Logistic Regression

CS6140

Predrag Radivojac

Khoury College of Computer Sciences

Northeastern University

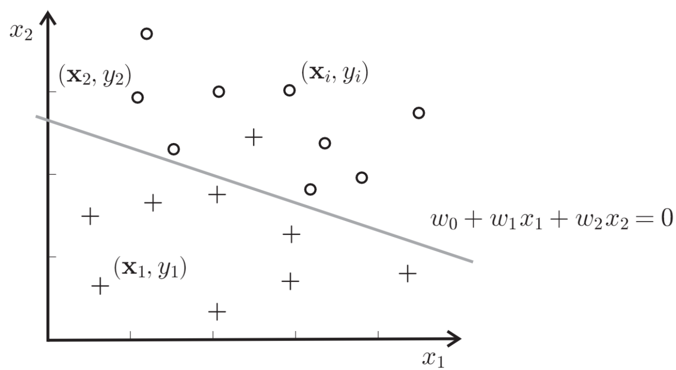
Spring 2021

Linear classification

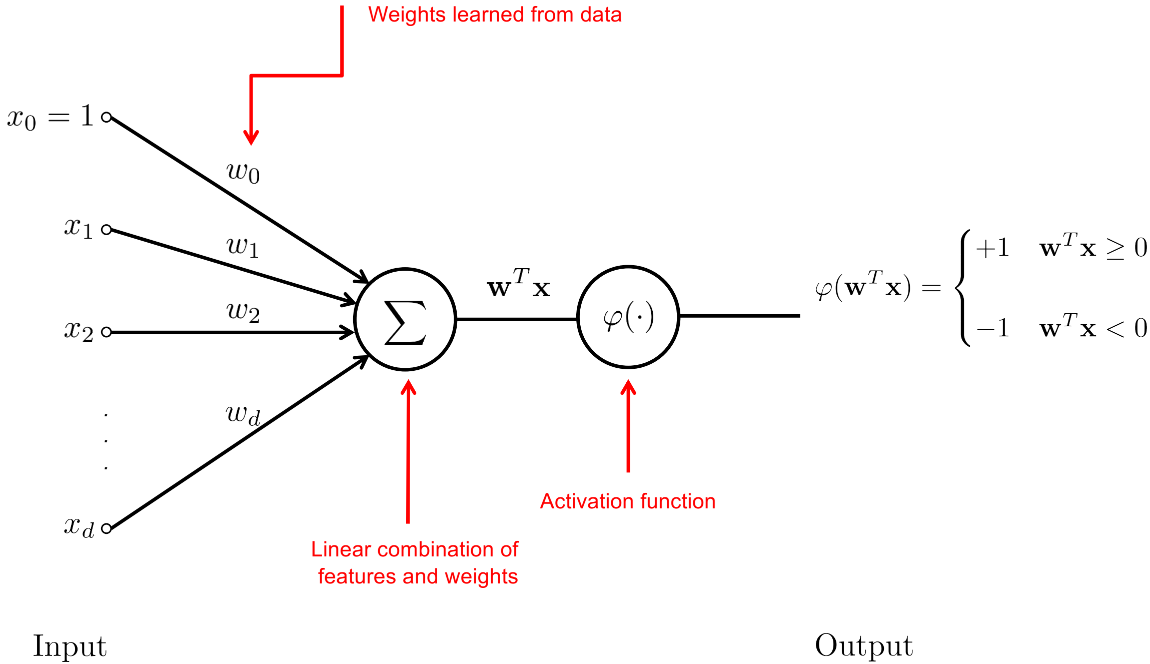
Given: a set of observations $\mathcal{D} = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^{n}$, $(\boldsymbol{x}_i, y_i) \in \mathbb{R}^d \times \{0, 1\}$

Objective: find best linear separator $f(\boldsymbol{x}) = 0$, where $f(\boldsymbol{x}) = w_0 + \sum_{j=1}^{d} w_j x_j$

