Lec 08: Logistic Regression

$$P(y_i \mid x_i, W)$$

$$\begin{bmatrix} P(y_i = 1 \mid x_i, w) \\ P(y_i = 0 \mid x_i, w) \end{bmatrix}$$

$$P(y_i = 1 \mid x_i, w) = \frac{e^{w_1^T x_i}}{e^{w_1^T x_i} + e^{w_2^T x_i}}$$

$$\begin{bmatrix} \varrho \, e^{w_1^T x_i} \\ \varrho \, e^{w_2^T x_i} \end{bmatrix} \qquad \varrho = \frac{1}{e^{w_1^T x_i} + e^{w_2^T x_i}}$$

$$f(w, x_i) = \underline{\sigma(w^T x_i)} = \frac{1}{1 + e^{-w^T x_i}}, \text{ where } \underline{w = w_1 - w_2}$$

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

Binary LR:

$$W^*_{LR} \in \mathbb{R}^d$$

$$W^*_{LR} = \arg\max \prod_{i=1}^{n} P(y_i \mid x_i, W)$$

$$= \arg\max \sum_{i=1}^{n} \log P(y_i \mid x_i, W)$$

$$= \arg\min \left\{ -\sum_{i=1}^{n} \log P(y_i \mid x_i, W) \right\}$$

$$-\log P(y_i \mid x_i, W)$$

$$= -y_i \log\left[\sigma(W^T x_i)\right] - (1-y_i)\log\left[1-\sigma(W^T x_i)\right]$$

$\boxed{\text{Cross-entropy loss}}$

$$H(p,q) = -\sum_{x \in X} p(x) \log q(x)$$

$$NLL_i(W) = -\log P(y_i \mid x_i, W)$$: negative log likelihood

$\hookrightarrow \log P(y_i = 1 \mid x_i, W)$, $y_i = 1$ or $\log(1 - P(y_i = 1 \mid x_i, W))$ if $y_i = 0$

$$NLL_i(w) = y_i \log\left(1+e^{-w^T x_i}\right) - (1-y_i)\log\left(\frac{e^{-w^T x_i}}{1+e^{-w^T x_i}}\right)$$

$$= y_i \log\left(1+e^{-w^T x_i}\right) + (1-y_i)\, w^T x_i + (1-y_i)\log\left(1+e^{-w^T x_i}\right)$$

$$= \log\left(1+e^{-w^T x_i}\right) + (1-y_i)\, w^T x_i$$

$$\nabla_w NLL_i(w) = -\frac{e^{-w^T x_i}}{1+e^{-w^T x_i}} \cdot x_i + (1-y_i)\cdot x_i \quad \in \mathbb{R}^d$$

$$= -\left(y_i - \sigma\left(w^T x_i\right)\right) x_i$$

Apply GD:

