Logistic Regression

Jeonghun Yoon

Terms

Odds

Log odds (logit)

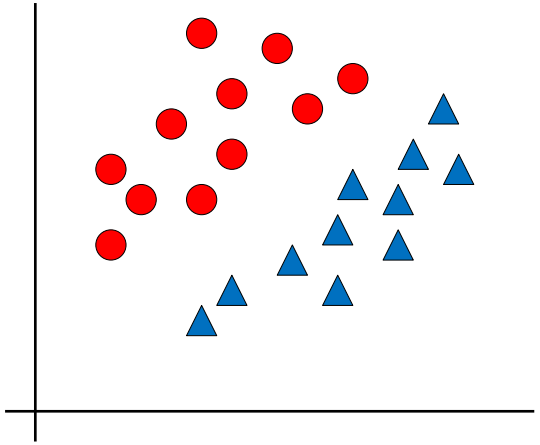
Sigmoid function

Logistic regression

Odds ratio

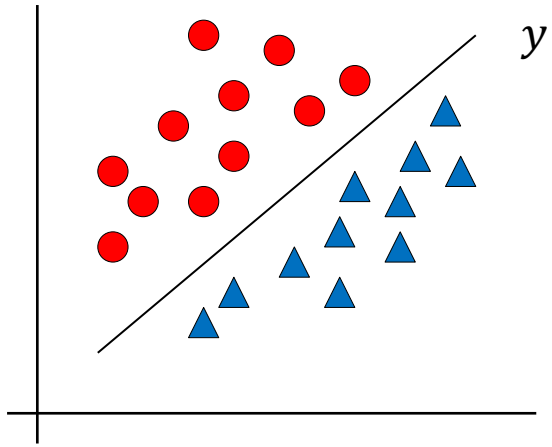
Intro

Question : 2개의 cluster로 나누고 싶다. How?



Intro

그렇다면, 우리는 GLM(Generalized Linear Models), 즉 선형모델을 classification 에 사용할 수 있을까?



$$y = \Theta^T \mathbb{x} = \theta_0 + \theta_1 x$$

(Fitted from data)

일반화 선형모형(Generalized Linear Model)

회귀분석이나 분산분석은 종속변수가 정규분포되어 있는 연속형 변수이다. 하지만 많은 경우에 있어서 종속변수가 정규분포되어 있다는 가정을 할 수 없는 경우도 있으며 범주형 변수가 종속변수인 경우도 있다. 다음과 같은 경우에 일반화 선형모형을 사용한다.

