데이터과학과 AI를 위한 파이썬

13강. 엔트로피

세종사이버대학교

김명배 교수

학습내용

- 엔트로피의 정의
- 조건부 엔트로피
- 교차엔트로피
- 상호정보량

학습목표

- 엔트로피란 무엇이고, 엔트로피를 추정하는 방법을 학습하고 파이썬으로 실행할 수 있다.
- 결합엔트로피와 조건부 엔트로피를 이해하고 파이썬으로 실행 할 수 있다.
- 교차엔트로피를 설명할 수 있고 파이썬으로 실행 할 수 있다.
- 상호정보량과 최대정보 상관계수를 설명할 수 있고 파이썬으로 실행할 수 있다.

1. 엔트로피의 정의

1) 엔트로피(entropy)란

- 확률분포가 가지는 정보의 확신도 혹은 정보량을 수치로 표현한 것
- 성공과 실패가 나올 각각의 확률의 차이가

  클 수록 → → 엔트로피 ↓

  작을 수록 → 엔트로피 ↑
- 물리학에서는 물질의 상태가 분산되는 정도를 의미함
- 이산확률변수의 엑트로피 :

$$H[Y] = -\sum_{k=1}^{K} p(y_k)\log_2 p(y_k), \qquad p(y): \text{확률질량함수}$$

- 연속형확률변수의 엑트로피 :

$$H[Y] = -\int_{-\infty}^{\infty} p(y)\log_2 p(y)\, dy, \qquad p(y): \text{확률밀도함수}$$

1. 엔트로피의 정의

1) 엔트로피(entropy)란

[예시] 다음 각각의 경우 엔트로피 산출

가) 확률분포 $Y_1: P(Y=0)=0.5,\ P(Y=1)=0.5$

나) 확률분포 $Y_2: P(Y=0)=0.8,\ P(Y=1)=0.2$

다) 확률분포 $Y_1: P(Y=0)=1.0,\ P(Y=1)=0.0$

[풀이]

$$H[Y_1] = -\frac{1}{2}\log_2\frac{1}{2} - \frac{1}{2}\log_2\frac{1}{2} = 1$$

$$H[Y_2] = -\frac{8}{10}\log_2\frac{8}{10} - \frac{2}{10}\log_2\frac{2}{10} \approx 0.72$$

$$H[Y_3] = -1\log_2 1 - 0\log_2 0 = 0$$

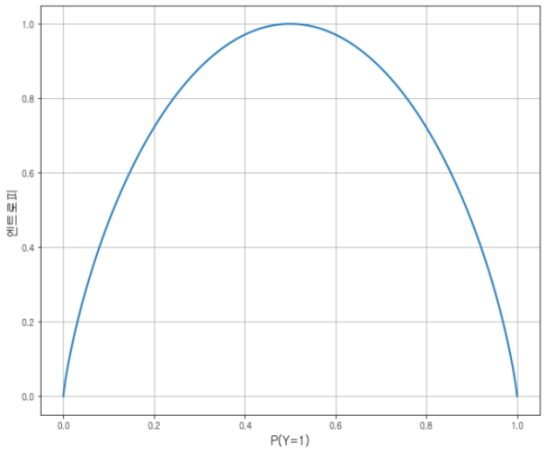
[파이썬] 0대신 np.finfo(float).eps 사용

1. 엔트로피의 정의

1) 엔트로피(entropy)란

[연습문제1] 베르누이 분포에서 확률값 P(Y=1)은 0부터 1까지의 값을 가질 수 있다. 각각의 값에 대해 엔트로피를 계산하여 가로축에 P(Y=1)이고 세로축이 H[Y]인 그래프를 그려라.

[풀이]



1. 엔트로피의 정의

1) 엔트로피(entropy)란

[연습문제2] 다음 확률분포의 엔트로피를 계산하세요.

