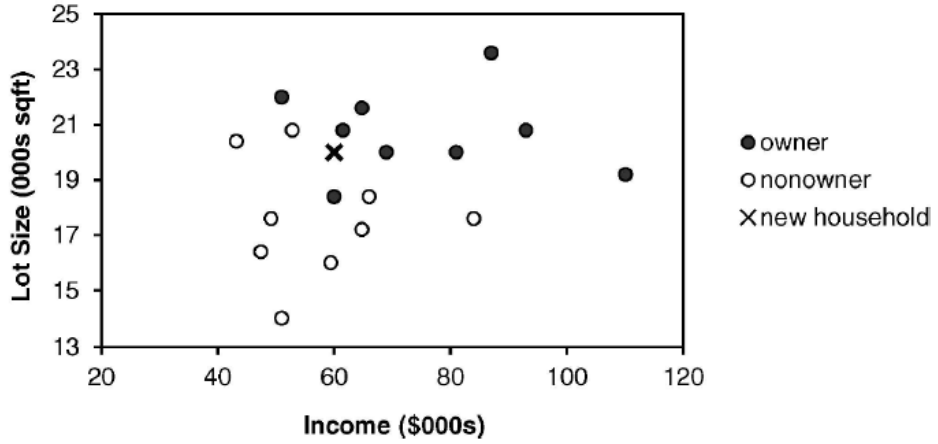
: 모델 구축을 통한 경우에는, 모델 구축시에 비용이 많이 소모되지만 그 후엔 산정된 모델/산술식에 따라 빠르게 계산
KNN의 경우 모델이 없으므로, 매 수행시마다 거리를 계산하여 분류해야 함.

예측 및 분류 방법

II. KNN (K-Nearest Neighbors, k-최근접이웃 알고리즘)

✓ **사례** : Income과 Lot-Size 정보를 통한 Riding Mower의 소유여부 예측

Income	Lot_Size	Ownership
60.0	18.4	owner
85.5	16.8	owner
64.8	21.6	owner
61.5	20.8	owner
87.0	23.6	owner
110.1	19.2	owner
108.0	17.6	owner
82.8	22.4	owner
69.0	20.0	owner
93.0	20.8	owner
51.0	22.0	owner
81.0	20.0	owner
75.0	19.6	non-owner
52.8	20.8	non-owner
64.8	17.2	non-owner
43.2	20.4	non-owner
84.0	17.6	non-owner
49.2	17.6	non-owner
59.4	16.0	non-owner
66.0	18.4	non-owner
47.4	16.4	non-owner
33.0	18.8	non-owner
51.0	14.0	non-owner
63.0	14.8	non-owner



Value of k	% Error Training	% Error Validation	
1	0.00	33.33	
2	16.67	33.33	
3	11.11	33.33	
4	22.22	33.33	
5	11.11	33.33	
6	27.78	33.33	
7	22.22	33.33	
8	22.22	16.67	<--- Best k
9	22.22	16.67	
10	22.22	16.67	
11	16.67	33.33	
12	16.67	16.67	
13	11.11	33.33	
14	11.11	16.67	
15	5.56	33.33	
16	16.67	33.33	
17	11.11	33.33	
18	50.00	50.00	

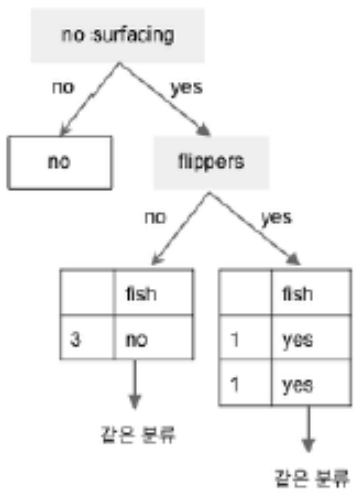
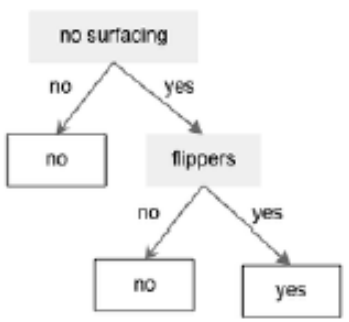
예측 및 분류 방법

IV. Decision Tree

✓ 개념

- 관찰된 데이터로부터 분할기준 속성을 판별하고, 분할기준 속성에 따라 트리 형태로 모델링한 분류, 예측 모델

✓ 의사결정 트리 수행 과정

1)최적 분할속성 선정	2)분할속성에 따른 분류	3)하위 분할속성 선정	4)동일한 분류결과 정리
			
분류를 가장 잘 분할하는 속성으로 최적 분할속성을 선택	선택된 분할속성에 따라 데이터를 분할, 트리 형성	트리의 가지에서 분할 속성을 선택하여 하위 가지를 형성	동일한 결과값을 가진 분류결과를 대표값으로 정리

- 데이터로부터 트리 형성과정을 완료한 후 타당성 평가과정을 거쳐 분류 및 예측모형으로 사용함

- purity check를 하며 split을 수행하고, maximum purity 달성 시에 다음 split을 위한 값을 찾음 (Recursive Partitioning)

- 지니계수 (Geni Index)

$$I(A) = 1 - \sum_{k=1}^{m} p_k^2$$

모든 케이스가 Rectangle A에 있을 때 0

- 엔트로피 (Entropy)

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

0일 때 most pure

(p : Rectangle A 안에 k case가 속하는 비중)

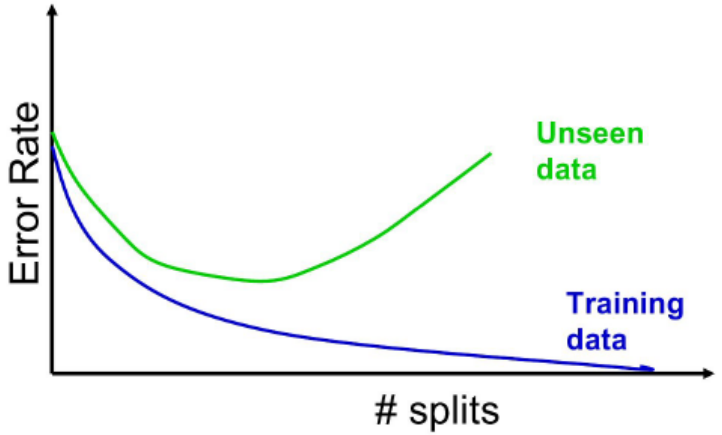
예측 및 분류 방법

IV. Decision Tree

✓ 구성요소

구성요소		내용
노드(Node)	뿌리노드(Root Node)	나무구조가 시작되는 지점
	부모노드(Parent Node)	자식지점의 상위지점을 의미
	자식노드(Child Node)	하나의 지점으로부터 분리되어진 2개 이상의 마디들을 의미
	잎(Leaf)	각 나무줄기의 끝에 위치하고 있는 마디
가지(Branch)		하나의 마디로부터 끝마디까지 연결된 줄기를 의미하며, 이때 가지를 이루는 개수는 깊이(Depth)

✓ Overfitting 방지 (Stopping Tree Growth)



- Overfitting이 발생하지 않도록 Validation Data로 검증

✓ 의사결정 트리의 장단점

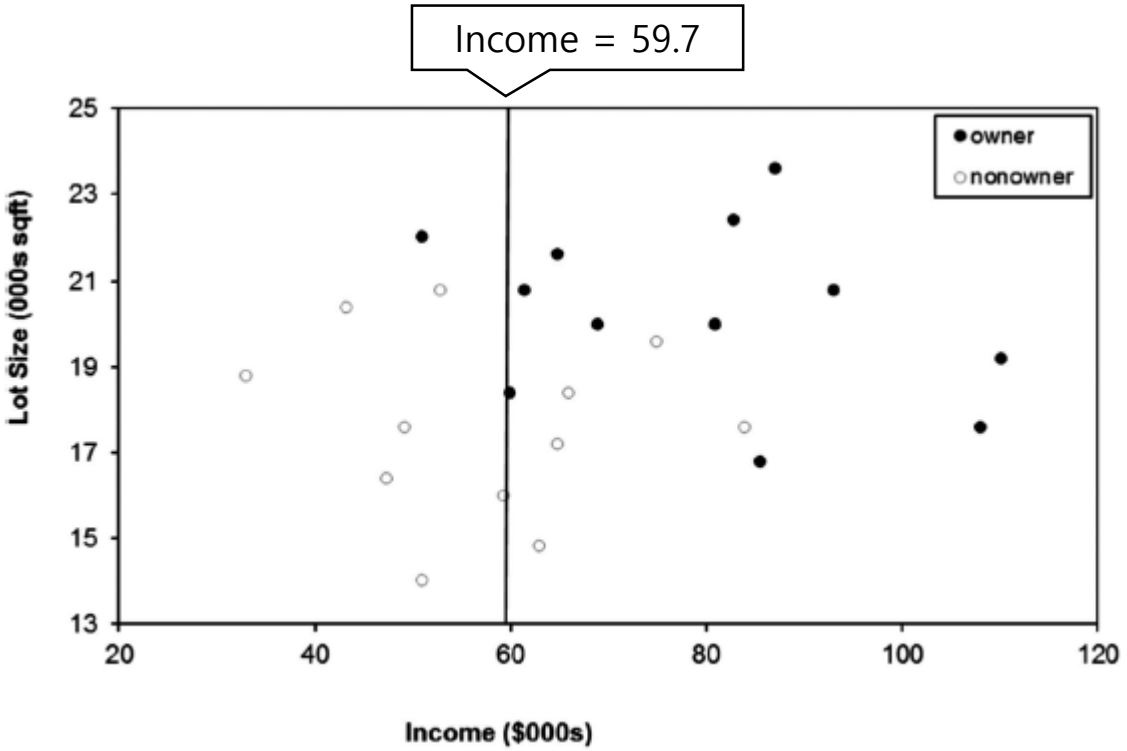
장점	해석 용이성	모형이 쉽게 파악
	가정 불필요	모델 형성을 위한 가정 불필요
	직관성	의사결정에 직접적 활용 가능
단점	불안정성	레코드 수의 차이에도 트리모양 변형
	최적해 보장 불가	Greedy 알고리즘 사용으로 최적해 보장 불가
	낮은 정확도	신경망, 회귀분석 등의 방식보다 정확도 낮음

