알고리즘 마케팅 5강

2023. 3. 23. (목)

서울과학기술대학교 데이터사이언스학과
김 종 대

오늘의 강의

- 4차시 review
- 가설검정
- 상관관계분석
- 선형회귀분석
- 일반선형모형
- 알고리즘 마케팅

인과관계

- 인과관계(causality)의 개념
  - “어떠한 변수(원인변수)가 다른 변수(결과변수)의 값에 영향을 미치는 관계”
  - “X가 원인이 되어 결과 Y가 나타난다는 의미”

  - 대부분의 연구들은 복잡한 변수들의 구성과 다양한 인과관계를 가설로 설정하고 있기 때문에 이러한 인과관계를 확률적인 가능성으로 표현하는 것이 필요

인과관계

- 인과관계의 필요조건 (NOT 충분조건)
  - 조건 1. 동시에 발생하는 변동(concomitant variation)
    - 원인변수와 결과변수 간의 인과관계가 자료분석을 통해 얻은 관계와 일치해야 한다.
  - 조건 2. 원인변수가 아닌 다른 변수에 의한 변동(variation due to other possible causal factor)
    - 연구자가 미처 생각하지 못했던 원인변수가 존재하며, 이에 따라 우연히 실제로 존재하지 않는 인과관계가 존재하는 것 같이 보일 수 있다는 가능성을 고려해야 한다.
  - 조건 3. 시간 순서에 따른 변동(time order of occurrence)
    - 원인변수의 값이 변화하였다면, 변화한 시점이나 그 이후에 결과변수의 값이 변화해야 한다.

인과관계

- 인과관계에 관한 주요 용어
  - 실험설계(experimental design):
    - (1) 실험 단위(연구 대상이 되는 개체)를 어떻게 정할 것인지?
      - 예: 소비자 단위, 기업 단위, 부서 단위 등
    - (2) 실험 단위를 비슷한 집단으로 어떻게 구분할 것인지?
      - 예: 소비자의 소득 수준, 교육 수준 등 다른 여건이 고르게 섞여있는 집단으로 구분
    - (3) 독립변수를 어떻게 정의하고 조작할 것인지?
    - (4) 종속변수를 어떻게 측정할 것인지?
    - (5) 혼란변수를 어떻게 통제할 것인지?

가설검정

- 가설검정의 원리
  - 기본적으로 모집단을 대표하는 표본의 “표본평균”에 대한 분포를 파악하여 문제를 해결
  - 어떠한 가설을 참이라고 가정하고, 일어날 가능성이 희박한 표본평균 값에 대한 수준을 정해 놓고, 표본 자료를 통하여 계산된 표본평균이 그 수준을 벗어나면, 참이라고 가정한 그 가설을 기각하게 되는 원칙

  - 절차
    - (1) 귀무가설 및 대립가설 설정과 유의 수준 결정 → (2) 검정통계량 결정 → (3) 기각역 결정 → (4) 검정통계량의 계산 → (5) 통계적 의사결정

가설검정

- 가설검정의 절차
  - (1) 귀무가설 및 대립가설 설정과 유의 수준 결정
    - 연구가설을 통계적 가설로서 구체화: 모집단의 특성을 나타내는 모수(parameter)를 이용하여 표현
    - 귀무가설(null hypothesis) ↔ 대립가설(alternative hypothesis)
    - 일반적으로 연구의 목표는 귀무가설(예: 효과가 없다)을 기각하는 것

가설검정

- 가설검정의 절차
  - (1) 귀무가설 및 대립가설 설정과 유의 수준 결정
    - 제1종 오류(Type 1 error): True인 귀무가설을 기각하는 오류
    - 제2종 오류(Type 2 error): False인 귀무가설을 기각하지 않고 받아들이는 오류

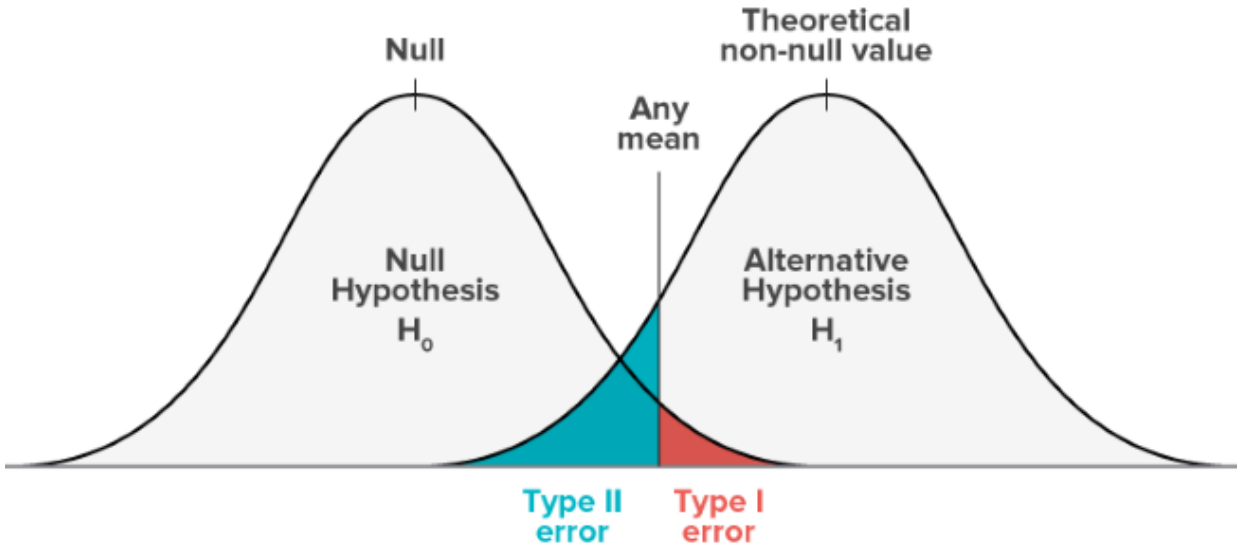
통계적 결정 \ 실제상황	H_0가 사실 (참)	H_0가 허위 (거짓)
H_0 채택	옳은결정 확률 $= 1-\alpha$	제II종오류 확률 $= \beta$
H_0 기각	제I종오류 확률 $= \alpha$	옳은결정 확률 $= 1-\beta$

유의수준

검정력

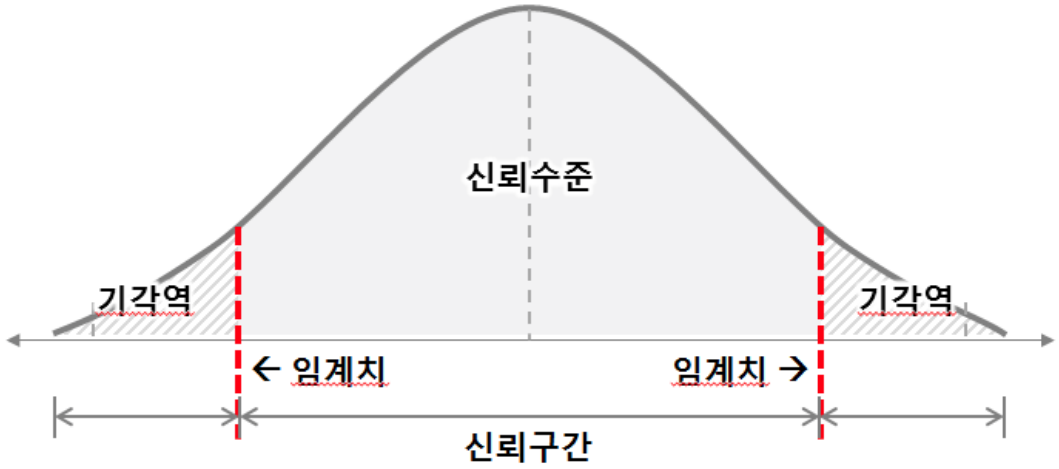
가설검정

- 가설검정의 절차
  - (1) 귀무가설 및 대립가설 설정과 유의수준 결정
    - 유의수준: 제1종 오류를 허용하는 최대 확률 (α)
      - 일어날 가능성이 희박하다고 생각하는 확률수준으로, 귀무가설을 기각하는 기준
      - 가설검정 이전에 연구자가 유의수준을 미리 결정