$\mathcal{X} = \mathbb{R} \times \mathbb{R}$, $\mathcal{Y} = \{0, 1\}$

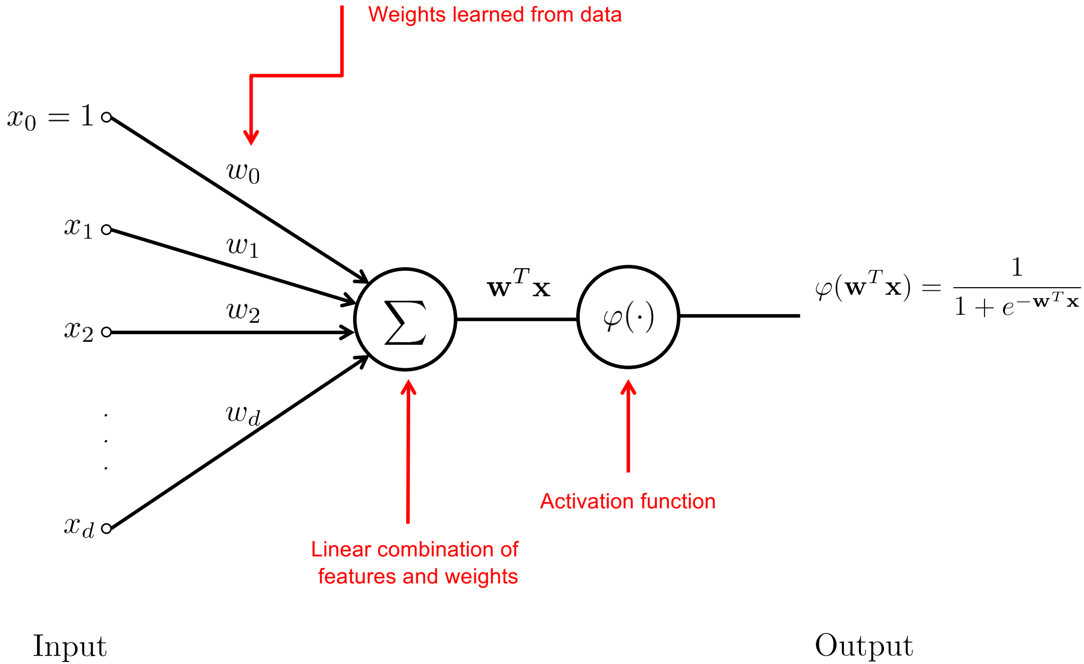


PERCEPTRON



Weights learned from data
$x_0 = 1$
x_1
x_2
x_d
w_0
w_1
w_2
w_d
$\sum$
$\mathbf{w}^T\mathbf{x}$
$\varphi(\cdot)$
$\varphi(\mathbf{w}^T\mathbf{x}) = \begin{cases} +1 & \mathbf{w}^T\mathbf{x} \geq 0 \\ -1 & \mathbf{w}^T\mathbf{x} < 0 \end{cases}$
Linear combination of features and weights
Activation function
Input
Output