예측 및 분류 방법

IV. Decision Tree

✓ 의사결정 트리 사례

Household Number	Income ($000s)	Lot Size (000s ft^2)	Ownership of Riding Mower
1	60.0	18.4	Owner
2	85.5	16.8	Owner
3	64.8	21.6	Owner
4	61.5	20.8	Owner
5	87.0	23.6	Owner
6	110.1	19.2	Owner
7	108.0	17.6	Owner
8	82.8	22.4	Owner
9	69.0	20.0	Owner
10	93.0	20.8	Owner
11	51.0	22.0	Owner
12	81.0	20.0	Owner
13	75.0	19.6	Nonowner
14	52.8	20.8	Nonowner
15	64.8	17.2	Nonowner
16	43.2	20.4	Nonowner
17	84.0	17.6	Nonowner
18	49.2	17.6	Nonowner
19	59.4	16.0	Nonowner
20	66.0	18.4	Nonowner
21	47.4	16.4	Nonowner
22	33.0	18.8	Nonowner
23	51.0	14.0	Nonowner
24	63.0	14.8	Nonowner



$$\text{Gini\_left} = 1 - (7/8)^2 - (1/8)^2 = 0.219.$$

$$\text{entropy\_left} = -(7/8)\log_2(7/8) - (1/8)\log_2(1/8) = 0.544.$$

$$\text{Gini\_right} = 1 - (11/16)^2 - (5/16)^2 = 0.430.$$

$$\text{entropy\_right} = -(11/16)\log_2(11/16) - (5/16)\log_2(5/16) = 0.896.$$

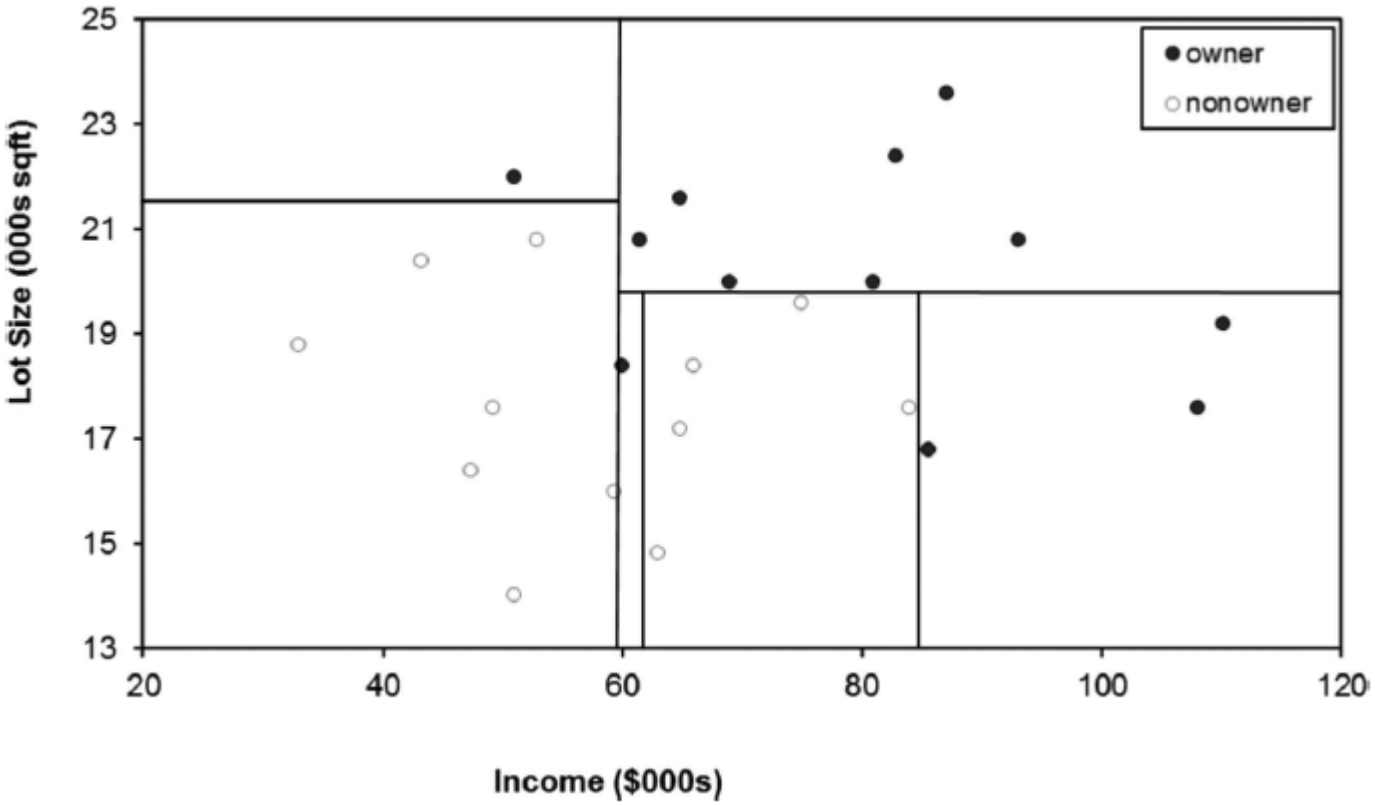
$$\text{Gini} = (8/24)(0.219) + (16/24)(0.430) = 0.359.$$

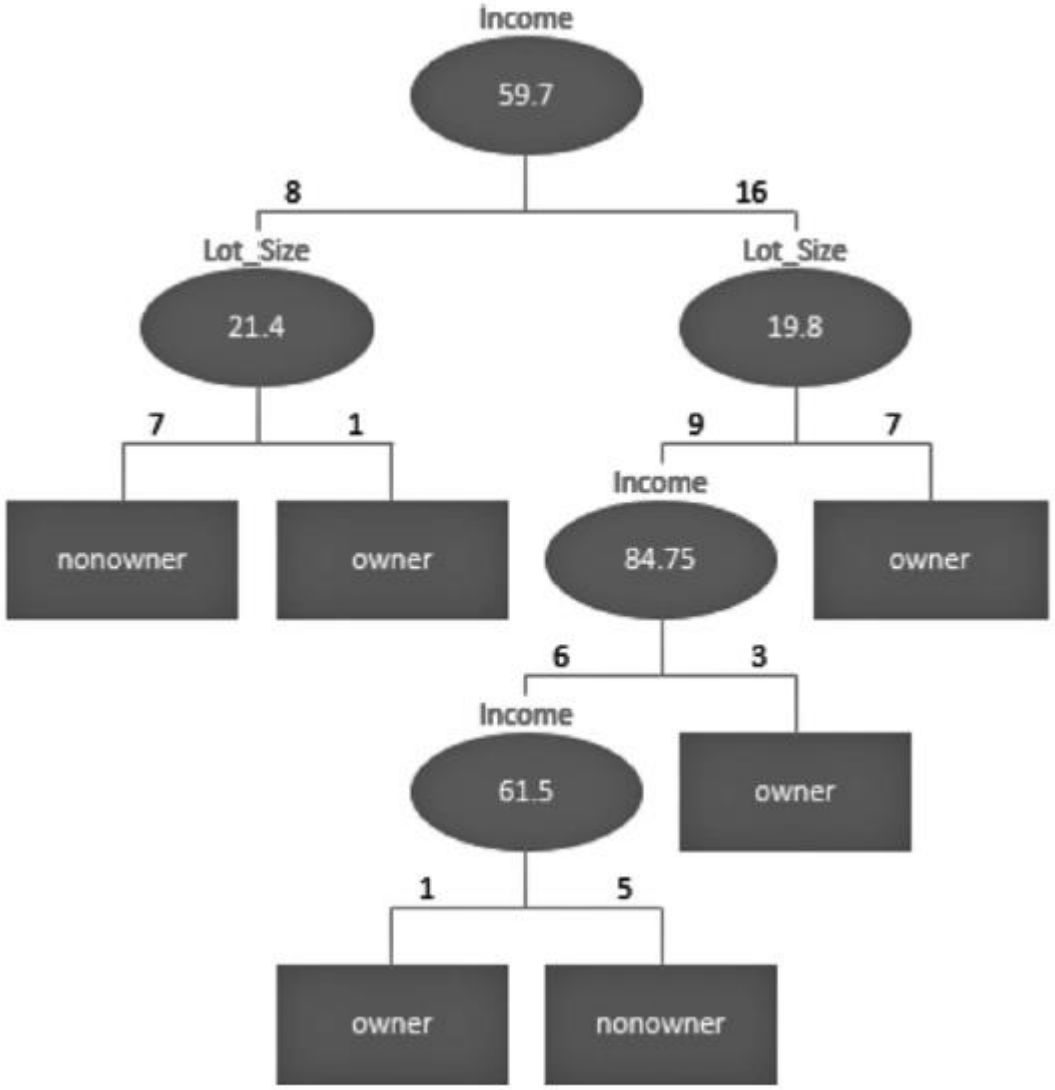
$$\text{entropy} = (8/24)(0.544) + (16/24)(0.896) = 0.779.$$