가설검정

- 가설검정의 절차
  - (2) 검정통계량 결정
    - 검정통계량: 수집된 자료로부터 계산된 통계량
    - 검정통계량 분포: 귀무가설이 참이라는 가정 하의 표본이 따르는 확률 분포 (예: 표준정규분포, t 분포)
  - (3) 기각역 결정
  - (4) 검정통계량의 계산
  - (5) 통계적 의사결정



가설검정

- P-value를 이용한 가설검정
  - P-value: 검정통계량 값이 귀무가설의 가정으로부터 얼마만큼 벗어나는지의 정도
    - 귀무가설이 True라고 가정하였을 때, 표본을 통해 계산된 검정통계량 값보다 귀무가설을 기각하는 방향으로 더 심하게 검정통계량 값이 관측될 확률
      - 예: 검정통계량 Z는 0으로부터 멀리 떨어져 있을수록 (Z의 절대값이 클수록) 귀무가설을 기각하는 방향에 가까워지는 것
  - P-value가 사전에 정한 유의수준보다 작을 경우, 검정통계량이 기각역에 위치 → 가설을 기각
    - 즉, 연구자는 p-value와 유의수준을 단순 비교함으로써 귀무가설 기각 여부를 판단 가능

상관관계분석

- 상관관계분석: 두 변수 X, Y의 연관도, 즉, 선형관계에 대한 정도
  - 공분산(covariance)

$$Cov(X,Y) = E[(X - E(X))(Y - E(Y))]$$

    - 변수 간 선형관계의 방향과 유무를 판단할 수 있으나, 선형관계의 정도는 파악할 수 없음.
    - 예: 측정단위에 따른 공분산 값의 변화

  - 상관계수(correlation coefficient)

$$Corr(X,Y) = \rho(X,Y) = \frac{Cov(X,Y)}{\sigma_X \sigma_Y}$$

    - X,Y의 표준편차로 나누어 준 형태로서 언제나 $-1 \leq \rho(X,Y) \leq 1$

상관관계분석

- 상관관계분석: 두 변수 X, Y의 연관도, 즉, 선형관계에 대한 정도
  - 상관관계의 계수가 절대값 1이라면?
    - X, Y는 완벽한 선형관계를 가지고 있는 것 $\rightarrow Y = aX + b$ 꼴로 표현 가능 (100% 예측 가능)

  - 공분산 또는 상관관계가 0이라면 X, Y는 무조건 독립인가?
    - 예: X, Y가 비선형관계일 경우
    - 즉, X, Y가 서로 독립이면 $Corr(X, Y) = \rho(X, Y) = 0$은 항상 성립하지만, 반대는 성립하지 않을 수 있다.

상관관계분석

- 상관관계의 검정
  - 피어슨 상관계수(Pearson’s Correlation Coefficient)
    - 상관계수가 t 분포를 따른다는 가정 하에 검정
    - X, Y가 이변량 정규분포를 따라야 한다는 가정이 필요
  - 스피어만 상관계수(Spearman’s Correlation Coefficient)
    - 비모수적 접근으로, X, Y에 대한 분포 가정이 불필요
  - 편상관계수 (Partial Correlation Coefficient)
    - 나머지 변수들에 대한 효과를 제거하고 X, Y의 상관관계를 분석

선형회귀분석

- 단순선형회귀분석
  - 두 변수 사이에 존재하는 상호의존관계를 함수 관계로 표현하여 연관성을 검정하는 통계적 분석

  - $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$
    - Y_i: i번째 관측값에 대한 종속변수의 값
    - X_i: i번째 관측값에 대한 독립변수의 값
    - β_0, β_1: 회귀 계수(regression coefficient)
    - ε_i: Y_i의 오차항을 나타내는 확률변수

선형회귀분석

- 단순선형회귀분석
  - 최소제곱법(least square method)을 통한 추정
  - $\min[\sum_{i=1}^{I} {\varepsilon_i}^2] = \min\left[\sum_{i=1}^{I}\left(Y_i - \widehat{Y_i}\right)^2\right] = \min\left[\sum_{i=1}^{I}\left(Y_i - (\widehat{\beta_0} + \widehat{\beta_1} X_i)\right)^2\right]$
    - 위 식을 최소화하는 $\widehat{\beta_0}, \widehat{\beta_1}$의 값을 구하는 것
    - 결과로서 추정된 $\widehat{\beta_0}, \widehat{\beta_1}$의 값은 독립변수가 취하는 범위에 제약을 받으며, 따라서 종속변수의 값을 추정할 때 데이터의 범위를 벗어나는 값을 독립변수의 값으로 사용하면 안 된다.
      - 참고: 데이터 수가 적을수록 예측구간이 넓어지는 성질

선형회귀분석

- 분산의 분해

$$\underbrace{\sum_{i=1}^{n}(y_i-\bar{y})^2}_{\text{SST (Total sum of squares)}}=\underbrace{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}_{\text{SSE (Error sum of squares)}}+\underbrace{\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2}_{\text{SSR (Regression sum of squares)}}$$

- 결정계수(coefficient of determination): $R^2=\frac{SST-SSE}{SST}=1-\frac{SSE}{SST}$
  - 독립변수의 수가 증가할수록 자연스럽게 커지는 경향
- 수정된 결정계수(adjusted R^2): $=1-\frac{\frac{SSE}{n-2}}{\frac{SST}{n-1}}$

선형회귀분석

- 회귀계수에 대한 t검정 ($\widehat{\beta_1}$)
  - 평균: $\mathrm{E}(\widehat{\beta_1}) = \beta_1$
  - 분산: $Var(\widehat{\beta_1}) = \frac{\sigma^2}{\sum(X_i - \bar{X})^2}$

  - 가설: $H_0: \beta_1 = 0 \ \ vs. \ H_0: \beta_1 \neq 0$
  - 검정통계량:

$$Z = \frac{\widehat{\beta_1} - \beta_1}{\sigma \Big/ \sqrt{\sum(X_i - \bar{X})^2}} \quad \rightarrow \quad \mathrm{t} = \frac{\widehat{\beta_1} - \beta_1}{s \Big/ \sqrt{\sum(X_i - \bar{X})^2}}$$

선형회귀분석

- 다중선형회귀분석
  - Y와 p개의 독립변수 $X_1,,,X_p$ 사이의 관계를 분석하는 통계적 방법론
  - 주요 이슈: 다중공선성(multicollinearity), 모형 선택(model selection)

