10.4 상호정보량 FOR FEATURE SELECTION

Mutual Information

선형적 상관관계 ONLY

이 절에서는 상관계수를 대체할 수 있는 확률변수의 특성, 상호정보량을 알아본다.

상호정보량 "비선형적 상관관계" 까지 포함

두 확률변수 X, Y가 독립이면 정의에 의해 결합확률밀도함수는 주변확률밀도함수의 곱과 같다.

$$p(x,y) = p(x)p(y)$$

쿨백-라이블러 발산은 두 확률분포가 얼마나 다른지를 정량적으로 나타낸다.
두 확률분포가 같으면 쿨백-라이블러 발산은 0이 되고, 다를수록 커진다.

$$KL(p\|q) = \sum_{k=1}^{K} p(y_i) \log_2 \left(\frac{p(y_i)}{q(y_i)} \right)$$

상호정보량 (mutual information)은 결합확률밀도함수 $p(x,y)$와 주변확률밀도함수의 곱 $p(x)p(y)$의 쿨백-라이블러 발산이다.
즉, 결합확률밀도함수와 주변확률밀도함수의 차이를 측정함으로써 두 확률변수의 상관관계를 측정하는 방법이다. 만약 두 확률변수가 독립이면 결합확률밀도함수는 주변확률밀도함수의 곱과 같으므로 상호정보량은 0이된다.
반대로, 상관관계가 있다면 그만큼 양(+)의 상호정보량을 갖는다.

$$MI[X,Y] = KL(p(x,y)\|p(x)p(y)) = \sum_{i=1}^{K} p(x_i,y_i) \log_2 \left(\frac{p(x_i,y_i)}{p(x_i)p(y_i)} \right)$$

상호정보량은 엔트로피와 조건부엔트로피의 차이와 같다.

$$MI[X,Y] = H[X] - H[X|Y]$$

$$= H[Y] - H[Y|X]$$

조건부엔트로피는 두 확률변수의 상관관계가 강할수록 원래의 엔트로피보다 더 작아지므로 상호정보량이 커진다.

이산확률변수의 상호정보량

상관관계가 있는 두 개의 카테고리 확률변수 X, Y에서 나온 표본 데이터 N개가 있다. 이 데이터를 이용하여 두 이산확률변수의 상호정보량을 추정하려면 우선 다음과 같은 기호를 정의해야 한다.

- I : X의 카테고리 갯수
- J : Y의 카테고리 갯수
- N_i : $X = i$인 데이터 갯수
- N_j : $Y = j$인 데이터 갯수
- N_{ij} : $X = i$, $Y = j$인 데이터 갯수

이 때 확률밀도함수는 다음과 같이 추정할 수 있다.

$$P_X(i) = \frac{N_i}{N}$$

$$P_Y(j) = \frac{N_j}{N}$$

$$P_{XY}(i,j) = \frac{N_{ij}}{N}$$

이를 대입하면,

$$MI[X,Y] = \sum_{i=1}^{I} \sum_{j=1}^{J} \frac{N_{ij}}{N} \log_2 \left(\frac{N N_{ij}}{N_i N_j} \right)$$

scikit learn . metrics . mutual_info_score

→ 이산확률변수의 상호정보량 계산

→ 각 데이터에 대해 X, Y 카테고리값을 표시한 2차원 배열을 인자로 입력.

예제

뉴스그룹 카테고리와 키워드 간 상호정보량

- 사이킷런 패키지가 제공하는 문서 카테고리 분류용 데이터
  - 'rec.autos' — 자동차
  - 'sci.med' — 의학
  - 'rec.sport.baseball' — 야구

'Bag of Words'

document class	X_1	X_2	X_3	... (단어)
rec.autos	3	0	1	...
sci.med	0	0	2	...
rec.sport.baseball	0	6	0	...

↳ 각각의 word가 모든 document에 대해 값 MI 계산

baseball
banks automotive
auto ball autos
batting atlanta
alomar bat

"상호정보량 높은 keyword ≒ 문서 분류에 효과적인 단어"

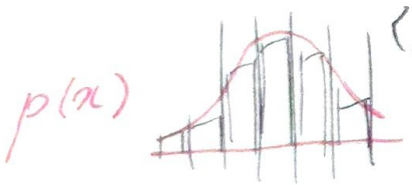
최대정보 상관계수 → Feature Selection 활용

연속확률변수의 표본 데이터에서 상호정보량을 측정하려면, 우선 확률분포함수를 알아야 한다. 확률분포함수는 보통 히스토그램을 사용하여 유한개의 구간(bin)으로 나누어 추정하게 되는데, 이때 구간의 갯수나 경계 위치에 따라 추정 오차가 커질 수 있다.

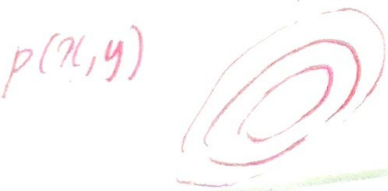
따라서 구간을 나누는 방법을 다양하게 시도한 다음에 그 결과로 구한 다양한 상호정보량 중에서 가장 큰 값을 선택하여 정규화한 것을 최대정보 상관계수 (maximal information coefficient, MIC) 라고 한다.

conda install minepy

minepy 패키지를 사용하면 최대정보 상관계수를 구할 수 있다.
아래는 선형상관계수 (피어슨 상관계수)로, 0이 나오지만
비선형적인 상관관계를 갖는 데이터들에 대해 최대정보 상관계수를 구한 결과.



(연속확률변수)
구간을 쪼개서 (bin) 카테고리값이라고 가정한다
→ bin의 갯수 (촘촘함의 정도)에 따라서 상호정보량 달라짐



비선형 상관관계