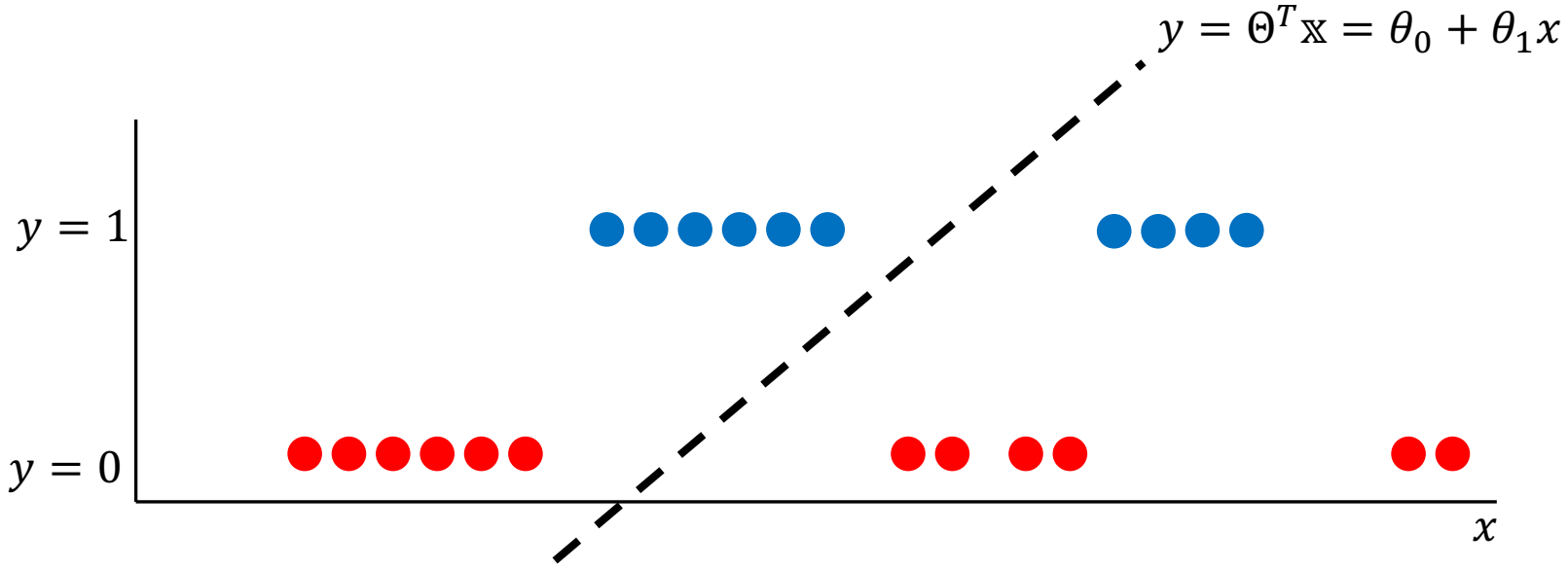
- 종속변수가 범주형변수인 경우 : 이항변수(0 또는 1, 합격/불합격, 사망.생존 등)인 경우도 있으며 다항변수(예를 들어 poor/good/excellent 또는 공화당/민주당/무소속 등)인 경우 정규분포를 따르지 않는다.
- 종속변수가 count(예를 들면 한 주간 교통사고 발생 건수, 하루에 마시는 물이 몇잔인지 등)인 경우. 이들 값은 매우 제한적이며 음수가 되지 않고 평균과 분산이 밀접하게 관련되어 있고 정규분포를 따르지 않는다.

일반화 선형 모형은 종속변수가 정규분포하지 않는 경우를 포함하는 선형모형의 확장이며 대표적으로 로지스틱회귀(Logistic regression)와 포아송회귀(Poisson regression)가 있다.

(http://cs229.stanford.edu/notes/cs229-notes1.pdf 참고)

Intro

Classification 에서는 y가 연속 값을 갖는 것이 아니고 불연속의 값을 갖는다. Binary classification일 경우를 살펴보자.



- $y \in \{0,1\}$ 인데 $\Theta^T \mathbb{X}$ 가 1보다 크거나, 0보다 작은 수를 가지게 될 수 있다.
- x의 값이 증가 할수록 에러의 크기도 증가하고 있다. (선형회귀분석의 가정에 어긋난다.)

Motivation

$y = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \cdots + \theta_n x_n$과 같은 선형 모델의 장점은 특성 x_i가 y에 미치는 영향을 설명하기 쉽다는 것이다.

그렇다면, GLM선형모델을 이진 분류(binary classification)에 사용할 방법은 없을까?

$$\Theta^T \mathbb{X} = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \cdots + \theta_n x_n = y \qquad y = \begin{cases} 0 \\ 1 \end{cases}$$

이렇게 maping rule을 바꾸면 어떨까?

$$\Theta^T \mathbb{X} = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \cdots + \theta_n x_n = \boldsymbol{P}$$

$$\boldsymbol{P} = \begin{cases} \boldsymbol{P} \geq thres : 1 \\ \boldsymbol{P} < thres : 0 \end{cases}$$

$$(-\infty \leq \boldsymbol{P} \leq \infty)$$

Motivation

$\Theta^T\mathbb{X}$의 값을 0 또는 1로 바로 사용하지 않고, 0 또는 1에 속할 **확률**을 나타내는 값으로 mapping 한다. 단 이 확률 값은 0과 1사이의 값이 아닌 $-\infty, \infty$ 사이의 값이다. 이것을 위해 다음 개념을 사용한다.

