Discriminative Model
Logistic/SVM/Tree
(no assumptions on p(x)

$$P(x, y) = P(Y{=}y \mid X{=}x)\, P(X{=}x)$$
$$= P(X{=}x \mid Y{=}y)\, P(Y{=}y)$$

Generative Model
Discriminant Analysis
LDA/QDA/NB

Logistic Regression (Binary)

range = (0, 1)

$$\eta(x) = P(Y=1|X=x)$$
$$1-\eta(x) = P(Y=0|X=x)$$

range = (-inf, inf)

$$\log\frac{\eta(x)}{1-\eta(x)} = x^t\boldsymbol{\beta}$$

Here beta contains the intercept

At each data point, we have $(x_i, y_i, \eta(x_i))$

How to estimate eta(x), or equivalently beta?

Choice of Loss function L(y, eta(x)):
(y - eta(x))^2, **log-likelihood**?

Too flat to train

MLE for Binary Logistic Regression

$$\log \frac{\eta(x)}{1-\eta(x)} = x^t \boldsymbol{\beta}$$

$$P(Y=1|X=x) = \eta(x) = \frac{\exp(x^t\boldsymbol{\beta})}{1+\exp(x^t\boldsymbol{\beta})}$$

$$P(Y=0|X=x) = 1-\eta(x) = \frac{1}{1+\exp(x^t\boldsymbol{\beta})}$$

$$P(Y=y|X=x) = \sigma(z)^y \big(1-\sigma(z)\big)^{(1-y)}$$

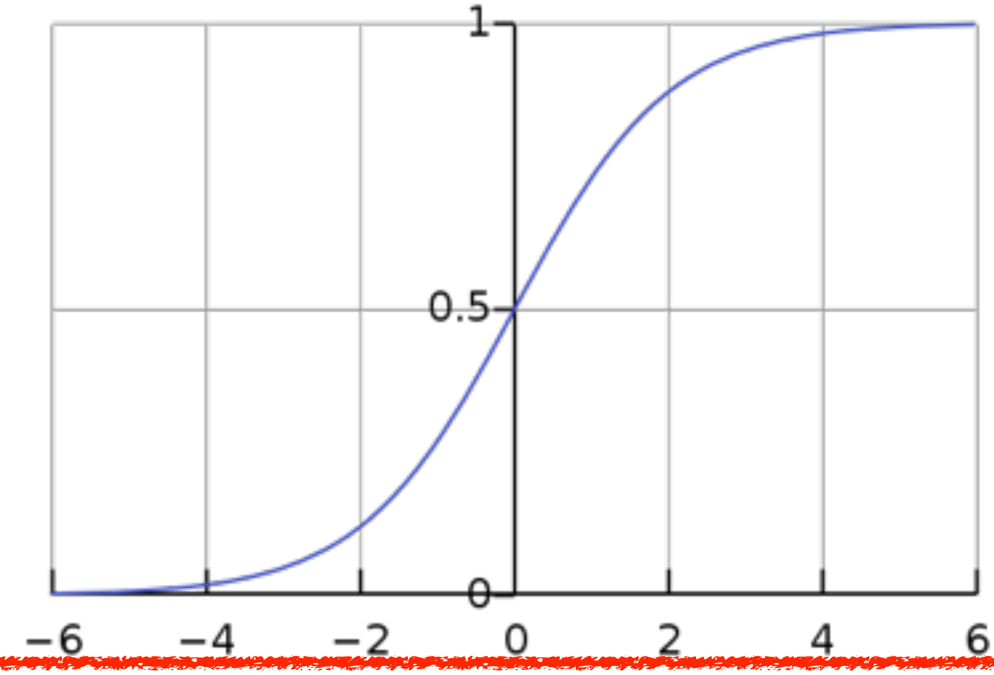
$$\sigma(z) = \frac{e^z}{1+e^z}, \quad z = x^t\boldsymbol{\beta}$$

$$\log P(Y=y|X=x) = y \cdot \log \sigma(z) + (1-y) \cdot \log\big(1-\sigma(z)\big)$$