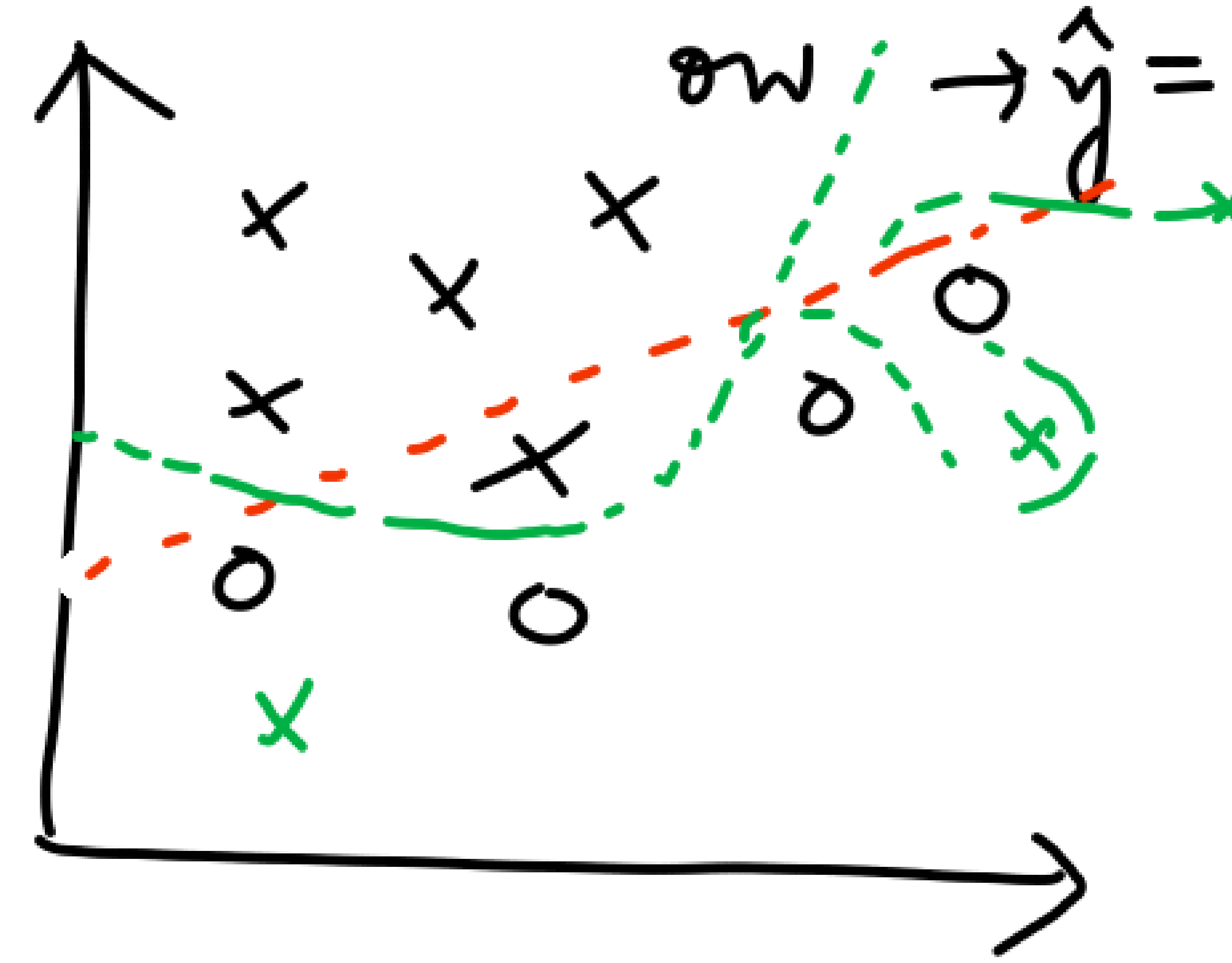
$$\boxed{w_{t+1} \leftarrow w_t - \eta \sum_{i=1}^{n} \nabla_w NLL_i(w)}$$

$$\frac{P(y=1 \mid \hat{x}, w)}{P(y=0 \mid \hat{x}, w)} > 1 \rightarrow \hat{y} = 1$$

ow $\rightarrow \hat{y} = 0$

$\Rightarrow w^T \hat{x} > 0 \rightarrow \hat{y} = 1$

ow $\rightarrow \hat{y} = 0$



Basis functions: $\Phi(x) \rightarrow$

$\sigma(w^T \phi)$

$w^T \phi > 0$

$$NLL_{Reg}(w) = NLL_{LR}(w) + \lambda \|w\|^2$$: minimize using GD.

