IR Basics:
Probability Theory

확률 (Probability)

- 예)
  - 동전 하나를 던져 앞면이 나올 확률은 얼마인가?
    - 임의 시행(experiment, random trial) = 하나의 동전을 던져 그 결과를 본다
    - 시행 결과(outcome) = 앞면 혹은 뒷면
    - 표본공간(sample space) = {앞면, 뒷면}
      - 모든 가능한 시행 결과들의 집합
    - 사건(event) = {앞면}
      - 확률이 부여되어야 할 시행 결과들의 집합
      - 일반적으로 표본공간의 부분집합이 됨
  - 주사위 하나를 던져 짝수가 나올 확률은 얼마인가?
    - 표본공간 = {1, 2, 3, 4, 5, 6}
    - 사건 = 주사위의 짝수 = {2, 4, 6}

표본공간, 사건

- 임의시행 (experiment, random trial)
  - 결과를 만드는 임의 실험
    - 예) 동전 두 개를 던진다
- 시행결과 (outcome)
  - 임의시행의 결과
    - 예) HH, TT 등
- 표본공간 (sample space)
  - 모든 가능한 시행결과들의 집합
    - 예) S = {HH, HT, TH, TT}
- 사건 (event)
  - 표본공간의 부분집합
    - 예) E = {HH, TT}
- 단위사건 (elementary event)
  - 단일 결과로만 이루어진 사건
    - 예) {HH}, {TT} 등

확률

확률

- 다음 공리 하에서 임의 사건을 [0, 1] 사이 값으로 사상하는 함수
  - Kolmogorov axiom 1
    - 임의 사건의 확률은 0에서 1사이의 값이다
  - Kolmogorov axiom 2
    - 모든 단위사건의 확률의 합은 1이다
    - 공사건의 확률은 0이다
      - 공사건 = 단위결과를 하나도 포함하지 않는 사건, 즉 공집합
  - Kolmogorov axiom 3
    - 서로 배반인 사건들의 확률은 각 배반사건의 확률들의 합과 같다

사건

- 곱사건
  - 두 사건 A, B의 공통 부분이 발생하는 사건
    - $P(A \cap B)$
- 합사건
  - 두 사건 A, B 중 적어도 하나가 발생하는 사건
    - $P(A \cup B)$
- 여사건
  - 어떤 사건 A가 발생하지 않는 사건
    - $1 - P(A)$

사건

- 예
  - A = 주사위를 던져 짝수가 나오는 사건 = {2, 4, 6}
  - B = 주사위를 던져 4이상이 나오는 사건 = {4, 5, 6}
  - $A \cap B$ = 주사위를 던져 짝수이면서 4이상 눈이 나오는 사건 = {4, 6}
    - $P(A \cap B) = P(\{4,6\}) = P(\{4\}) + P(\{6\}) = 1/6+1/6 = 2/6$
  - $A \cup B$ = 주사위를 던져 짝수 혹은 4이상 눈이 나오는 사건 = {2,4,5,6}
    - $P(A \cup B) = P(\{2,4,5,6\}) = P(\{2\})+P(\{4\})+P(\{5\})+P(\{6\})=1/6+1/6+1/6+1/6=4/6$
  - A^C = 주사위를 던져 짝수가 나오지 않는 사건 = $\{1,2,3,4,5,6\} – \{2,4,6\} = \{1,3,5\}$
    - $P(A^C) = P(\{1,3,5\}) = P(\{1,2,3,4,5,6\}) – P(\{2,4,6\}) = 1 – P(A)$

조건부 확률 (conditional probability)

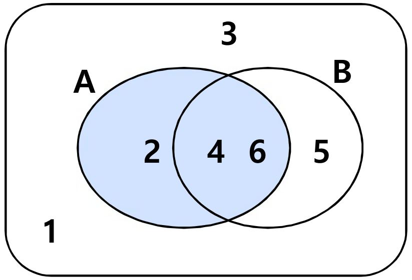
조건부 확률

- 사건 A가 발생한 조건 하에 사건 B가 발생할 확률
  - A = 주사위를 던져 짝수가 나오는 사건 = {2,4,6}
  - B = 주사위를 던져 4이상이 나오는 사건 = {4,5,6}
  - B|A = 주사위를 던져 짝수가 나온 상황에서 4이상이 나오는 사건
    - $P(B \mid A) = P(A \cap B) / P(A)$