- Odds$_{오즈}$: 성공(1)과 실패(0)의 비율. $Odds(Y=1) = \frac{p}{1-p}$ p : label 값이 1이 될 확률
- Logit$_{로짓}$ (Log odds$_{로그\ 오즈}$) : $-\infty \sim \infty$의 범위에서 어떤 클래스에 속할 확률을 결정하는 함수

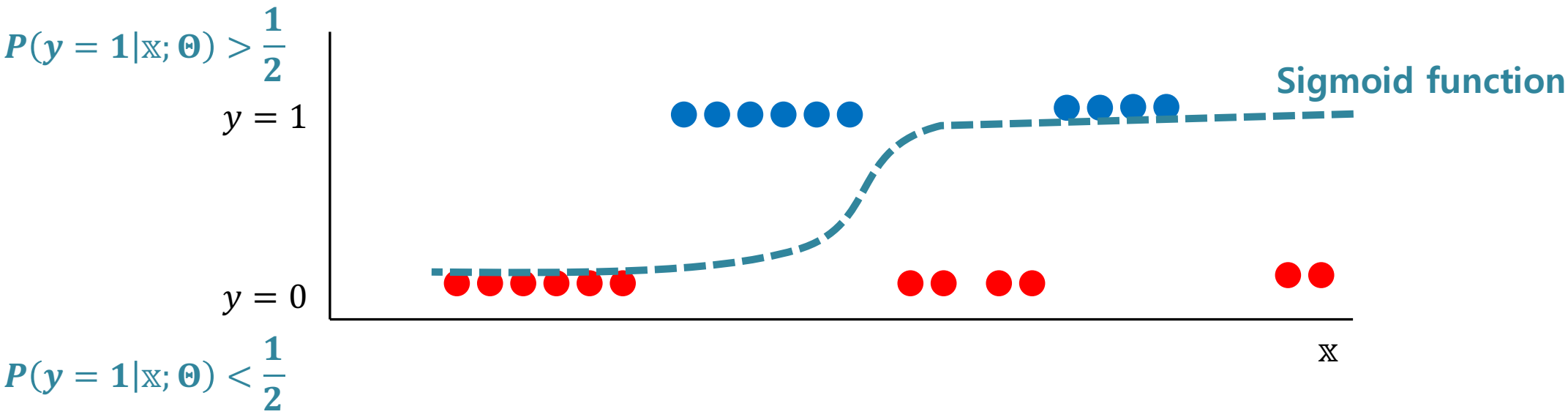
Rule of mapping : $\log\frac{p}{1-p} = \Theta^T\mathbb{X}$

- 여기서 p는 label 값이 1이 될 확률을 의미한다. $\log\frac{p}{1-p} > 0$ 이면 1로 분류하고, $\log\frac{p}{1-p} < 0$ 이면 0으로 분류한다.
- 사실 다른 mapping 도 존재하지만 logit(log odds)을 사용하는 것이 가장 보편적인 방법이다.

따라서, $\boldsymbol{p(y=1|\mathbb{X};\Theta) = \frac{1}{1+e^{-(\Theta^T\mathbb{X})}}}$ 로 식을 다시 쓸 수 있고,

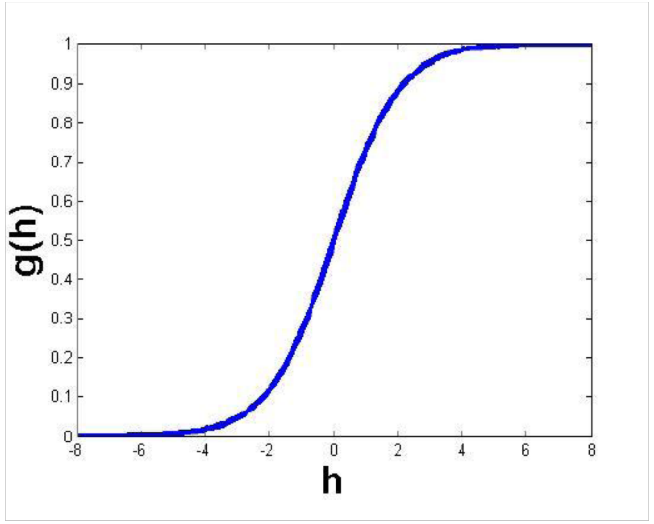
- $p(y=1|\mathbb{X};\Theta) > \frac{1}{2}$ 이면 1로 분류하고
- $p(y=1|\mathbb{X};\Theta) < \frac{1}{2}$ 이면 0으로 분류한다.

