Logistic Regression

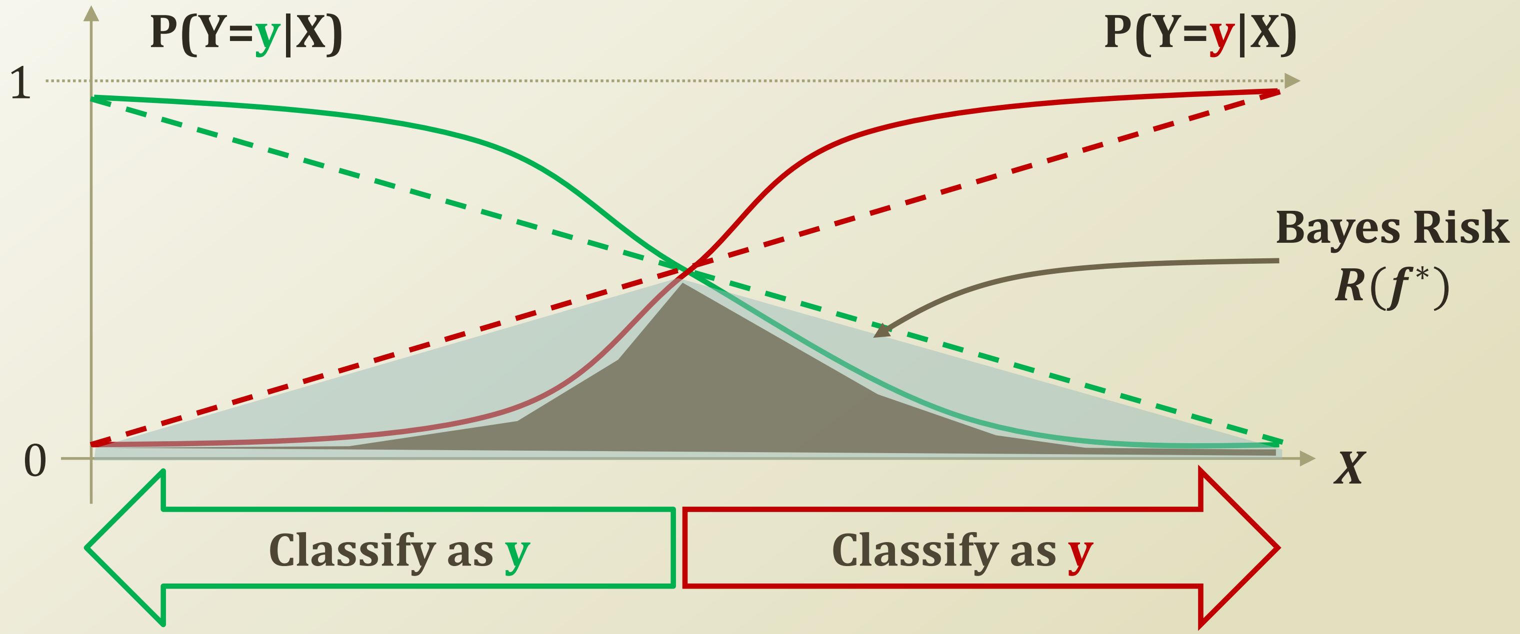
Il-Chul Moon
Dept. of Industrial and Systems Engineering
KAIST
icmoon@kaist.ac.kr

Weekly Objectives

- Learn the logistic regression classifier
  - Understand why the logistic regression is better suited than the linear regression for classification tasks
  - Understand the logistic function
  - Understand the logistic regression classifier
  - Understand the approximation approach for the open form solutions
- Learn the gradient descent algorithm
  - Know the tailor expansion
  - Understand the gradient descent/ascent algorithm
- Learn the different between the naïve Bayes and the logistic regression
  - Understand the similarity of the two classifiers
  - Understand the differences of the two classifiers
  - Understand the performance differences