예측 및 분류 방법

IV. Decision Tree

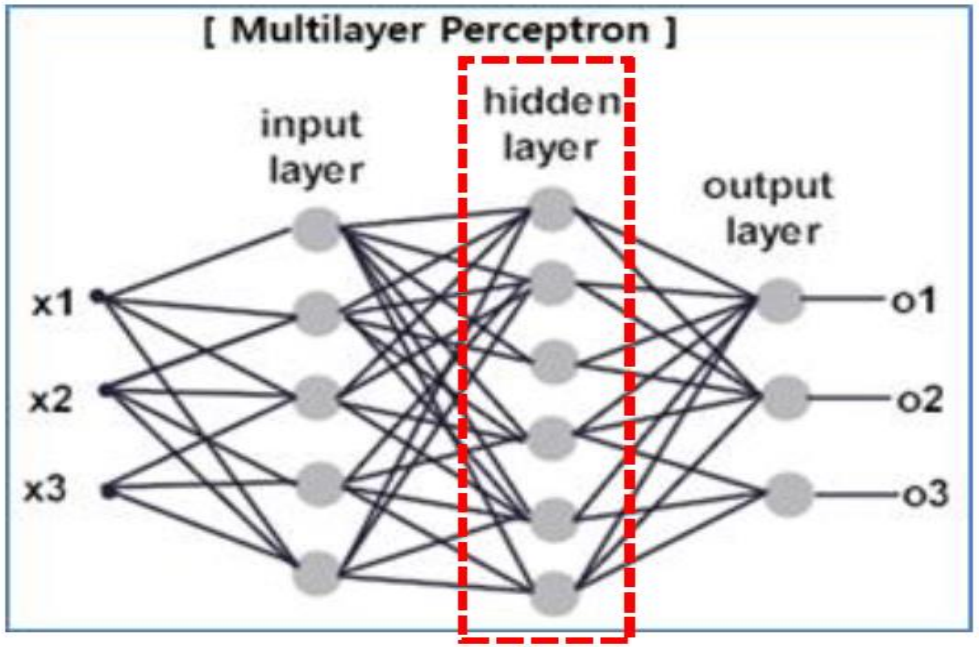
✓ 의사결정 트리 사례

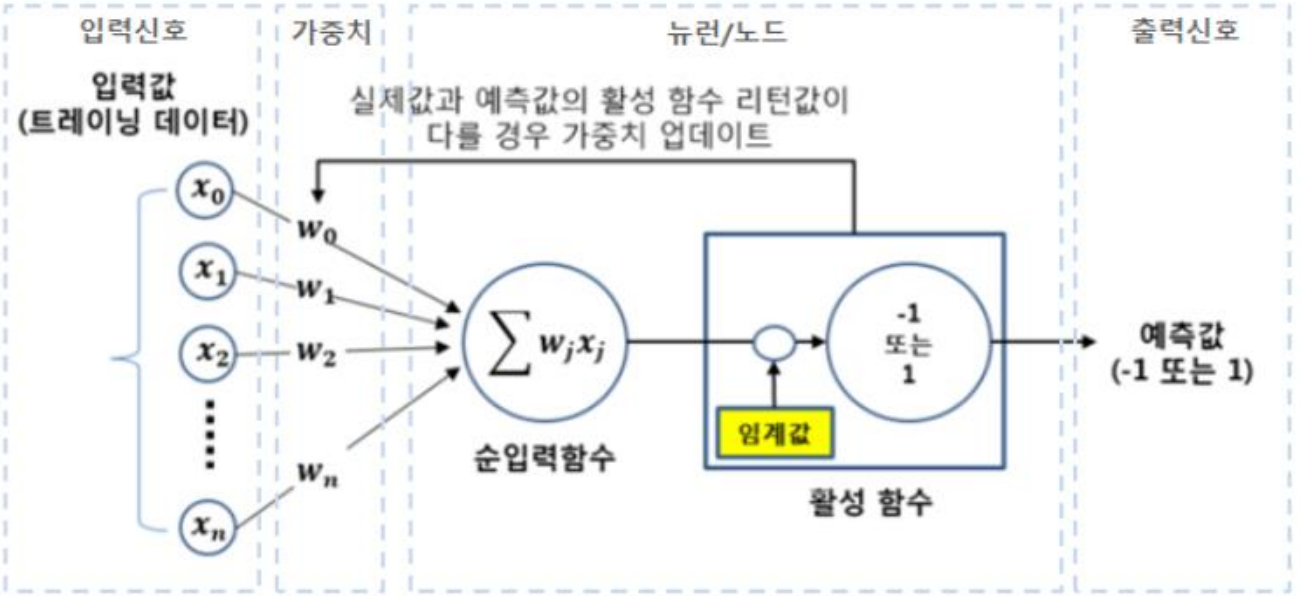




예측 및 분류 방법

V. 신경망





구성 요소	내용
입력층	- 학습을 위한 기초데이터 입력계층
출력층	- 학습을 통해 도출된 결과값 출력 계층
은닉층	- 다중신경회로망에서 입력층과 출력층 사이에 존재 - 정보를 전파, 학습, 활성화
활성화 함수	- 임계값을 이용하여 뉴런의 활성화 여부를 결정하기 위해 사용되는 함수
가중치 (연결강도)	- 활성화 함수의 입력값으로 사용되는 뉴런간의 연결계수

- 인간두뇌 세포를 모방한 개념으로 뉴런들의 상호작용하고 경험을 통해 배우는 생물학적 활동을 반복적인 학습과정으로 모형화 하는 분석 기법

예측 및 분류 방법

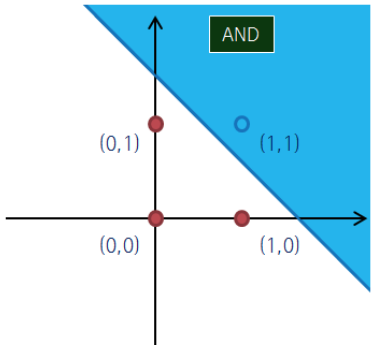
V. 신경망

✓ 단층 퍼셉트론

① AND 연산 가능

AND 진리표

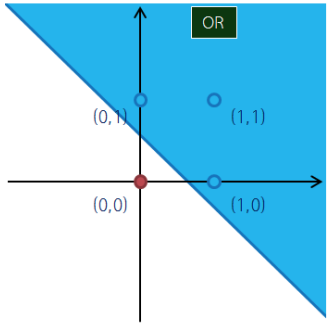
A	B	Output
0	0	0
0	1	0
1	0	0
1	1	1



② OR 연산 가능

OR 진리표

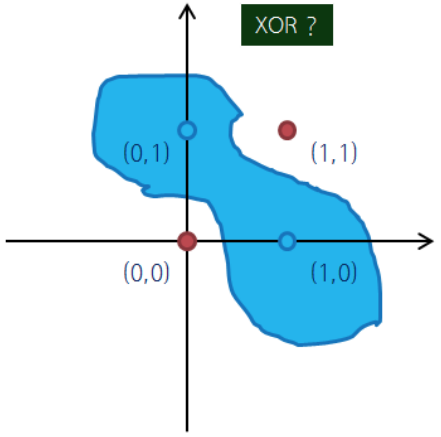
A	B	Output
0	0	0
0	1	1
1	0	1
1	1	1



③ XOR 연산 불가능

XOR 진리표

A	B	Output
0	0	0
0	1	1
1	0	1
1	1	0



→ 은닉층(hidden layer)이라는 하나의 층을 더 만들면 신경망이 그리는 모양이 직선이 아니라 다양한 모양으로 변할 수 있음 (다층 퍼셉트론)

예측 및 분류 방법

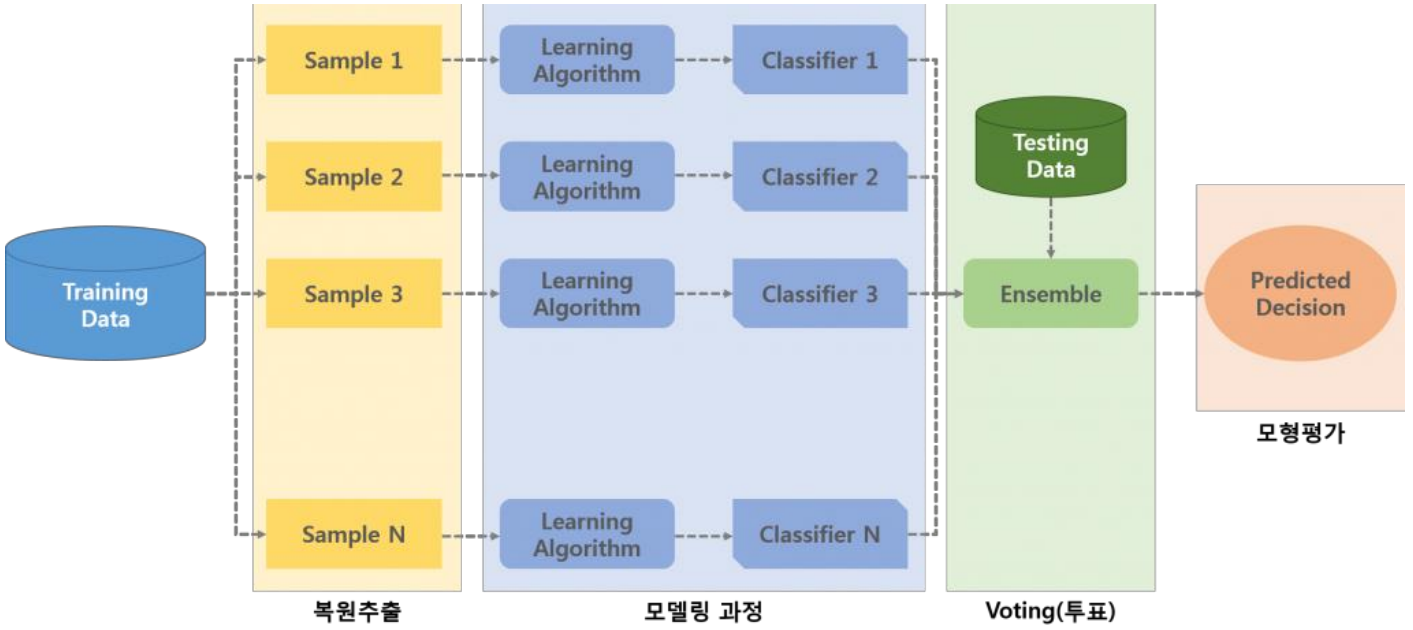
VII. 방법론 결합 : 앙상블

✓ 개념

- 간단한 알고리즘으로 학습을 수행하되, 복수개의 학습결과를 결합함으로써 결과적으로 보다 좋은 성능을 내고자 하는 방법
- 다양한 분류기의 예측 결과를 결합함으로써 단일 분류기보다 신뢰성이 높은 예측값을 얻는 것이 앙상블 학습의 목표
- 설명보다 예측이 중요할 때 사용하는 방법
- 앙상블 학습 방법 : 부스팅(Boosting), 배깅(Bagging)