Logistic Regression



Sigmoid Function

$$g(h) = \frac{1}{1 + e^{-t}} \qquad 0 \leq g(h) \leq 1$$



$$P(y = 1|\mathbb{x}; \Theta) = w_\Theta(\mathbb{x}) = g(\Theta^T \mathbb{x}) = \frac{1}{1 + e^{-\Theta^T \mathbb{x}}}$$

$$P(y = 0|\mathbb{x}; \Theta) = 1 - w_\Theta(\mathbb{x}) = 1 - g(\Theta^T \mathbb{x}) = \frac{e^{-\Theta^T \mathbb{x}}}{1 + e^{-\Theta^T \mathbb{x}}}$$

Coefficient for Logistic Regression

로지스틱 회귀의 계수를 해석하기 위해서는 odds ratio오즈비를 알아야 한다.

$$odds\ ratio = \frac{odds(Y=1|X=1)}{odds(Y=1|X=0)}$$

(분자: 이진변수가 존재할 때, 분모: 이진변수가 존재하지 않을 때)

- 이진 요인 변수 X에 대하여, 요인 X가 존재할 때 $Y=1$인 odds와 요인 X가 존재하지 않을 때 $Y=1$인 odds 를 비교한 비율
- 로지스틱 회귀분석에서 계수 θ_i는 x_i에 대한 odds ratio의 로그 값이다. ($\log\frac{p}{1-p} = \Theta^T\mathbb{X}$ 임을 기억)

$$\frac{1}{1+\exp(-(\theta_1x_1+\theta_2x_2+\theta_3x_3+\cdots+\theta_nx_n))}$$

$$\log\frac{odds(Y=1|X=1)}{odds(Y=1|X=0)}$$

- (x_i가 있을 때의 $\log\frac{p}{1-p}$) − (x_i가 없을 때의 $\log\frac{p}{1-p}$)
- $\log a - \log b = \log\frac{a}{b}$

Training

그러면, 회귀 식

$$P(y = 1|\mathbb{x}; \Theta) = w_{\Theta}(\mathbb{x}) = g(\Theta^T \mathbb{x}) = \frac{1}{1 + e^{-\Theta^T \mathbb{x}}}$$

어떻게 구할까?

즉, 회귀 식의 모수(parameter)를 어떻게 구할까?

Maximum Likelihood Estimator (M.L.E.)

Maximum Likelihood Estimator

① $P(y=1|\mathbb{x};\Theta) = w_\Theta(\mathbb{x})$

② $P(y=0|\mathbb{x};\Theta) = 1 - w_\Theta(\mathbb{x})$

③ $w_\Theta(\mathbb{x}) = \frac{1}{1+e^{-\Theta^T\mathbb{x}}}$

$$\Big\} \; p(y|\mathbb{x};\Theta) = \big(w_\Theta(\mathbb{x})\big)^y \big(1 - w_\Theta(\mathbb{x})\big)^{1-y}$$

$\mathbb{x} = (x_1, \dots, x_n)^T \in R^n, \quad \Theta = (\theta_0, \theta_1, \dots, \theta_n),$

training data points $X = (\mathbb{x}_1, \dots, \mathbb{x}_m)$ 과 각 data points에 대응하는 label

$Y = (y_1, \dots, y_m)$이 주어졌을 때, likelihood를 구하는 공식은 아래와 같다. 단 $y_i \in \{0,1\}$

$L(\Theta) = p(Y|X;\Theta) = p(y_1, \dots, y_m \mid \mathbb{x}_1, \dots, \mathbb{x}_m\,;\, \theta_0, \dots, \theta_m)$

$= \prod_{i=1}^m p(y_i|\mathbb{x}_i;\Theta)$

$= \prod_{i=1}^m \big(w_\Theta(\mathbb{x}_i)\big)^{y_i} \big(1 - w_\Theta(\mathbb{x}_i)\big)^{1-y_i}$

Maximum Likelihood Estimator

우리는, $L(\Theta)$를 최대값이 나오도록 하는 모수 Θ를 찾는 것이 목표이다.