Multi-class classification:

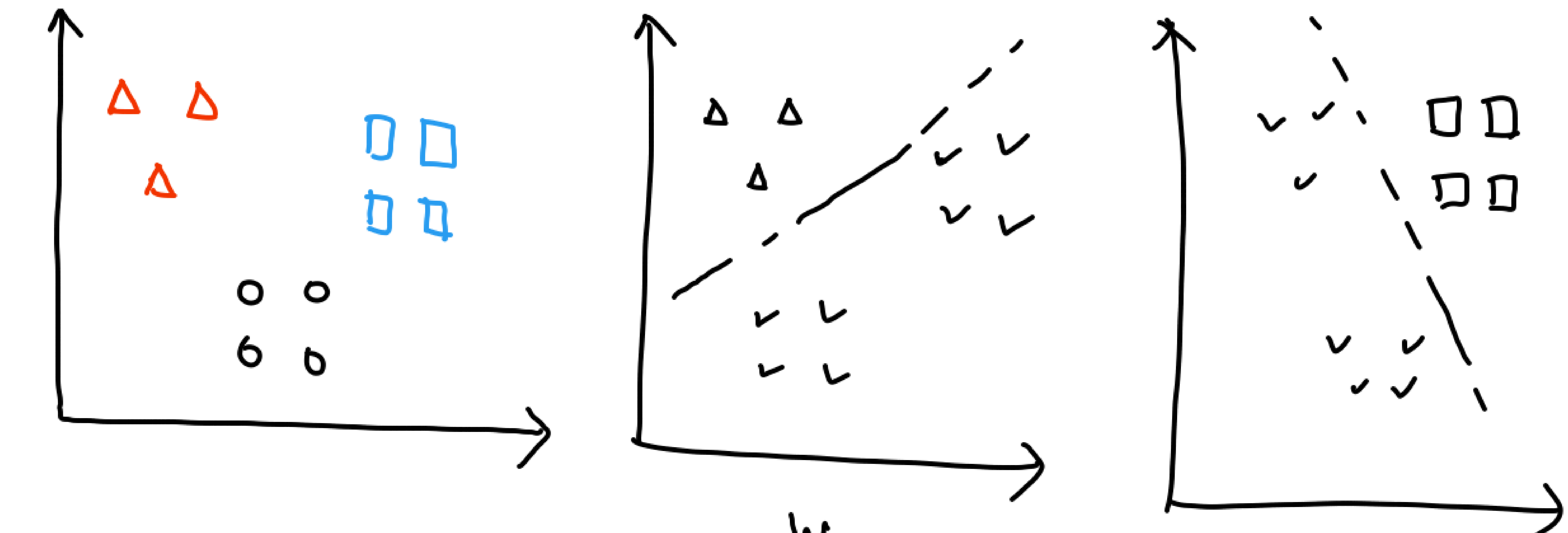
test point $\hat{x}$

$\sigma(w_1^T \hat{x}) \rightarrow$ class 1

$\sigma(w_2^T \hat{x}) \rightarrow$ class 2

$\sigma(w_3^T \hat{x}) \rightarrow$ class 3

① One-vs-rest classifier



w_1 $\quad$ w_2 $\quad$ w_3

Any binary classifier can be used.

Final prediction

$$\hat{y} = \underset{k}{\operatorname{argmax}}\, \sigma(w_k^T \hat{x})$$

Softmax Regression

$$
\begin{aligned}
&1 \quad \varphi e^{w_1^T x} \longrightarrow P(y=1 \mid x, w_1) \\
&2 \quad \varphi e^{w_2^T x} \longrightarrow P(y=2 \mid x, w_2) \\
&3 \quad \varphi e^{w_3^T x} \\
&\vdots \qquad\qquad\qquad\qquad \vdots \\
&K \quad \varphi e^{w_K^T x} \qquad\quad P(y=K \mid x, w_K)
\end{aligned}
$$

$$\varphi = \frac{1}{\sum_{j=1}^{K} e^{w_j^T x}}$$

$$\boxed{P(y=j \mid x, w) = \frac{e^{w_j^T x}}{\sum_{i=1}^{K} e^{w_i^T x}}}$$

$$f(x, W) = \begin{bmatrix} P(y=1 \mid x, w) \\ P(y=2 \mid x, w) \\ \vdots \\ P(y=K \mid x, w) \end{bmatrix} = \begin{bmatrix} \varphi e^{w_1^T x} \\ \vdots \\ \varphi e^{w_K^T x} \end{bmatrix}$$