가) $P(Y=0)=\frac{1}{8}, P(Y=1)=\frac{1}{8}, P(Y=2)=\frac{1}{4}, P(Y=3)=\frac{1}{2}$

나) $P(Y=0)=1, P(Y=1)=0, P(Y=2)=0, P(Y=3)=0$

다) $P(Y=0)=\frac{1}{4}, P(Y=1)=\frac{1}{4}, P(Y=2)=\frac{1}{4}, P(Y=3)=\frac{1}{4}$

[풀이]

가) $H=-\frac{1}{8}\log_2\frac{1}{8}-\frac{1}{8}\log_2\frac{1}{8}-\frac{1}{4}\log_2\frac{1}{4}-\frac{1}{2}\log_2\frac{1}{2}=\frac{3}{8}+\frac{3}{8}+\frac{2}{4}+\frac{1}{2}=\frac{14}{8}=\frac{7}{4}$

나) $H=-1\cdot\log_2 1-0\cdot\log_2 0-0\cdot\log_2 0-0\cdot\log_2 0=0$

다) $H=-\frac{1}{4}\log_2\frac{1}{4}-\frac{1}{4}\log_2\frac{1}{4}-\frac{1}{4}\log_2\frac{1}{4}-\frac{1}{4}\log_2\frac{1}{4}=4\times\left(-\frac{1}{4}\right)\times(-2)=2$

1. 엔트로피의 정의

2) 엔트로피의 성질

- 엔트로피의 최솟값은 특정 값이 나올 확률이 1인 경우로, 0이다.
- 엔트로피의 최댓값은 모든 확률변수 값의 확률이 동일한 경우로, 이산확률변수의 클래스 개수에 따라 달라짐

확률변수 값이 2^k개면서 각 확률이 $\frac{1}{2^k}$로 모두 같을 때 최대값을 갖음

$$H = -2^k \cdot \frac{1}{2^k} \log_2 \frac{1}{2^k} = k$$

1. 엔트로피의 정의

3) 엔트로피의 추정

- 이론적으로 확률밀도함수가 없고 실제 데이터 주어진 경우
  1) 데이터에서 확률질량함수를 추정
  2) 이를 기반으로 엔트로피를 계산

[파이썬] scipy의 entropy 함수 제공

scipy.entropy(확률분포, base=확률변수 개수)

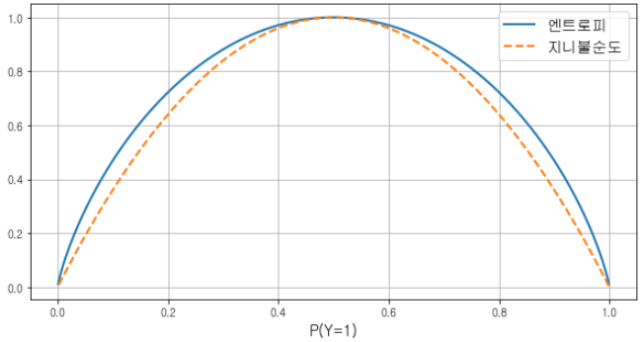
1. 엔트로피의 정의

4) 지니불순도(Gini impurity)

- 엔트로피처럼 확률분포가 어느 쪽에 치우쳤는가를 재는 척도
- 로그를 사용하지 않아 계산량이 더 적음
- 엔트로피 대용으로 많이 사용됨

$$G[Y] = \sum_{k=1}^{K} P(y_k)(1 - P(y_k))$$

[엔트로피와 지니불순도의 비교]



2. 조건부 엔트로피

1) 결합 엔트로피(joint entropy)

- 두 확률변수에 대해 결합확률분포를 사용하여 정의한 엔트로피를 말함

가) 이산확률변수 X, Y에 대한 결합엔트로피(p : 확률질량함수)

$$H[X,Y] = -\sum_{i=1}^{K_x}\sum_{j=1}^{K_y} p(x_i, y_j)\log_2 p(x_i, y_j)$$

나) 연속확률변수 X, Y에 대한 결합엔트로피(p : 확률밀도함수)

$$H[X,Y] = -\int_x \int_y p(x,y)\log_2 p(x,y)\,dxdy$$

2. 조건부 엔트로피

2) 조건부 엔트로피(conditional entropy)

- 어떤 확률변수 X가 다른 확률변수 Y값을 예측(설명)하는 데 도움이 되는지를 측정하는 방법(인과관계)
- 1:1의 함수관계라면 X로 Y를 예측할 수 있음
- 이산확률변수의 조건부엔트로피(p : 확률질량함수)

$$H[Y|X] = -\sum_{i=1}^{K_x} p(x_i) H[Y|X = x_i] - \sum_{i=1}^{K_x}\sum_{j=1}^{K_y} p(x_i, y_j) \log_2 p(y_j|x_i)$$