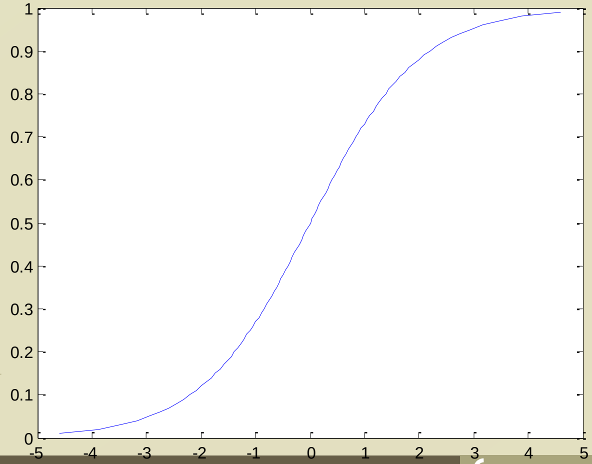
LOGISTIC REGRESSION

Optimal Classification and Bayes Risk



- Linear function vs. Non-linear function of P(Y|X)
  - Which is better?
- Problems of linear function
  - Range
  - Risk optimization
- Which function to use?
  - Need S-curve!

S-curve
a.k.a. Sigmoid
function

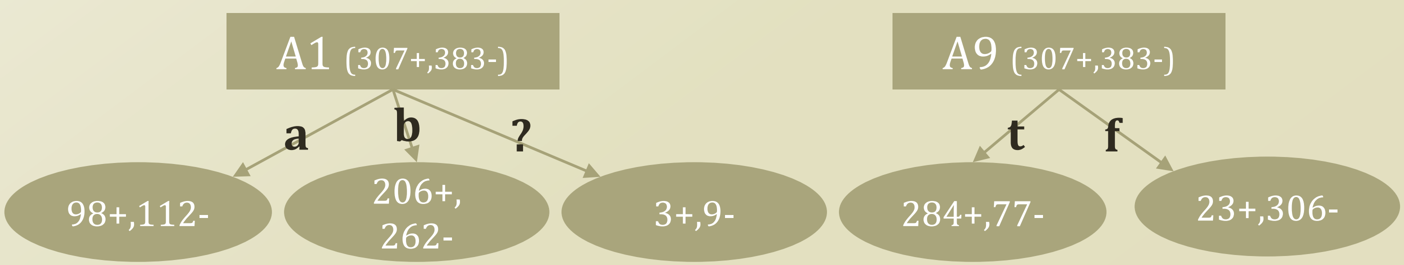


Detour: Credit Approval Dataset

- http://archive.ics.uci.edu/ml/datasets/Credit+Approval
- To protect the confidential information, the dataset is anonymized
  - Feature names and values, as well
- A1: b, a.
  A2: continuous.
  A3: continuous.
  A4: u, y, l, t.
  A5: g, p, gg.
  A6: c, d, cc, i, j, k, m, r, q, w, x, e, aa, ff.
  A7: v, h, bb, j, n, z, dd, ff, o.
  A8: continuous.
  A9: t, f.
  A10: t, f.
  A11: continuous.
  A12: t, f.
  A13: g, p, s.
  A14: continuous.
  A15: continuous.
  C: +,- (class attribute)

Some Counting Result

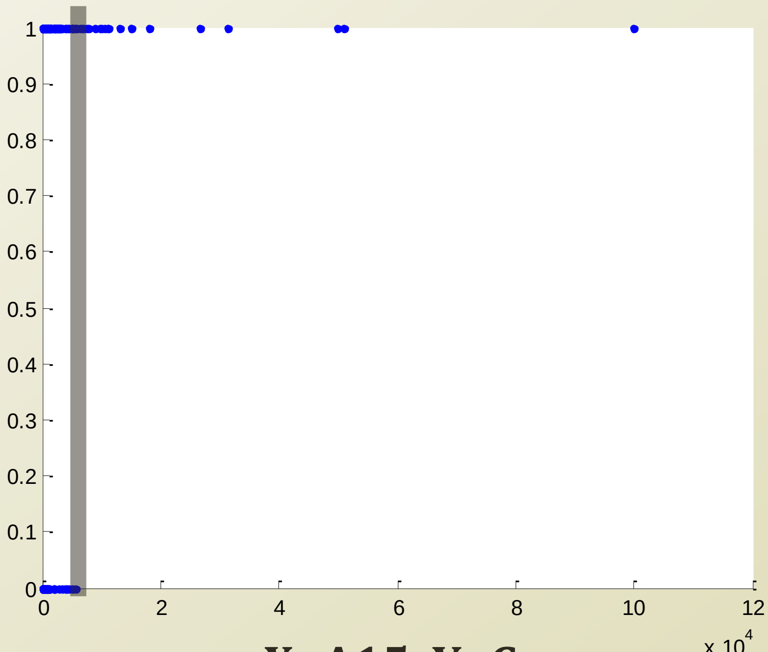
- 690 instances total
- 307 positive instances
- Considering A1
  - 98 positive when a
  - 112 negative when a
  - 206 positive when b
  - 262 negative when b
  - 3 positive when ?
  - 9 negative when ?
- Considering A9
  - 284 positive when t
  - 77 negative when t
  - 23 positive when f
  - 306 negative when f