  - $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_p X_{pi} + \varepsilon_i$
    - 단순회귀분석에서 독립변수 여러 개가 추가하여 확장된 형태
    - 검정 방법 등에서 근본적인 차이는 없음

선형회귀분석

- 선형회귀분석의 기본 가정
  - 선형성(linearity)
  - 정규성(normality)
  - 외생성(exogeneity)
  - 조건부 독립성(conditional independence)
  - 등분산성(equal-variance)

선형회귀분석

- 선형회귀분석의 기본 가정
  - 선형성: X, Y의 관계를 선형으로 가정
    - 검정: Q-Q plot 등 시각화
    - 해결 방안: 다른 형태의 모형을 고려
  - 정규성: $\varepsilon_i \sim N(0, \sigma^2)$
    - 검정: Q-Q plot, Shapiro-Wilks test, Kolmogorove-Smirnov test
    - 해결 방안: Y 등의 데이터 분포에 맞는 모형을 고려 (예: 로지스틱 회귀분석)

선형회귀분석

- 선형회귀분석의 기본 가정
  - 외생성(exogeneity) / 내생성(endogeniety)

    $Cov(\varepsilon_i, X_i) = 0$

    $Cov(\varepsilon_i, X_i) = E[\varepsilon_i X_i] - E[\varepsilon_i]E[X_i] = 0$

    $\mathrm{E}[\varepsilon_i|X_i] = 0 \quad \rightarrow \quad \mathrm{E}\big[\mathrm{E}[\varepsilon_i|X_i]\big] = E[\varepsilon_i] = 0$

    - 예: 임금(y)을 설명하는 모형에서 설명변수(x)로서 교육연수를 활용 → 이 경우, 모형에서 누락된 능력, 경험, 열의(동기부여) 등이 오차에 포함되는데, x와 강한 연관성을 가질 수도 있음.
    - 해결 방안: 도구 변수(instrumental variable), 누락 변수(omitted variable), Heckman's approach

선형회귀분석

- 선형회귀분석의 기본 가정
  - 조건부 독립성(conditional independence) / 자기상관성(auto-correlation): $Cov(\varepsilon_i, \varepsilon_j | X) = 0$
    - 예: 시계열 데이터, 패널 데이터
    - 검정: Durbin-Watson test
    - 해결 방안: 시계열 모형 또는 패널 분석 모형으로 대체
  - 등분산성: $\mathrm{Var}(\varepsilon_i) = \sigma^2$, $\mathrm{Cov}(\varepsilon_i) = \sigma^2 I$
    - 검정: Bartlett test, Fligner test, Levene test
    - 해결 방안: 가중최소제곱법(weighted least square method), 변수 변환을 통한 분산 안정화(variance stabilization)

선형회귀분석

- 다중공선성(multicollinearity)
  - 독립변수 간의 강한 상관관계로 인해 회귀분석의 결과가 왜곡되는 경우
    - 예: X_1, X_2 모두 Y에 유의한 영향을 미치는 변수이지만, X_1, X_2가 강한 상관관계를 가지고 있어 Y와의 관계가 제대로 포착되지 않는 경우
  - 다중공선성의 검정: 분산확대인자(Variance inflation factor; VFI)
    - 특정 독립변수를 대상으로 나머지 독립변수로 다중회귀분석을 실시한 후 나타나는 결정계수(R^2)의 값을 기준으로 판단

$$VIF = \frac{1}{1 - R_x^2}$$

    - 즉, 결정계수가 높으면 다른 독립변수에 의해 충분히 설명된다는 의미이므로 높은 다중공선성
    - 일반적으로 VFI 값이 5보다 높으면 문제가 의심되고, 10 이상이면 문제가 심각한 상황으로 판단

선형회귀분석

- 모형의 선택(Model selection)
  - 1. 어떤 형태의 함수를 가정하여 모형을 구성할 것인가
  - 2. 어떤 독립변수를 회귀모형에 포함시킬 것인가 (Variable selection)
    - (1) Forward selection
    - (2) Backward elimination
    - (3) Stepwise selection
      - (1)과 (2)의 혼합 형태로, (1)로 한 개 변수를 추가한 뒤, (2)에 의해 제거될 것이 있는지 판단
    - (4) All possible regression
    - (5) Best subset regression

선형회귀분석

- 마케팅과 머신 러닝
  - 마케팅 연구에서 머신 러닝 또는 딥러닝을 어떻게 활용할 것인가?
    - 유의성 검정: 계량 분석에 있어서의 확률 분포 부여 문제
    - 모형의 이론적 근거
    - 기본적으로 기존 통계학/계량경제학 모형과 어떻게 다른가?

선형회귀분석

- 마케팅과 머신 러닝
  - 송인성 (2020), 마케팅 애널리틱스에서 머신 러닝 기법 활용의 한계, *경영논집*, 54, 39-57.
    - 공학/자연과학 문제: 투입 변수의 충분성이 보장
    - 마케팅 애널리틱스 문제: 현상에 영향을 미치는 모든 변수가 자료에 포함되어 있지 않은 경우가 대부분
    - 이러한 “구조적 변수의 누락”은 모형 선택(model selection)과 예측력 향상에 부정적인 영향
    - 마케팅 문제 해결에 있어 머신 러닝 방법론의 선택은 마케팅 현상에 대한 질적 이해와 안목을 갖추고 접근해야 할 것

일반선형모형: 로지스틱 회귀분석

- 로지스틱 회귀분석
  - 반응변수가 더미변수(dummy variable)일 경우
  - 일종의 classification으로 해석 가능
- 오즈(odds)
  - 그룹 1에 속할 확률을 p, 그룹 2에 속할 확률을 1-p라고 했을 때,
  - $Odds = \frac{p}{1-p}$
  - 로지스틱 회귀분석에서 오즈를 활용하는 이유는 오즈가 취할 수 있는 값의 범위가 0~무한대
    - 음수의 값을 가지지 않는 등의 문제가 있기 때문에, 최종적으로는 Log(Odds)를 사용

일반선형모형: 로지스틱 회귀분석

- 로지스틱 회귀분석

$$Log(Odds) = \log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2$$

  - 일종의 비선형회귀분석: p와 predictor들의 관계
  - 로지스틱 함수

$$p = \frac{1}{1 + \exp[-(\beta_0 + \beta_1 X_1 + \beta_2 X_2)]}$$

  - 연결 함수(link function): p와 predictor의 관계를 이어주는 사전에 정의된 함수
    - 로지스틱 회귀분석에서는 로지스틱 함수가 연결 함수의 역할

일반선형모형: 로지스틱 회귀분석

- Independence of Irrelevant Alternatives (IIA)
  - 로지스틱 회귀분석을 할 때 반드시 확인하여야 하는 가정
    - 특히 다중로짓 모형 등에서는 반드시 확인하여야 함.
  - “선택지 k에 비해 선택지 j를 선택할 오즈(odds)가 다른 선택지의 존재 여부에 영향을 받지 않아야 한다.”
  - “범주(category)의 추가나 제외가 남아있는 범주에 대한 설명변수의 상대적 위험도(relative risks)에 영향을 주지 않아야 한다.”