- 연속형 확률변수의 조건부엔트로피(p : 확률밀도함수)

$$H[Y|X] = -\int_x p(x) H[Y|X = x] dx = -\int_x \int_y p(x, y) \log_2 p(y|x)\, dxdy$$

2. 조건부 엔트로피

2) 조건부 엔트로피(conditional entropy)

[X가 변별력이 있는 경우]

	$Y=0$	$Y=1$
$X=0$	0.4	0.0
$X=1$	0.0	0.6

$P(Y=0|X=0)=1, P(Y=1|X=0)=0$

$P(Y=0|X=1)=0, P(Y=1|X=1)=1$

$H(Y|X=0)=H(Y|X=1)=0$

$\therefore H(Y|X)=0$

[X가 변별력이 없는 경우]

	$Y=0$	$Y=1$
$X=0$	1/9	2/9
$X=1$	2/9	4/9

$P(Y=0|X=0)=1/3, P(Y=1|X=0)=2/3$

$P(Y=0|X=1)=1/3, P(Y=1|X=1)=2/3$

$H(Y|X=0)=H(Y|X=1)=-\frac{1}{3}\log_2\frac{1}{3}-\frac{2}{3}\log_2\frac{2}{3}\approx 0.92$

$\therefore H(Y|X)\approx 0.92$

2. 조건부 엔트로피

2) 조건부 엔트로피(conditional entropy)

[예제] 스펨메일 분류 문제에서 X_1과 X_2 키워드 존재여부로 판단한다고 할 때, X1과 X2중 어떤 키워드가 효과적인가? → 조건부 엔트로피

[스펨메일여부(Y)와 X_1]

X_1 \ Y	$Y=0$	$Y=1$	소계
$X_1=0$	30	10	40
$X_1=1$	10	30	40
소계	40	40	80

[스펨메일여부(Y)와 X_2]

X_2 \ Y	$Y=0$	$Y=1$	소계
$X_2=0$	20	40	60
$X_2=1$	20	0	20
소계	40	40	80

$$P(Y|X_1)$$
$$= P(X_1=0)H[Y|X_1=0] + P(X_1=1)H[Y|X_1=1]$$
$$= \frac{40}{80}\cdot 0.81 + \frac{40}{80}\cdot 0.81 = 0.81$$

$$P(Y|X_2)$$
$$= P(X_2=0)H[Y|X_2=0] + P(X_2=1)H[Y|X_2=1]$$
$$= \frac{60}{80}\cdot 0.92 + \frac{20}{80}\cdot 0 = 0.69$$

3. 교차엔트로피

1) 교차엔트로피(cross information)

- 확률변수가 아닌 확률분포를 인수로 받음으로써, 분류모형의 성능을 측정하는 데 사용됨.

[이산확률분포일 때 두 확률분포 p, q의 교차엔트로피]

$$H[p,q] = -\sum_{k=1}^{K} p(y_k) \log_2 q(y_k)$$

[연속확률분포일 때 두 확률분포 p, q의 교차엔트로피]

$$H[p,q] = -\int_{y} p(y) \log_2 q(y)\, dy$$

3. 교차엔트로피

1) 교차엔트로피(cross information)

- 확률분포 p, q가 다음과 같을 때

  분포 p : X값이 정해졌을 때 정답인 Y의 확률분포

  분포 q : X값이 정해졌을 때 예측값($\hat{Y}$)의 확률분포

- 확률분포 p, q 의 교차엔트로피는

  정답 $Y = 1$일 때, $H[p,q] = -\log_2 \mu$

  정답 $Y = 0$일 때, $H[p,q] = -\log_2(1-\mu)$

- 교차엔트로피 값은 예측의 틀린 정도를 나태 내는 오차함수 역할을 함

  $Y = 1$일 때 : μ가 작아질 수록, 즉 예측이 틀릴 수록 $\log_2 \mu$가 커짐

  $Y = 0$일 때 : μ가 커질 수록, 즉 예측이 틀릴 수록 $\log_2(1-\mu)$가 커짐

3. 교차엔트로피

1) 교차엔트로피(cross information)

