Probability basics

Sang Yup Lee

Review

- 분류 모형에서 사용되는 비용함수인 교차 엔트로피는 MLE를 통해 도출될 수 있음
- MLE
  - 비선형모형 (예, 로지스틱 회귀 모형)의 파라미터 값을 계산할 때 사용되는 추정 방법
  - 학습 데이터에 존재하는 종속변수의 확률 (likelihood)을 최대화하는 파라미터의 값
  - Likelihood
    - 학습 데이터에 존재하는 각 관측치의 종속변수에 대한 결합확률로 표현
    - $P(Y_1 = y_1, Y_2 = y_2, \dots, Y_n = y_n)$
    - 각 관측치의 종속변수는 독립이라고 가정
    - $\prod_{i=1}^{n} P_i(Y_i = y_i)$
    - 이산변수인 경우 $P_i(Y_i = y_i)$ 는 pmf를, 연속변수인 경우 pdf를 사용해서 표현
    - 종속변수가 베르누이 변수인 경우, 베르누이 분포의 pmf 사용, $\prod_{i=1}^{n} p^{y_i}(1-p)^{1-y_i}$
  - Log-likelihood
    - $\sum_{i=1}^{n} \{y_i \log p + (1-y_i)\log(1-p)\}$

Cross Entropy

Cross entropy

- 교차 엔트로피 비용함수
  - MLE를 가지고 도출 가능
  - 정보이론에서의 교차 엔트로피 개념을 이용해서도 표현 가능
    - 교차 엔트로피라고 하는 이름은 정보이론에서 온 것임

Entropy

- Definition
  - 정보이론에서 사용되는 개념으로 한 변수의 엔트로피를 의미
  - 변수의 불확실성 (uncertainty)을 의미
  - 확률분포 p를 갖는 어떠한 변수 X에 대해서 엔트로피는 다음과 같이 정의
    - 여기서는 설명을 위해서 변수 X가 이산변수, 즉, 한정된 수(K개)의 값을 취하는 변수, 라고 가정
    - $\mathrm{H(p)} = -\sum_{k=1}^{K} p(X=k) \log p(X=k)$
      - 연속변수의 경우, $-\int_X f(x) \log(fx)\, dx$
    - 해당 값은 X가 특정한 값 하나를 가질 확률이 1인 경우에 최소 (그 때의 값은 0, 즉, 불확실성이 없다는 것을 의미). 반대로 각 값을 가질 확률이 동일한 경우, 즉 위의 경우는 p(X=k) = 1/K 인 경우 최대 (이러한 경우 불확실성이 제일 크다는 것을 의미)
      - 예) $x \in \{1,2,3,4,5\}$

Cross entropy

- Definition
  - 엔트로피가 하나의 변수 혹은 해당 변수가 갖는 (하나의) 확률 분포의 불확실성을 의미한다면, 교차 엔트로피는 하나의 변수 (즉, X)가 가질 수 있는 서로 다른 분포*들 간의 차이를 의미

    * 동일한 분포라도 형태가 다를 수 있음
  - 아래와 같이 정의
    - X가 갖는 분포가 p와 q 인 경우
    - $\mathrm{H(p,q)} = -\sum_{k=1}^{K} p(X=k)\log q(X=k)$
    - 두 분포의 차이가 클수록 해당 값 증가

Cross entropy

- 기계학습에서의 분류 문제의 경우
  - 분류의 문제: 종속변수에 대한 교차 엔트로피 사용
  - 분류의 문제(예, 로지스틱 회귀 모형을 사용하는 경우)에서는 p를 y의 실제값 그리고 q를 모형을 통해서 예측된 값을 사용
  - $y \in \{0,1\}$ 인 경우 $\Rightarrow \mathrm{p} \in \{\mathrm{y}, 1-\mathrm{y}\}$ 그리고 $\mathrm{q} \in \{\hat{\mathrm{y}}, 1-\hat{\mathrm{y}}\}$
  - 하나의 관측치에 대해서 교차 엔트로피는 다음과 같이 표현
  - $-[y \ln \hat{y} + (1-y) \ln(1-\hat{y})]$, $\hat{y}$은 P(y=1)을 의미
  - 즉, y = 1 인 경우, $-\log P(y=1)$; 0인 경우 $-\log P(y=0)$
  - 종속변수가 취할 수 있는 값이 3개 이상인 경우 (하나의 관측치에 대해) 다음과 같이 표현
  - $-\boldsymbol{y_i} \bullet \boldsymbol{ln\_p}$, where $\boldsymbol{y_i}$ = one hot vector, $\boldsymbol{ln\_p}$ is also a vector
  - When y takes K values, $\boldsymbol{ln\_p} = \big(\ln(P(y=1)), \ln(P(y=2)), \dots, \ln(P(y=K))\big)$
  - $y \in \{0,1,2\}$, 특정 관측치의 종속변수 실제값 = 0 인 경우 아래와 같이 표현
    - $-(1,0,0) \bullet \big(\ln(P(y=0)), \ln(P(y=1)), \ln(P(y=2))\big) = -\ln(P(y=0))$

Q & A