우도의 정의

⇒ 확률변수 X가 매개변수 θ에 대한 확률분포 $P_\theta(X)$를 가지고, X를 시행하여 특정 데이터 x를 얻었다고 하자.
θ에 따라 변화.

- θ의 가능도(우도)함수는 $L(\theta|x)$로 쓰고, 다음과 같음.

$$L(\theta|x) = P_\theta(x) = P_\theta(X=x)$$

$\underline{w}$ (가중치 벡터)　　확률밀도함수

⇒ 확률변수 X가 벡터라면, 전체 우도는 각 성분별 우도의 곱.

- $L(\theta|x) \neq P(\theta|x)$
- $P(\theta|x)$: x가 주어져 있을 때 θ의 확률.
  (Posterior probability as θ given x)

우도와 확률 비교 예시

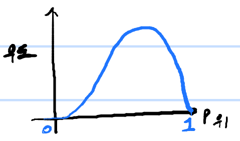
- 확률은 어떤 모델(확률분포)이 먼저 주어져 있는 상황에서 특정 데이터셋에 대해 얻을 수 있음.

ex) (주어진 확률분포) 동전던지기의 앞/뒷면이 나올 확률은 각각 $\frac{1}{2}$.

(특정 데이터셋의 확률 얻기)

→ 특정 데이터셋 {위, 위, 아래}을 얻을 확률은 $\left(\frac{1}{2}\right)^3 = \frac{1}{8}$

→ 우도는 어떤 데이터셋이 먼저 주어진 상황에서 특정 모델이 대해 얻을 수 있음.



ex) 특정 데이터셋 {위, 위, 아래}

→ 1. (위 $\frac{2}{3}$, 아래 $\frac{1}{3}$)인 모델. ⇒ 우도 $= \frac{2}{3}\cdot\frac{2}{3}\cdot\frac{1}{3} = \frac{4}{27}$ → 우도가 더 큰 모델.

→ 2. (위 $\frac{1}{2}$, 아래 $\frac{1}{2}$)인 모델 ⇒ 우도 $= \frac{1}{2}\cdot\frac{1}{2}\cdot\frac{1}{2} = \frac{1}{8}$

Logistic Regression (비용함수를 만들기)

1. $h(\underline{x}) = \sigma(\underline{w}^T\underline{x})$

$$\sigma(s) = \frac{e^s}{1+e^s}$$

2. 학습하고자 하는 대상

$$P(y|\underline{x}) = \begin{cases} f(\underline{x}), & y=+1 \\ 1-f(\underline{x}), & y=-1 \end{cases} \rightarrow P(y|\underline{x}) = \begin{cases} h(\underline{x}), & y=+1 \\ 1-h(\underline{x}), & y=-1 \end{cases}$$

$$P(y|\underline{x}) = \sigma(y\cdot\underline{w}^T\underline{x})$$

$1-\sigma(s) = \sigma(-s) \sim \frac{e^s}{1+e^s} = \frac{1}{1+e^{-s}}$

$\hookrightarrow 1-\sigma(s) = \frac{1+e^s-e^s}{1+e^s} = \frac{1}{1+e^s} = \frac{e^{-s}}{e^{-s}+1} = \frac{e^{-s}}{1+e^{-s}} = \sigma(-s)$

3. 훈련데이터셋에 포함된 $x = \{(\underline{x}_1, y_1), \cdots, (\underline{x}_N, y_N)\}$들이 모두 독립적으로 생성된 것이라 가정한다면, 해당 입력데이터 $\underline{x}_1, \cdots, \underline{x}_N$들이 주어졌을 때 출력 $y_1, \cdots y_N$들이 나올 확률은 모두 곱해서 얻어짐 (독립사건의 확률)

⇒ $\prod_{n=1}^{N} P(y_n|\underline{x}_n) = L(\underline{w}|x)$ → 최대화하는 Dataset에 맞는 model.
(1)

4. 로지스틱 회귀에서의 최적모델은, 이를 최대화하는 model.

5. (1)에 $-\frac{1}{N}\ln(\cdot)$를 적용.

⇒ $-\frac{1}{N}\ln\left(\prod_{n=1}^{N} P(y_n|\underline{x}_n)\right) = \frac{1}{N}\sum_{n=1}^{N}\ln\left(\frac{1}{\theta(y_n\underline{w}^T\underline{x}_n)}\right)$ ⇒ (1)번 최대화는 (2)번 최소화문제와 동일

$= \frac{1}{N}\sum_{n=1}^{N}\ln\left|1+e^{-y_n\underline{w}^T\underline{x}_n}\right| \triangleq E_{in}$ − (2) (오그우드)

6. (1)을 최대화하는 문제는 (2)를 최소화하는 문제와 수학적으로 같음.

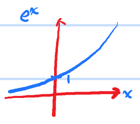
각 항들의 곱으로 표현되던 식이 합으로 표현되는 식으로 바뀜.

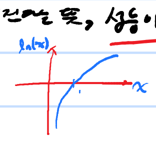
⇒ 개별 항은 각 데이터 샘플 $(\underline{x}_n, y_n)$에서 대응하는 오차값으로 생각

7. 개별 데이터 샘플에 대한 오차.

$\underset{\text{error}}{e}(\sigma(\underline{x}_n), y_n) = \ln(1+e^{-y_n\underline{w}^T\underline{x}_n})$

- $y_n\underline{w}^T\underline{x}_n$ 값이 큰 양수일수록 오차 $e(\cdot)$값은 작을 값.
  정분류한 확률이 1 가까워진다는 뜻, 성능이 좋은 모델이라는 의미





Cross-Entropy Error Measure

- 두 확률변수 P, q의 Cross-Entropy error는 다음과 같이 정의

ex)
P: Gaussian 분포
q: Logistic Regression

⇒ $p\log\frac{1}{q} + (1-p)\log\frac{1}{1-q}$

- 어떤 확률의 오차를 최소화하는 모델을 학습하고자 하는 경우에는 비용함수로 두 확률의 Cross-Entropy error를 사용하는 것을 고려.