Naïve Bayes Classifier

Unit 1 확률기초.

• 조건부 확률 $P(B|A) = \frac{P(A \cap B)}{P(A)}$

- 독립 : $P(A \cap B) = P(A)P(B)$
- 조건부 독립 : $P(A, B|C) = P(A|C)P(B|C)$
  - ↳ 한 사건이 일어났다는 가정 하에서, 서로 다른 두 사건은 독립인 상황.

Unit 2 베이즈정리.

$$P(H|D) = \frac{P(D|H)P(H)}{P(D)}$$

likelihood, Prior, Posterior, Normalizing constant

↳ 두 확률변수의 prior와 posterior 사이의 관계를 나타내는 정리

: prior로부터 posterior를 구하고자 한다.

- Prior : 사전확률, 과거의 경험을 토대로 나름대로 추정한 parameter H의 확률
- Posterior : 사후확률. 관측결과 사건 D가 일어난 조건 하의 H의 확률
- Likelihood : 모델 파라미터 H를 바탕으로하는 관측결과 사건 D의 확률
- Normalizing Constant : 사건 D의 발생가능성. 보통 상수로 취급.

Unit 3. Naive Bayes Classification

• 가정 : 종속변수 Y가 주어졌을 때, 입력변수들이 모두 독립이다 (조건부 독립 가정)

$$f^*(x) = \underset{Y=y}{\operatorname{argmax}} P(X=x|Y=y)P(Y=y)$$

$$\approx \underset{Y=y}{\operatorname{argmax}} P(Y=y) \prod_{1 \le i \le d} P(X=x_i|Y=y_i)$$

↳ 장점 ① 알아야 할 파라미터의 수가 줄어든다.
② feature들의 곱으로 바뀌면서 계산이 수월해진다.
③ text에서 강점.
④ input이 연속형일 때도 사용가능 (가우시안 나이브베이즈)

↳ 단점 ① 희귀한 확률이 나왔을 때
② 조건부 독립 가정 자체가 비현실적.

• 라플라스 스무딩 $P_{LAP} = \frac{C(x)+1}{\sum x[C(x)+1]}$

↳ likelihood $P(D|H)$가 0이 되는것을 방지하도록 최소한의 확률을 정해줌.