정보검색 응용

- $P(\text{Relevant} \mid \text{Document}) = ?$
- $P(\text{Document} \mid \text{Query}) = ?$

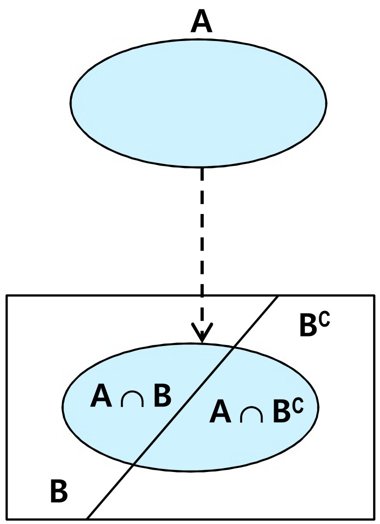
Product rule of probability



Sample space S

확률 공식

- 곱사건과 조건부 확률
  - $P(A \cap B) = P(A, B) = P(A)P(B|A)$
- 독립사건
  - $P(A \cap B) = P(A)P(B)$ 이면 A, B는 독립사건
- 체인법칙
  - $P(A, B, C)=P(A,B)P(C|A,B)=P(A)P(B|A)P(C|A,B)$
- $P(A) = P(A, B) + P(A, B^C)$
  - B와 B^C는 서로 배반이고 $B \cup B^C$는 표본공간이 됨
  - $P(A)=P(A \cap S)=P(A \cap (B \cup B^C))=P((A \cap B) \cup (A \cap B^C))= P(A \cap B)+P(A \cap B^C)$
    $=P(A, B) + P(A, B^C)$
- $P(A) = P(A, B_1) + P(A, B_2) + P(A, B_3) + \dots + P(A, B_n)$
  - 조건
    - 서로 배반인 n개의 사건들의 합집합 $B_1 \cup B_2 \cup B_3 \cup \dots \cup B_n$이 표본공간이 됨



정보검색 응용: 문서의 확률

문서의 확률

D=[동계 체전 개최]

P(D)

문서 내 개별 용어의 출현을 사건으로 고려 (문서는 개별 용어 출현 사건들의 곱사건)

= P(동계 체전 개최)

개최라는 용어가 출현한 사건

체인법칙 적용

= P(동계) × P(체전|동계) × P(개최|동계,체전)

동계, 체전, 개최의 출현이 (실제로는 독립이 아니지만) 상호 독립이라고 가정하면

= P(동계) × P(체전) × P(개최)

Rules of probability

- **Product rule** $p(X,Y) = p(X)p(Y \mid X)$
  - Chain rule
    - P(A, B, C) = P(A)P(B|A)P(C|A,B)

- **Partition rule** $p(X) = \sum_{Y} p(X,Y)$
  - $P(A) = P(A, B) + P(A, B^C)$
  - $P(A) = P(A \cap \text{SampleSpace}) = P(A \cap (B \cup B^C)) = P((A \cap B) \cup (A \cap B^C)) = P(A \cap B) + P(A \cap B^C) = P(A, B) + P(A, B^C)$

베이스 정리 (Bayes' theorem, Bayes' rule, Bayes' law)

- 베이스규칙 (Bayes' rule)
  - 사건 X, Y에 대해서,

Likelihood

사전확률
Prior probability
(before observing X)

$$p(Y \mid X) = \frac{p(X \mid Y)p(Y)}{p(X)}$$

사후확률
Posterior probability of Y
(after observing X)

Normalizing constant

$$p(X) = \sum_{Y} p(X \mid Y)p(Y)$$

$$P(X,Y) = P(X)P(Y \mid X) = P(Y)P(X \mid Y)$$

베이스 정리 (Bayes' theorem, Bayes' rule, Bayes' law)

- 예
  - 전체 인구 중 0.8%가 암에 걸린다고 한다. 이 병원의 기존 의료검진 결과에 따르면, 실제 암에 걸린 사람의 98%가 양성결과를 보였고, 실제 암에 걸리지 않은 사람의 97%가 음성결과를 보였다고 한다. 이 병원에서 어떤 새로운 환자가 암으로 검진을 받았다고 할 때, 이 환자가 실제로 암일 확률은 얼마인가? P(Cancer|양성)?

