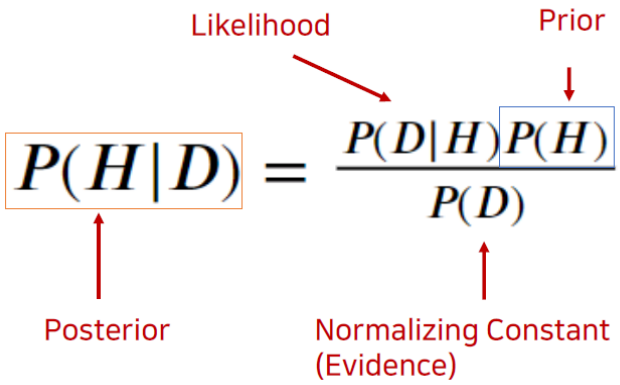
Naïve Bayes 정리

나이브 베이즈 정리를 알기 전에 베이즈 정리가 무엇인지 알아야한다. 베이즈 정리란 사전 확률과 사후 확률 사이의 관계를 나타내는 정리로 사전 확률을 이용해 사후 확률을 구하고자 한다. 다른 말로, 이전에 알던 지식과 이론을 이용해 확률을 구하는 것이다. 이것을 더 잘 이해하기 위해서는 베이즈 확률론의 반대인 frequentist 빈도주의 확률론을 알아보는 것도 좋다.

Frequentist 확률론이란 직접 실행을 통해 확률을 구하는 것을 말한다. 주사위를 6 번 굴려 1 이 1 번 나오는 것을 보고 '1 이 나올 확률이 1/6 이다' 라고 말하는 것이다. 이러한 확률론은 실제와 실험에 기반했기 때문에 독립적이고 대용량 데이터여도 처리하기가 쉽다. 이와 반대되는 베이즈 확률론은 사전 확률 또는 이론을 이용해 사후 확률을 구하는 확률이다. 이러한 성질 때문에 머신러닝 학습할 때 '새로운 데이터를 기존의 모델에 업데이트 시키는 방법' 관련해서 베이즈 확률이 접목이 되기 때문에 인공지능에서도 매우 중요한 이론이다.



베이즈 정리에 의한 식은 위와 같다. 조건부 확률을 이용한 식으로 H 는 알고 싶은 정보, D 는 알고 있는 정보이다. 위에서 4 개의 항은 각각 Posterior (사후 확률, D 가 일어날 때 H 가 일어날 확률), Likelihood (과거 경험을 잘 설명하는 정도) , Prior (사전 확률, 일반 확률 H), 그리고 Normalizing Constant (사건 D 의 발생 가능성) 이다.

이 베이즈 정리가 확률론과 인공지능에서 매우 중요하고 기초적인 이론이지만 하나의 단점이 있다면 변수가 많아질수록 계산량이 지수적으로 늘어난다는 점이다. 이를 막기 위해서는 입력 변수들이 모두 독립이다라는 가정을 하면 된다. 조금 막연하고 억지스러운 가정이라 'naïve' (조금 덜떨어진) bayes 이라 불린다. 하지만 이 나이브 베이즈는 큰 데이터셋에서는 놀라울 정도로 좋은 효율적이고 정확한 확률값들을 만들고는 한다. 이러한 나이브 베이즈는 특히 텍스트에서 좋은 성능을 보인다.