- 전체 학습 데이터 N개에 대해 교차엔트로피 평균을 구한 것을 로그손실(log-loss)라고 함

$$\log \quad loss = -\frac{1}{N}\sum_{k=1}^{K}(y_i \log_2 \mu_i + (1-y_i)\log_2(1-\mu_i))$$

- 다중분류에서도 교차엔트로피를 오차함수로 사용할 수 있음
- 다중분류 문제의 교차엔트로피 손실함수를 카테고리 로그손실(categorical log-loss)이라고 함

$$categorical\log \quad loss = -\frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{K}(\mathbb{I}(y_i = k)\log_2 p(y_i = k))$$

[파이썬] sklearn 패키지의 metrics의 log_loss 함수 제공

3. 교차엔트로피

2) 쿨백-라이블러 발산(Kullback-Leibler divergence; KLD)

- 두 확률분포 $p(y), q(y)$의 모양이 얼마나 다른지를 숫자로 계산한 값

[이산확률분포]

$$KL(p \parallel q) = H[p,q] - H[p] = \sum_{i=1}^{K} p(y_1) \log_2 (\frac{p(y_i)}{q(y_i)})$$

[연속확률분포]

$$KL(p \parallel q) = H[p,q] - H[p] = \int p(y) \log_2 \left(\frac{p(y)}{q(y)}\right) dy$$

- 상대엔트로피(relative entropy)라고도 함
- 동일한 분포인 경우 $KL(p \parallel p)$는 0
- 거리개념이 아니며, $KL(p \parallel q) \neq KL(q \parallel p)$

4. 상호정보량

1) 상호정보량(mutual information)

- 상호정보량은 결함확률밀도함수 $p(x,y)$와 주변확률밀도함수의 곱 $p(x)p(y)$의 쿨벡-라이블러 발산임

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

- 상호정보량은 엔트로피와 조건부엔트로피의 차이와 같음

$$MI[X,Y] = H[X] - H[X|Y] \text{ or } H[Y] - H[Y|X]$$

- 두 확률변수가 독립이면 상호정보량은 0임

[파이썬] 이산확률변수의 경우 metrics의 mutual_info_score 함수를 사용

4. 상호정보량

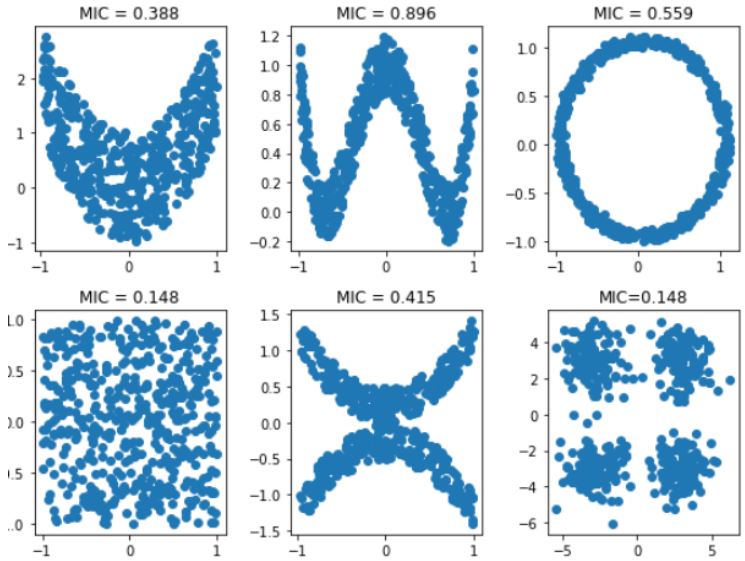
2) 최대정보 상관계수(maximal information coefficient; MIC)

- 연속확률변수의 경우 상호정보량 측정하려면 확률분포함수를 알아야 함.
- 확률분포함수는 히스토그램을 사용하여 유한 개의 구간(bin)으로 나눔
- 구간의 개수나 경계 위치에 따라 추정오차가 달라질 수 있음
- 구간을 나누는 방법을 다양하게 시도하여 상호정보량이 가장 큰 것을 선택하여 정규화 한 것을 최대정보 상관계수라고 함
- 피어슨의 상관계수는 선형상관인 반면, MIC는 비선형적 관계를 찾을 수 있는 장점이 있음
- 의사결정나무분석에서 연속형 설명변수를 분류할 때 사용 됨

