10.2 조건부 엔트로피

- 이 절에서는 두 확률변수의 결합엔트로피와 조건부 엔트로피를 정의하는 방법을 공부하고, 분류문제에 어떻게 활용할 수 있는지 살펴본다.

결합엔트로피 (joint entropy)

결합엔트로피는 결합확률분포를 사용하여 정의한 엔트로피를 말한다.

이산확률변수 X, Y에 대해 결합엔트로피는 다음처럼 정의한다.

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

(이 식에서 K_X, K_Y는 각각 X와 Y가 가질 수 있는 값의 갯수고, p는 확률질량함수다.)

연속확률변수 X, Y에 대해 결합엔트로피는 다음처럼 정의한다.

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

(이 식에서 p는 확률밀도함수다.)

결합엔트로피도 결합확률분포라는 점만 제외하면 일반적인 엔트로피와 같다.
모든 경우에 대해 확률이 골고루 분포되어 있으면 엔트로피값이 커지고,
특정한 한 가지 경우에 대해 확률이 모여 있으면 엔트로피가 0에 가까워진다.

조건부엔트로피 conditional entropy

조건부엔트로피는 어떤 확률변수 X가 다른 확률변수 Y의 값을 예측하는데 도움이 되는지를 측정하는 방법중의 하나다. 만약 확률변수 X의 값이 어떤 특정한 하나의 값을 가질 때, 확률변수 Y도 마찬가지 규칙/방식/정도로 특정한 값이 된다면, X로 Y를 예측할수 있다.

반대로, 확률변수 X의 값이 어떤 특정한 하나의 값을 가져도, 확률변수 Y가 여러값으로 골고루 분포되어 있다면 X는 Y의 값을 예측하는데 도움이 되지 않는다.

- 조건부 엔트로피의 유도.

확률변수 X, Y가 모두 이산확률변수라고 가정하고 X가 특정한 값 x_i을 가질때의 Y 엔트로피 $H[Y|X=x_i]$는 다음처럼 조건부 확률분포의 엔트로피로 정의한다.

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

조건부 엔트로피는 확률변수 X가 가질수 있는 모든 경우에 대해 $H[Y|X=x_i]$를 가중평균한 값으로 정의한다.

$$\begin{aligned} H[Y|X] &= \sum_{j=1}^{K_X} p(x_i) H[Y|X=x_i] \\ &= -\sum_{i=1}^{K_X}\sum_{j=1}^{K_Y} p(y_j|x_i)p(x_i)\log_2 p(y_j|x_i) \\ &= -\sum_{i=1}^{K_X}\sum_{j=1}^{K_Y} p(x_i,y_j)\log_2 p(y_j|x_i) \end{aligned}$$

연속확률변수의 경우 다음과 같다.

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

$$\begin{aligned} H[Y|X] &= -\int_x p(x)\,H[Y|X=x]\,dx \\ &= -\int_x p(x)\left(\int_y p(y|x)\log_2 p(y|x)\,dy\right)dx \\ &= -\int_x\int_y p(y|x_i)p(x)\log_2 p(y|x)\,dx\,dy \\ &= -\int_x\int_y p(x,y)\log_2 p(y|x)\,dx\,dy \end{aligned}$$

따라서 조건부 엔트로피의 최종적인 수학적 정의는 다음과 같다.

(이산확률변수)

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

(연속확률변수)

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

① 예측에 도움이 되는 경우

(ex)

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

X=0, X=1일 때의 조건부확률분포

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

이 때, Y의 엔트로피는 모두 0이다.

$$H[Y|X=0]=0$$
$$H[Y|X=1]=0$$

따라서 조건부엔트로피도 0이 된다.

$$H[Y|X]=0$$

- Y를 가지고 X를 구분할 수 있음
- X를 가지고 Y를 구분할 수 있음

→ X 정해지면 Y도 정해진다.