MLE for Binary Logistic Regression

$$\log \frac{\eta(x)}{1-\eta(x)} = x^t \boldsymbol{\beta}$$

Sigmoid function



$$P(Y=y|X=x) = \sigma(z)^y \big(1-\sigma(z)\big)^{(1-y)}$$

$$\boxed{\sigma(z) = \frac{e^z}{1+e^z},} \quad z = x^t \boldsymbol{\beta}$$

$$\log P(Y=y|X=x) = y \cdot \log \sigma(z) + (1-y) \cdot \log \big(1-\sigma(z)\big)$$

MLE for Binary Logistic Regression

$$P(Y = y|X = x) = \sigma(z)^y \big(1 - \sigma(z)\big)^{(1-y)}$$

$$\sigma(z) = \frac{e^z}{1 + e^z}, \quad z = x^t \boldsymbol{\beta}$$

$$\log P(Y = y|X = x) = y \cdot \log \sigma(z) + (1 - y) \cdot \log \big(1 - \sigma(z)\big)$$

The Log-likelihood function is given by

$$\ell(\beta) = \sum_{i=1}^{n} y_i \cdot \log \sigma(z_i) + (1 - y_i) \cdot \log \big(1 - \sigma(z_i)\big)$$

$$\frac{\partial \ell(\beta)}{\partial \beta} = ???$$

gradient (vector)

$$\frac{\partial^2 \ell(\beta)}{\partial \beta \partial \beta} = ???$$

Hessian matrix

MLE for Binary Logistic Regression

$$P(Y=y|X=x)=\sigma(z)^y\big(1-\sigma(z)\big)^{(1-y)}$$

$$\sigma(z)=\frac{e^z}{1+e^z},\qquad z=x^t\boldsymbol{\beta}$$

$$\log P(Y=y|X=x)=y\cdot\log\sigma(z)+(1-y)\cdot\log\big(1-\sigma(z)\big)$$

$$\frac{\partial}{\partial z}\sigma(z)=\frac{\partial}{\partial z}\frac{1}{1+e^{-z}}$$

$$=\frac{e^{-z}}{(1+e^{-z})^2}=\frac{e^{-z}}{1+e^{-z}}\frac{1}{1+e^{-z}}$$

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

MLE for Binary Logistic Regression

$$P(Y=y|X=x) = \sigma(z)^y \big(1-\sigma(z)\big)^{(1-y)}$$

$$\sigma(z) = \frac{e^z}{1+e^z}, \quad z = x^t\boldsymbol{\beta}$$

$$\log P(Y=y|X=x) = y \cdot \log\sigma(z) + (1-y)\cdot\log\big(1-\sigma(z)\big)$$

$$\frac{\partial}{\partial\beta}\log P(Y=y|X=x)$$

$$= \frac{y}{\sigma(z)}\frac{\partial\sigma(z)}{\partial z}\frac{\partial z}{\partial\beta} - \frac{1-y}{1-\sigma(z)}\frac{\partial\sigma(z)}{\partial z}\frac{\partial z}{\partial\beta}$$

$$= \frac{y}{\sigma(z)}\sigma(z)(1-\sigma(z))x - \frac{1-y}{1-\sigma(z)}\sigma(z)(1-\sigma(z))x$$

$$= \big(y-\sigma(z)\big)x$$

MLE for Binary Logistic Regression

$$\boxed{\begin{gathered} P(Y=y|X=x) = \sigma(z)^y \big(1-\sigma(z)\big)^{(1-y)} \\ \sigma(z) = \frac{e^z}{1+e^z}, \qquad z = x^t \boldsymbol{\beta} \\ \log P(Y=y|X=x) = \underline{y \cdot \log \sigma(z)} + (1-y)\cdot \log\big(1-\sigma(z)\big) \end{gathered}}$$