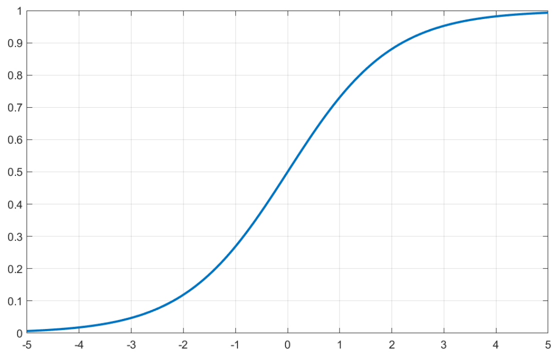
LOGISTIC REGRESSION



ACTIVATION FUNCTIONS

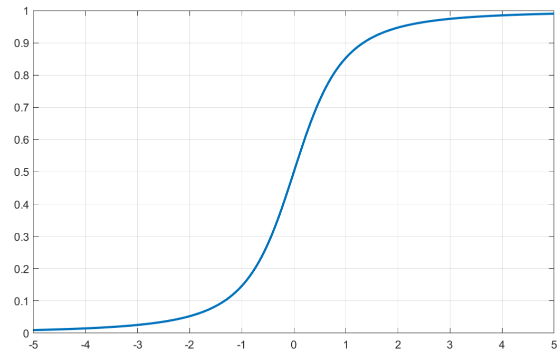
$$\varphi(t) = \frac{1}{1+e^{-t}}$$

$\leftarrow$ sigmoid function



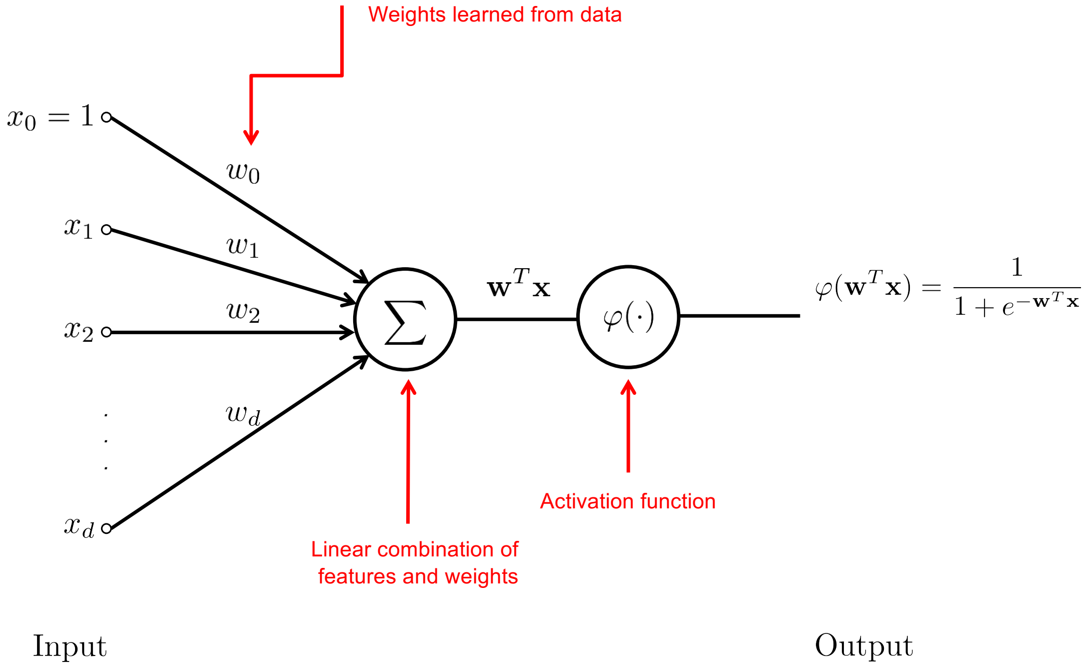
t

$$\varphi(t) = \frac{1}{2} \cdot \left(1 + \frac{t}{\sqrt{1+t^2}}\right)$$



t

Logistic regression



Thinking about Logistic Regression

Can we model the (X, Y) dependence using a linear combination? $\mathcal{X} = \{1\} \times \mathbb{R}^d$, $\mathcal{Y} = \{0, 1\}$

$$P(Y = 1|\mathbf{x}) = \mathbf{w}^T\mathbf{x}$$

How about the odds function?

$$\underbrace{\frac{P(Y = 1|\mathbf{x})}{1 - P(Y = 1|\mathbf{x})}}_{\text{odds function}} = \mathbf{w}^T\mathbf{x}$$

How about the log odds function?

$$\log \frac{P(Y = 1|\mathbf{x})}{1 - P(Y = 1|\mathbf{x})} = \mathbf{w}^T\mathbf{x}$$

Maximizing Likelihood

$$P(Y=1|\mathbf{x},\mathbf{w}) = \frac{1}{1+e^{-\mathbf{w}^T\mathbf{x}}}$$

$$\mathcal{X} = \{1\} \times \mathbb{R}^d,\ \mathcal{Y} = \{0,1\}$$

Let's express the probability mass function as:

$$p(y|\mathbf{x},\mathbf{w}) = \begin{cases} \left(\frac{1}{1+e^{-\mathbf{w}^T\mathbf{x}}}\right)^y & \text{for } y=1 \\ \left(1-\frac{1}{1+e^{-\mathbf{w}^T\mathbf{x}}}\right)^{1-y} & \text{for } y=0 \end{cases}$$

And make it more compact:

$$p(y|\mathbf{x},\mathbf{w}) = \left(\frac{1}{1+e^{-\mathbf{w}^T\mathbf{x}}}\right)^y \left(1-\frac{1}{1+e^{-\mathbf{w}^T\mathbf{x}}}\right)^{1-y}$$