일반선형모형: 로지스틱 회귀분석

- Independence of Irrelevant Alternatives (IIA)
  - 사례: 파란 버스 – 빨간 버스 딜레마
    - 1. 출근할 때 이용할 수 있는 교통 수단으로 “자가용”과 “파란 버스” 두 가지가 있다고 가정
    - 2. “자가용”을 이용할 확률이 0.8일 때, “파란 버스”에 대한 오즈 = 4
    - 3. “파란 버스”와 노선이 똑같은 “빨간 버스” 도입 (새로운 범주의 추가)
    - 4. 버스의 색깔은 사실 별 상관이 없기 때문에, 각 버스를 이용할 확률이 0.1씩으로 쪼개질 것
    - 5. “자가용”과 “파란 버스”의 오즈는 8이 되므로 IIA 가정을 위배

일반선형모형: 프로빗 회귀분석

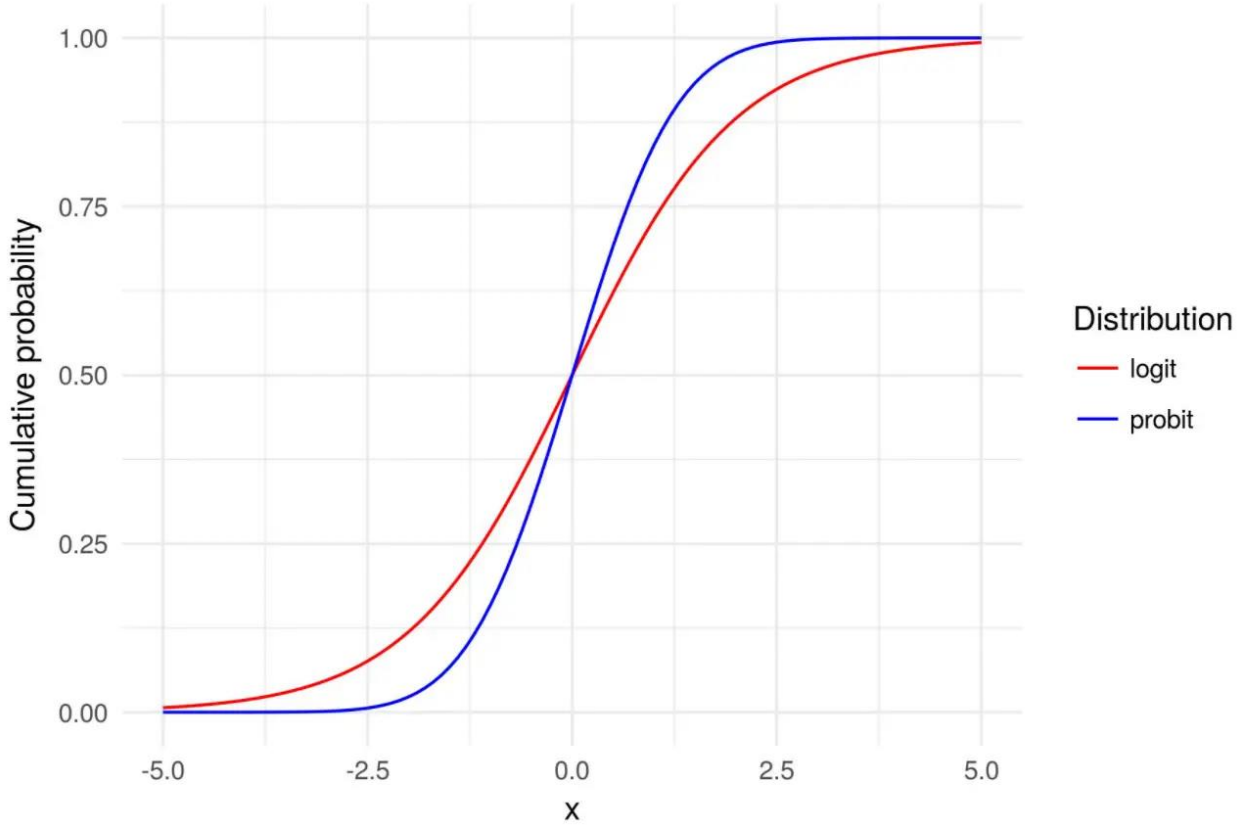
- 프로빗 모형(Probit model)
  - 로지스틱 모형의 한계를 극복하기 위해 대안적 연결 함수로서 프로빗 함수를 제안

$$p = \Phi(\beta_0 + \beta_1 X_1 + \beta_2 X_2)$$

  - 표준정규분포의 누적분포함수를 이용하여 모형화
  - 종속변수 p에 미치는 한계 효과(marginal effect)는 설명변수에 대해 편미분함으로써 도출 가능

일반선형모형: 프로빗 회귀분석

- 프로빗 모형(Probit model)
  - 로지스틱 함수와 프로빗 함수의 형태는 유사하지만, p=0.5 근처에서 로지스틱 함수가 더 완만
  - 데이터가 p=0.5에 몰려 있으면 두 모형의 차이가 거의 없지만, 그보다 멀리 떨어진 곳에 많이 있을 경우 두 모형의 분석 결과가 상이할 수 있음.



일반선형모형: 포아송 회귀분석

- 포아송 회귀분석
  - 종속변수가 0, 1, … 와 같은 가산자료(count data)일 경우
  - 사건(event)의 기대 빈도(expected frequency)를 모형화
  - 가정
    - 1. 동일한 길이의 두 구간에서 사건 발생의 확률은 동일
    - 2. 어떤 구간에서 사건이 발생할 확률은 다른 구간에서 사건이 발생할 확률과 독립
    - 3. 매우 짧은 시간 내에 두 개 이상의 사건이 발생할 확률은 0
  - 포아송 분포: $\Pr(\mathrm{Y} = \mathrm{y}) = \frac{\lambda^{y}\exp(-\lambda)}{y!}$

인과 추론 (Causal Inference)

- 인과 추론이란?
  - 일반적인 회귀 분석 등만으로는 변수 간의 인과관계를 추정하는 것이 불가능
    - 랜덤화 여부, 누락 변수에 따른 오차 존재 등
  - 인과 추론이란 일반적인 데이터(예: cross-sectional data)를 이용하여 준실험(quasi-experiment) 환경을 조성/모형화함으로써 변수 간 인과관계를 추론하고자 하는 접근 방식
    - 랜덤화 실험설계와 같은 통제(대조군을 두되 여러 외생 변수들을 조절)는 불가능하겠지만, 유사하게 조작(Treatment를 받은 그룹과 그렇지 않은 그룹)하여 실험을 설계

인과 추론 (Causal Inference)

- 인과 추론이란?

$$Y_i = D_i Y_i^1 + (1 - D_i) Y_i^0$$

  - Y_i: 실제 결과 (actual outcome)
  - D_i: Treatment를 받았는지 여부, 받았으면 1, 받지 않았으면 0
  - Y_i^1: $D_i = 1$일 경우의 potential outcome
  - Y_i^0: $D_i = 0$일 경우의 potential outcome

인과 추론 (Causal Inference)

- 인과 추론이란?

$$Y_i = D_i Y_i^1 + (1 - D_i) Y_i^0$$

  - Unit i 단위의 treatment effect (= unit-specific causal effect): $\delta_i = Y_i^1 - Y_i^0$
  - Average treatment effect (ATE)
    - 위와 같은 unit 단위의 효과를 계산하는 것이 불가능
    - $ATE = E[\delta_i] = E\left[Y_i^1 - Y_i^0\right] = E\left[Y_i^1\right] - E\left[Y_i^0\right]$