0 에제에 도움이 되지 않는 경우

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

$X=0$, $X=1$ 일 때의 조건부확률분포는 다음과 같다.

$$P(Y=0|X=0)=\frac{1}{3},\quad P(Y=1|X=0)=\frac{2}{3}$$

$$P(Y=0|X=1)=\frac{1}{3},\quad P(Y=1|X=1)=\frac{2}{3}$$

두 경우, 모두 Y의 엔트로피는 약 0.92다.

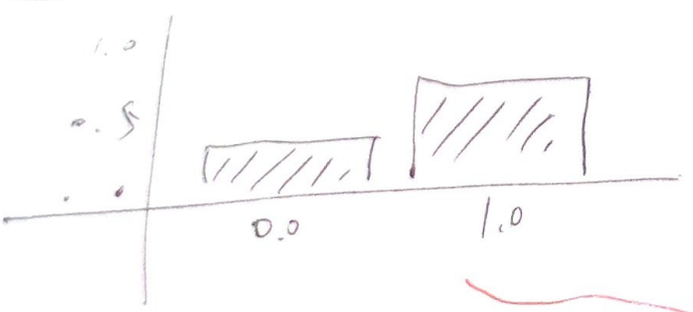
$$H[Y|X=0]=H[Y|X=1]=-\frac{1}{3}\log_2\frac{1}{3}-\frac{2}{3}\log_2\frac{2}{3}\approx 0.92$$

위 값들을 가중평균하면, 조건부 엔트로피값은 똑같이 약 0.92다.

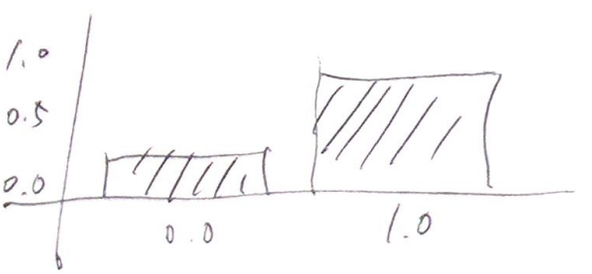
$$H[Y|X]=\frac{1}{3}H[Y|X=0]+\frac{2}{3}H[Y|X=1]\approx 0.92$$

+ 지식in

조건부확률분포 $p(Y|X=0)$



조건부확률분포 $p(Y|X=1)$



// X의 값이 Y에 대해 아무런 추가정보를 주지 못함.

III 조건부 엔트로피를 사용한 스팸메일 분류문제

<학습용 메일 데이터 80개>

- 40개 정상 (Y=0)
- 40개 스팸 (Y=1)

스팸메일 여부를 "특정 키워드가 존재하는지" ($X=1$) 혹은 "특정 키워드가 존재하지 않는지" ($X=0$) 여부로 판단하고자 함.

↳ 키워드후보: X_1, X_2, X_3

(X_1, Y의 관계)

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

(X_2, Y의 관계)

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

(X_3, Y의 관계)

	Y=0	Y=1	
$X_3=0$	0	40	40
$X_3=1$	40	0	40
	40	40	80

* 누가 더 좋은 키워드? (X_1, X_2, X_3)

⇒ 조건부 엔트로피 값이 최소화되는 X

① X_1, Y의 조건부 엔트로피

$$H[Y|X_1] = p(X_1=0)\,H[Y|X_1=0] + p(X_1=1)\,H[Y|X_1=1]$$

$$= \frac{40}{80} \cdot 0.81 + \frac{40}{80} \cdot 0.81 = 0.81$$

② X_2, Y의 조건부 엔트로피

$$H[Y|X_2] = p(X_2=0)\,H[Y|X_2=0] + p(X_2=1)\,H[Y|X_2=1]$$

$$= \frac{60}{80} \cdot 0.92 + \frac{20}{80} \cdot 0 = 0.69$$

③ X_3, Y의 조건부 엔트로피

$$H[Y|X_3] = p(X_3=0)\,H[Y|X_3=0] + p(X_3=1)\,H[Y|X_3=1] = 0$$

⇒ X_3, X_2, X_1 순으로 좋은 키워드.

⇒ 의사결정나무 (decision tree) 분류모형은
조건부엔트로피를 사용하여 가장 좋은 특징값과 기준을 선택함.