$$\begin{aligned}
&\frac{\partial}{\partial \beta} \log P(Y=y|X=x) \\
=&\frac{y}{\sigma(z)} \frac{\partial \sigma(z)}{\partial z} \frac{\partial z}{\partial \beta} - \frac{1-y}{1-\sigma(z)} \frac{\partial \sigma(z)}{\partial z} \frac{\partial z}{\partial \beta} \\
=&\frac{y}{\cancel{\sigma(z)}} \cancel{\sigma(z)}(1-\sigma(z))x - \frac{1-y}{\cancel{1-\sigma(z)}} \sigma(z)\cancel{(1-\sigma(z))}x \\
=&\big(y-\sigma(z)\big)x
\end{aligned}$$

MLE for Binary Logistic Regression

$$P(Y = y|X = x) = \sigma(z)^y \big(1 - \sigma(z)\big)^{(1-y)}$$

$$\sigma(z) = \frac{e^z}{1 + e^z}, \qquad z = x^t \boldsymbol{\beta}$$

$$\log P(Y = y|X = x) = \underline{y \cdot \log \sigma(z)} + (1 - y) \cdot \log\big(1 - \sigma(z)\big)$$

$$\frac{\partial}{\partial \beta} \log P(Y = y|X = x)$$

$$= \frac{y}{\sigma(z)} \frac{\partial \sigma(z)}{\partial z} \frac{\partial z}{\partial \beta} - \frac{1 - y}{1 - \sigma(z)} \frac{\partial \sigma(z)}{\partial z} \frac{\partial z}{\partial \beta}$$

$$= \frac{y}{\cancel{\sigma(z)}} \cancel{\sigma(z)}(1 - \cancel{\sigma(z)})x - \frac{1 \cancel{- y}}{\cancel{1 - \sigma(z)}} \sigma(z)\cancel{(1 - \sigma(z))}x$$

$$= \big(y - \sigma(z)\big)x$$

MLE for Binary Logistic Regression

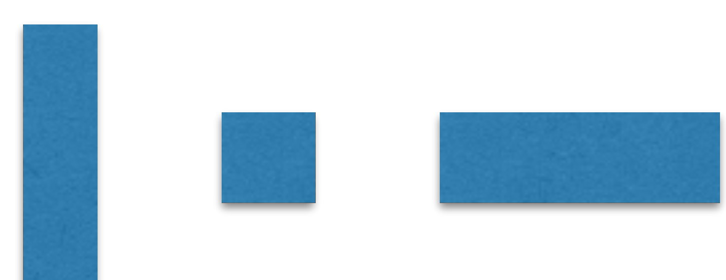
$$P(Y = y|X = x) = \sigma(z)^y \big(1 - \sigma(z)\big)^{(1-y)}$$

$$\sigma(z) = \frac{e^z}{1 + e^z}, \qquad z = x^t \boldsymbol{\beta}$$

$$\log P(Y = y|X = x) = \underline{y \cdot \log \sigma(z)} + (1 - y) \cdot \log \big(1 - \sigma(z)\big)$$

$$\boxed{\frac{\partial}{\partial \beta} \log P(Y = y|X = x) = \big(y - \sigma(z)\big) x}$$

$$\frac{\partial^2}{\partial \beta \partial \beta} \log P(Y = y|X = x) = -x\sigma(z)(1 - \sigma(z))x^t$$