인과 추론 (Causal Inference)

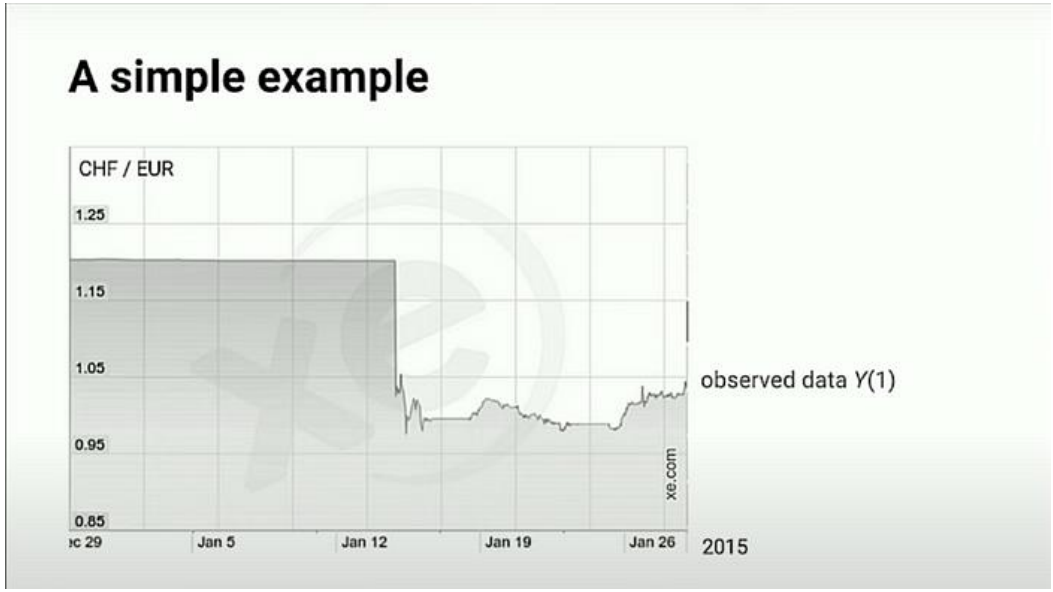
- 인과 추론이란?
  - Average treatment effect for the treatment group (ATT)
    - $ATT = E[\delta_i | D_i = 1] = E\left[Y_i^1 - Y_i^0 | D_i = 1\right] = E\left[Y_i^1 | D_i = 1\right] - E\left[Y_i^0 | D_i = 1\right]$
  - Average treatment effect for the untreatment (control) group (ATU)
    - $ATT = E[\delta_i | D_i = 0] = E\left[Y_i^1 - Y_i^0 | D_i = 0\right] = E\left[Y_i^1 | D_i = 0\right] - E\left[Y_i^0 | D_i = 0\right]$

인과 추론 (Causal Inference)

- 이중차분법(Difference-in-Differences; DID)
  - 정책 시행 또는 사건 발생의 전후를 비교하는 기본적인 모형
  - 가정
    - 1. OLS 가정이 기본적으로 모두 충족 필요
    - 2. Parallel trend assumption: Treatment group과 control group의 종속변수가 treatment가 기본적으로 평행한 추세를 보일 것
      - Treatment가 없을 경우, 두 그룹의 추세가 거의 동일 → 관측되지 않는 차이의 최소화
    - 3. Stable composition of groups: 두 그룹의 구성이 시간에 따라 변동하지 않을 것
    - 4. No spillover effects: Treatment의 효과가 control group으로 전이되지 않을 것 (예: 소셜 네트워크)

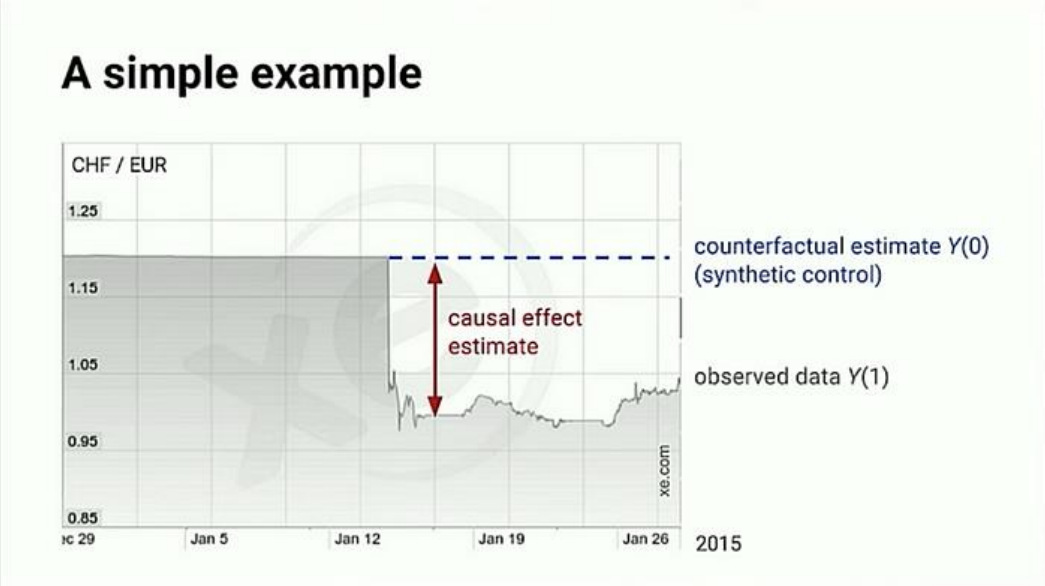
인과 추론 (Causal Inference)

- Synthetic Control Method (SCM)
  - 특정 사건의 효과를 어떻게 추정할 것인가?: Treatment가 없을 경우의 counterfactual estimate 예측!



인과 추론 (Causal Inference)

- Synthetic Control Method (SCM)
  - 특정 사건의 효과를 어떻게 추정할 것인가?: Treatment가 없을 경우의 counterfactual estimate 예측!



인과 추론 (Causal Inference)

- Synthetic Control Method (SCM)
  - 특정 사건의 효과를 어떻게 추정할 것인가?: Treatment가 없을 경우의 counterfactual estimate 예측!

Causal inference and potential outcomes

Unit	Treatment status T_i	Outcome under treatment $Y_i(1)$	Outcome under no treatment $Y_i(0)$	Covariates X_i
1	1	✓	estimate	✓
2	1	✓	estimate	✓
3	0	estimate	✓	✓
4	0	estimate	✓	✓

⇒ causal effect estimate

알고리즘 마케팅

- 알고리즘 마케팅의 중요성
  - 마케팅의 역사: 특정 비즈니스를 최적화하기 위한 원리, 기술, 프랙티스의 진화 과정
  - 수학적, 통계학적, 공학적 접근 방식을 통해 마케팅 문제를 해결하고자 하지만,
    - 한계: 데이터의 불완전성, 현실 마케팅 환경의 복잡성, 비즈니스 프로세스의 융통성 부족 등
  - 그럼에도 또 한 가지 트렌드의 등장: 디지털 마케팅 채널의 진화
    - 타깃 광고, 온라인/오프라인 매장에서의 동적 가격 책정, 추천 서비스, 온라인 광고 등
    - 고객은 개인화된 경험을 원하고, 따라서 수백만 개의 서로 다른 의사 결정이 필요
  - 전례 없는 수준의 자율성, 규모, 깊이를 갖춘 의사 결정과 이를 위한 마케팅 시스템 구축 필요
  - 마케팅 자동화와 통합적 관리를 위한 다양한 비즈니스 솔루션 필요