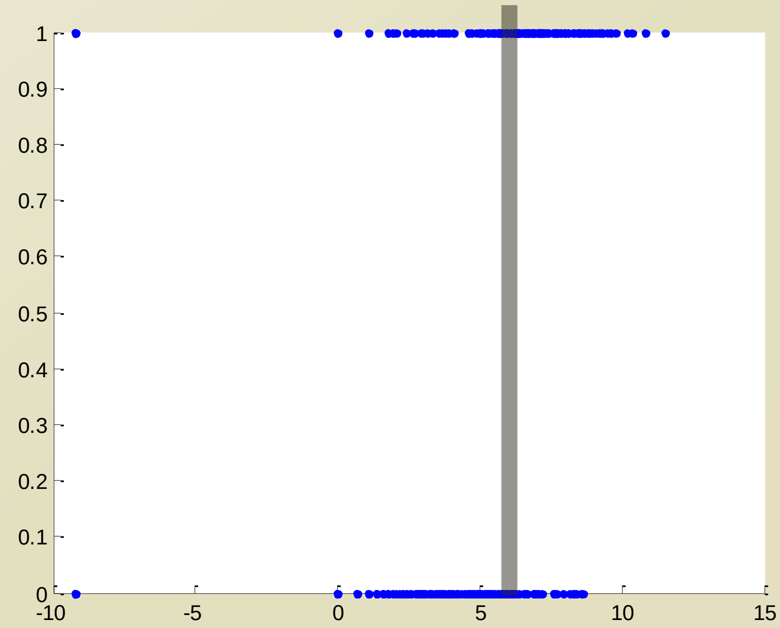
Which is a better attribute to include in the feature set of the hypothesis?

Classification with One Variable

- Let's predict the class, C, with an attribute, A15
  - Imagine that the Y axis shows P(Y|X)
  - There is a decision boundary
    - You can see it intuitively
- Then, How to find the boundary?