4. 상호정보량

2) 최대정보 상관계수(maximal information coefficient; MIC)

[두 확률변수의 형태별 MIC]



정리하기

1. 엔트로피의 정의

- 확률분포가 가지는 정보의 확신도 혹은 정보량을 수치로 표현한 것

$$H[Y] = -\sum_{k=1}^{K} p(y_k) \log_2 p(y_k)$$

- 성공과 실패가 나올 각각의 확률의 차이가 클 수록 엔트로피 ↓, 차이가 작을 수록 엔트로피 ↑
- 엔트로피의 최솟값은 특정 값이 나올 확률이 1인 경우로, 0이다.
- [파이썬] scipy의 entropy() 함수 사용
- 엔트로피와 비슷한 개념으로 지니불순도가 있음

$$G[Y] = \sum_{k=1}^{K} P(y_k)(1 - P(y_k))$$

정리하기

2. 조건부 엔트로피

- 두 확률변수에 대해 결합확률분포를 사용하여 정의한 엔트로피를 결합엔트로피라고 함

$$H[X,Y] = -\sum_{i=1}^{K_x}\sum_{j=1}^{K_y} p(x_i, y_j)\log_2 p(x_i, y_j)$$

- 어떤 확률변수 X가 다른 확률변수 Y값을 예측(설명)하는 데 도움이 되는지를 측정하는 방법(인과관계)을 조건부 엔트로피라고 함

$$H[Y|X] = -\sum_{i=1}^{K_x} p(x_i)H[Y|X=x_i] - \sum_{i=1}^{K_x}\sum_{j=1}^{K_y} p(x_i, y_j)\log_2 p(y_j|x_i)$$

정리하기

3. 교차엔트로피

- 확률변수가 아닌 확률분포를 인수로 받음으로써, 분류모형의 성능을 측정하는 데 사용함

$$H[p,q] = -\sum_{k=1}^{K} p(y_k)\log_2 q(y_k)$$

- 전체 학습 데이터 N개에 대해 교차엔트로피 평균을 구한 것을 로그손실이라고 함

$$\log \quad loss = -\frac{1}{N}\sum_{k=1}^{K}(y_i \log_2 \mu_i + (1-y_i)\log_2(1-\mu_i))$$

- 다중분류 문제의 교차엔트로피 손실함수를 카테고리 로그손실이라고 함

$$categorical\log \quad loss = -\frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{K}(\mathbb{I}(y_i = k)\log_2 p(y_i = k))$$

- [파이썬] sklearn 패키지 metrics의 log_loss 함수 사용

정리하기

4. 쿨백-라이블러 발산

- 두 확률분포 $p(y), q(y)$의 모양이 얼마나 다른지를 숫자로 계산한 값

$$KL(p \parallel q) = H[p, q] - H[p] = \sum_{i=1}^{K} p(y_1) \log_2 (\frac{p(y_i)}{q(y_i)})$$

- 동일한 분포인 경우 $KL(p \mathbin{/\!/} p)$는 0
- 거리개념이 아니며, $KL(p \mathbin{/\!/} q) \neq KL(q \mathbin{/\!/} p)$

정리하기

5. 상호정보량

- 상호정보량은 결함확률밀도함수 $p(x,y)$와 주변확률밀도함수의 곱 $p(x)p(y)$의 쿨벡-라이블러 발산임
- 상호정보량은 엔트로피와 조건부엔트로피의 차이와 같음
- 두 확률변수가 독립이면 상호정보량은 0임

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

- [파이썬] 이산확률변수의 경우 metrics의 mutual_info_score 함수를 사용

정리하기

6. 최대정보 상관계수

- 연속확률변수의 경우 상호정보량 측정을 위해 확률분포함수의 히스토그램을 사용하여 유한 개의 구간(bin)으로 나눔
- 구간을 나누는 방법을 다양하게 시도하여 상호정보량이 가장 큰 것을 선택하여 정규화 한 것을 최대정보 상관계수라고 함
- 의사결정나무분석에서 연속형 설명변수를 분류할 때 사용 됨
- [파이썬] minepy 패키지의 comppute_score() 함수 사용

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