① 배깅(Bagging = Bootstrap Aggregating)

- Raw Data에서 다수 복원 추출 (Bootstrap Sampling)
- 각 데이터를 모델링하여 모델 생성
- 단일 모델들을 결합하여 배깅 모델 생성
  (다수 투표 또는 평균치 사용)

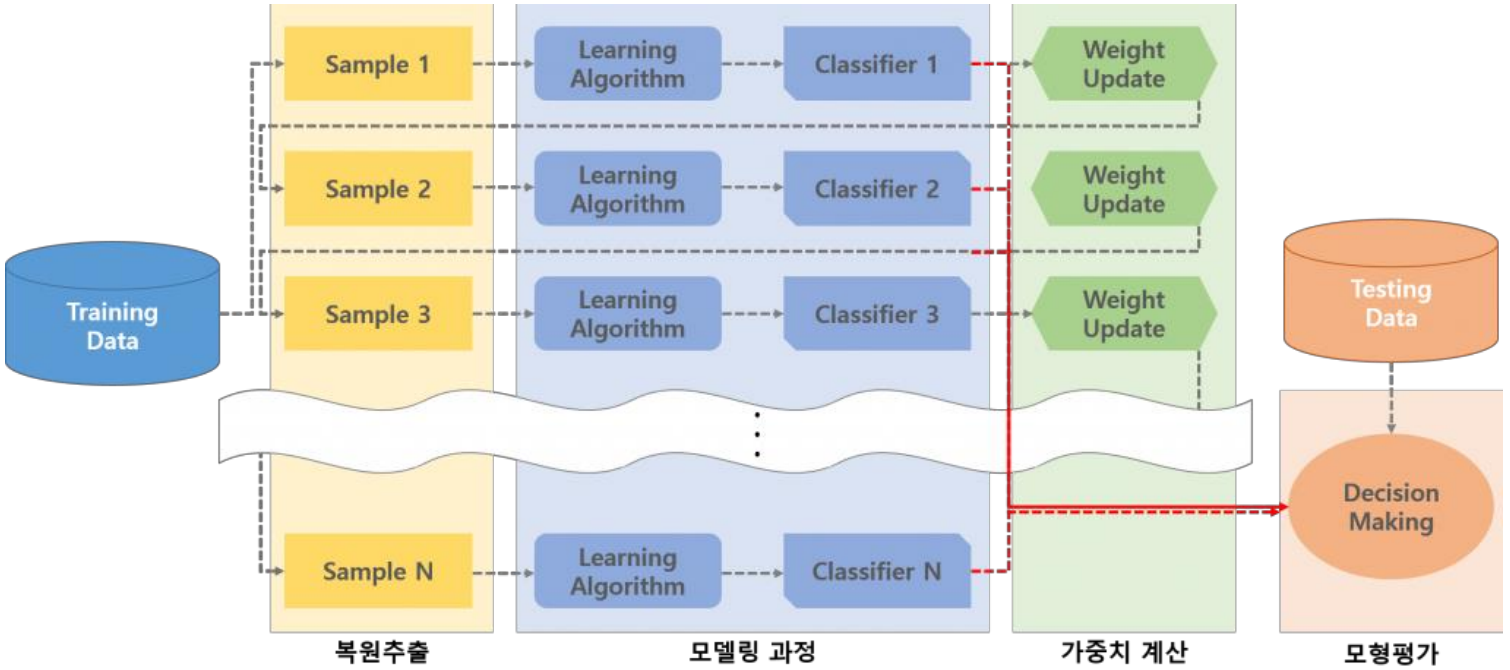


예측 및 분류 방법

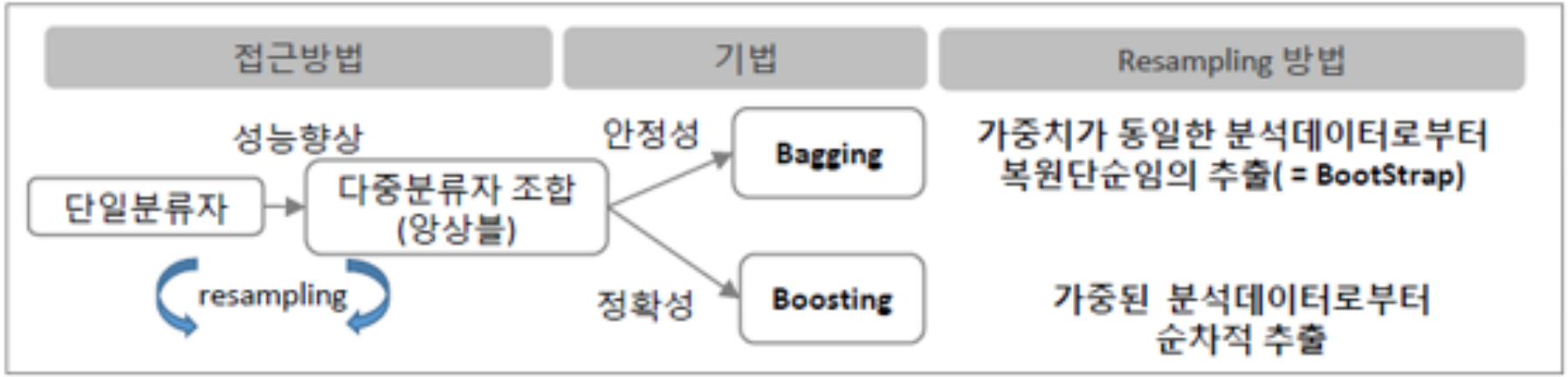
VII. 방법론 결합 : 앙상블

② 부스팅(Boosting)

- Raw Data에서 동일 가중치로 모델 생성
- 생성된 모델로 인한 오분류 데이터 수집
- 오분류 데이터에 높은 가중치 부여
- 과정 반복을 통한 정확도 향상



[참고] Bagging과 Boosting

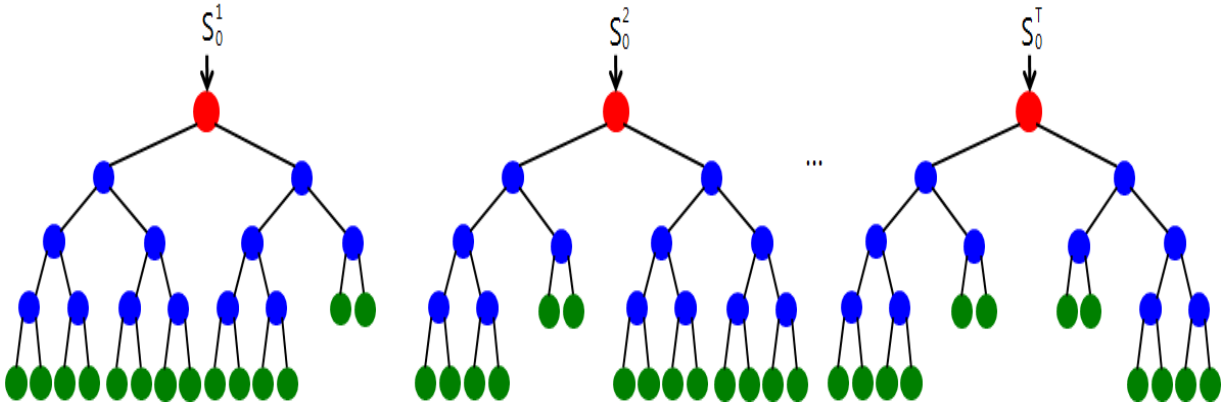


예측 및 분류 방법

VII. 방법론 결합 : 앙상블

[참고] **랜덤 포레스트 (Random Forest)**

- 여러 개의 결정 트리들을 임의적으로 학습하는 방식의 앙상블 방법으로서, 배깅(bagging)보다 더 많은 임의성을 주어 학습기들을 생성한 후 이를 선형 결합하여 최종 학습기를 만드는 방법



단계	내용
데이터집합생성	- 부트스트랩(bootstrap)을 통해 T개의 훈련데이터 집합 생성
훈련	- T개의 기초분류기(tree)들을 훈련시킨다
결합	- 기초분류기(tree)들을 하나의 분류기(random forest)로 결합 (평균 또는 과반수투표 방식을 이용)

- 예측 성능은 높아질 수 있으나 설명하기 어려워진다는 단점 존재

레코드들 간의 관계 마이닝

I. 연관규칙 (Association Rules)

✓ 개념

- 특정 사건(상품 구매)들이 동시 발생하는 빈도로 상호간 연관성을 표현하는 규칙
- 여러 개의 트랜잭션들 중에서 동시에 발생하는 트랜잭션의 연관관계를 발견하는 규칙 (시장 바구니 분석)

✓ 지지도(Support), 신뢰도(Confidence), 향상도(Lift)

구분	수식	설명
지지도 (Support)	$S = P(X \cap Y) = \frac{\text{품목X와 품목Y를 포함하는 거래수}}{\text{전체 거래 수(N)}}$	전체 거래 중 항목 X와 항목 Y를 동시에 포함하는 거래가 어느 정도 인가를 나타내며 전체적 구매도 경향을 파악
신뢰도 (Confidence)	$C = P(Y\|X) = \frac{P(X \cap Y)}{P(X)} = \frac{\text{품목X와 품목Y를 포함하는 거래수}}{\text{품목X를 포함하는 거래수}}$	항목 X를 포함하는 거래 중에서 항목 Y가 포함될 확률은 어느 정도인가를 나타내 주며 연관성의 정도를 파악
향상도 (Lift)	$L = \frac{P(Y\|X)}{P(Y)} = \frac{P(X \cap Y)}{P(X)P(Y)}$	항목 X를 구매한 경우 그 거래가 항목 Y를 포함하는 경우와 항목 Y가 임의로 구매되는 경우의 비

- 향상도가 1보다 크면 양의 관계(상호 보완), 1이면 독립적 관계, 1보다 작으면 음의 상관 관계(대체 관계)

레코드들 간의 관계 마이닝

I. 연관규칙 (Association Rules)

✓ **예제** - (모의. 컴시응. 201412. 4교시)

연관분석과 Apriori 알고리즘에 대해 다음 질문에 답하시오.