Maximizing Likelihood

Given: a set of observations $\mathcal{D} = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^{n}$, $(\boldsymbol{x}_i, y_i) \in \mathcal{X} \times \mathcal{Y}$ $\mathcal{X} = \{1\} \times \mathbb{R}^d$, $\mathcal{Y} = \{0, 1\}$

Let's express the conditional likelihood as:

$$l(\mathbf{w}) = \prod_{i=1}^{n} p(y_i | \mathbf{x}_i, \mathbf{w}) \qquad \longrightarrow \qquad \mathbf{w}^* = \arg\max_{\mathbf{w}} \left\{ \prod_{i=1}^{n} p(y_i | \mathbf{x}_i, \mathbf{w}) \right\}$$

UPDATE RULES TO MAXIMIZE LIKELIHOOD

$$\mathbf{w}^{(0)} = \text{something}$$

$$\mathbf{w}^{(t+1)} = \mathbf{w}^{(t)} + \Delta\mathbf{w}^{(t)}$$

Maximum Likelihood:

$$\mathbf{w}^{(0)} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$$

$$\mathbf{w}^{(t+1)} = \mathbf{w}^{(t)} + \eta\left(\mathbf{X}^T\mathbf{P}^{(t)}\left(\mathbf{I} - \mathbf{P}^{(t)}\right)\mathbf{X}\right)^{-1}\mathbf{X}^T\left(\mathbf{y} - \mathbf{p}^{(t)}\right)$$

$\mathbf{p} = (p_1, p_2, \ldots, p_n)$ $\quad\mathbf{P} = \text{diag}\{\mathbf{p}\}$ $\quad p_i = P(Y_i = 1|\mathbf{x}_i, \mathbf{w}) = \frac{1}{1+e^{-\mathbf{w}^T\mathbf{x}_i}}$

$\mathbf{y} = (y_1, y_2, \ldots, y_n)$ $\quad\mathbf{I} = \text{identity matrix}$

MINIMIZING SUM OF SQUARED ERRORS

Given: a set of observations $\mathcal{D} = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^{n}$, $(\boldsymbol{x}_i, y_i) \in \mathcal{X} \times \mathcal{Y}$ $\qquad \mathcal{X} = \{1\} \times \mathbb{R}^d$, $\mathcal{Y} = \{0, 1\}$

Let's express the sum of squared errors as:

$$E(\mathbf{w}) = \sum_{i=1}^{n} (y_i - p_i)^2 = \sum_{i=1}^{n} e_i^2$$

$$p_i = P(Y_i = 1 | \mathbf{x}_i, \mathbf{w}) = \frac{1}{1+e^{-\mathbf{w}^T \mathbf{x}_i}}$$

Update Rules to Minimize Sum of Squared Errors

$$\mathbf{w}^{(0)} = \text{something}$$

$$\mathbf{w}^{(t+1)} = \mathbf{w}^{(t)} + \Delta\mathbf{w}^{(t)}$$

Minimum Sum of Squared Errors:

$$\mathbf{w}^{(0)} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$$

$$\mathbf{w}^{(t+1)} = \mathbf{w}^{(t)} + \eta\left(\mathbf{J}^{(t)T}\mathbf{J}^{(t)} + \mathbf{J}^{(t)T}\mathbf{E}^{(t)}(2\mathbf{P}^{(t)} - \mathbf{I})\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{P}^{(t)}\left(\mathbf{I} - \mathbf{P}^{(t)}\right)\left(\mathbf{y} - \mathbf{p}^{(t)}\right)$$

$\mathbf{p} = (p_1, p_2, \ldots, p_n)$ $\quad$ $\mathbf{P} = \text{diag}\{\mathbf{p}\}$ $\quad$ $\mathbf{e} = (y_1 - p_1, y_2 - p_2, \ldots, y_n - p_n)$

$\mathbf{y} = (y_1, y_2, \ldots, y_n)$ $\quad$ $\mathbf{J} = \mathbf{P}(\mathbf{I} - \mathbf{P})\mathbf{X}$ $\quad$ $\mathbf{E} = \text{diag}\{\mathbf{e}\}$ $\quad$ $\mathbf{I} = \text{identity matrix}$

