1. MLE란?

확률변수 $X_1, X_2, \cdots, X_n$이 $X_1 = x_1, X_2 = x_2, \cdots, X_n = x_n$으로 관측되었을 때, 우도함수는 $L(\theta|x_1, x_2, \cdots, x_n) = f_\theta(x_1, x_2, \cdots, x_n)$으로 정의된다. 여기에서 $f_\theta(x_1, x_2, \cdots, x_n)$는 확률분포 함수(연속형일 때는 pdf, discrete일 때는 pmf)이다. $x_1, x_2, \cdots, x_n$이 구체적으로 관측되었으므로 $L(\theta|x_1, x_2, \cdots, x_n)$은 오직 θ만의 함수이다. 즉, 고정된 $x_1, x_2, \cdots, x_n$에서 θ가 변하면 $L(\theta|x_1, x_2, \cdots, x_n)$도 변하며, $L(\theta_1|x_1, x_2, \cdots, x_n) \geq L(\theta_2|x_1, x_2, \cdots, x_n)$이면 θ_1이 θ_2보다 좋은 추정치라고 한다. 만약 모든 θ에 대해 $L(\theta|x_1, x_2, \cdots, x_n) \leq L(\theta^*|x_1, x_2, \cdots, x_n)$를 만족하는 추정치 θ^*가 존재하면 이를 우도함수를 이용한 optimal 추정치라고 말하고 통계학에서는 이를 MLE라고 한다. $\log L(\theta|x_1, x_2, \cdots, x_n)$과 $L(\theta|x_1, x_2, \cdots, x_n)$는 1:1 대응되는 함수이므로 로그우도함수도 동일한 해석이 된다.

$-\log L(\theta_1|x_1, x_2, \cdots, x_n) \leq -\log L(\theta_2|x_1, x_2, \cdots, x_n)$이므로

$-\log L(\theta|x_1, x_2, \cdots, x_n)$은 loss함수가 된다.

2. 우도함수와 결합분포의 차이

$L(\theta|x_1, x_2, \cdots, x_n) = f_\theta(x_1, x_2, \cdots, x_n)$이다. $f_\theta(x_1, x_2, \cdots, x_n)$는 고정된 θ에서 $x_1, x_2, \cdots, x_n$의 함수이므로 $x_1, x_2, \cdots, x_n$이 변하면 $f_\theta(x_1, x_2, \cdots, x_n)$값도 변하며 결합확률분포함수이므로 $\int \cdots \int f_\theta(x_1, x_2, x_n) dx_1 dx_2 \cdots dx_n = 1$을 만족해야 한다. 그러나 $L(\theta|x_1, x_2, \cdots, x_n)$은 오직 θ만의 함수이므로 결합확률분포가 아니다. 그러므로 $\int \log L(\theta|x_1, x_2, x_n) d\theta$는 $-\infty \sim \infty$의 값을 가지게 된다.

3. 각 분포의 로그우도 함수 도출

Bernoulli: $Y_1, Y_2, \cdots, Y_n$ 을 서로 간에 독립이며 $P(Y_i = 1) = p$, $P(Y_i = 0) = 1 - p$ for all i (이것이 identical 임)

$$P(Y_1 = y_1, Y_2 = y_2, \cdots, Y_n = y_n) = \prod_{i=1}^{n} p^{y_i}(1-p)^{1-y_i}$$

이다. 그러므로 $\log L(p) = \sum_{i=1}^{n} (y_i \log p + (1-y_i)\log(1-p))$가 된다.

Binomial 분포: 이항분포의 확률변수는 Bernoulii 확률변수의 합이다. 즉, $X=\sum_{i=1}^{n} Y_i$이므로 $X=0,1,\cdots,n$의 값을 가지게 된다.

$$P(X=k)=\binom{n}{c}p^k(1-p)^{n-k},\ \text{여기에서}\ k=\sum_{i=1}^{n}y_i\text{이다.}$$

그러므로 $\log L(p)=\log\binom{n}{c}+\sum_{i=1}^{n}y_i\log p+\sum_{i=1}^{n}(1-y_i)\log(1-p)$

$$=\log\binom{n}{c}+\sum_{i=1}^{n}(y_i\log p+(1-y_i)\log(1-p))$$

즉, Bernoulli 로그우도함수와 이항분포의 우도함수는 모수 p의 관점에서 같다는 것을 알 수 있다.