가. Apriori 알고리즘의 개념과 선험적 규칙을 설명하시오.

나. 연관분석을 위한 3 가지 척도를 설명하시오.

다. 다음 data 를 이용해 3 가지 척도를 계산하시오.(2 개 항목으로 구성된 후보 항목집합을 대상으로 최소지지도 50% 적용)

Transaction	Products
1	우유, 빵, 버터
2	우유, 버터, 콜라
3	빵, 버터, 콜라
4	우유, 콜라, 라면
5	빵, 버터, 라면

레코드들 간의 관계 마이닝

I. 연관규칙 (Association Rules)

✓ **예제** - (모의. 컴시응. 201412. 4교시)

가. Apriori 알고리즘의 개념

- 데이터베이스에서 후보 항목집합(Candidate Itemset)을 생성하고, 이를 데이터베이스 트랜잭션과 비교하여 후보 항목집합들의 발생 빈도를 계산하고, 사용자가 정의한 최소지지도를 기준으로 빈발 항목집합(Large Itemset)을 결정하는 알고리즘
- Apriori 알고리즘을 통해 빈발항목집합을 찾아낸 후, 3 가지 척도(지지도, 신뢰도, 향상도)를 계산해 연관분석을 수행함

가. Apriori 알고리즘의 선험적 규칙

규칙	설명
규칙 1	- 한 항목집합이 빈발 하다면, 이 항목집합의 모든 부분집합은 빈발항목 집합 예) 모든 항목집합 {a, b, c, d}, 빈발항목집합 {b, c, d}라면, 이 집합의 부분집합 {b, c}, {b, d}, {c, d}, {b}, {c}, {d}는 빈발항목 집합
규칙 2	- 한 항목집합이 비 빈발 하다면, 이 항목집합을 포함하는 모든 집합은 비 빈발항목 집합 예) 모든 항목집합 {a, b, c, d}, 비 빈발항목집합 {a, b}라면, 이 집합을 포함하는 {a, b, c}, {a, b, d}, {a, b, c, d}는 비 빈발항목 집합

- Apriori 알고리즘을 적용해 빈발집단을 결정하여 3 가지 척도를 계산해 연관분석을 수행함

레코드들 간의 관계 마이닝

I. 연관규칙 (Association Rules)

✓ **예제** - (모의. 컴시응. 201412. 4교시)

가. 최소지지도 50% 항목집합 생성

항목 집합	지지도	최소지지도 만족여부
우유, 빵	1 / 5 = 20%	X
우유, 버터	2 / 5 = 40%	X
우유, 콜라	1 / 5 = 20%	X
우유, 라면	1 / 5 = 20%	X
빵, 버터	3 / 5 = 60%	O
빵, 콜라	1 / 5 = 20%	X
빵, 라면	1 / 5 = 20%	X
버터, 콜라	2 / 5 = 40$	X
버터, 라면	1 / 5 = 20%	X
콜라, 라면	1 / 5 = 20%	X

- 최소 지지도 50%를 만족하는 항목집합은 {빵, 버터}이므로, 이를 빈발항목집합으로 선정함

레코드들 간의 관계 마이닝

I. 연관규칙 (Association Rules)

✓ **예제** - (모의. 컴시응. 201412. 4교시)

나. 빈발항목집합에 대한 지지도, 신뢰도, 향상도 계산

연관 규칙	지지도	신뢰도	향상도
빵 → 버터	3 / 5 = 0.6	0.6 / 0.6 = 1	0.6 / (0.6 * 0.8) = 1.25
버터 → 빵	3 / 5 = 0.6	0.6 / 0.8 = 0.75	0.6 / (0.8 * 0.6) = 1.25

- 빈발항목집합에 대한 2가지 연관규칙은 모두 1보다 큰 향상도 값을 나타냄

4. 연관분석 결과

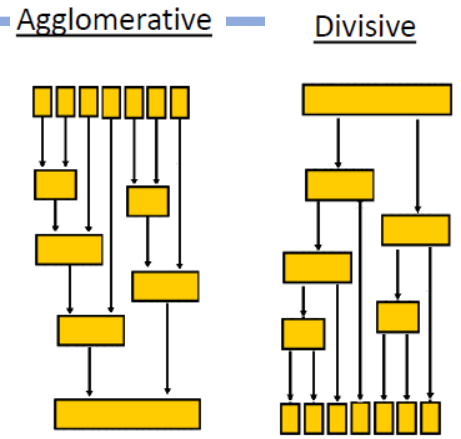
항목	결과 설명
빈발항목 집합	{빵, 버터} 항목집합이 유일하게 최소지지도 50%를 만족하여 빈발항목집합으로 선정됨
연관 규칙	{빵, 버터}는 '빵 → 버터', '버터 → 빵'의 2가지 연관규칙을 정의할 수 있음
향상도 분석	- 빵 → 버터 : 지지도(0.6), 신뢰도(1.0), 향상도(1.25) - 버터 → 빵 : 지지도(0.6), 신뢰도(0.75), 향상도(1.25) 2가지 연관규칙 모두 향상도 값이 1보다 크므로, 양의 상관관계에 있어 빵과 버터는 상호 보완관계로 볼 수 있음
신뢰도 분석	'빵 → 버터'의 신뢰도(1.0)가 '버터 → 빵'의 신뢰도(0.75)보다 크므로, '빵을 구입한 고객이 버터를 구입한다'는 규칙이 '버터를 구입한 고객이 빵을 구입한다'는 규칙보다 정확함
결과 활용	- 버터의 판매를 늘리기 위해 빵의 가격을 할인해서 더 많은 버터의 판매를 유도 - 버터와 빵의 진열위치를 인접시켜 두 상품의 더 많은 판매를 유도

레코드들 간의 관계 마이닝

III. 군집 분석

✓ 개념

- 통찰력을 생성할 목적으로 데이터를 동질적인 군집들로 세분화
- 활용 : 마케팅(고객세분화), 금융(균형 포트폴리오), 검색엔진(검색어 군집화)



✓ 종류

계층적 방법 (Hierarchical Methods)	**응집분석 (Agglomerative analysis)**	- N개의 레코드를 각각 하나의 군집으로 간주하고 레코드의 특성이 유사한 군집끼리 순차적으로 결합해 나가는 방법
	분할분석 (Divisive analysis)	- 모든 레코드를 하나의 군집으로 보고, 순차적으로 특성이 먼 레코드를 분할하는 방법
비계층적 방법 (Nonhierarchical Methods)	**k-means**	- 계층적 방법과 달리, 미리 군집의 수를 정해, 레코드를 각 군집에 할당하는 방법

- 계층적 방법의 경우 군집간 거리 계산 방법 존재

- 단일연결법(Single Linkage Clustering) : 최단거리
- 완전연결법(Complete Linkage Clustering) : 최장거리
- 평균연결법(Average Linkage Clustering) : 평균거리
- 중심연결법(Centroid Linkage Clustering) : 중심거리

[참고] 유클리디안 거리 (Euclidean distance)

$$d_{ij} = \sqrt{(x_{i1} - x_{j1})^2 + (x_{i2} - x_{j2})^2 + \cdots + (x_{ip} - x_{jp})^2}$$