즉, Maximum Likelihood Estimate는

$$\Theta = \arg\max_{\Theta} L(\Theta)$$

$$= \arg\max_{\theta} \prod_{i=1}^{m} \left(w_{\Theta}(\mathbb{x}_i)\right)^{y_i} \left(1 - w_{\Theta}(\mathbb{x}_i)\right)^{1-y_i}$$

Maximum Likelihood Estimator

유도 된 MLE식을 풀기 위해서 log를 사용한다.

- $P(y=1|\mathbb{x};\Theta) = w_\Theta(\mathbb{x}) = \frac{1}{1+e^{-\Theta^T\mathbb{x}}} = \frac{e^{\Theta^T\mathbb{x}}}{1+e^{\Theta^T\mathbb{x}}}$
- $P(y=0|\mathbb{x};\Theta) = 1 - w_\Theta(\mathbb{x}) = \frac{e^{-\Theta^T\mathbb{x}}}{1+e^{-\Theta^T\mathbb{x}}} = \frac{1}{1+e^{\Theta^T\mathbb{x}}}$

$$l(\Theta)$$

$$= logL(\Theta)$$

$$= \sum_{i=1}^{m}\{y_i \log w_\Theta(\mathbb{x}_i) + (1-y_i)\log(1-w_\Theta(\mathbb{x}_i))\}$$

$$= \sum_{i=1}^{m}\left\{y_i \log\frac{w_\Theta(\mathbb{x}_i)}{(1-w_\Theta(\mathbb{x}_i))} + \log(1-w_\Theta(\mathbb{x}_i))\right\}$$

$$= \sum_{i=1}^{m}\left\{y_i\Theta^T\mathbb{x}_i - \log(1+e^{\Theta^T\mathbb{x}_i})\right\}$$

Concave 함수의 성질에 의하여 $l(\Theta)$는 극대 값을 갖는다.

Maximum Likelihood Estimator

MLE의 극대 값을 구하기 위하여 gradient ascent를 사용할 것이다. $l(\Theta)$를 θ_j에 관하여 미분한 식을 구하자.

- $\mathbb{x}_i = \left(x_1^{(i)}, \dots, x_n^{(i)}\right)^T \in R^n, \quad \Theta = (\theta_0, \theta_1, \dots, \theta_n)$

$$\frac{\partial}{\partial \theta_j} l(\Theta)$$

$$= \sum_{i=1}^{m} \{ y_i x_j^{(i)} - \frac{e^{\Theta^T \mathbb{x}_i}}{\left(1+e^{\Theta^T \mathbb{x}_i}\right)} x_j^{(i)} \}$$

$$= \sum_{i=1}^{m} \{ y_i x_j^{(i)} - P(y_i = 1 | \mathbb{x}_i; \Theta) x_j^{(i)} \}$$

$$= \sum_{i=1}^{m} x_j^{(i)} \{ y_i - P(y_i = 1 | \mathbb{x}_i; \Theta) \}$$

Maximum Likelihood Estimator

Gradient ascent 공식에 의하여, $\theta_j = \theta_j + \alpha \frac{\partial}{\partial \theta_j} l(\Theta)$

$$\frac{\partial}{\partial \theta_j} l(\Theta) = \sum_{i=1}^{m} \left\{ y_i x_j^{(i)} - \frac{e^{\Theta^T \mathbb{x}_i}}{\left(1 + e^{\Theta^T \mathbb{x}_i}\right)} x_j^{(i)} \right\} = \sum_{i=1}^{m} x_j^{(i)} \{ \underbrace{y_i - P(y_i = 1 | \mathbb{x}_i; \Theta)}_{\text{Prediction error}} \}$$

- Prediction error : 관찰 된 y_i와, y_i의 예측 된 확률의 차이

Maximum Likelihood Estimator

1. α를 선택한다.