알고리즘 마케팅

- 알고리즘 마케팅의 주제
  - 전통적인 마케팅: 마케팅 믹스(marketing mix; 4P) 조합의 변형이 중심
    - 전략: 회사가 제공하는 가치를 정의하고 마케팅 프로세스를 위한 방향을 정하는 최상위 단계의 장기적인 비즈니스 의사 결정
    - 프로세스: 회사의 연속적인 운영을 지원하기 위한 전술적 결정에 집중하는 전략의 실행
  - 알고리즘 마케팅: 마케팅 소프트웨어 시스템 안에서 비즈니스 목표를 자동으로 결정할 수 있을 정도로 자동화된 마케팅 프로세스
    - 마케팅 전략 수립과 프로세스 실행을 위한 데이터 기반 방법론의 제공이 목적

알고리즘 마케팅의 역사

- 마케팅 과학
  - 사례: 온라인 광고
    - 스팸 메일, 단순한 배너 광고

    → 최적화 알고리즘을 바탕으로 한 개인화 광고

경향신문

PICK

야놀자·여기어때·부킹닷컴…맨 위 숙박상품 추천 아닌 광고

입력 2023.03.21. 오후 2:28 · 수정 2023.03.21. 오후 3:00 기사원문

13 13

[숙박플랫폼별 광고 상품 표시 실태]

구분	광고 유무	광고 비율*	광고 표시	광고 유형
야놀자	O	93개/100개(93%)	'AD' 클릭 시 광고 안내 화면이 나타남	'야놀자초이스', '지역초이스 플러스', '지역초이스' 등

알고리즘 마케팅의 역사

- 마케팅 과학
  - 사례: 항공사 매출 관리
    - 1978년 미국 내 항공 관련 규제가 풀리면서 항공사들이 가격과 항로를 자유롭게 조정 가능
    - 각 항공사들이 고객 데이터베이스 관리에서의 최적화를 통해 서비스를 제안
    - 성수기 고가 전략/비성수기 저가 전략 등을 이용해 매출 극대화

  - 이처럼 마케팅 과학은 여러 산업에 걸쳐 다양하게 응용 가능
  - 1960년대 마케팅 믹스의 조합이 강조된 것도 마케팅 과학의 발달이 그 배경

프로그램 기반 서비스

- 프로그램 기반 마케팅 시스템
  - 가격 책정이나 프로모션 관리와 같은 특정 비즈니스 프로세스를 구현하는 하나 이상의 서비스 제공자
  - 여섯 가지 분야의 프로그램 기반 서비스에 대해 다룰 것
    - 1. 판매 촉진(promotion)
    - 2. 광고(advertising)
    - 3. 검색(search)
    - 4. 추천(recommendation)
    - 5. 가격 책정(pricing)
    - 6. 상품 구성(assortment)

기술적, 예측적, 처방적 분석

- 기술적 분석
  - 데이터 요약, 데이터 품질 측정, 관련성 분석
  - 예: 판매량 데이터 분석, 마켓 바스켓 분석(특정 제품과 같이 구매되는 다른 제품을 찾아내는 분석)
- 예측적 분석
  - 관찰된 데이터 또는 결과 이전의 확률을 사용해 가능한 결과들을 예측
  - 예: 수요 예측, 프로모션에 따른 고객의 구매 확률 예측
- 처방적 분석
  - 최적 의사 결정을 위해 의사 결정과 미래의 경과 사이의 의존성을 모델링
  - 예: 가격 할인 전략이 가져다 줄 이익의 예측을 통해 최적 가격 책정

경제적 최적화

- 경제적 모델 설계
  - 경제적 모델: 비즈니스 목표에 관한 목표 함수와 제약 조건 등을 수식화
    - 예: $p_{opt} = argmax\, \pi(Q, p, x)$
  - 비즈니스 목표: 최적화할 수 있는 수리적 지표로 표현
    - 예: 회사의 이익과 고객의 효용 사이에서 절충해야 하는 경우
  - 데이터 수집: 모형의 복잡성 등에 영향
  - 모델의 세분화 수준

경제적 최적화

- 라그랑지안 함수와 최적화
  - 제한조건(constraint)이 있는 최적화 문제
    - M개의 제한조건식을 모두 만족시키면서 $f(x)$를 극대화시키는 x를 찾는 문제

$$x^* = argmin\, f(x)$$
$$s.t.\, g_j(x) = 0 \quad (j = 1, \dots, M)$$

경제적 최적화

- 라그랑지안 함수와 최적화
  - N+M개의 연립방정식을 풀어 아래의 미지수를 계산

$$h(x,\lambda) = h(x_1, x_2, \ldots, x_N, \lambda_1, \ldots, \lambda_M) = f(x) + \sum_{j=1}^{M} \lambda_j g_j(x)$$

$$\frac{\partial h}{\partial x_1} = \frac{\partial f}{\partial x_1} + \sum_{j=1}^{M} \lambda_j \frac{\partial g_j}{\partial x_1} = 0$$

$$\frac{\partial h}{\partial x_2} = \frac{\partial f}{\partial x_2} + \sum_{j=1}^{M} \lambda_j \frac{\partial g_j}{\partial x_2} = 0$$

$$\vdots$$

$$\frac{\partial h}{\partial x_N} = \frac{\partial f}{\partial x_N} + \sum_{j=1}^{M} \lambda_j \frac{\partial g_j}{\partial x_N} = 0$$

$$\frac{\partial h}{\partial \lambda_1} = g_1 = 0$$

$$\vdots$$

$$\frac{\partial h}{\partial \lambda_M} = g_M = 0$$

$x_1, x_2, \ldots, x_N, , \lambda_1, \ldots, \lambda_M$

머신 러닝: 지도학습

- 지도학습 (감독학습; Supervised learning)
  - X를 이용해 Y 값을 예측해주는 함수의 학습, 즉, $P(Y|X)$의 분포를 학습하는 것
  - 학습 과정을 안내해주는 응답 변수 Y가 존재한다는 의미에서 지도/감독 학습
    - 분류(Classification): 응답 변수가 유한한 범주/등급인 경우
    - 회귀분석(Regression): 응답 변수가 무한한 연속 변수인 경우
  - 모수 모형 (Parametric model): 데이터의 분포가 몇 개의 모수에 의해 함수의 형태로 정의
    - 예: 정규 분포에 기반을 둔 선형회귀분석
  - 비모수 모형 (Nonparametric model)
    - 예: k-nearest Neighbor (kNN) 알고리즘

머신 러닝: 지도학습

- 최대 가능성 추정 (Maximum Likelihood Estimation; MLE)
  - 가능성 함수 (Likelihood function)

$$L(\theta) = \Pr(y|X, \theta)$$

    - 모수 θ로 구성된 모형에 의해 정의된 분포에서 학습 데이터가 관찰될 확률(= 가능성)
    - 가능성 함수를 극대화하는 모수의 추정치를 찾는 것이 목표
    - Log 값을 취한 log-likelihood가 계산하기가 더 용이

$$\theta_{ML} = argmax \log[\Pr(y|X, \theta)] = argmin \ -\log[\Pr(y|X, \theta)]$$