$$= \text{Softmax}(x, w)$$

$$W = \begin{bmatrix} - & w_1^T & - \\ - & w_2^T & - \\ & \vdots & \\ - & w_K^T & - \end{bmatrix}_{K \times d}$$

$$NLL(W) = -\sum_{i=1}^{n} \log P(y_i \mid x_i, W)$$

$$= -\sum_{i=1}^{n} \sum_{k=1}^{K} \mathbb{I}\{y_i = k\} \log \frac{e^{w_k^T x}}{\sum_{j=1}^{K} e^{w_j^T x}}$$

indicator $\{ \quad \}$

$$NLL_i(W) = -\sum_{k=1}^{K} \mathbb{I}\{y_i = k\} \left[w_k^T x - \log \sum_{j=1}^{K} e^{w_j^T x} \right]$$

$$-\nabla_{w_k} NLL_i(W) = \begin{cases} x - \dfrac{e^{w_k^T x}}{\sum e^{w_j^T x}} \cdot x \ ; & y_i = k \\ -\dfrac{e^{w_k^T x}}{\sum e^{w_j^T x}} \cdot x \ ; & y_i \neq k \end{cases}$$

$$\nabla_{w_k} NLL_i(W) = -\Big[\underbrace{\mathbb{I}\{y_i = k\}}_{y_i^{(k)}} - f_k(x, W)\Big] \cdot x$$

$$\nabla_{w_k} NLL(W) = \sum_{i=1}^{n} \nabla_{w_k} NLL_i(W)$$

$$y_i \in \{1, 2, 3, \ldots, K\}$$

$$y_i = \begin{array}{c} \begin{bmatrix} 0 & 0 & \cdots & 1 & 0 & 0 \end{bmatrix} \\ \quad\quad\quad\quad\;\; \uparrow \\ \quad\quad\quad\quad\;\; y_i^{(k)} \end{array}$$

: one-hot representation

<u>GD Step</u>

$$\begin{bmatrix} - & w_1^T & - \\ & \vdots & \\ - & w_K^T & - \end{bmatrix}_{t+1} \leftarrow \begin{bmatrix} w_1^T \\ \vdots \\ w_K^T \end{bmatrix}_t - \eta \begin{bmatrix} \cdots & \nabla_{w_1} NLL(W) & \cdots \\ & \vdots & \\ \cdots & \nabla_{w_K} NLL(W) & \cdots \end{bmatrix} \rightarrow W^*$$

NB vs LR

$$\left\{ \arg\max_{y} \underbrace{P(Y=y)}_{\text{infer class distribution}} \prod_{i=1}^{d} \underbrace{P(x_i \mid Y=y)}_{\text{infer data distribution}} \right. \quad : \text{Naive Bayes}$$

Generative models

Logistic regression $\longrightarrow$ $\arg\max\limits_{y} P(y \mid x, w)$ $\longrightarrow$ Discriminative models.

Gaussian Naive Bayes : Special case of LR

$$\underset{y_k}{\operatorname{argmax}} \; P(Y = y_k) \prod_{i=1}^{d} P(x_i \mid Y = y_k)$$

continuous

$$P(x_j \mid y_k) \sim N(\mu_{jk}, \sigma_{jk}^2)$$

1. x_i, x_j are conditionally indept.
2. $P(y=1) = \pi$, $P(y=0) = 1 - \pi$
3. $P(x_i \mid y=0) \sim N(\mu_{i0}, \sigma_i^2)$
   $P(x_i \mid y=1) \sim N(\mu_{i1}, \sigma_i^2)$

HW: find w_0, w_j's

$$P(y_i = 1 \mid x) = \frac{P(x \mid y_i = 1) P(y_i = 1)}{P(x \mid y_i = 1) P(y_i = 1) + P(x \mid y_i = 0) P(y_i = 0)}$$

$$\vdots$$

$$= \frac{1}{1 + \exp\{w_0 + \sum w_j x_j\}}$$