Think Bayes

2017.08.25 토

Chapter 1 베이즈 이론

1.1 조건부 확률

- 베이지안 통계의 기본 개념은 베이즈 이론 : 조건부 확률 -> 베이즈 이론 -> 베이지안 통계 순서로 학습
- $p(A \mid B)$: B라는 조건이 주어졌을 때의 A가 참일 확률

1.2 결합 확률

- 두 가지에 대한 조건이 참이라고 말하는 방법
- $p(A\ and\ B)$: A와 B가 모두 참인 확률
- $p(A\ and\ B) = p(A) * p(B)$ #항상 참은 아니다
- (ex.1) 두 개의 동전 던지기 예시
  - $p(A)$: 천 번째 동전이 앞면이 나올 확률, $p(B)$: 두 번째 동전이 앞면이 나올 확률
  - $p(A) = p(B) = 0.5$ 이고 이에 따라 $p(A\ and\ B) = p(A) * p(B) = 0.25$
  - 하지만 이 식은 A와 B가 독립일 경우에만 만족.
  - 즉, 첫 번째 사건의 결과를 안다고 해도 구 번째 사건의 확률이 바뀌지 않아야 함
  - 수식으로 $p(A \mid B) = p(B)$
- (ex.2) 독립적이지 않은 예시.
  - $p(A)$ 오늘 비가 올 확률, $p(B)$ 내일 비가 올 확률
  - 만약 오늘 비가 왔다는 것을 알고 있다면 내일도 비가 올 확률이 좀 더 높다고 가정하면
  - $p(A \mid B) > p(B)$ 이다.
  - 일반적으로 결합 확률은 어떤 A와 B의 경우에도 다음과 같이 나타난다.
  - $p(A\ and\ B) = p(A) * p(B \mid A)$
  - 따라서 어떤 주어진 날에 비가 올 확률이 0.5라고 해도 연달아 이틀간 비가 올 확률은 0.25가 아니라 좀 더 높을 것

1.3 쿠키 문제

- 쿠키 두 그릇. 첫 번째 그릇에는 바닐라 쿠키 30개, 초콜렛 쿠키 10개. 두 번째 그릇에는 두 가지 쿠키 20개씩.
- 어떤 그릇인지 보지 않고 한 그릇에서 임의로 쿠키를 집었는데 바닐라 쿠키였다.
- 이때 이 바닐라 쿠키가 그릇 1에서 나왔을 가능성은? $p\left(그릇1 \mid 바닐라쿠키\right)$
- "그릇 1에서 바닐라 쿠키가 나올 확률은?" -> $p\left(바닐라쿠키 \mid 그릇1\right) = 3/4$
- 하지만 $p(A \mid B)$와 $p(B \mid A)$ 는 같지 않다.
- 이 중 하나로 나머지 다른 하나를 구할 수 있는 방법이 베이즈 이론

1.4 베이즈 이론

- 어떤 사건 A와B는 교환 가능 : $p(A\ and\ B) = p(B\ and\ A)$

- 결합확률 : $p(A\ and\ B) = p(A) * p(B \mid A)$,A, B교환가능, $p(B\ and\ A) = p(B) * p(A \mid B)$
- 이 둘을 결합하면 $p(B) * p(A \mid B) = p(A) * p(B \mid A)$
- 양변을 $p(B)$로 나눈다 $p(A \mid B) = \frac{p(A)*p(B|A)}{p(B)}$: 베이즈 정리
- 쿠키 문제 해결에 응용
  - 1번 그릇에서 쿠키를 꺼내는 사건 B_1 , 바닐라를 꺼내는 사건 V
  - $p(B_1 \mid V) = \frac{p(B_1)*p(V|B_1)}{p(V)}$: 바닐라 쿠키를 꺼냈는데 그릇1에서 꺼냈을 확률
  - $p(B_1) = 1/2$: 그릇 1을 골랐을 확률
  - $p(V \mid B_1) = 3/4$: 그릇1에서 바닐라 쿠키를 고를 확률
  - $p(V) = 5/8$: 각 그릇에서 바닐라 쿠키를 고를 확률. 각 그릇 고를 확률 동일(1/2)하고 그릇에 동일한 수의 쿠키(40개씩)가 들어 있으므로 각 쿠키를 고를 확률도 동일. 두 그릇에 바닐라쿠키 50개, 초콜렛 쿠키 30개가 들어 있으므로 5/8
  - $p(B_1 \mid V) = \frac{(1/2)*(3/4)}{5/8} = 3/5$
  - 따라서 바닐라 쿠키를 골랐다는 것은 그릇 1을 선택했다는 가정에 대한 증거
  - 바닐라 쿠키는 그릇 1에서 나올 가능성이 더 높기 때문

