# Logistic regression (binary classification)

feature/design matrix $\bar{\bar{X}}$

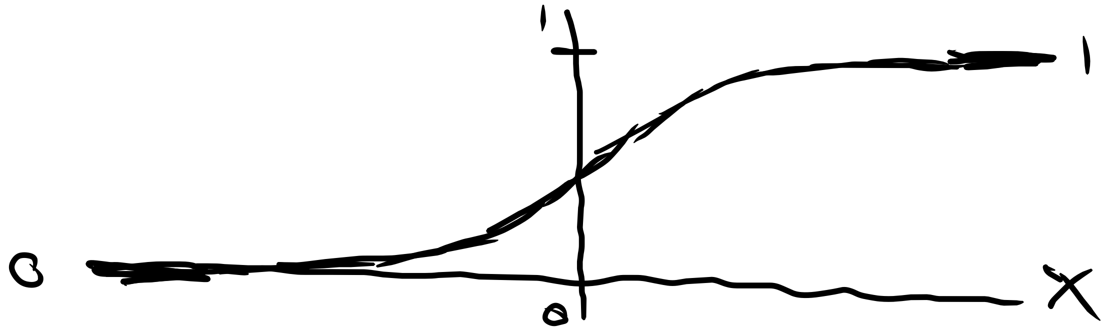
$$\begin{array}{c|cccc} & 1 & & \cdots & n \\ \hline 1 & x_{11} & & & \\ \vdots & & & & \\ m & & & & x_{mn} \end{array}$$

target $\bar{y}$

$$\begin{array}{c|c} 0 & 0{,}00 \\ 1 & 1{,}00 \\ 1 & 1{,}00 \\ \vdots & \\ 0 & 0{,}00 \end{array}$$

model predicts the probability of class 1 :

$$\hat{y}_1 = \sigma(x_{11}\theta_1 + x_{12}\theta_2 + \ldots + x_{1n}\theta_n) = \sigma(\bar{\bar{X}} \cdot \bar{\theta})$$




"sigmoid" function

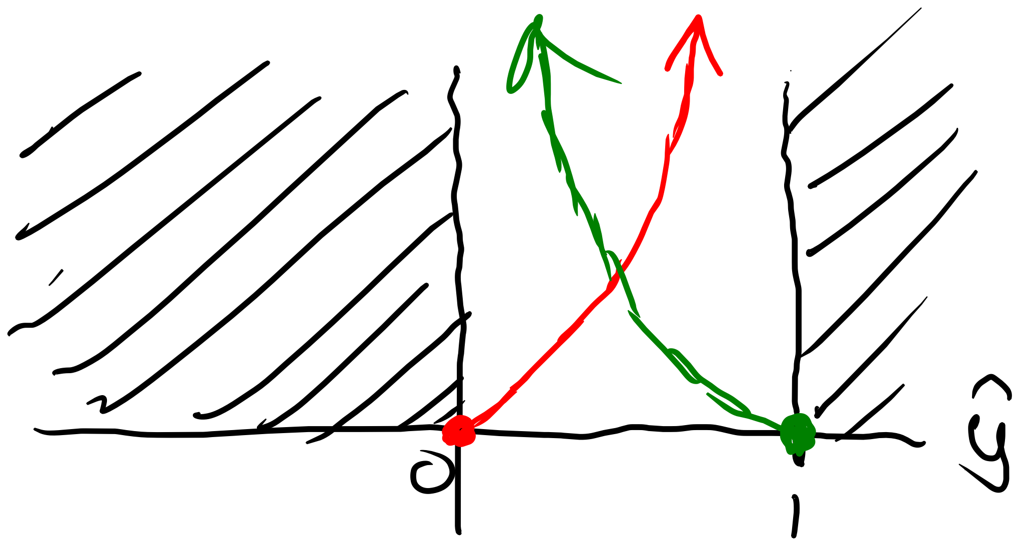
logistic function : $\sigma(x) = \frac{1}{1+e^{-x}}$

Cost function :

regression : $$J = \frac{1}{2m}\sum_i \overbrace{(\hat{y}_i - y_i)}^{e_i}{}^2$$


$e_i = \hat{y}$ if $y=0$

classification : $$\begin{cases} -\ln(1-\hat{y}) & \text{if } y=0 \\ -\ln(\hat{y}) & \text{if } y=1 \end{cases}$$




if $y=0$

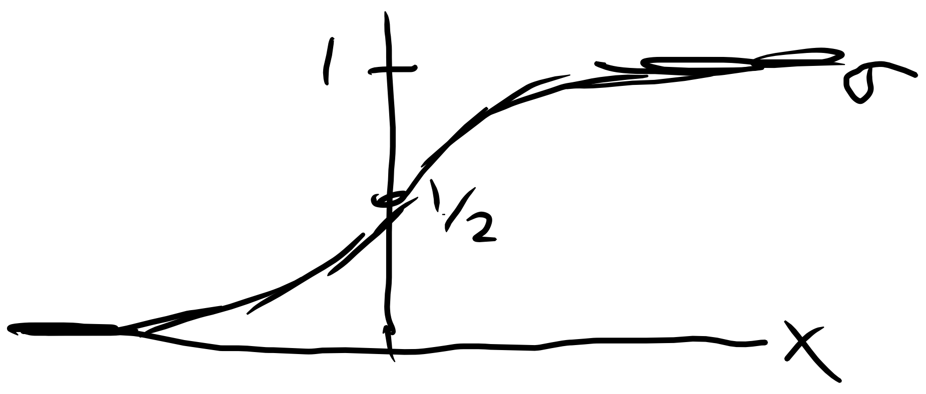
$y=1$

"cross-entropy"

alternative : $$J = \frac{1}{m}\sum_i \left(-y_i \ln(\hat{y}_i) - (1-y_i)\ln(1-\hat{y}_i)\right)$$