	Cancer	~Cancer
양성	0.98	0.03
음성	0.02	0.97

P(Cancer)=0.8%=0.008
P(~Cancer)=99.2%=0.992
P(양성|Cancer)=98%=0.98
P(음성|~Cancer)=97%=0.97

Source: Tom M. Mitchell. (1997). Machine Learning. The McGraw-Hill Companies.

베이스 정리 (Bayes' theorem, Bayes' rule, Bayes' law)

사전확률
(prior probability)

likelihood

사후확률
(posterior probability)

$$P(\text{암} \mid \text{양성}) = \frac{P(\text{암})P(\text{양성} \mid \text{암})}{P(\text{양성})}$$

$$= \frac{P(\text{암})P(\text{양성} \mid \text{암})}{P(\text{양성},\text{암}) + P(\text{양성},\neg\text{암})}$$

$$= \frac{P(\text{암})P(\text{양성} \mid \text{암})}{P(\text{암})P(\text{양성} \mid \text{암}) + P(\neg\text{암})P(\text{양성} \mid \neg\text{암})}$$

$$= \frac{0.008 \times 0.98}{0.008 \times 0.98 + 0.992 \times 0.03} = \frac{0.0078}{0.0078 + 0.0298} = 0.21$$

$$P(\neg\text{암} \mid \text{양성}) = 0.79$$

Source: Tom M. Mitchell. (1997). Machine Learning. The McGraw-Hill Companies.

정보검색 응용

Q = [부산 경제]
D = [부산 자갈치 시장 부산 관광 명소 부산 경제 견인]

문서 집합 (Collection)

- cf(부산)=124
- cf(경제)=578
- 문서집합 내 전체 용어 출현 회수 = 100,000

$$P(D|Q)$$

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

$$\approx P(Q|D)$$

$$= P(\text{부산 경제}|D)$$

$$= \underline{P(\text{부산}|D) \times P(\text{경제}|D)}$$

$$\frac{3}{9} \times \frac{1}{9}$$

$$\left(0.7\frac{3}{9} + 0.3\frac{124}{100000}\right) \times \left(0.7\frac{1}{9} + 0.3\frac{578}{100000}\right)$$

$$log\left(0.7\frac{3}{9} + 0.3\frac{124}{100000}\right) + log\left(0.7\frac{1}{9} + 0.3\frac{578}{100000}\right)$$

정보검색 응용

- 예
  - 정보검색에서 사용자의 질의 Q가 주어졌을 때, 문서 D의 확률은 얼마인가?

각 문서에 대해 P(D)의 값은 모두 동일하다고 가정
즉, P(D)는 uniform하다고 가정

$$P(D \mid Q) = \frac{P(D)P(Q \mid D)}{P(Q)}$$

서로 다른 문서를 순위화하는 관점에서
P(Q)는 서로 다른 문서에 대해 공통이므로
문서 순위화의 결과에 영향을 미치지 않음

$$P(Q \mid D) = P(q_1, q_2, \ldots, q_n \mid D) = \prod_{i=1}^{n} P(q_i \mid D) \approx \sum_{i=1}^{n} \log P(q_i \mid D)$$

정보검색 응용

- 예
  - 정보검색에서 사용자의 질의 Q에 대해 문서 D가 검색되었다(즉, 관찰되었다)고 할 때, 이 문서가 적합문서일 확률은?

$$P(Relevant \mid Document) = \frac{P(Relevant)P(Document \mid Relevant)}{P(Document)}$$

$$P(\text{적합}|D) > P(\text{부적합}|D)$$

$$\frac{P(\text{적합})P(D|\text{적합})}{P(D)} > \frac{P(\text{부적합})P(D|\text{부적합})}{P(D)}$$

$$P(\text{적합})P(D|\text{적합}) > P(\text{부적합})P(D|\text{부적합})$$

$$\frac{\boldsymbol{P(D|}\text{적합})}{\boldsymbol{P(D|}\text{부적합})} > \frac{P(\text{부적합})}{P(\text{적합})}$$

확률 분포 (probability distribution)