X=A15, Y=C

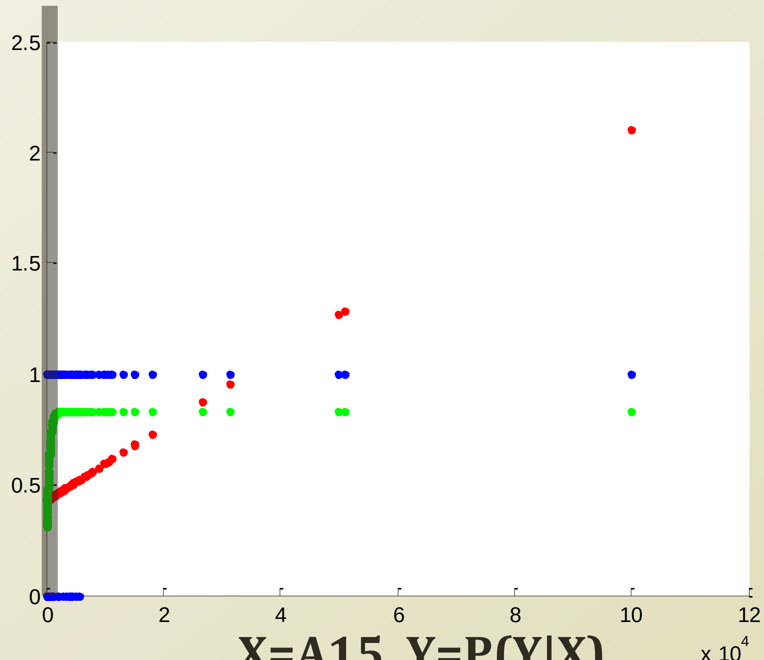


X=log(A15), Y=C

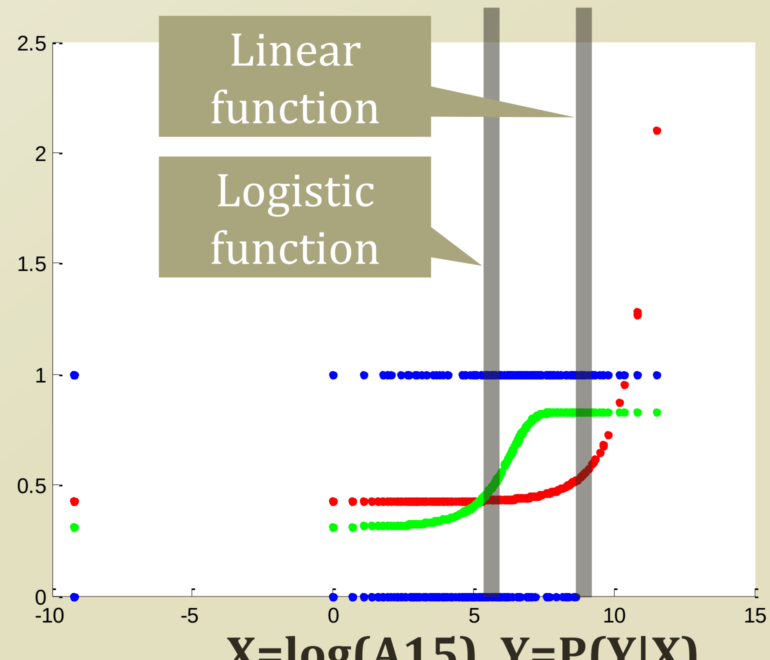
Linear Function vs. Non-Linear Function

- Problem of fitting to the linear function
  - Violate the probability axiom
  - Slow response to the examples
- Better to fit to the logistic function
  - Keep the probability axiom
  - Quick response around the decision boundary
- Which function to use?
  - Logistic function – a special case of sigmoid function

Blue = (X, Y_{true})
Red = $(X, P_{lin}(Y|X))$
Green = $(X, P_{log}(Y|X))$



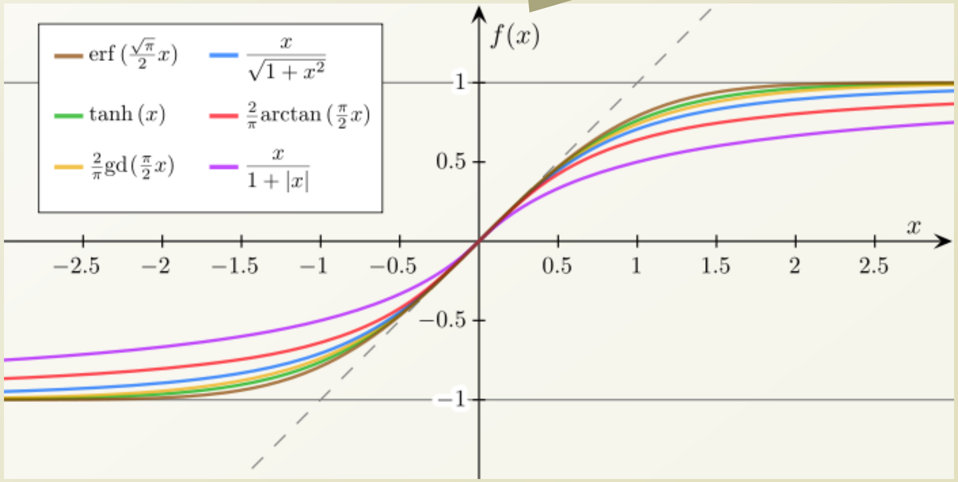
X=A15, Y=P(Y|X)



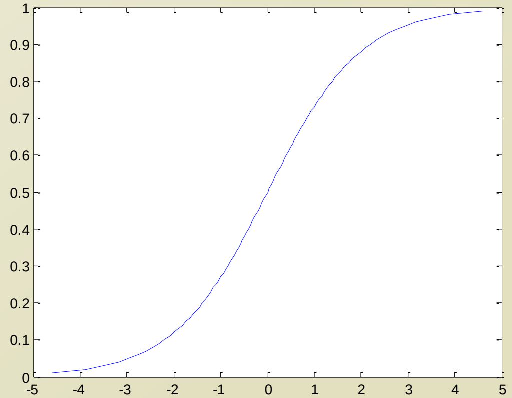
X=log(A15), Y=P(Y|X)