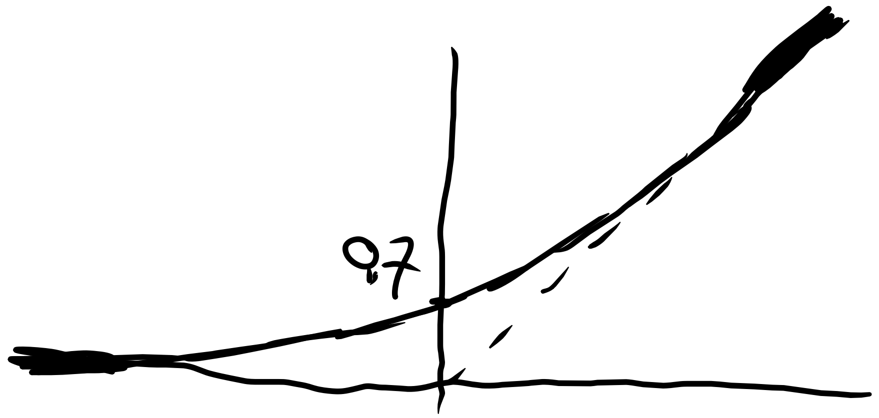
Intermezzo:

logistic function $\sigma(x)$




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

Softplus function $S(x)$




$$S(x) = \ln(1+e^x)$$

A $\quad \sigma(-x) = 1 - \sigma(x) \Leftrightarrow \sigma(x) + \sigma(-x) = 1$

B
$$\ln(\sigma(x)) = \ln\left(\frac{1}{1+e^{-x}}\right) = \ln(1) - \ln(1+e^{-x})$$
$$= -\ln(1+e^{-x}) = -S(-x)$$

C $\quad S'(x) = \frac{d}{dx}\ln(1+e^x) = \frac{e^x}{1+e^x} = \sigma(x)$

Optimize the model $\hat{y} = \sigma(\bar{\bar{X}}\bar{\theta})$ by means of gradient descent: $\theta_k \leftarrow \theta_k - \alpha \cdot \frac{\partial J}{\partial \theta_k}$

if $y=0$: $\theta_k \leftarrow \theta_k - \alpha \cdot \frac{\partial}{\partial \theta_k} -\frac{1}{m} \ln(1-\hat{y}_i) = \theta_k + \frac{\alpha}{m} \frac{\partial}{\partial \theta_k} \ln\left(1-\sigma\left(\sum_j X_{ij}\theta_j\right)\right) \overset{A}{=}$

$\overset{A}{=} \theta_k + \frac{\alpha}{m} \frac{\partial}{\partial \theta_k} \ln\left(\sigma\left(-\sum_j X_{ij}\theta_j\right)\right) \overset{B}{=} \theta_k - \frac{\alpha}{m} \frac{\partial}{\partial \theta_k} S\left(\sum_j X_{ij}\theta_j\right) \overset{C}{=} \theta_k - \frac{\alpha}{m} \sigma\left(\sum_j X_{ij}\theta_j\right) \cdot X_{ik}$

$= \theta_k - \frac{\alpha}{m}(\hat{y}_i - 0) \cdot X_{ik}$

if $y=1$: $\theta_k \leftarrow \theta_k - \alpha \cdot \frac{\partial}{\partial \theta_k} -\frac{1}{m} \ln(\hat{y}_i) = \theta_k + \frac{\alpha}{m} \frac{\partial}{\partial \theta_k} \ln\left(\sigma\left(\sum_i X_{ij}\theta_i\right)\right) \overset{B}{=} \theta_k - \frac{\alpha}{m} \frac{\partial}{\partial \theta_k} S\left(-\sum X_{ij}\theta_j\right) \overset{C}{=}$

$\overset{C}{=} \theta_k + \frac{\alpha}{m} \sigma\left(-\sum_j X_{ij}\theta_j\right) X_{ik} \overset{A}{=} \theta_k + \frac{\alpha}{m}\left(1-\sigma\left(\sum_j X_{ij}\theta_j\right)\right) X_{ik} = \theta_k + \frac{\alpha}{m}(1-\hat{y}_i) X_{ik} =$

$= \theta_k - \frac{\alpha}{m}(\hat{y}_i - 1) X_{ik}$

In general: $\theta_k \leftarrow \theta_k - \frac{\alpha}{m}(\hat{y}_i - y_i) X_{ik} = \theta_k - \frac{\alpha}{m} X_{ik} e_i \Rightarrow \underline{\underline{\bar{\theta} \leftarrow \bar{\theta} - \frac{\alpha}{m} \bar{\bar{X}}^T \bar{e}}}$

This is identical to the update-rule for linear regression