투빅스 3주차 정리 Naïve Bayes Classifier 정리

Unit 02 : 베이즈 정리

베이즈 정리란?: 두 확률 변수 (prior) 과 사후확률 (posterior) 사이의 관계를 나타내는 정리

↳ 사전확률 (prior) 로부터 사후확률 (posterior) 을 계산하려한다.

(사건에 A에 대한 정보를 바탕으로 $P(A)$와 $P(B|A)$를 알고있다면, 이를 활용해 $P(A|B)$를 추론할 수 있게 해주는 것.)

↳ : 조건확률!

$P(B)$를 생각해보면, B라는 결과는 A의 부분집합으로 볼수있고, $P(B)=P(A\cap B)$ 이므로,

A의 근원사건을 A_i 라 한다면 $P(B)=P(A\cap B)=P(A_1\cap B)+P(A_2\cap B)+P(A_3\cap B)$ $(i=1,2,3)$

$\therefore$ 이는 이렇게 정리할 수 있다: $P(B)=P(B|A_1)P(A_1)+P(B|A_2)P(A_2)+P(B|A_3)P(A_3)$

$$\therefore P(A_i|B)=\frac{P(A_i\cap B)}{P(B)}=\frac{P(B|A_i)P(A_i)}{P(B|A_1)P(A_1)+P(B|A_2)P(A_2)+P(B|A_3)P(A_3)}$$

사전확률의 '과거 경험을 잘 설명한정도' : likelihood

모델 파라미터 (H)를 바탕으로 하는 관측결과 사건 (D)의 확률.

$$\underbrace{P(H|D)}_{\text{Posterior}}=\frac{\overbrace{P(D|H)}^{\text{likelihood}}\overbrace{P(H)}^{\text{Prior}}}{\underbrace{P(D)}_{\text{Normalizing Constant}}}$$

Prior : 사후확률.
- 관측결과 D가 일반 조건하의 파라미터의 확률

Posterior
- 사후확률
- 과거의 경험을 토대로 나름대로 추정한 파라미터의 확률.
- 특정적 (데이터, 실험 등)으로 관찰 및 표현이 불가능 하지만, 계산을 통해 '사후적으로' 알 수 있는 확률.

Normalizing Constant : 사건 D의 발생가능성.
- 모델 파라미터 (H)와 무관한 관측결과 (D) 자체의 확률.
- 보통 상수로 생각하고 무시하고 계산.

⇒ but 이런식으로하면 ... 변수가 늘어날수록 연산량이 기하급수적으로 많아짐. $(2^d-1)^k$ 개 (너무 많기하네...)

⇒ ∴ solution : 조건부독립을 가정 ↳ 그게바로...

Unit 3: Naïve Bayes Classification

Naïve Bayes Classification

↳ 계산량이 기하급수적으로 많아지는것을 막기위해 종속변수(Y)가 주어졌을때, 입력 변수들이 모두 독립이라고 가정한다.

예측한다의 정확한 조건부확률은 각 조건부 확률의 곱으로 충분히 추정할 수 있다는 단순한 가정을 의미.

$$\hat{y}^*(x)=\arg\max_{y=q} P(X=x|Y=q)P(Y=q)$$

$$\propto \arg\max_{y=q} P(Y=q)\prod_{1\le i\le d} P(X=x_i|Y=q)$$

} 알아야할 파라미터 수가 대폭 줄어들게 된다. $(2^d-1)^k$ 개 → dk 개
feature들의 곱으로 바뀌면서 계산이 수월해진다.

라플라스 스무딩

↳ 사후확률이 0일때, likelihood 가 0이 되면서 결과값이 0이 되는 경우발생

이러한 경우를 방지하기위해 최소한의 확률을 정해주자.

$$P_{LAP}=\frac{c(x)+1}{\sum_x[c(x)+1]},\quad P(x|c)=\frac{count(x,c)+1}{\sum_{x\in V} count(x,c)+V}$$

나이브베이즈 장/단점.

장점.	단점
- 입력공간의 차원이 높을때 유리	- 희귀한 확률이 나왔을때 (라플라스 스무딩)
- Text에서 강점.	- 조건부독립이라는 가정이 비현실적.
- 가우시안 나이브 베이즈를 활용하면 input이 연속형? ok	