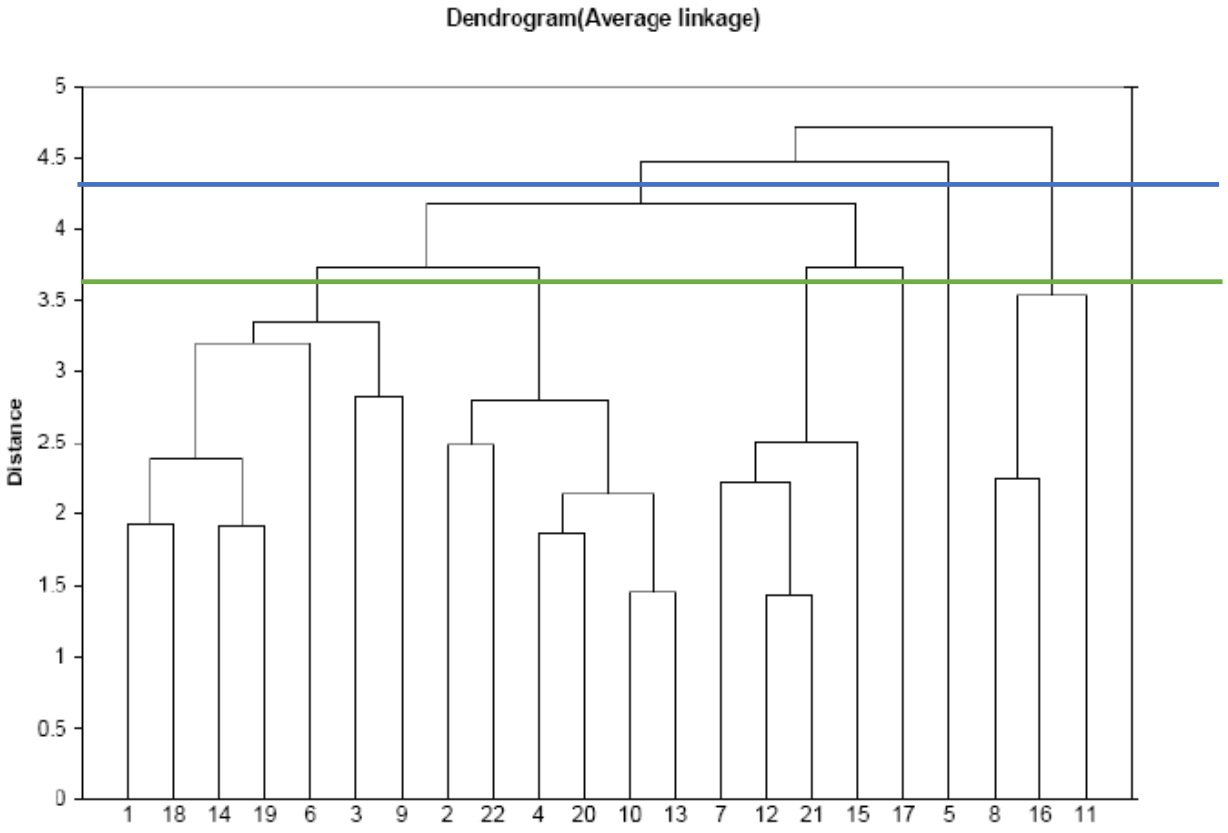
→ 단위 및 큰 값에 민감하므로, 정규화(표준화) 필요

레코드들 간의 관계 마이닝

III. 군집 분석

✓ 계층적 방법 – Dendrogram (덴드로그램)

- 군집화 과정을 요약하는 나무형태의 도표
- X축에 레코드들을 표시하고, y축에 레코드들 사이의 거리를 표현 (거리측정 방식에 따라서 달라짐)
- Y축에서 컷오프 거리를 선택해서 군집 집합을 생성함



컷오프 = 4.3일 경우, 세 개의 군집으로 형성

컷오프 = 3.6일 경우, 여섯 개의 군집으로 형성

✓ 계층적 방법 장단점

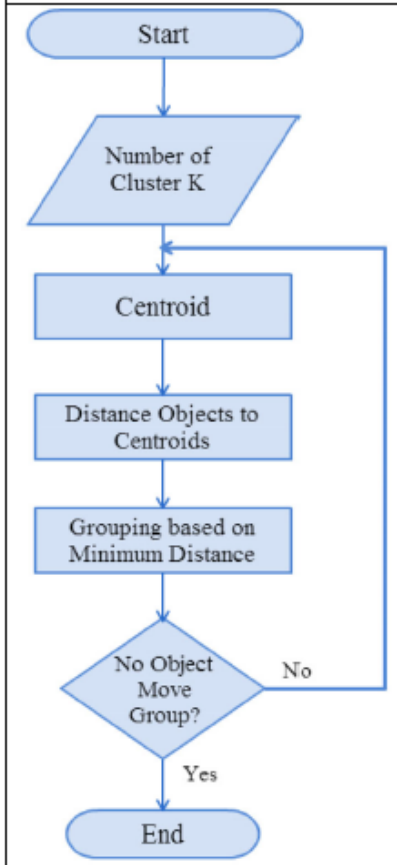
장점	군집수 명시 불필요	- 초기 군집수 명시 불필요
	데이터 기반 판별	- 거리 기반으로 군집 판별
	직관성	- 덴드로그램 활용하여 이해 용이
단점	계산 속도 느림	- nxn 거리 행렬 계산 필요
	이상치/거리 척도에 따라 달라짐	- 이상치에 민감하며, 거리 척도 방식에 따라 군집 변경 가능

레코드들 간의 관계 마이닝

III. 군집 분석

✓ 비계층적 방법 – k means 군집화

- n개의 데이터를 K개의 군집으로 분류하기 위해 거리 기반으로 반복적으로 계산해 나가는 Clustering 알고리즘

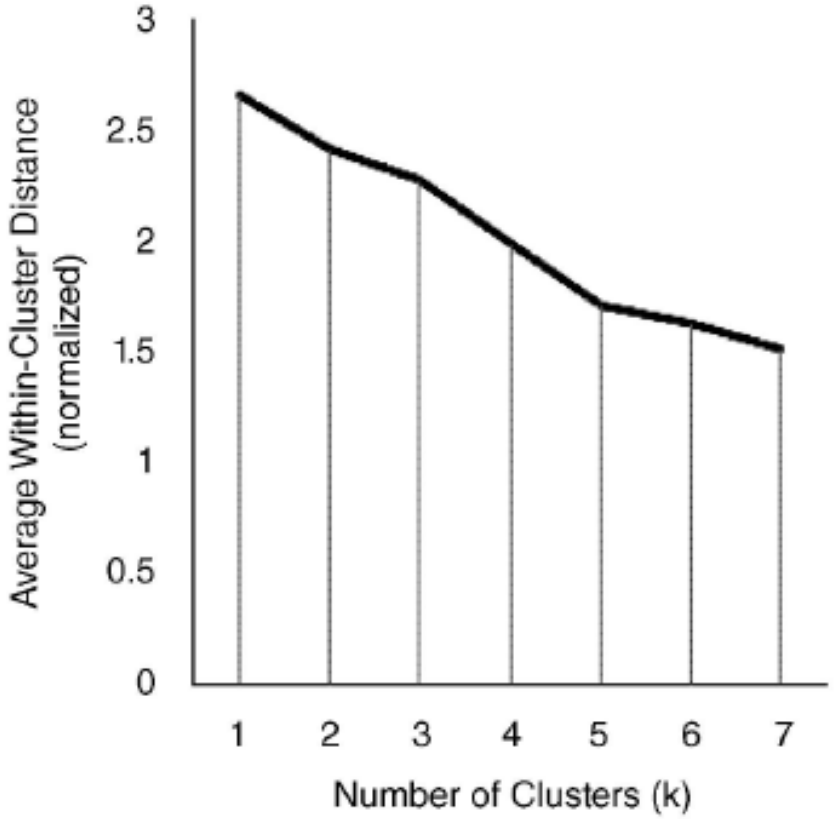
수행절차	단계	설명
	1) 시작	데이터를 모두 받아들임(lazy learning)
	2) Cluster K개 지정	파라미터 값으로 K개의 cluster개수를 사전에 입력 받음
	3) 초기 평균값 선정	초기 평균값은 데이터 오브젝트 중 무작위로 뽑음
	4) 초기 평균값 기준으로 데이터 선별	K의 각 데이터 오브젝트들은 가장 가까이에 있는 평균값 기준으로 묶임
	5) 최소 거리를 가진 데이터들로 그룹핑	최소 거리에 기반하여 grouping수행
	6) 평균값 재조정	k개의 클러스터 중심점을 기준으로 평균값 재조정 (수렴할때까지 3~5 단계 반복수행)
	7) 알고리즘 종료	더 이상 평균값이 변경되지 않는 경우, 그룹핑을 완료

