Naïve Bayes Classifier

unit 01) 확률 기초

- 확률 : 특정한 사건이 일어날 확률
- 조건부 확률 : 어떤 사건이 일어난 조건 하에서, 다른 사건이 일어날 확률

$$P(B|A) = \frac{P(A \cap B)}{P(A)} \rightarrow \text{사건 A가 일어났을 때, 사건 B가 일어날 확률}$$

$$P(B|A)P(A) = P(A|B)P(B)$$

- 독립 : 한 사건이 일어날 확률에 영향을 미치지 않는 상태

$$P(A \cap B) = P(A)P(B)$$

- 조건부독립 : 한 사건이 일어났다는 가정하에 서로 다른 두 사건은 독립인 상황

$P(A, B|C) = P(A|C)P(B|C)$ → C 사건이 일어났을 때, 사건 A가 일어날 확률은 사건 B확률에 영향 x

unit 02) 베이즈 정리

- 두 확률 변수의 사전 확률(Prior)과 사후확률(Posterior) 사이의 관계를 나타내는 정리
  → 사전 확률(Prior)로 부터 사후확률(Posterior)을 구하고자 한다.

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

Likelihood, Prior, Posterior, Normalizing Constant

1) Posterior : 사후확률
→ 직접적으로 관찰이 불가능하지만, 계산을 통해 사후적으로 알수 있는 확률
→ 관측 결과 사건 D가 일어난 조건 하의 파라미터(H) 확률

2) Prior : 사전 확률
→ 과거의 경험을 토대로 나름대로 추정한 파라미터(H) 확률

3) Likelihood
→ 사전 확률의 과거 경험을 잘 설명하는 정도
→ 모델 파라미터(H)를 바탕으로 하는 관측결과 사건(D)의 확률

4) Normalizing Constant
→ 사건 D의 발생 가능성
→ 우리가 관심있는 H와 무관한 값, 보통 상수로 생각하고 무시하고 계산

* 베이즈 정리의 한계점 : 계산량이 너무 많아짐
  ⇒ 해결책으로 조건부 독립을 가정

unit 03) Naive Bayes classification

- 가정 : 종속변수 (Y)가 주어졌을 때, 입력 변수들이 모두 독립이다. (조건부 독립 가정)
  ↳ 결과가 주어졌을 때, 예측 변수 벡터의 정확한 조건부 확률은 각 조건부 확률의 곱으로 충분히 잘 추정할 수 있다는 단순한 가정을 기초로 한다.

$$f^*(x) = \text{argmax}_{Y=y} P(X=x \mid Y=y) P(Y=y)$$

$$\approx \text{argmax}_{Y=y} P(Y=y) \prod_{1 \le i \le d} P(X=x_i \mid Y=y) \quad \rightarrow \text{Naive Bayes를 적용한 식}$$

* 특징

- 알아야 할 파라미터의 수가 대폭 줄어들게 된다. $P(X=x \mid Y=y) \Rightarrow dK$
- 피쳐들의 곱으로 바뀌면서 계산이 수월해진다.

* 라플라스 스무딩

→ Likelihood 가 0 이 되는 것을 방지하기 위해 최소한의 확률을 정해주는 것

$$P(x \mid c) = \frac{\text{count}(x,c)+1}{\sum_{x \in v} \text{count}(x,c) + V}$$; 분자에 1 더하고 분모에 입력변수 개수 V 더하는 꼴

$$P_{LAP} = \frac{C(x)+1}{\sum x [C(x)+1]}$$: 실제보다 한 번씩 더 관찰되었다고 가정하기

* Naive Bayes 장단점

· 장점

1) 입력 공간의 차원이 높을 때 유리
2) 텍스트에서 강점
3) input이 연속형일 때도 사용가능 (가우시안 나이브 베이즈 활용)

· 단점

1) 희귀한 확률이 나왔을 때 (라플라스 스무딩)
2) 조건부 독립이라는 가정 자체가 비현실적