회귀분석: 회귀분석 소개

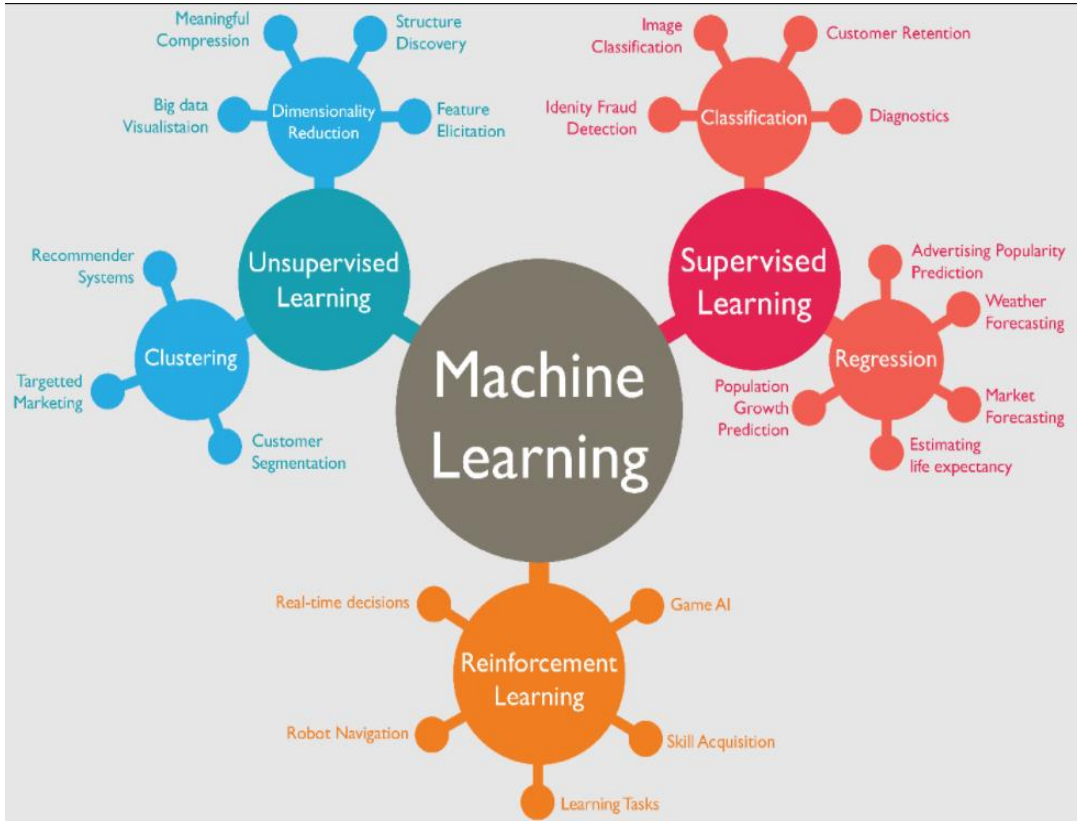
숙명여자대학교 경영학부 오중산

회귀분석 소개

- 회귀분석(regression analysis)의 목적
  - 인과관계 규명
    - 종속변수(dependent variable: DV)는 '결과', 독립변수(independent variable: IV)는 '원인'을 의미
    - DV는 한 개 이상의 IV로 설명됨($k \geq 1$)
  - 예측
    - DV는 예측치(predictor), IV는 예측치에 영향을 미치는 요인을 의미
  - 목적에 따라 회귀분석 유형이 다른 것은 아니고, 분석 초점이 다름
    - 예시: 에어컨 판매량(DV)과 기온/습도/소득수준/가격(이상 IV) 관련 회귀분석
      - 인과관계 규명: 에어컨 판매량에 영향을 미치는 IV는 무엇인가?
      - 예측치 추정: 올해 2/4분기 IV 예측치를 알고 있을 때 DV는 얼마가 될 것인가?

회귀분석 소개

- 변수의 척도
  - ◆ DV: 계량형 척도로 측정
    - ▪ 로짓회귀분석에서는 이진 범주형 척도로 측정
  - ◆ IV: 계량형 척도(일부 범주형 척도)로 측정
    - ▪ 더미(dummy) 변수는 범주형 척도로 측정
- 회귀분석과 머신러닝
  - ◆ 머신러닝의 세 가지 유형
    - ▪ 지도/비지도(자율)/강화학습
      - ❖ 지도학습(SL)은 DV가 존재하며, 비지도학습(UL)은 DV가 존재하지 않음
    - ▪ (예측목적의) 회귀분석은 SL의 대표적인 유형



Source: hocul.net

회귀분석 소개

- 회귀분석에서의 표본크기(sample size, n) 문제
  - 단순회귀분석(simple linear regression)
    - IV가 한 개이며($k = 1$), 표본 크기는 최소한 30개 이상이어야 함($n \geq 30$)
  - 다중회귀분석(multiple linear regression)
    - IV가 두 개 이상이며($k \geq 2$), IV 개수 대비 n비율($k : n$)이 최소 1:15 (가능하면 1:30) 이상이어야 함
    - 자유도(degree of freedom, df) = 표본크기(n) – 추정치(estimator) 개수($= k + 1$) = $n - k - 1$
      - 자유도가 클수록 1) 인과관계가 명확하고, 2) 연구결과 일반화에도 유리함
      - 따라서 표본크기가 크고, 독립변수가 적을수록 좋은 모형임

다중회귀분석 소개

- 다중회귀분석이란?
  - ◆ 모집단 대상 회귀모형: $Y_i = \alpha + \beta_j X_{ji} + e_i$
  - ◆ 표본 대상 회귀식: $\hat{Y}_i = a + b_j X_{ji}$
    - Y_i: DV / X_{ji}: 독립변수 / $\hat{Y}_i$: DV 추정치(예측치)
      - ❖ j는 IV 번호(j = 1, 2, …., k) / i는 case 번호 (i = 1, 2, …., n)
    - α :상수(모수) ↔ a: 상수 추정치 / β_j: 회귀계수(모수) ↔ b_j : 회귀계수 추정치
      - ❖ 회귀분석은 표본으로부터 회귀식(즉, a와 b_j)을 도출하여 α와 β_j를 추정하는 것
  - ◆ e_i: 오차(error) vs. ε_i: 잔차(residual = $Y_i - \hat{Y}_i$)
    - 잔차는 DV 실제값에서 예측치를 뺀 값으로 이것이 작아야 좋은 회귀식

다중회귀분석 소개

- 단순회귀분석 예시: 체중(DV)과 신장(IV) 간의 인과관계 분석

체중과 신장 간의 인과관계 $\hat{Y}_i = 1.044X_i - 112.48$