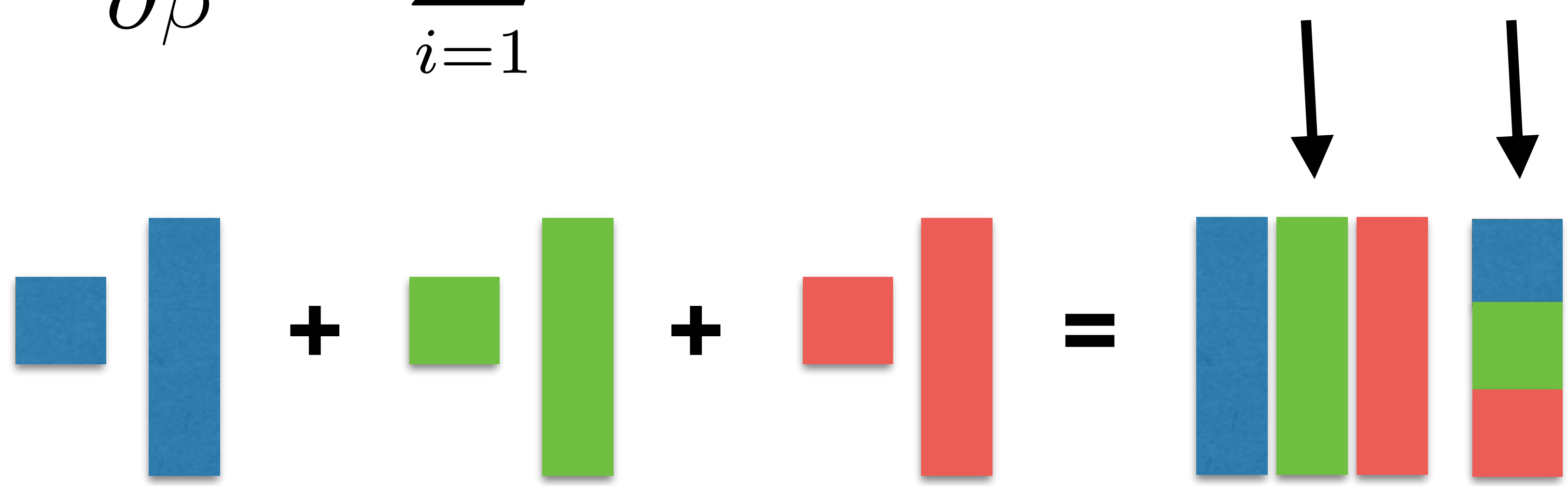


MLE for Binary Logistic Regression

The Log-likelihood function is given by

$$\ell(\beta) = \sum_{i=1}^{n} \underline{y_i \cdot \log \sigma(z_i) + (1 - y_i) \cdot \log\left(1 - \sigma(z_i)\right)}$$

$$\frac{\partial \ell(\beta)}{\partial \beta} = \sum_{i=1}^{n} \left(y_i - \sigma(z_i)\right) x_i = \mathbf{X}^t(\mathbf{y} - \mathbf{p})$$



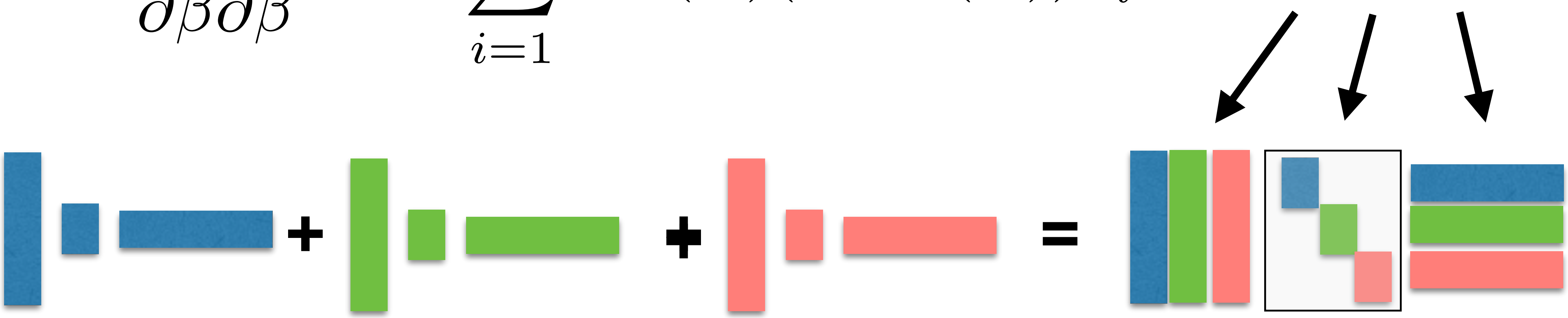
$$\boxed{p_i = \sigma(z_i) = \frac{\exp(x_i^t \beta)}{1 + \exp(x_i^t \beta)} = \eta(x_i) = P(Y = 1 | X = x_i)}$$