레코드들 간의 관계 마이닝

III. 군집 분석

✓ 비계층적 방법 – k means 군집화

- Elbow chart를 활용한 적절한 클러스터 수 책정



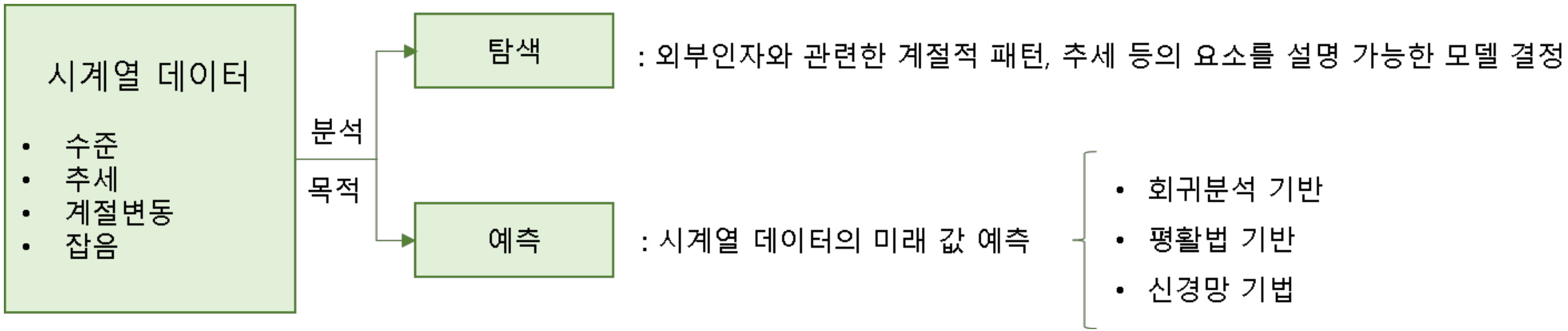
✓ k means 수행 시 고려사항

- 초기 k값 선정 시 도메인 지식이 필요
- 휴리스틱 방법

: 초기 Centroid 위치에 따라 다른 클러스터 형성 가능

시계열 예측

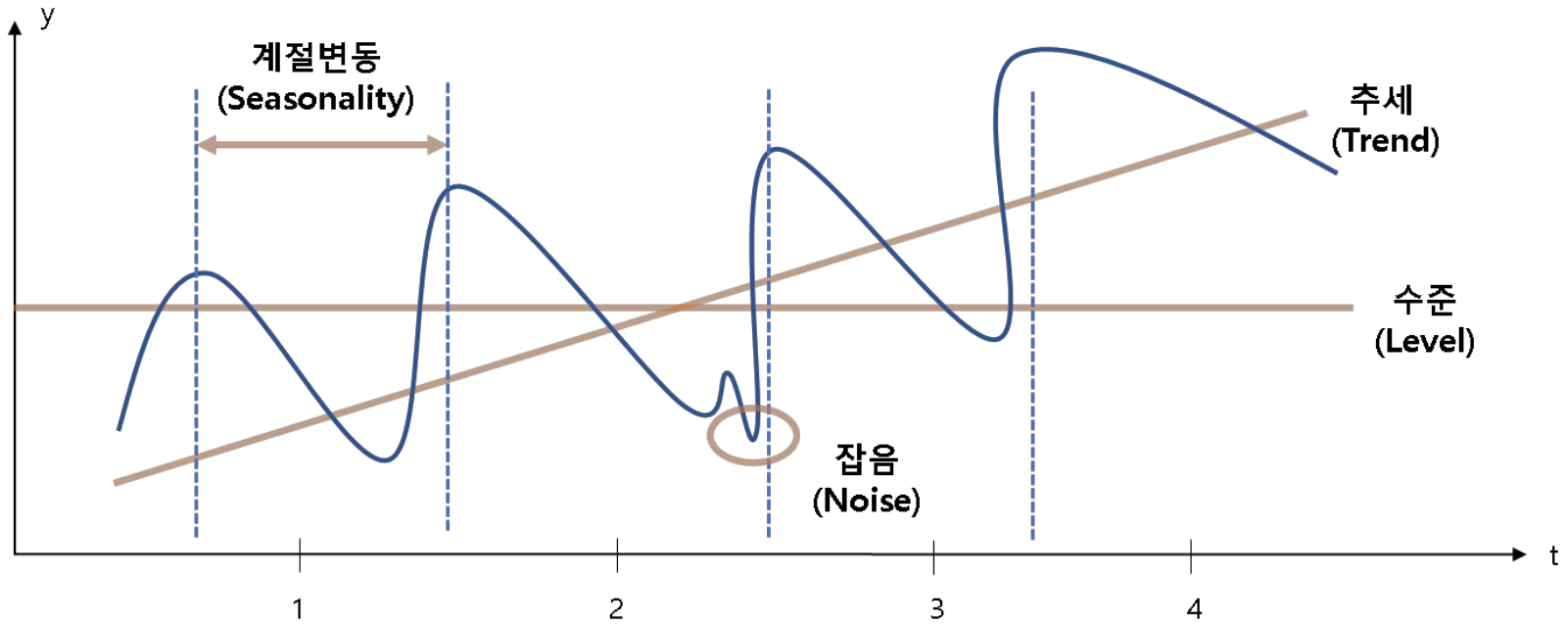
✓ 개요



- 시계열 자료들 간의 계열 상관을 이용하여 동태적인 관계를 분석하는 기법.
- 자료 간의 인과관계나 시차분포형태에 대한 사전적인 제약이 최소화된 모형을 추정해서 데이터의 의미를 도출하는 방법
- 시계열 (시간의 흐름에 따라 기록된) 자료 (data)를 분석하고 여러 변수들 간의 인과관계를 분석하는 방법론
- 종합 주가지수, 유가 변동사항, 환율 등 데이터들은 시계열 데이터로 볼 수 있으므로, 시계열 해석은 미래를 예측하는 데에 중요한 도구

시계열 예측

✓ 시계열 데이터의 구성요소



시계열 요소	핵심요소	설명
수준 (Level)	시계열의 평균값	- 시계열 데이터의 전체적 평균
추세 (Trend)	전반적인 패턴 변화	- 어떤 현상이 일정한 방향으로 증가/하락하는 경향 - 선형, 2차 함수, 지수 성장, S-곡선 등의 유형 존재
계절변동 (Seasonality)	짧은 기간 동안의 주기적인 패턴	- 특정 기간 동안 유사한 형태의 관측치가 반복되는 구간
잡음 (Noise)	무작위적 변동	- 일반적으로 알 수 없는 이유로 발생하는 무작위적 관측치

시계열 예측

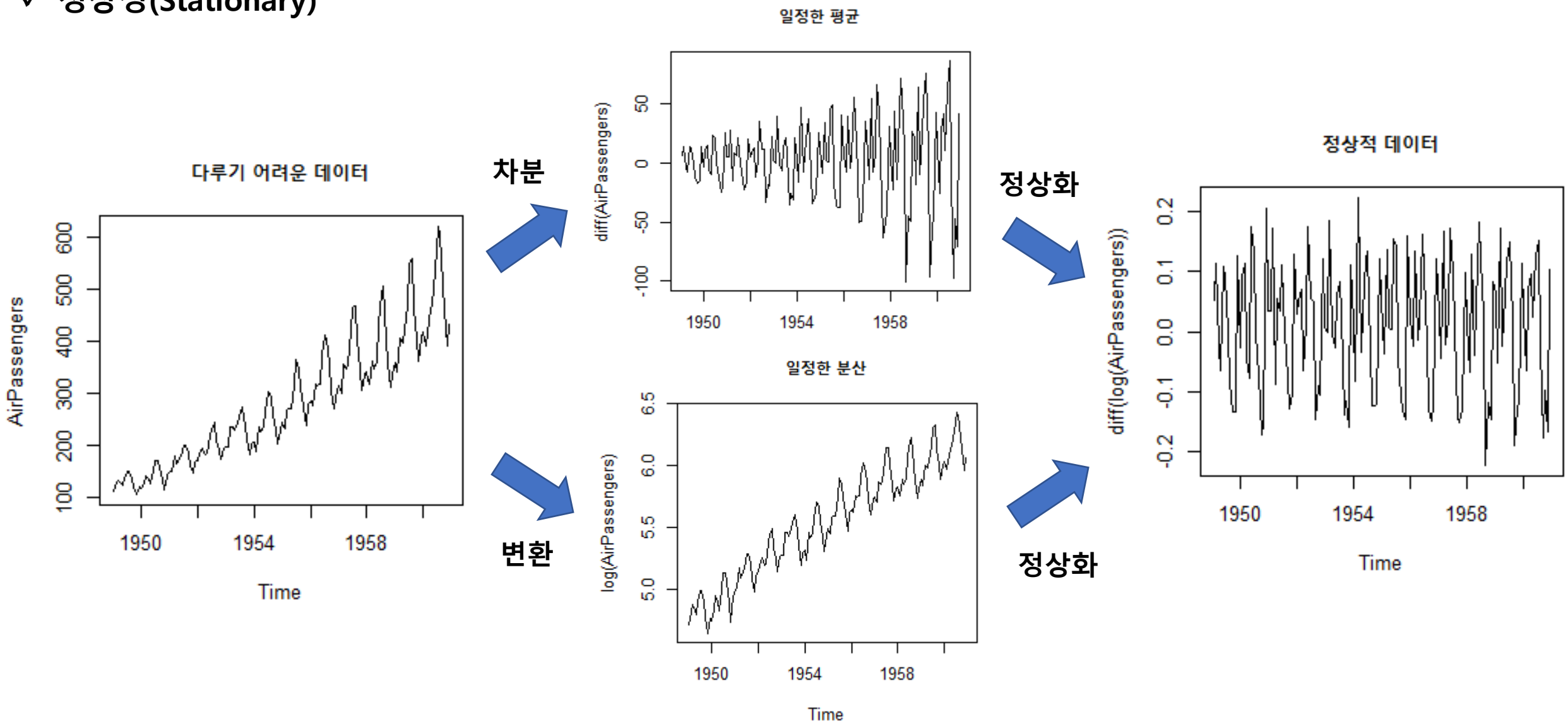
✓ 정상성(Stationary)

구분	상세 설명	
개념	시계열의 수준과 분산에 체계적인 변화가 없고 엄밀하게 추가적 변동이 없다는 것으로, 미래는 확률적으로 과거와 동일하다는 성질	
조건	일정한 평균	- 시계열 자료의 모든 시간 t에 대하여 평균이 일정함
	분산이 시점에 의존 안함	- 시계열 자료의 모든 시간 t에 대하여 분산이 일정함
	공분산이 시점에 의존 안함	- 시계열 자료의 자기상관함수 및 편자기상관함수는 시간 t1, t2에만 의존함
정상/비정상 분류	비정상 시계열	- 정상성 조건을 하나라도 만족하지 않는 경우의 시계열 자료
	정상 시계열	- 정의한 정상성 조건을 모두 만족하는 데이터로 시계열 분석 수행

- 일반적으로 평균이 일정하지 않을 때 차분을 수행하고, 분산이 일정하지 않을 때 변환을 수행하여 정상성을 갖추도록 함

시계열 예측

✓ 정상성(Stationary)



시계열 예측

✓ 회귀분석 기반 시계열 예측

<table>
<tr><td>개념</td><td colspan="2">- 추세와 계절변동이 있는 시계열 데이터를 선형, 지수, 다항, 회기회귀 등의 성격을 통해 회귀 모델을 구축하여 데이터를 예측하는 기법</td></tr>
<tr><td rowspan="4">기법</td><td>추세 반영 모델</td><td>① 선형 추세
: 시간을 독립변수(X), 관측값을 종속변수(Y)로 사용하여 선형 모델의 회귀 계수를 산정하는 예측 기법
$Y_t = b_0 + b_1 t + e$
(데이터의 일부를 Training Data로 학습시킨 후, Validation Set에서 실제값과 예측값의 차이 비교 필요)

② 지수 추세
: 시간을 독립변수(X), 관측값 Y를 log(Y)로 대체하여 선형회귀모델을 구축하는 예측 기법
$\log(Y_t) = b_0 + b_1 t + e$