UPDATE RULES, BATCH MODE

$$\mathbf{w}^{(0)} = \text{something}$$

$$\mathbf{w}^{(t+1)} = \mathbf{w}^{(t)} + \Delta\mathbf{w}^{(t)}$$

Maximum Likelihood:

$$\mathbf{w}^{(0)} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$$

$$\mathbf{w}^{(t+1)} = \mathbf{w}^{(t)} + \eta\left(\mathbf{X}^T\mathbf{P}^{(t)}\left(\mathbf{I}-\mathbf{P}^{(t)}\right)\mathbf{X}\right)^{-1}\mathbf{X}^T\left(\mathbf{y}-\mathbf{p}^{(t)}\right)$$

Minimum Sum of Squared Errors:

$$\mathbf{w}^{(0)} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$$

$$\mathbf{w}^{(t+1)} = \mathbf{w}^{(t)} + \eta\left(\mathbf{J}^{(t)T}\mathbf{J}^{(t)} + \mathbf{J}^{(t)T}\mathbf{E}^{(t)}(2\mathbf{P}^{(t)}-\mathbf{I})\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{P}^{(t)}\left(\mathbf{I}-\mathbf{P}^{(t)}\right)\left(\mathbf{y}-\mathbf{p}^{(t)}\right)$$

$\mathbf{p} = (p_1, p_2, \ldots, p_n)$ $\mathbf{P} = \text{diag}\{\mathbf{p}\}$ $\mathbf{e} = (y_1 - p_1, y_2 - p_2, \ldots, y_n - p_n)$

$\mathbf{y} = (y_1, y_2, \ldots, y_n)$ $\mathbf{J} = \mathbf{P}(\mathbf{I}-\mathbf{P})\mathbf{X}$ $\mathbf{E} = \text{diag}\{\mathbf{e}\}$ $\mathbf{I} = \text{identity matrix}$

Update rules, batch mode

$$\mathbf{w} \leftarrow \text{something}$$

$$\mathbf{w} \leftarrow \mathbf{w} + \Delta\mathbf{w}$$

Gradient Descent:

$$\Delta\mathbf{w} = \eta \mathbf{X}^T \mathbf{P} \left(\mathbf{I} - \mathbf{P}\right) \left(\mathbf{y} - \mathbf{p}\right)$$

$$\Delta\mathbf{w} = \eta \cdot \begin{array}{l} \left[\begin{array}{cccc} p_1(1-p_1)x_{11} & p_2(1-p_2)x_{21} & \cdots & p_n(1-p_n)x_{n1} \\ p_1(1-p_1)x_{12} & p_2(1-p_2)x_{22} & & \\ \vdots & & \ddots & \\ p_1(1-p_1)x_{1d} & & & p_n(1-p_n)x_{nd} \end{array}\right] \\ \qquad\qquad\qquad \mathbf{x}_1 \end{array} \cdot \left[\begin{array}{c} e_1 \\ e_2 \\ \vdots \\ e_n \end{array}\right]$$

$$\Delta\mathbf{w} = \eta \sum_{i=1}^{n} e_i p_i (1 - p_i) \mathbf{x}_i,$$

Update rules, stochastic mode

$$\mathbf{w} \leftarrow \text{something}$$

$$\mathbf{w} \leftarrow \mathbf{w} + \Delta\mathbf{w}$$

Gradient Descent:

$$\Delta\mathbf{w} = \eta \sum_{i=1}^{n} e_i p_i (1 - p_i) \mathbf{x}_i$$

Stoschastic Gradient Descent:

$$\Delta\mathbf{w} = \eta e_i p_i (1 - p_i) \mathbf{x}_i$$

Prediction

Given: a predictor defined by $\mathbf{w}$ and a new data point $\mathbf{x}$

$$P(Y = 1|\mathbf{x}, \mathbf{w}) = \frac{1}{1+e^{-\mathbf{w}^T\mathbf{x}}}$$

$$r = \frac{w_0 + \sum_{j=1}^{d} w_j x_j}{\sqrt{\sum_{j=1}^{d} w_j^2}} \qquad \longleftarrow \qquad \text{distance from } \mathbf{x} \text{ to hyperplane}$$

$$P(Y = 1|\mathbf{x}) = \frac{1}{1+e^{-r\cdot\sqrt{\sum_{j=1}^{d} w_j^2}}}$$