MLE for Binary Logistic Regression

The Log-likelihood function is given by

$$\ell(\beta) = \sum_{i=1}^{n} y_i \cdot \log \sigma(z_i) + (1 - y_i) \cdot \log \big(1 - \sigma(z_i)\big)$$

$$\frac{\partial^2 \ell(\beta)}{\partial\beta\partial\beta} = -\sum_{i=1}^{n} x_i \sigma(z_i)(1 - \sigma(z_i)) x_i^t = -\mathbf{X}^t\mathbf{W}\mathbf{X}$$



$$p_i = \sigma(z_i) = \frac{\exp(x_i^t\beta)}{1 + \exp(x_i^t\beta)} = \eta(x_i) = P(Y = 1 | X = x_i)$$

MLE for Binary Logistic Regression

The Log-likelihood function is given by

$$\ell(\beta) = \sum_{i=1}^{n} y_i \cdot \log \sigma(z_i) + (1 - y_i) \cdot \log \big(1 - \sigma(z_i)\big)$$

$$\boxed{\frac{\partial \ell(\beta)}{\partial \beta} = \mathbf{X}^t(\mathbf{y} - \mathbf{p})}$$

gradient (vector)

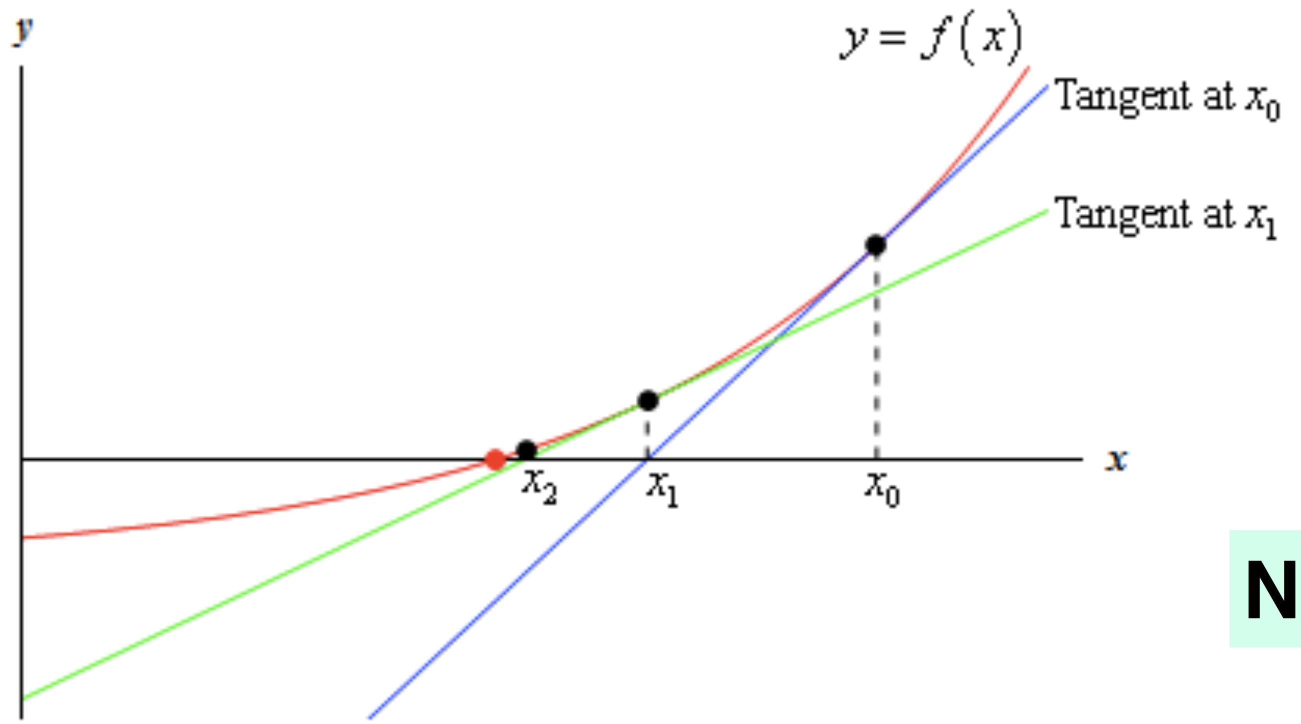
$$\boxed{\frac{\partial^2 \ell(\beta)}{\partial \beta \partial \beta} = -\mathbf{X}^t \mathbf{W} \mathbf{X}}$$