1.5 통시적 해석

- 베이즈 이론을 다르게 해석할 수 있는 방법 : 데이터 D의 관점에서 봤을 때 가설 H의 확률을 수정해준다
- 이런 방식의 베이즈 이론 해석법 : **통시적(diachronic) 해석**
- 통시적 : 무언가 시간에 따라 일어나는 것을 의미, 이 경우에 가설에 대한 확률이 시간에 따라 새로운 데이터를 접하게 되며 달라진다는 의미
- 베이즈 이론을 H와 D에 대하여 정리
  - $p(H \mid D) = \frac{p(H)*p(D|H)}{p(D)}$
  - $p(H)$: 데이터를 보기 전의 가설의 확률, **사전확률(prior probability)**
  - $p(H \mid D)$: 계산하고자 하는 데이터를 확인한 이후의 가설의 확률, **사후확률(posterior probability)**
  - $p(D \mid H)$: 데이터가 가설에 포함될 확률, **우도(가능도, likelihood)**
  - $p(D)$: 어떤 가설에든 포함되는 데이터의 비율, **한정 상수**
- 사전확률
  - 배경지식을 기반으로 확률 계산하는 경우 : 쿠키문제 (동일한 확률로 그릇 선택한다고 가정)
  - 사전확률이 주관적인 경우 : 서로 다른 배경지식을 적용하거나 동일한 정보를 서로 다르게 해석할 수 있음
- 우도 : 쿠키문제 예시 - 쿠키가 어느 그릇에서 나왔는지 안다면 바닐라 쿠키일 확률은 세어보면 된다.
- 한정 상수 : 어떤 가정에서든 데이터를 볼 수 있는 확률, 하지만 이 상수가 무엇인지 정의하기 어려움. 다음의 가정집합을 단순화하여 정의
  - ***상호 배제*** : 집합 중 하나의 가설만 참일 경우
  - ***전체 포괄*** : 다른 가능성이 전혀 없는 경우. 단 하나의 가설이라도 참일 경우
  - 이런 성격의 가설 집합 : **스윗,suite**
- 쿠키 문제 예시
  - 쿠키가 그릇 1이나 2에서 나왔다는 것. 이 두 그릇은 상호 배제 및 전체 포괄적이라는 두 개의 가설
  - 전체 확률 법칙을 사용하여 $p(D)$ 계산 가능, 두 개의 배타적 사건이 있다면 다음과 같이 표현 가능
  - $p(D) = p(B_1) * p(D \mid B_1) + p(B_2) * p(D \mid B_2) = (1/2) * (3/4) + (1/2) * (1/2) = 5/8$

    - 앞서 두 값을 합친다고 가정한 계산과 동일 (바닐라쿠키 고를확률 $p(V)$)

1.6 M&M 문제

- Mars사는 1995년 M&M 초콜렛에 파란색 추가
- 그 전까지 기본 M&M 봉지의 색 조합 : 갈색30%, 노랑20%, 빨강20%, 녹색10%, 주황10%, 황갈색10%
- 파란색 추가 이후 : 파랑24%, 녹색20%, 주황16%, 노랑14%, 빨강13%, 갈색13%
- M&M을 두 봉지 구입. 각각 1994년, 1996년 생산
- 생산년도를 알려주지 않고 각 봉지에서 M&M 하나씩 꺼냈을 때 한 알은 노란색, 한 알은 녹색
- 이 때 노랑 초콜렛이 1994년에 생산한 봉지에서 나왔을 확률은?
    - 첫 단계 : 가설을 수치화
        - 노란 초콜렛은 봉지1에서 녹색 초콜렛은 봉지2에서 꺼냈다고 가정. 이 때의 가설
        - A : 봉지1이 1994년에 생산되었고 봉지2는 1996년에 생산
        - B : 봉지1이 1996년에 생산되었고 봉지2는 1994년에 생산
        - 각 가설을 행으로 두고 베이즈 이론의 항목을 열로 놓은 표 그리기
        -

-	사전확률$p(H)$	우도$p(D \mid H)$	$p(H)p(D \mid H)$	사후확률$p(H \mid D)$
A	1/2	(20)(20)	200	20/27
B	1/2	(10)(14)	70	7/27

        - 첫 번째 열 : 사전확률, $p(A) = p(B) = 1/2$
        - 두 번째 열 : 문제에 제시된 정보를 따라 파악된 우도
            - A가 참 : 노랑색이 1994년 봉지 확률 20%, 녹색이 1996년 봉지 확률 20%
            - B가 참 : 노랑색이 1996년 봉지 확률 14%, 녹색이 1994년 봉지 확률 10%
            - 각 선택은 독립적이므로 이 둘을 곱하여 결합확률 계산 : 20% * 20%
        - 세 전째 열 : 앞 두 값의 곱, 이 열의 합 270은 한정 상수
        - $p(D \mid H)$ 값은 확률 값이 아닌 상수 10,000을 곱한 퍼센트 값 사용. 한정 상수로 나누면서 이 표기는 무의미해져 결과에 영향주지 않는다.