LOGISTIC REGRESSION

CS6140

Predrag Radivojac
KHOURY COLLEGE OF COMPUTER SCIENCES
NORTHEASTERN UNIVERSITY
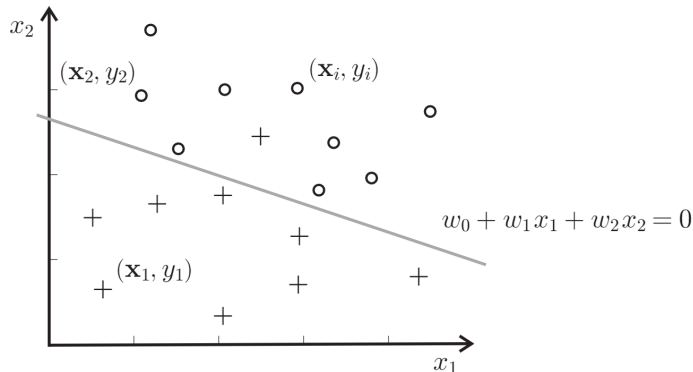
Fall 2024

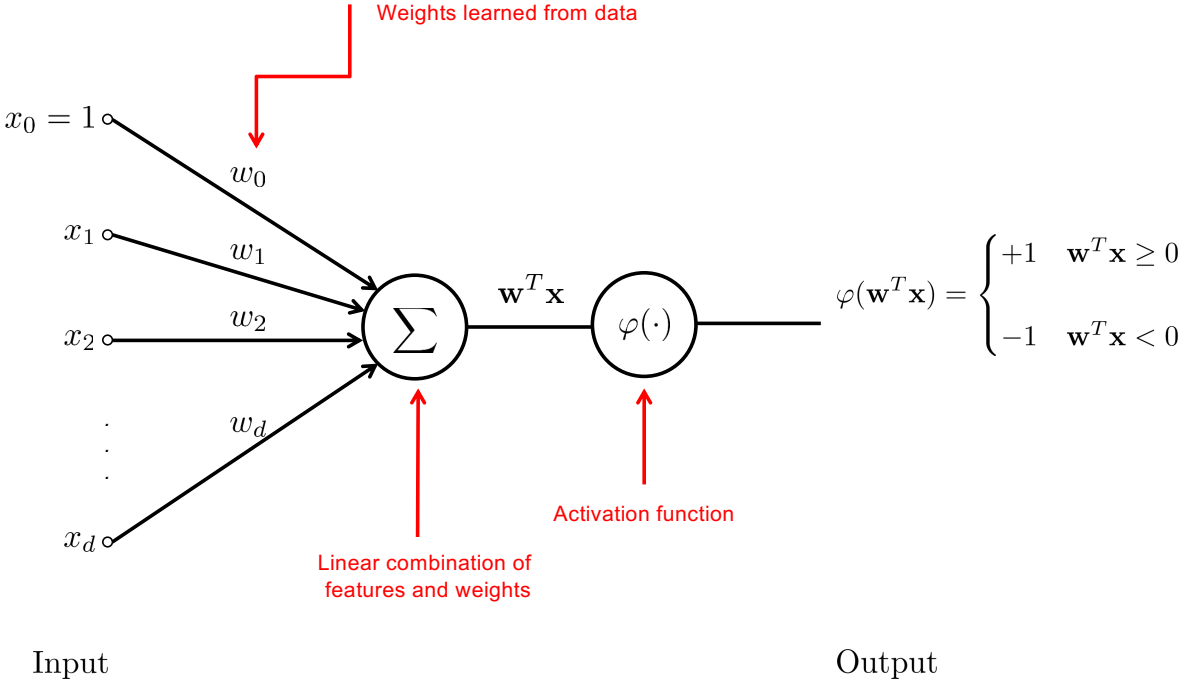# LINEAR CLASSIFICATION

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 $f(\boldsymbol{x}) = w_0 + \sum_{j=1}^{d} w_j x_j$
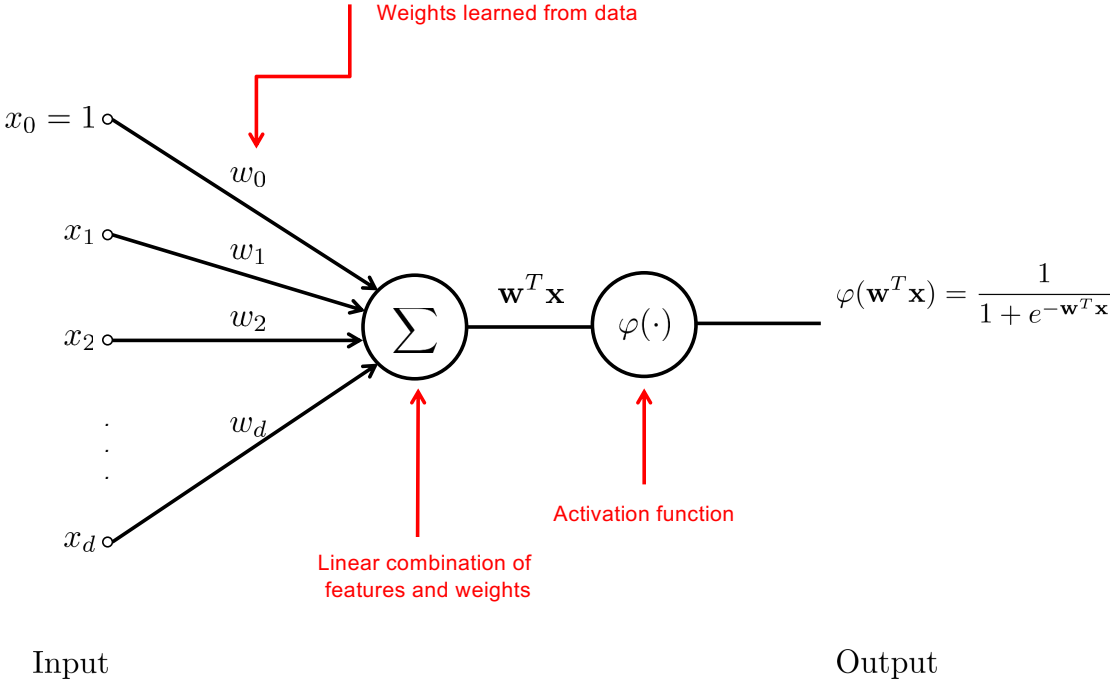
$\mathcal{X} = \mathbb{R} \times \mathbb{R}, \ \mathcal{Y} = \{0,1\}$

PERCEPTRON

Weights learned from data

$x_0 = 1$

w_0

x_1

w_1

x_2

w_2

$\sum$

$\mathbf{w}^T\mathbf{x}$

$\varphi(\cdot)$

w_d

x_d

Linear combination of features and weights

Activation function

$$\varphi(\mathbf{w}^T\mathbf{x}) = \begin{cases} +1 & \mathbf{w}^T\mathbf{x} \geq 0 \\ -1 & \mathbf{w}^T\mathbf{x} < 0 \end{cases}$$