Hessian matrix

Hessian is negative semi-definite, so log-likelihood function is **concave**, i.e., any local maximum is global maximum.

MLE is obtained via iterative **Newton-Raphson** to find the root of the derivative of log-likelihood function.

Newton-Raphson Method



New value

Old value

$$f(x) \approx f(x_0) + f'(x_0)(x - x_0) \Longrightarrow x = x_0 - \frac{f(x_0)}{f'(x_0)}$$

$$\begin{aligned}\beta &= \beta_0 + (\mathbf{X}^t\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^t(\mathbf{y} - \mathbf{p}) \\ &= (\mathbf{X}^t\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^t\mathbf{W}(\mathbf{X}\beta_0 + \mathbf{W}^{-1}(\mathbf{y} - \mathbf{p}))\end{aligned}$$

Compute MLE for Logistic Regression

The MLE $\hat{\beta}$ can be obtained by the following Reweighted LS Algorithm:

- Start with some initial values β^0
- Calculate the corresponding $p_i^0 = \eta^0(x_i)$ for $i = 1, \ldots, n$; define $W = \text{diag}(p_i^0(1 - p_i^0))_{i=1}^n$.
- Calculate

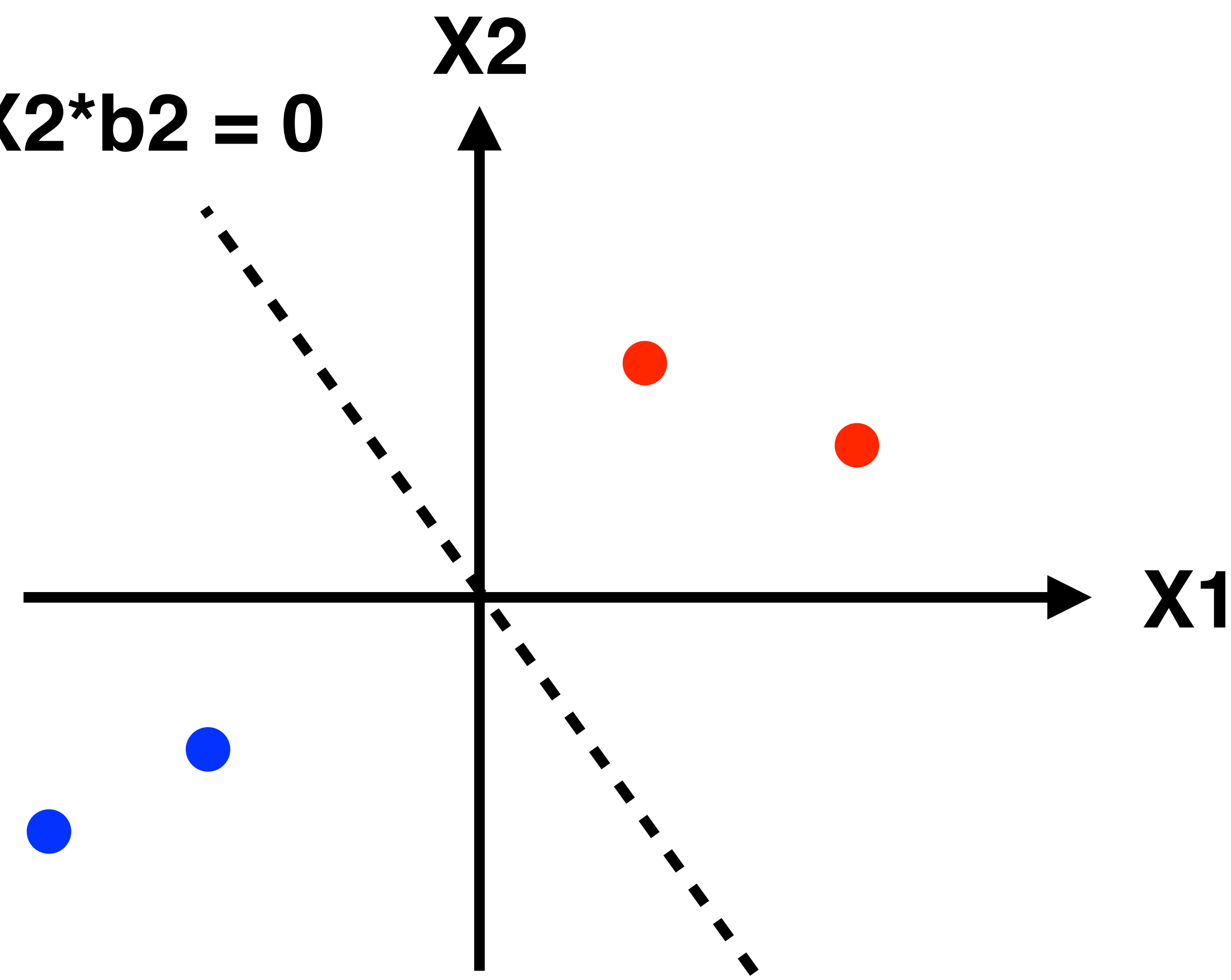
$$\mathbf{z} = X\beta^0 + W^{-1}(y - \mathbf{p}^0).$$

- Update $\beta^0 \leftarrow \beta^1$ with

$$\beta^1 = (X^t W X)^{-1} X^t W \mathbf{z}.$$

and iterative the above steps until convergence.