머신 러닝: 지도학습

- 선형 모델
  - 선형회귀분석
  - 로지스틱 회귀분석 또는 이진 분류 (binary classification)
  - 로지스틱 회귀분석 또는 다항 분류 (multi classification)

$$\Pr(Y = c|x) = \frac{\exp(X_c\beta)}{\sum_i \exp(X_i\beta)}$$

머신 러닝: 지도학습

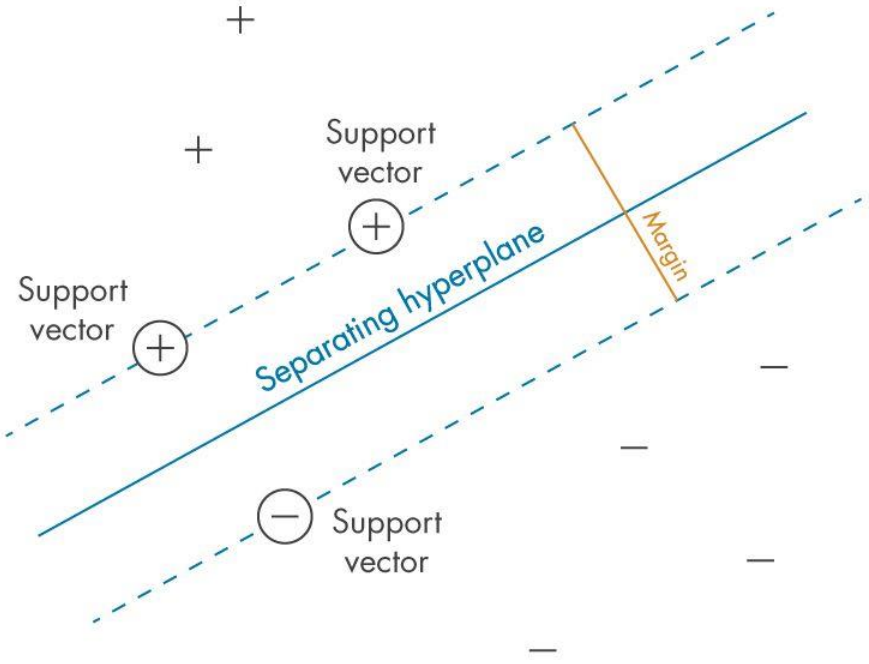
- 나이브 베이즈 분류 (Naïve Bayes Classification)
  - 텍스트 분류 등에서 유용하게 활용

$$\Pr(Y = c|x) = \frac{\Pr(x|Y = c)\Pr(Y = c)}{\Pr(x)}$$

$$Y^* = argmax \Pr(Y = c|x) = argmax \Pr(x|Y = c)\Pr(Y = c)$$
$$= argmax \Pr(Y = c)\prod_{i=1}^{m}\Pr(x_i|Y = c)$$

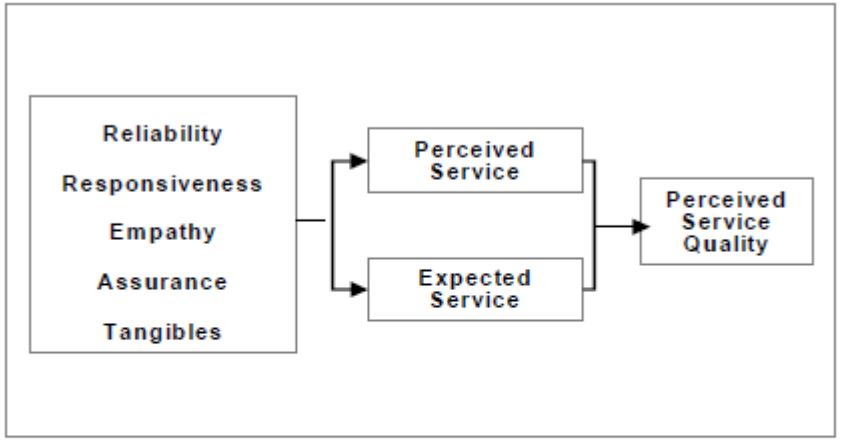
머신 러닝: 지도학습

- 커널 기법 추정
  - 비선형 모형의 하나
  - 특징 공간(feature space)들을 더 높은 차원의 특징 공간으로 변환시키는 방식으로 하나 또는 그 이상의 기존 특징의 비선형 함수로 표현되는 차원을 추가
  - 분류 문제 등에 있어 특징(feature)의 조합에 관하여 상당한 유연성을 제공
  - 예: 서포트 벡터 머신 (Support vector machine; SVM)



발표 논문

- 서비스 품질의 측정: SERVQUAL 모형 (Parasuraman et al., 1988)
  - 지각된 서비스 품질의 개념을 '서비스의 우수성과 관련된 소비자의 전반적인 판단이나 태도'로 정의
  - 포커스 그룹 인터뷰(FGI)를 통해 고객이 서비스 품질을 평가하는 10가지 차원을 추출한 뒤, 이중 중복되는 차원 등을 정리하여 최종적으로 5가지 차원 22개 항목을 추출
  - 신뢰성(reliability), 응답성(responsiveness), 공감성(empathy), 확신성(assurance), 유형성(tangibles)



자료원: Parasuraman, Zeithaml, and Berry(1988)

〈그림 1〉 SERVQUAL 모형

발표 논문

- “Emotion and Decision-making”
  - Lerner, J. S., Li, Y., Valdesolo, P., & Kassam, K. S. (2015). Emotion and decision making. Annual review of psychology, 66, 799-823.
  - The literature reveals that emotions constitute potent, pervasive, predictable, sometimes harmful and sometimes beneficial drivers of decision-making.
  - This paper proposes the emotion-imbued choice model, which accounts for inputs from traditional rational choice theory and from newer emotion research, synthesizing scientific models.
    - Bounded rationality (Herbert Simon, 1967): to refine existing normative models of rational choice to include cognitive and situational constraints

발표 논문

- Lerner, J. S., Li, Y., Valdesolo, P., & Kassam, K. S. (2015). Emotion and decision making. Annual review of psychology, 66, 799-823.

