6주차

조건부확률(Conditional Probability)

* E 라는 사실/근거가 있을 때 🡪 미래를 예측 or H 라는 가설을 주장
* P(H|E) = P(H∩E) / P(E) = P(E|H) \* P(H) / P(E)
* P(H|E) ≠ P(E|H)

\*H와 E가 상호배타적 🡪 P(H|E) = P(H∩E) / P(E) = 0 / P(E) = **0**

E⊂H 🡪 P(H|E) = P(H∩E) / P(E) = P(E) / P(E) = **1**

결합확률(Joint Probability)

* 두 사건이 동시에 발생할 확률
* 조건부확률과 개별 사건의 발생 확률을 구할 수 있어 유용함
* P(A∩B) = P(B∩A)
* 사건 A와 B가 독립사건일 경우🡪 확률의 곱셈법칙: P(A∩B) = P(A) \* P(B)

주변확률(Marginal Probability)

* 개별 사건이지만 결합사건들의 합으로 표시될 수 있는 확률
* P(H|E) = P(E∩H) / P(E) = P(E∩H) / P(E∩H)+P(E∩HC) = P(E∩H) /

\*P(E): 정규화 상수(Normalization Constant)

베이즈 정리(Bayes’ Rules)

* 사전 확률과 사후 확률의 관계를 구함
* 사전 확률이 있었고 어떤 사건/근거를 관찰한 다음 사후 확률을 계산
* 공식: P(H|E) = P(E|H) \* P(H) / P(E)
* P(H|E): 사후확률
* P(E|H): 역조건부확률
* P(H): 사전확률
* P(E|H) \* P(H): 가능도(우도)
* P(E): 주변확률

\*실제 생활에서는 사후 확률만 알고 있는 경우가 많은데, 이때 사전 확률을 유추할 수 있음(역확률: 결과를 관측한 뒤 원인을 추론할 수 있음)

* n개의 사건(H1, H2, H3, … Hn-1, Hn)이 표본공간 S의 분할일 때, 사건 D가 표본공간 S의 임의의 사건이라면

H=과거의 경험

D=현재의 경험

P(Hi|D) = P(Hi D) / P(D) = P(Hi) \* P(D|Hi) /

예제) 질병 A의 발병률이 0.1%이고 실제로 질병이 있을 때 질병이 있다고 검진할 확률(민감도)은 99%이다. 질병이 없을 때 없다고 검진할 확률(특이도)은 98%일 때, 질병에 걸렸다고 진단받았을 때 정말로 병에 걸렸을 확률은?

풀이) Hypothesis(H): True(실제로 병이 있다)

Evidence(E): Positive(병이 있다고 진단받다)

질병에 걸렸을 확률: P(H)=0.001

실제로 질병이 있을 때 질병이 있다고 검진할 확률(TP): P(E|H) = 0.99

질병이 없을 때 질병이 없다고 검진할 확률(TN): P(EC|HC) = 0.98

질병에 걸리지 않았을 확률: P(HC) = 1 - P(H) = 1 – 0.001 = 0.999

질병이 없을 때 질병이 있다고 검진할 확률(FN): P(E|HC) = 1 - P(EC|HC) = 1 - 0.98 = 0.02

질병에 걸렸다고 진단받았을 때 정말로 병에 걸렸을 확률

: P(H|E) = P(E|H) \* P(H) / P(E) = P(E|H) \* P(H) / P(E|H) \* P(H) + P(E|HC) \* P(HC)

= 1 + P(E|H) \* P(H) / P(E|HC) \* P(HC) = 1 + 0.99 \* 0.001 / 0.02 \* 0.999 ≒ 0.047

나이브 베이즈(Naïve Bayes)

* 각 특징이 독립적이라는 가정 하에 데이터 또는 코퍼스에서의 출현 빈도를 쉽게 활용하여 데이터의 희소성 문제를 해결하는 방식
* 사후확률을 최대화하는 클래스를 예측
* Sentimental Analysis: 문장 데이터의 각 단어를 긍정/부정 두 클래스로 구분
* 장점
* 단순히 코퍼스에서 각 단어의 클래스당 출현빈도를 계산하는 것만으로도 분석 가능
* Add-one smoothing: 어떠한 단어의 출현 횟수가 0일 때 해당 샘플의 실제 확률을 0으로 추정하는 것은 위험 🡪 출현 횟수에 1을 더해주는 것으로 조정
* 단점
* 단어간 순서로 인해 발생하는 관계, 정보 고려 어려움

사이킷런 나이브 베이즈 알고리즘

* BernoulliNB 🡪 이진 데이터(레이블 당 문장 수가 적은 경우)
* MultinomNB 🡪 카운트 데이터(특성이 비교적 많은 데이터셋)
* GaussianNB 🡪 연속적인 데이터(고차원 데이터셋)
* 딥러닝 🡪 단어 간 순서, entity recognition 등의 깊은 분석

규제(Regularization)

* 과대적합이 되지 않도록 모델을 강제로 제한
* 방식

1. 매개변수 조절

* 훈련 데이터를 잘 예측하기 위함
* 추가 제약 조건을 만족시키기 위함
* 예) 가중치(w) 선택 🡪 가중치의 절댓값을 최소화, 기울기를 작게 만듦(=특성이 출력에 미치는 영향을 최소화)

1. 매개변수 고정, 훈련 데이터셋의 크기 변화