Logistic Regression with Separable Data



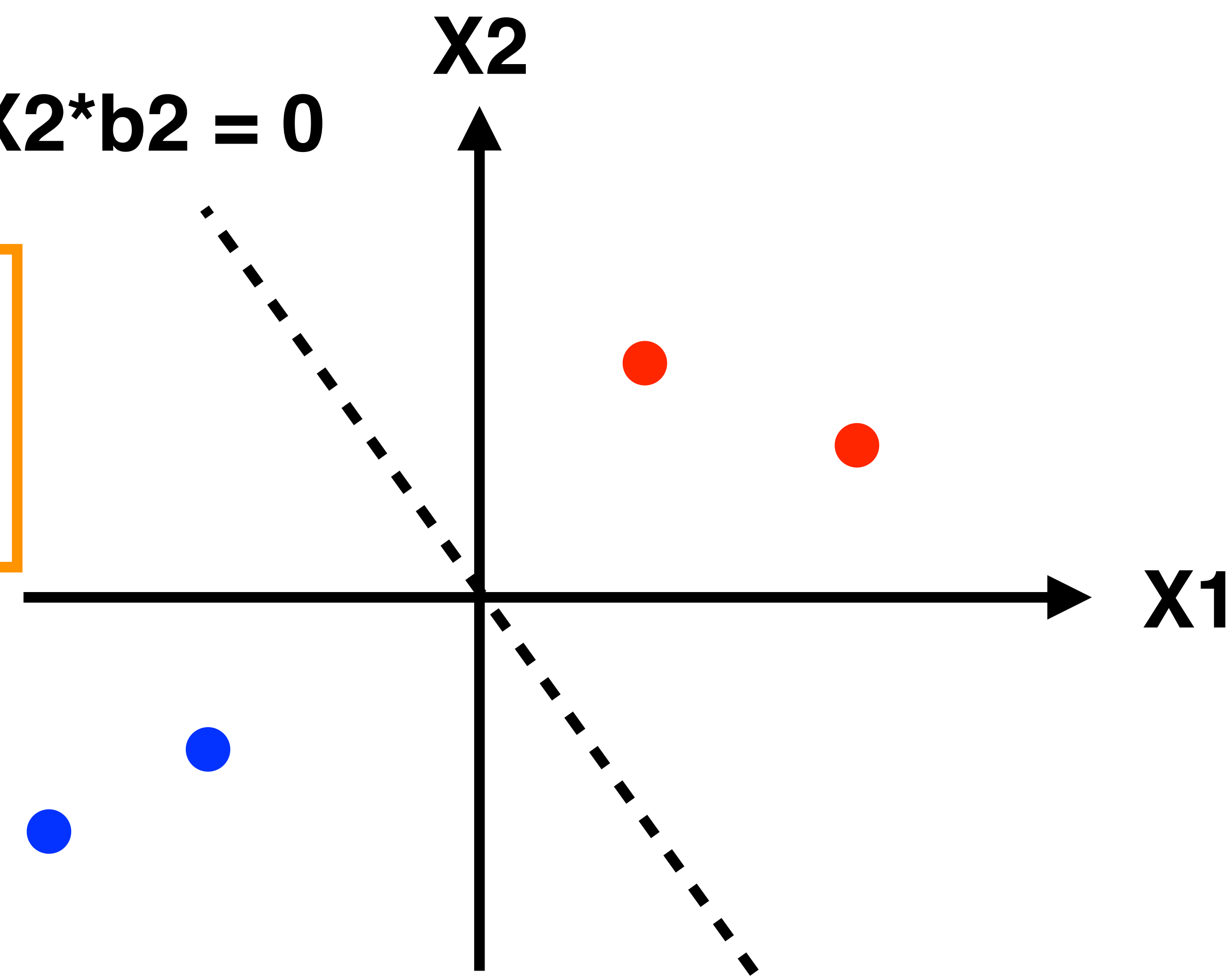
- b1=1, b2=1
- b1=10, b2=10
- b1=500, b2=500

$$\frac{e^{(\beta_1 x_1 + \beta_2 x_2)}}{1 + e^{(\beta_1 x_1 + \beta_2 x_2)}} = \boxed{\frac{1}{1 + e^{-(\beta_1 x_1 + \beta_2 x_2)}}} \qquad \boxed{\frac{1}{1 + e^{(\beta_1 x_1 + \beta_2 x_2)}}}$$

Logistic Regression with Separable Data

For separable data, the MLE doesn't converge. However, the separating line is well-define.

- b1=1, b2=1
- b1=10, b2=10
- b1=500, b2=500



$$\frac{e^{(\beta_1 x_1+\beta_2 x_2)}}{1+e^{(\beta_1 x_1+\beta_2 x_2)}} = \boxed{\frac{1}{1+e^{-(\beta_1 x_1+\beta_2 x_2)}}} \qquad \boxed{\frac{1}{1+e^{(\beta_1 x_1+\beta_2 x_2)}}}$$