- 확률분포
  - 표본공간 내의 각 근원사건에 확률을 대응시킨 것
    - X축 → 각 근원사건에 대응되는 수
      - 각 근원사건을 수에 대응시키는 방법이 필요함
        - 확률변수(random variable) → 근원사건에 대응하는 수를 갖는 변수 or 표본공간을 실수(R)에 대응시키는 함수
    - Y축 → 각 근원사건의 확률

파라미터 추정 (Parameter estimation): Maximum Likelihood Estimation, MLE

- 어떤 윷가락 하나를 총 5번 던져 '앞면(둥근 면)'이 3번, '뒷면(평평한 면)'이 2번 나왔다. 이 윷가락의 앞면 확률 p는 얼마인가?

$$P(p \mid H,H,H,T,T) = \frac{\overbrace{P(p)}^{\text{Prior}}\,\overbrace{P(H,H,H,T,T \mid p)}^{\text{Likelihood}}}{P(H,H,H,T,T)} \approx P(H,H,H,T,T \mid p) = L(p \mid H,H,H,T,T)$$

$$\begin{aligned} &L(p \mid H,H,H,T,T) \\ &= P(H \mid p)P(H \mid p)P(H \mid p)P(T \mid p)P(T \mid p) \\ &= P(H \mid p)^3 P(T \mid p)^2 \\ &= P(H \mid p)^3 (1 - P(H \mid p))^2 \\ &= p^3 (1-p)^2 \end{aligned}$$

Likelihood를 maximize하는 파라미터 p를 찾는다

$$\hat{p} = \arg\max_p L$$

$$\frac{\partial L}{\partial p} = \frac{\partial \log L}{\partial p} = \frac{\partial \log\left(p^3(1-p)^2\right)}{\partial p} = \frac{\partial\left(\log p^3 + \log(1-p)^2\right)}{\partial p} = \frac{\partial(3\log p + 2\log(1-p))}{\partial p} = \frac{3}{p} - \frac{2}{1-p} = 0$$

$$\hat{p} = \frac{3}{5}$$

참조: Gregor Heinrich. (2008). Parameter Estimation for Text Analysis.

파라미터 추정 (Parameter estimation): Maximum A Posteriori Estimation, MAP estimation

- 이 윷가락은 제작 당시 앞면, 뒷면의 확률이 동일하도록 만들어졌다고 한다
  - 이 윷가락에 대해 알려진 기존 사실(prior)을 반영할 필요가 있음

$$P(p \mid H,H,H,T,T) = \frac{P(p)P(H,H,H,T,T \mid p)}{P(H,H,H,T,T)} \approx P(p)P(H,H,H,T,T \mid p)$$

$$\hat{p} = \underset{p}{arg\,max}\, P(p \mid H,H,H,T,T) = \underset{p}{arg\,max} \frac{P(p)P(H,H,H,T,T \mid p)}{P(H,H,H,T,T)} = \underset{p}{arg\,max}\, P(p)P(H,H,H,T,T \mid p)$$

$$= \underset{p}{arg\,max}\, log\left(P(p)P(H,H,H,T,T \mid p)\right) = \underset{p}{arg\,max}\, log\, P(p) + log\, P(H,H,H,T,T \mid p)$$

$$\frac{\partial\left(log\, P(p) + log\left(p^3(1-p)^2\right)\right)}{\partial p} = \frac{\partial\, log\, P(p)}{\partial p} + \frac{3}{p} - \frac{2}{1-p} = 0$$

$$P(p) \equiv P(p \mid \alpha, \beta) = \frac{1}{B(\alpha,\beta)} p^{\alpha-1}(1-p)^{\beta-1} \longleftarrow$$

Beta distribution (베타분포)
이 예의 likelihood에 대한 conjugate prior가 됨
즉 prior와 posterior가
같은 함수 형태를 취하도록 하는 prior임

$$\frac{\partial\left(log\, P(p) + log\left(p^3(1-p)^2\right)\right)}{\partial p} = \frac{\alpha-1}{p} - \frac{\beta-1}{1-p} + \frac{3}{p} - \frac{2}{1-p} = 0$$

$$\hat{p} = \frac{3+\alpha-1}{5+\alpha-1+\beta-1} \longleftarrow$$

α=β=5이면 이 값은 7/13이 되어 원래 값 3/5보다 ½에 더 가까워 진다

참조: Gregor Heinrich. (2008). Parameter Estimation for Text Analysis.