Logistic function

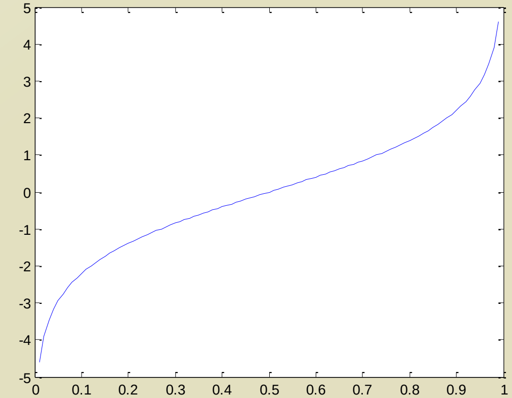
Many types of sigmoid functions



- Sigmoid function is
  - Bounded
  - Differentiable
  - Real function
  - Defined for all real inputs
  - With positive derivative
- Logistic function is
  - $f(x) = \frac{1}{1+e^{-x}}$
  - In relation to the population growth
  - Why is this good?
    - Sigmoid function
    - Particularly, easy to calculate the derivative…



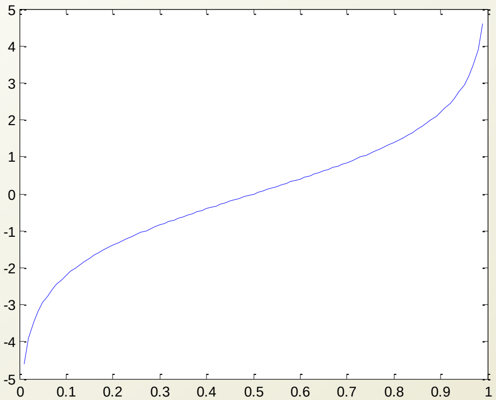
Logistic Function

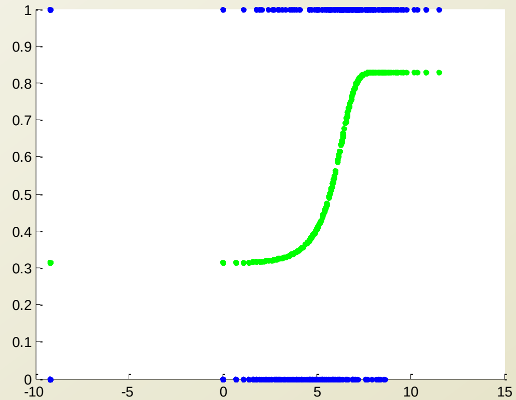


Logit Function

$$f(x) = \log\left(\frac{x}{1-x}\right)$$

Logistic Function Fitting





Linear Regression:

$$\hat{f} = X\theta \quad \theta = (X^T X)^{-1} X^T Y$$

Very similar to the linear regression.
Turning to the multivariate case

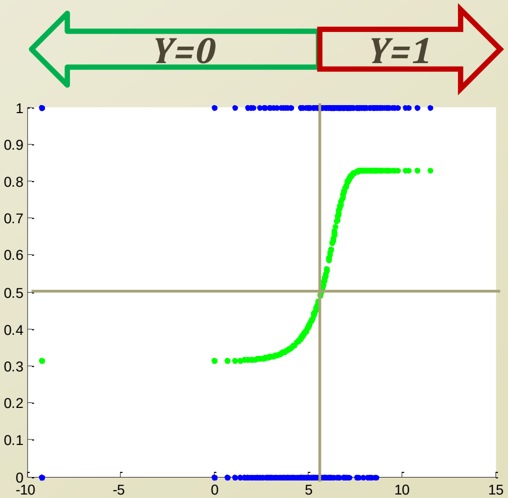
$$f(x) = \log\left(\frac{x}{1-x}\right) \rightarrow x = \log\left(\frac{p}{1-p}\right) \rightarrow ax + b = \log\left(\frac{p}{1-p}\right) \rightarrow X\theta = \log\left(\frac{p}{1-p}\right)$$

Logit→Logistic
Inverse of X and Y
X in Logit is the probability

Linear shift for a better function fitting

- When we are fitting the linear regression to approximate P(Y|X)
  - $X\theta = P(Y|X)$
  - Though, this is not going to keep the probability axiom
- Now we are fitting to the logistic function to approximate P(Y|X)
  - $X\theta = \log\left(\frac{P(Y|X)}{1-P(Y|X)}\right)$
  - From linear to logistic