③ 다항 추세
: 독립변수(X)인 시간의 다차항을 이용하여 결과값 Y를 계산하는 예측 기법
$Y_t = b_0 + b_1 t + b_2 t^2 + e$ (이차 추세식)
(일반적으로 어떤 형태의 추세라도 수학적으로 표현 가능한 것으로 알려짐)</td></tr>
<tr><td>계절변동 반영 모델</td><td>- Dummy Coding을 통한 새로운 범주형 변수 생성
: n개 기간의 Seasonality가 있을 경우 (n-1)개의 가변수(이진변수)로 변환하여 회귀모델 구축
예)12개월의 Seasonality가 있는 경우, 11개의 Dummy 변수를 만들어 이진변수로 변환 후 회귀 모델을 구축</td></tr>
<tr><td>추세와 계절 변동 반영 모델</td><td>- 수학적으로 표현된 추세식과 Dummy Coding을 포함하여 회귀 모델 구축 및 예측 수행</td></tr>
<tr><td>자기상관 반영 모델</td><td>- 자기상관 모델(AR 모델)
: 시계열과, 지연을 둔 자신 시계열 사이의 상관계수를 반영하여, 독립변수의 과거 기간 값들이 예측변수로 사용되는 모델
$Y_t = b_0 + b_1 Y_{t-1} + b_2 Y_{t-2} + e$ (AR(2) 모델)</td></tr>
</table>

시계열 예측

✓ 평활법 기반 시계열 예측

<table>
<tr><td>개념</td><td colspan="2">- 시계열 데이터에서 여러 관측치의 평균을 취함으로써 데이터의 잡음을 제거하여 미래 값을 예측하는 데이터 기반 추정 예측 기법</td></tr>
<tr><td rowspan="2">기법</td><td>이동평균법
(Moving Average)</td><td>① 전후이동평균법 (Centered Moving Average)
: t시점을 중심으로 w개 값의 평균을 취하는 시계열 데이터 시각화 기법
$MA_t = (Y_{t-(w-1)/2}+\dots+Y_t+\dots+Y_{t+(w-1)/2}) / w$

② 이전 이동평균법 (Trailing Moving Average)
: t시점을 포함한 w개 값의 평균을 취하는 시계열 데이터 예측 기법
$F_{t+1} = (Y_t+Y_{t-1}+\dots+Y_{t-w+1}) / w$

전후 이동평균법 (w=5) 이전 이동평균법 (w=5)
t-2 t-1 t t+1 t+2 t-4 t-3 t-2 t-1 t

(w에 따라 전역/국소적 추세가 확인되므로, w값 변화를 통해 검증 필요)</td></tr>
<tr><td>단순지수평활법</td><td>- 과거 데이터들의 가중평균을 이용하여 시계열 데이터를 예측하는 기법
- 가중치들은 지수분포의 형태로 분포
- 아무리 오래된 관측치라도 작게나마 가중치가 주어져서 계산됨

$F_{t+1} = aY_t + a(1-a)Y_{t-1} + a(1-a)^2Y_{t-2} + \dots$

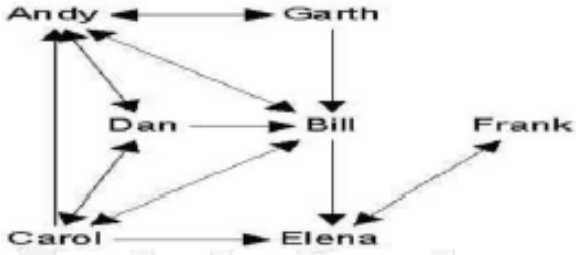
(a값은 예측하는 데에 얼마만큼의 과거 데이터가 관련되어 있는지에 따라 달라짐)</td></tr>
</table>

사회 연결망 네트워크 애널리틱스(SNA : Social Network Analysis)

✓ 개념

- 개인과 집단들 간의 관계를 노드와 링크로서 모델링 해 그것의 위상구조와 확산 및 진화과정을 계량적으로 분석하는 기법

✓ SNA 표현방법

<table>
<tr>
<td>그래프 표현</td>
<td> 
</td>
<td>- 개체는 노드, 연결망은 선(Edge)로 표현
- 방향성/무방향성 그래프 표현 가능
- x축, y축은 큰 의미가 없음
- 노드의 크기, 선의 굵기, 레이블, 화살표 방향 등 그래프 특성을 보여주는 요소들이 중요한 역할을 함</td>
</tr>
<tr>
<td>행렬 이용</td>
<td>
<table>
<tr><th></th><th>And</th><th>Bil</th><th>Car</th><th>Dan</th><th>Ele</th><th>Fra</th><th>Gar</th></tr>
<tr><td>Andy</td><td></td><td>1</td><td>0</td><td>1</td><td>0</td><td>0</td><td>1</td></tr>
<tr><td>Bill</td><td>1</td><td></td><td>1</td><td>0</td><td>1</td><td>0</td><td>0</td></tr>
<tr><td>Carol</td><td>1</td><td>1</td><td></td><td>1</td><td>1</td><td>0</td><td>0</td></tr>
<tr><td>Dan</td><td>1</td><td>1</td><td>1</td><td></td><td>0</td><td>0</td><td>0</td></tr>
<tr><td>Elena</td><td>0</td><td>0</td><td>0</td><td>0</td><td></td><td>1</td><td>0</td></tr>
<tr><td>Frank</td><td>0</td><td>0</td><td>0</td><td>0</td><td>1</td><td></td><td>0</td></tr>
<tr><td>Garth</td><td>1</td><td>1</td><td>0</td><td>0</td><td>0</td><td>0</td><td></td></tr>
</table>
</td>
<td>- 방향성/무방향성 그래프에 대한 인접행렬 표현 가능
- 단순 연결을 표현할 수도 있으며, 팔로우와 같이 방향성 있는 연결도 표현 가능</td>
</tr>
<tr>
<td>집합 표현</td>
<td>$A=\{(X_1, X_2), (X_2, X_1), (X_4, X_2), (X_3, X_1)\}$
$B=\{(X_1, X_2), (X_2, X_1), (X_3, X_4), (X_4, X_3)\}$</td>
<td>- 각 객체들 간의 관계를 관계 쌍으로 표현</td>
</tr>
</table>

사회 연결망 네트워크 애널리틱스(SNA : Social Network Analysis)

✓ 분석절차 및 기법

No	분석절차	설명
1	그래프 생성	특정 단어 빈도수 이상 데이터 추출 Term-Document Matrix 생성 후 네트워크 그래프 생성
2	그래프 목적에 따라 가공 및 분석	그래프에 연결된 노드의 위치를 분석 후 용도에 맞게 조절
3	커뮤니티 탐지 및 역할 정의	네트워크 그래프에서 인접한 노드 관계를 커뮤니티로 식별
4	데이터 변환 및 데이터 마이닝 기법 활용	커뮤니티 사용자 리스트에 대해 팔로어 정보를 추가로 검색. 네트워크 그래프를 생성하여 유력자 현황 분석

- 사회연결망 분석은 구매품목 추천 서비스 등의 Social CRM, 온톨로지 기반 사회 연결망 내의 전문가 추천 시스템 등 비정형 데이터 분석을 위해 활용

분석기법	유형	설명
중심성	연결정도 중심성	- 한 노드에 직접적으로 연결된 노드 합으로 얻어진 중심성
	근접(인접) 중심성	- 간접적으로 연결된 모든 노드 간 거리 합산한 지표
	매개(사이) 중심성	- 매개자 혹은 중재자 역할의 정도를 측정한 지표
	위세 중심성	- 연결된 노드의 중요성에 가중치를 부여한 지표
밀도	연결정도(degree)	- 한 노드와 직접적으로 연결된 노드들의 수
	포괄성	- 한 연결망 내에 서로 연결된 행위자들의 수

텍스트 마이닝 (Text Mining)

✓ 개념

- 다양한 문서형태의 비정형 데이터를 가져와 문서의 단어별 행렬을 만들어 추가적인 분석이나 데이터 마이닝 기법을 적용하여 의사결정을 지원해주는 방법

✓ 기능

문서 요약	문서의 주요 내용을 추출하여 요약
문서 분류	문서의 내용을 주어진 키워드에 따라 자동으로 구조화 , 분류
문서 군집	문서들을 분석하여 동일한 내용의 문서들을 묶는 기법
특성 추출	문서내 사용자가 원하는 정보 , 특성을 자동으로 추출

[참고]

This is the first sentence.

This is the second sentence.

The third sentence is here.

	Docs		
Terms	1	2	3
first	1	0	0
here	0	0	1
second	0	1	0
sentence	1	1	1
the	1	1	1
third	0	0	1
this	1	1	0

텍스트 마이닝 (Text Mining)

✓ 분석 절차

절차	기술요소	설명
수집	HTML Parsing, API	- 분석에 사용 가능한 텍스트 데이터 수집
전처리	말뭉치(Corpus) 형성	- 문서집합으로 정제, 통합, 변환 등의 구조화 작업 수행
	토큰화	분석의 기본 단위인 토큰(용어)으로 자동화 분리 공백, 콤마, 콜론 등의 구분기호 기준
	텍스트 축소	- 불용어 리스트 기반의 어휘 축소
	Stemming	- 단어의 서로 다른 변이를 줄여서 공통 어원 추출
	빈도 필터	- 극대/극소의 빈도를 가진 용어 제거
	정규화	같은 범주의 다양한 특수 용어들은 해당 범주명으로 대체
TDM 구축	TermDocumentMatrix	-분석 대상의 문장, 단어를 열과 행의 매트릭스로 표현
분석/시각화	데이터 마이닝 방법 구현	- 숫자의 행렬 형식으로 변환 후 데이터 마이닝 기법 적용
	워드 클라우드	- 단어 분석을 통해 중요도, 빈도, 인기도 등을 고려하여 시각적으로 표시
	감성분석	- 단어의 긍정, 부정 여부에 따른 추이 분석

- 단어주머니(bag of words) 접근법 : 순서, 문법, 신택스 상관없이 단어들의 집합으로 취급