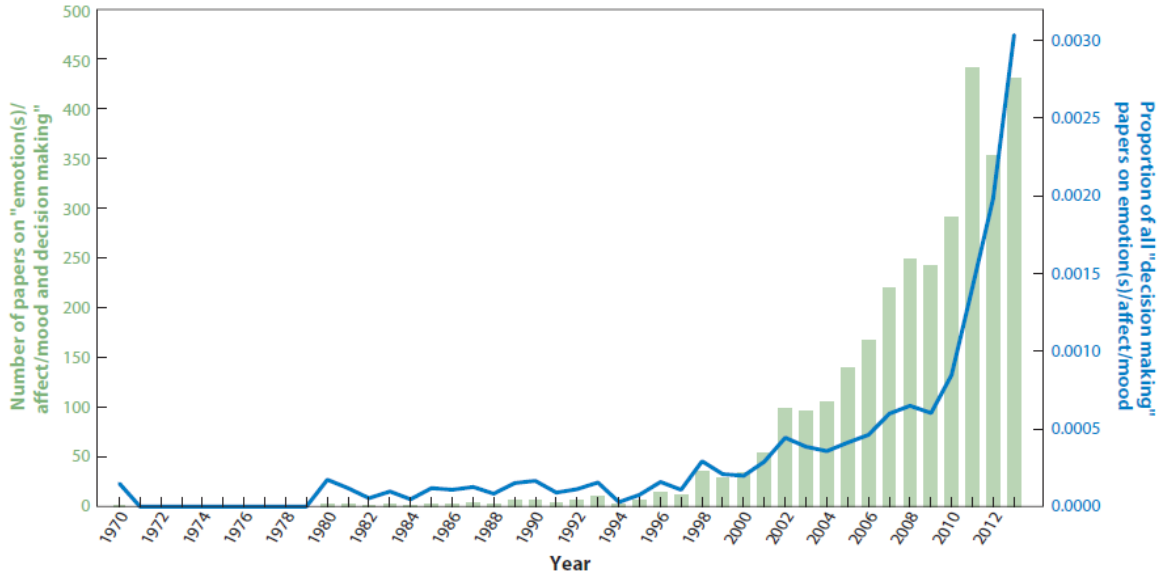
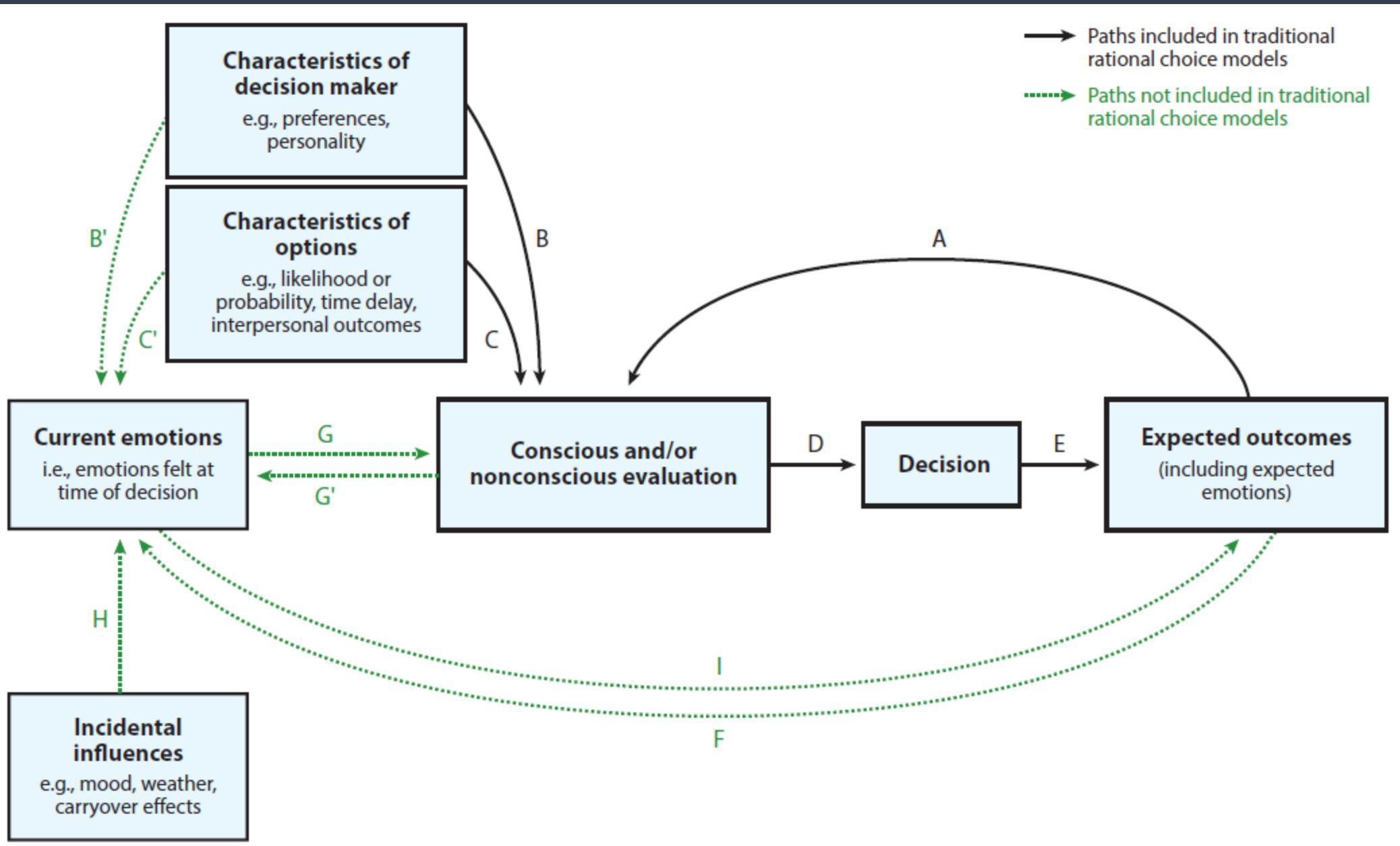


Figure 1

Number of scholarly publications from 1970 to 2013 that refer to "emotion(s)/affect/mood and decision making" (*green bars*) and proportion of all scholarly publications referring to "decision making" that this number represents (*blue line*).

발

- of


Paths included in traditional rational choice models
Paths not included in traditional rational choice models
Characteristics of decision maker
e.g., preferences, personality
Characteristics of options
e.g., likelihood or probability, time delay, interpersonal outcomes
B'
C'
B
C
A
Current emotions
i.e., emotions felt at time of decision
G
G'
Conscious and/or nonconscious evaluation
D
Decision
E
Expected outcomes
(including expected emotions)
H
I
F
Incidental influences
e.g., mood, weather, carryover effects

개인 프로젝트 안내

- **과제 목적**
  - 석사 과정: 수업 내용과 추가적인 문헌 연구를 바탕으로 구체적인 연구 주제 제안
  - 박사 과정: (석사 과정 목적) + 데이터를 통한 분석 결과 제시

- **연구 주제**: 본인이 관심 있는 분야
  - 단, [알고리즘 마케팅]의 수업 내용이 어느 단계에서든 반영되어야 할 것
- **데이터**: 자유
- **방법론**: 자유

개인 프로젝트 안내

- **무엇이 흥미로운 연구 주제인가?**
  - **Theoretical implication**
    - 산업공학, 경영학 등 관련 문헌에 이론적으로 기여할 수 있는가?
    - 해당 주제가 최근 학계로부터 관심을 받고 있는가?
    - **Key literature의 제시 + Research gap의 제시**
  - **Practical implication**
    - 기업이 실질적으로 수익을 높이는 데 기여할 수 있는가?
    - 기업 등이 실무적으로 활용할 가치가 높은가?

개인 프로젝트 안내

- **프로젝트 주제 발표**
  - 일시: **4월 6일(목)**
  - 내용
    - 석사 과정: 문헌 연구 요약 발표 (예: 본인이 follow할 key paper의 요약 발표)
    - 박사 과정: 주제 개요 + (예상되는) 시사점 + 주요 선행 연구 + 데이터 + 방법론
  - 분량: 발표 시간 10~15분

Lecture 6 Introduction

- **알고리즘 마케팅 2~3장**
  - 머신 러닝: 표현 학습 (Representation Learning)
    - 주성분분석
    - 클러스터링
  - 고객 선택 모형 (Choice Model)
    - 다항 로짓 모형
    - 생존 분석 모형
  - 프로모션과 광고에 대한 마케팅 모형