Logistic Regression



- Logistic regression is a probabilistic classifier to predict the binomial or the multinomial outcome
  - by fitting the conditional probability to the logistic function.
- You can see the problem from the different view.
  - This way is actually closer to the formal definition.
- Given the Bernoulli experiment
  - $P(y|x) = \mu(x)^y(1-\mu(x))^{1-y}$
  - $\mu(x) = \frac{1}{1+e^{-\theta^T x}} = P(y=1|x)$
  - Here, $\mu(x)$ is the logistic function
- From the previous slide,
  - $X\theta = \log\left(\frac{P(Y|X)}{1-P(Y|X)}\right) \rightarrow P(Y|X) = \frac{e^{X\theta}}{1+e^{X\theta}}$

Logistic Function

$$f(x) = \frac{1}{1+e^{-x}}$$

The goal, finally, becomes finding out $\boldsymbol{\theta}$, again

$$P(y=1|x)=\mu(x)=\frac{1}{1+e^{-\dot{\theta}^T x}}=\frac{e^{X\theta}}{1+e^{X\theta}}$$

Finding the Parameter, θ

$$X\theta=\log\left(\frac{P(Y|X)}{1-P(Y|X)}\right)$$

- **Maximum Likelihood Estimation (MLE) of $\boldsymbol{\theta}$**
  - Choose θ that maximizes the probability of observed data

$$\hat{\boldsymbol{\theta}}=\boldsymbol{argmax_{\theta}P(D|\theta)}$$

- **This is Maximum Conditional Likelihood Estimation (MCLE)**
- $\hat{\theta}=argmax_{\theta}P(D|\theta)=argmax_{\theta}\prod_{1\le i\le N}P(Y_i|X_i;\theta)$

$$=argmax_{\theta}log(\prod_{1\le i\le N}P(Y_i|X_i;\theta))=argmax_{\theta}\sum_{1\le i\le N}log(P(Y_i|X_i;\theta))$$

- $P(Y_i|X_i;\theta)=\mu(X_i)^{Y_i}(1-\mu(X_i))^{1-Y_i}$
- $log\left(P(Y_i|X_i;\theta)\right)=Y_i\log\left(\mu(X_i)\right)+(1-Y_i)\log\left(1-\mu(X_i)\right)$

$$=Y_i\{\log\left(\mu(X_i)\right)-\log\left(1-\mu(X_i)\right)\}+\log\left(1-\mu(X_i)\right)$$

$$=Y_i\log\left(\frac{\mu(X_i)}{1-\mu(X_i)}\right)+\log\left(1-\mu(X_i)\right)$$

$$=Y_iX_i\theta+\log\left(1-\mu(X_i)\right)=Y_iX_i\theta-\log\left(1+e^{X_i\theta}\right)$$

Finding the Parameter, θ, contd.

Linear Regression (Closed Form):

$\hat{f} = X\theta$

$\nabla_\theta(\theta^T X^T X\theta - 2\theta^T X^T Y) = 0$

$2X^T X\theta - 2X^T Y = 0$

$\theta = (X^T X)^{-1} X^T Y$

- $\hat{\theta} = argmax_\theta \sum_{1\le i\le N} log(P(Y_i|X_i;\theta))$
- $= argmax_\theta \sum_{1\le i\le N}\{Y_i X_i\theta - \log(1 + e^{X_i\theta})\}$
- Partial derivative to find a certain element in θ
- $\frac{\partial}{\partial\theta_j}\{\sum_{1\le i\le N} Y_i X_i\theta - \log(1 + e^{X_i\theta})\}$

$P(y = 1|x) = \frac{e^{X\theta}}{1 + e^{X\theta}}$

$$= \left\{\sum_{1\le i\le N} Y_i X_{i,j}\right\} + \left\{\sum_{1\le i\le N} -\frac{1}{1 + e^{X_i\theta}} \times e^{X_i\theta} \times X_{i,j}\right\}$$

$$= \sum_{1\le i\le N} X_{i,j}(Y_i - \frac{e^{X_i\theta}}{1 + e^{X_i\theta}}) = \sum_{1\le i\le N} X_{i,j}(Y_i - P(Y_i = 1|X_i;\theta)) = 0$$

Cannot be easily solved in the closed form because of the logistic function

- There is no way to derive further
  - There is no closed form solution!
  - Open form solution → approximate!