파라미터 추정 (Parameter estimation): Bayesian Estimation

Binary case:
Success or Fail
(Head or Tail)

$$P(p \mid H,H,H,T,T) = \frac{P(p)P(H,H,H,T,T \mid p)}{P(H,H,H,T,T)}$$

$$E(p) = \frac{3+\alpha}{5+\alpha+\beta} = \frac{n^H + \alpha}{n^H + n^T + \alpha + \beta}$$

Prior belief

MLE
Observation likelihood

α=β=5이면 이 값은 8/15이 됨

참조: Gregor Heinrich. (2008). Parameter Estimation for Text Analysis.

파라미터 추정 (Parameter estimation)

Multiple case:

White, Red, Blue의 세 가지 색상을 가진 공들로 가득 채워진 보자기에서
공을 하나씩 꺼내어 색상을 확인하는 실험을 10회 반복한 결과,
White 3회, Red 5회, Blue 2회 관찰되었다.
이 보자기에서 하나의 공을 꺼낼 때 Red 공이 나올 확률은 얼마인가?

$$P(p_R \mid WWWBBRRRRR) = \frac{P(p_R)P(WWWBBRRRRR \mid p_R)}{P(WWWBBRRRRR)}$$

$$E(p_R) = \frac{5+\alpha_R}{3+5+2+\alpha_W+\alpha_R+\alpha_B}$$

이 보자기에 대해 이전에 10000번 샘플한 아래 결과를 이용할 수 있다고 가정하고 이 정보를 이용하여 P(Red)를 계산
- P(White)=1/10, P(Red)=3/10, P(Blue)=6/10

이 보자기에 대해 이전에 100번 샘플한 아래 결과를 이용할 수 있다고 가정하고 이 정보를 이용하여 P(Red)를 계산
- P(White)=1/10, P(Red)=3/10, P(Blue)=6/10

파라미터 추정 (**Parameter estimation**)

Multiple case:

문서 D에서 용어 t가 tf(t,D)회 출현했다고 할 때, Bayesian estimation에 의한 P(t|D)는 얼마인가?

$$P\left(p\left(t_k \mid D\right) \mid D\right) = \frac{P\left(p\left(t_k \mid D\right)\right) P\left(D \mid p\left(t_k \mid D\right)\right)}{P\left(D\right)}$$

$$E\left(p\left(t_k \mid D\right)\right) = \frac{tf\left(t_k, D\right) + \alpha_{t_k}}{\mid D \mid + \alpha_{t_1} + \alpha_{t_2} + \ldots + \alpha_{t_n}}$$

$$= \frac{tf\left(t_k, D\right) + \mu p\left(t_k \mid C\right)}{\mid D \mid + \mu p\left(t_1 \mid C\right) + \mu p\left(t_2 \mid C\right) + \ldots + \mu p\left(t_n \mid C\right)}$$

$$= \frac{tf\left(t_k, D\right) + \mu p\left(t_k \mid C\right)}{\mid D \mid + \mu}$$

References

- Christopher D. Manning, Prabhakar Raghavan, Hinrich Schutze (2008). Introduction to Information Retrieval. Cambridge University Press.
- Bruce Croft, Donald Metzler, Trevor Strohman. (2009). Search Engines: Information Retrieval in Practice. Addison-Wesley Publishing Company.
- Ricardo Baeza-Yates, Berthier Ribeiro-Neto. (1999). Modern Information Retrieval. Addison-Wesley Publishing Company.
- https://en.wikipedia.org/
- Gregor Heinrich. (2008). Parameter Estimation for Text Analysis. (http://www.arbylon.net/publications/text-est.pdf)
- Sean Massung, (2014). A Note on Bayesian Statistics (http://massung1.web.engr.illinois.edu/~massung1/su14-cs410/files/inference.pdf)
- Avinash Kak, (2017). ML, MAP, and Bayesian — The Holy Trinity of Parameter Estimation and Data Prediction (https://engineering.purdue.edu/kak/Tutorials/Trinity.pdf)
- Stuart Russell (Author), Peter Norvig. Artificial Intelligence: A Modern Approach (3rd Edition), Prentice Hall.