다항분포: 다항분포는 이항분포의 일반화로, 이항분포가 두 개의 클래스로 분류하는 것이 목적이고 다항분포는 3개이상의 클래스로 분류하는 것이 목적이다. 다항분포는 n개의 객체를 $1,2,\ \cdots,c$의 클래스로 분류했을 때, 각 클래스의 개수가 확률변수이다. 이를 위해 i번째 객체는 $Y_{i1}, Y_{i2}, \cdots, Y_{ic}$으로 표현되고 i번째 객체가 j번째 클래스에 속하면 $Y_{ij}=1$이고 $Y_{ij'}=0,\ j'\neq j$가 된다. 즉 $\sum_{j=1}^{c}Y_{ij}=1$이며 이는 one-hot encoding과 동일하다. 예를 들어, $c=3$일 때, i번째 객체가 두 번째 클래스에 속하면 $Y_{i1}=0,\ Y_{i2}=1,\ Y_{i3}=0$이 된다.

$X_j=\sum_{i=1}^{n}Y_{ij}$으로 정의하면 X_j는 n개의 객체에서 j번째 클래스에 속한 객체의 수를 말한다.

$[Y_{11}, Y_{12}, \cdots, Y_{1c}]$, $\cdots$ $[Y_{i1}, Y_{i2}, \cdots, Y_{ic}]$, $\cdots$, $[Y_{n1}, Y_{n2}, \cdots, Y_{nc}]$이 서로 간에 독립이고 $P(Y_{ij}=1)=p_j$이면 for all i (이것이 identical임),

$$P(X_1=x_1,X_2=x_2,\cdots,X_c=x_c)=\binom{n}{x_1,x_2,\cdots,x_c}\prod_{j=1}^{c}p_i^{x_j}$$

$$=\binom{n}{x_1,x_2,\cdots,x_c}\prod_{j=1}^{c}p_i^{\sum_{i=1}^{n}y_{ij}}$$

이다. 그러므로

$$\log L(p_1,p_2,\ \cdots,p_c) = \log\binom{n}{x_1,x_2,\cdots,x_n} + \sum_{j=1}^{c}\sum_{i=1}^{n} y_{ij}\log p_j$$
$$\approx \sum_{i=1}^{n}\sum_{j=1}^{c} y_{ij}\log p_j$$

정규분포: $Y_1, Y_2, \cdots, Y_n$을 서로 간에 독립이며 평균이 μ이고 분산이 σ^2인 정규분포(이것이 identical임) 확률변수일 때

$$\log L(\mu) \approx -\frac{\sum_{i=1}^{n}(y_i - \mu)}{\sigma^2}$$

이다.

4. 전통적인 통계학과 머신러닝에서의 가정의 차이

머신러닝의 이항분류는 $P(Y_i = 1) = p_i$로, 이항분포(또는 동일하게 Bernoulli)에서

$P(Y_i = 1) = p$로 가정한 것이 근본적인 차이이다. 즉, 머신러닝에서는 identical 가정을 없앤 일반화한 중요한 차이점을 가지고 있다. 그러나 머신러닝의 이항분류 $p_i = p_\theta(\boldsymbol{x}_i)$으로 정의하여 모수가 관측치 i에 의존하지 않도록 하고 있다. 예를 들어, $p_\theta(\boldsymbol{x}_i) = \dfrac{1}{1+e^{-\theta\ \cdot\ \boldsymbol{x}_i}}$ (이를 통계학에서는 logistic regression이라고 한다), 또는 $p_\theta(\boldsymbol{x}_i)$는 MLP, 또는 CNN, 또는 RNN으로 출력된 이항 확률값이다.

5. 음의 로그우도함수와 loss function

이미 3번에서 설명했듯이 베루누이 로그우도함수는

$$\log L(p) = \sum_{i=1}^{n}(y_i \log p + (1-y_i)\log(1-p))$$

이고 binary crossentropy는

$$-\sum_{i=1}^{n}(y_i \log p_\theta(\boldsymbol{x}_i) + (1-y_i)\log(1-p_\theta(\boldsymbol{x}_i)))$$

이다. 그러므로 이는 음의 베르누이 우도함수를 일반화한 것이다. 여기에서 $p_\theta(\boldsymbol{x}_i)$는 4번에서 설명한 것과 동일하다.

다항분포의 로그우도함수는

$$\log L(p_1,p_2,\ \cdots,p_c)\ \approx \sum_{i=1}^{n}\sum_{j=1}^{c} y_{ij}\log p_j$$

이고 categorical crossentropy는

$$-\sum_{i=1}^{n}\sum_{j=1}^{c} y_{ij}\log p_{\theta_j}(\boldsymbol{x}_i)$$

으로 일반화한 음의 다항분포 로그우도함수이다.

정규분포 우도함수는 $\log L(\mu) \approx -\frac{\sum_{i=1}^{n}(y_i-\mu)}{\sigma^2} \approx -\sum_{i=1}^{n}(y_i-\mu)$이다. 머신러닝의 SSE는 $\sum_{i=1}^{n}(y_i-\mu_\theta(\boldsymbol{x}_i))$으로 정의하여 일반화된(즉 identical 가정을 제거한) 음의 정규분포 로그우도함수가 된다.