2. $\Theta = (\theta_0, \theta_1, \ldots, \theta_n)$의 적당한 초기 값을 설정한다.

3. 모든 j에 대하여, $\theta_j \leftarrow \theta_j + \alpha \frac{\partial}{\partial \theta_j} l(\Theta) = \theta_j + \alpha \sum_{i=1}^{m} x_j^{(i)} \{y_i - P(y_i = 1 | \mathbb{x}_i; \Theta)\}$

4. if, 모든 j에 대하여 $\sum_{i=1}^{m} x_j^{(i)} \{y_i - P(y_i = 1 | \mathbb{x}_i; \Theta)\}$의 값의 변화가 없으면 멈춘다.

   otherwise, 3번으로 간다.

Maximum A Priori Estimator

우리는 과적합(overfitting)을 방지하기 위하여 MLE 대신에 MAP를 사용한다. 즉, Penalized log likelihood function을 이용하여 Θ의 큰(large)값에 제약을 주는 것이다.

MLE : $\Theta = \arg\max_{\Theta} L(\Theta) = \arg\max_{\Theta} \prod_{i=1}^{m} p(y_i|\mathbb{x}_i;\Theta)$

MAP : $\Theta = \arg\max_{\Theta} L(\Theta)p(\Theta) = \arg\max_{\Theta} \prod_{i=1}^{m} p(y_i|\mathbb{x}_i;\Theta)\, p(\Theta)$

MLE(log) : $\Theta = \arg\max_{\Theta} \sum_i \log p(y_i|\mathbb{x}_i;\Theta)$

MAP(log) : $\Theta = \arg\max_{\Theta} \sum_i \log p(y_i|\mathbb{x}_i;\Theta) + \log p(\Theta)$

Maximum Apriori Estimator

$p(\Theta)$는 여러 가지 분포가 사용될 수 있으나, $\theta_i \sim N(0, \sigma^2)$를 사용하자.

사전 분포를 $f = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}$ 라하면 MAP estimate는,

$$l_{MAP}(\Theta) = logL_{MAP}(\Theta) = \sum_{i=1}^{m} \left\{ y_i \Theta^T \mathbb{x}_i - \log(1 + e^{\Theta^T \mathbb{x}_i}) \right\} - \sum_{j=1}^{m} \frac{\theta_j^2}{2\sigma^2}$$

Gradient ascent는,

$$\theta_j \leftarrow \theta_j + \alpha \frac{\partial}{\partial \theta_j} l_{MAP}(\Theta) = \theta_j + \alpha \sum_{i=1}^{m} x_j^{(i)} \{ y_i - P(y_i = 1 | \mathbb{x}_i; \Theta) \} - \alpha \frac{\theta_j}{\sigma^2}$$

Logistic Regression을 regularize하는 방법은 이 외에도 다양하다.

Multiple class Logistic Regression

Class의 개수가 2보다 클 경우, 즉 Y가 $\{y_1, \dots, y_n\}$의 값을 가질 경우의 logistic regression은

- $P(Y = y_k|\mathbb{X}; \Theta) = \frac{\exp(\theta_{k0}+\sum_{i=1}^{n} \theta_{ki}X_i)}{1+\sum_{j=1}^{K-1} \exp(\theta_{j0}+\sum_{i=1}^{n} \theta_{ji}X_i)}$
- $P(Y = y_k|\mathbb{X}; \Theta) = \frac{1}{1+\sum_{j=1}^{K-1} \exp(\theta_{j0}+\sum_{i=1}^{n} \theta_{ji}X_i)}$

Gradient ascent는

- $\theta_{ji} \leftarrow \theta_{ji} + \alpha \sum_{i=1}^{m} x_j^{(i)}\{\delta(y_i = j) - P(y_i = j|\mathbb{x}_i; \Theta)\}$ where $\delta(y_i = j) : y_i = j$이면 1, 그렇지 않으면 0