Input

Output

LOGISTIC REGRESSION

Weights learned from data

$x_0 = 1$

w_0

x_1

w_1

x_2

w_2

$\cdot$
$\cdot$
$\cdot$

w_d

x_d

$\sum$

$\mathbf{w}^T\mathbf{x}$

$\varphi(\cdot)$

$\varphi(\mathbf{w}^T\mathbf{x}) = \dfrac{1}{1 + e^{-\mathbf{w}^T\mathbf{x}}}$

Linear combination of
features and weights

Activation function

Input

Output
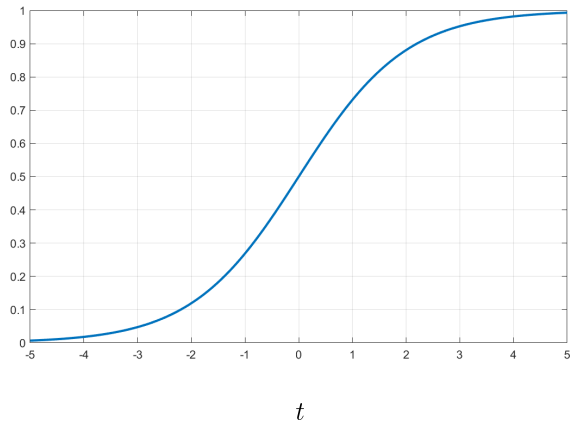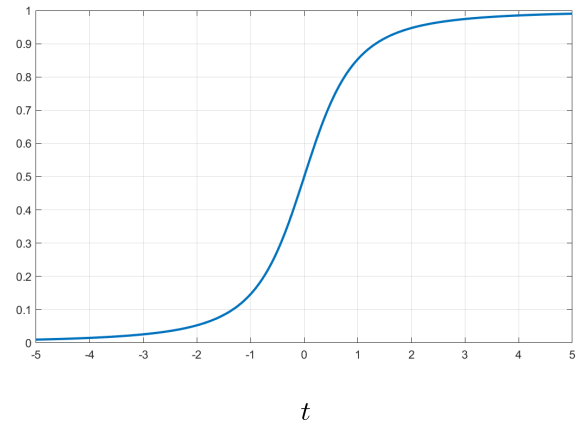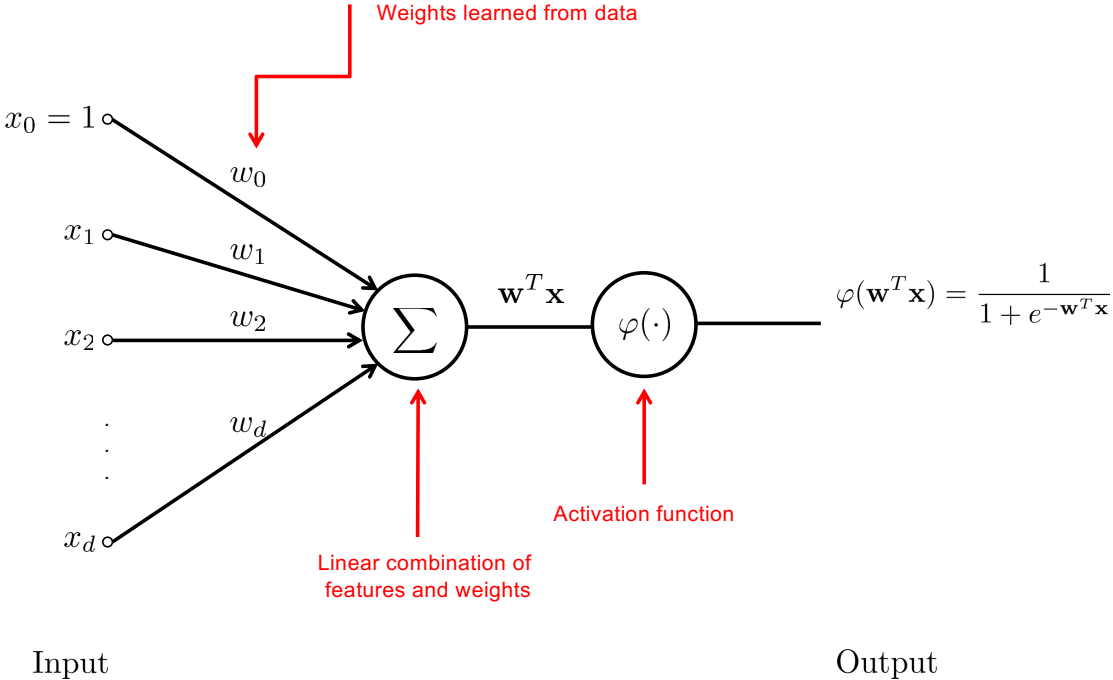
ACTIVATION FUNCTIONS

$$\varphi(t) = \frac{1}{1 + e^{-t}} \qquad \leftarrow \text{sigmoid function}$$

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

t

t

LOGISTIC REGRESSION

Thinking about Logistic Regression

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

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

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

Let's express the conditional likelihood as:

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

UPDATE RULES TO MAXIMIZE LIKELIHOOD

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

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

Maximum Likelihood:

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

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

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

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

Minimizing Sum of Squared Errors

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

Let's express the sum of squared errors as:

$$E(\mathbf{w}) = \sum_{i=1}^n (y_i - p_i)^2 = \sum_{i=1}^n e_i^2 \qquad\qquad p_i = P(Y_i = 1 | \mathbf{x}_i, \mathbf{w}) = \frac{1}{1 + e^{-\mathbf{w}^T \mathbf{x}_i}}$$

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:

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

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

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

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

UPDATE RULES, BATCH MODE

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

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

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:

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

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

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

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

UPDATE RULES, BATCH MODE

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

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

Gradient Descent:

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

$$\Delta\mathbf{w} = \eta \cdot \begin{bmatrix} p_1(1-p_1)x_{11} & p_2(1-p_2)x_{21} & \cdots & p_n(1-p_n)x_{n1} \\ p_1(1-p_1)x_{12} & p_2(1-p_2)x_{22} & & \\ \vdots & & \ddots & \\ p_1(1-p_1)x_{1d} & & & p_n(1-p_n)x_{nd} \end{bmatrix} \cdot \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix}$$

$\mathbf{x}_1$

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

UPDATE RULES, STOCHASTIC MODE

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

Gradient Descent:

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

Stochastic 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 $\mathbf{x}$

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

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

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