실습1.베이스 확률의 정리를 이용하여 스팸메일이 스팸으로 판정될 가능성을 높게 하는 과정을 이해하기

---

1. 이론1:

(1) 베이즈 확률의 정리(조건부 확률) : 스팸메일로 판정할 확률를 계속 높여나갈 때 사용하는 정리.
어떤 키워드 들을 사용할 때 스팸메일로 판정될 확률이 높은지 인공지능이 스스로 학습해서 알아내게 한다.

(2) 공식유도; 벤다이어그램 활용

2.이론 2. 연속 시행과 독립시행

(1) 시행1 : '스팸'인지 '일반'인지 판정

(2) 시행2 : '스팸메일' 안에서 '기회'가 나올 확률, '스팸메일' 안에서 '할인'이 나올 확률을 각각 독립시행의 확률로 구한다.

**유사 시행

(1) 시행1 : 주머니에서 흰공(스팸메일)또는 검은공(일반메일)을 뽑는다.

(2) 시행2 : 주머니에서 흰공(스팸)을 뽑으면 동전을 2번 던지고,주머니에서 검은공(일반)을 뽑으면 주사위를 2번 던진다.

3.이론 3.

(1) 베이즈 확률을 예제(키워드가 '기회'와 '할인'일 때)에 적용하여 계산

$$P(\text{스팸} \cap (\text{기회} \cap \text{할인})) = P(\text{스팸}) \times P((\text{기회} \cap \text{할인})/\text{스팸}) = P(\text{스팸}) \times P(\text{기회}/\text{스팸}) \times P(\text{할인}/\text{스팸}) = P1$$

$$P(\text{일반} \cap (\text{기회} \cap \text{할인})) = P(\text{일반}) \times P((\text{기회} \cap \text{할인})/\text{일반}) = P(\text{스팸}) \times P(\text{기회}/\text{일반}) \times P(\text{할인}/\text{일반}) = P2$$

$$P(\text{기회} \cap \text{할인}) = P(\text{스팸} \cap (\text{기회} \cap \text{할인})) + P(\text{일반} \cap (\text{기회} \cap \text{할인}) = P1 + P2$$

$$P(\text{스팸}/(\text{기회} \cap \text{할인})) = P(\text{스팸} \cap (\text{기회} \cap \text{할인}))/P(\text{기회} \cap \text{할인}) = P1/(P1 + P2) = 0.69$$

실습1.베이스 확률의 정리를 이용하여 스팸메일이 스팸으로 판정될 가능성을 높게 하는 과정을 이해하기

(2) 베이즈 확률을 예제(키워드가 '기회','할인','수익'일 때)에 적용하여 계산

$$P(\text{스팸} \cap (\text{기회} \cap \text{할인} \cap \text{수익})) = P(\text{스팸}) \times P((\text{기회} \cap \text{할인} \cap \text{수익})/\text{스팸}) = P(\text{스팸}) \times P(\text{기회}/\text{스팸}) \times P(\text{할인}/\text{스팸}) \times P(\text{수익}/\text{스팸}) = P1$$

$$P(\text{일반} \cap (\text{기회} \cap \text{할인} \cap \text{수익})) = P(\text{일반}) \times P((\text{기회} \cap \text{할인} \cap \text{수익})/\text{일반}) = P(\text{일반}) \times P(\text{기회}/\text{일반}) \times P(\text{할인}/\text{일반}) \times P(\text{수익}/\text{일반}) = P2$$

$$P(\text{기회} \cap \text{할인} \cap \text{수익}) = P(\text{스팸} \cap (\text{기회} \cap \text{할인} \cap \text{수익})) + P(\text{일반} \cap (\text{기회} \cap \text{할인} \cap \text{수익}) = P1 + P2$$

$$P(\text{스팸}/(\text{기회} \cap \text{할인} \cap \text{수익})) = P(\text{스팸} \cap (\text{기회} \cap \text{할인} \cap \text{수익}))/P(\text{기회} \cap \text{할인} \cap \text{수익}) = P1/(P1 + P2) = 0.910$$

**** *따라서 키워드를 세 개로 했을 때 그 세 개의 키워드가 다 있는 메일이 스팸일 가능성이 매우 더 높다.

인공지능은 이 세 키워드외에 스팸메일이 스팸으로 나올 확률을 더 높여가기 위하여 데이터를 분석해서 스스로 학습하면서 새로운 키워드를 찾아서 스팸메일일 가능성을 계속 높여나갈 것이다.