线性分类

线性分类

线性分类：线性回归+激活函数

$$y = w^T x + b \longrightarrow y = f(w^T x + b)$$

硬分类：$y \in \{0,1\}$，　　感知机等

软分类：$y \in [0,1]$，逻辑回归、高斯判别分析

$$p(y|x) = \frac{p(x|y)p(y)}{p(x)} \propto p(x|y)p(y)$$

感知机

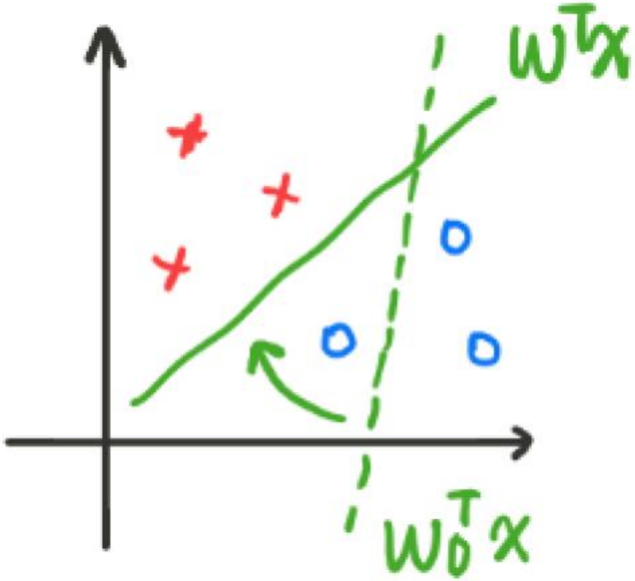
感知机

- □ 感知机模型是一类错误驱动的模型，它的中心思想也就是”错误驱动”：也就是哪些数据点分类错误了，那么我们就进行调整权值系数w，直到分类正确为止。

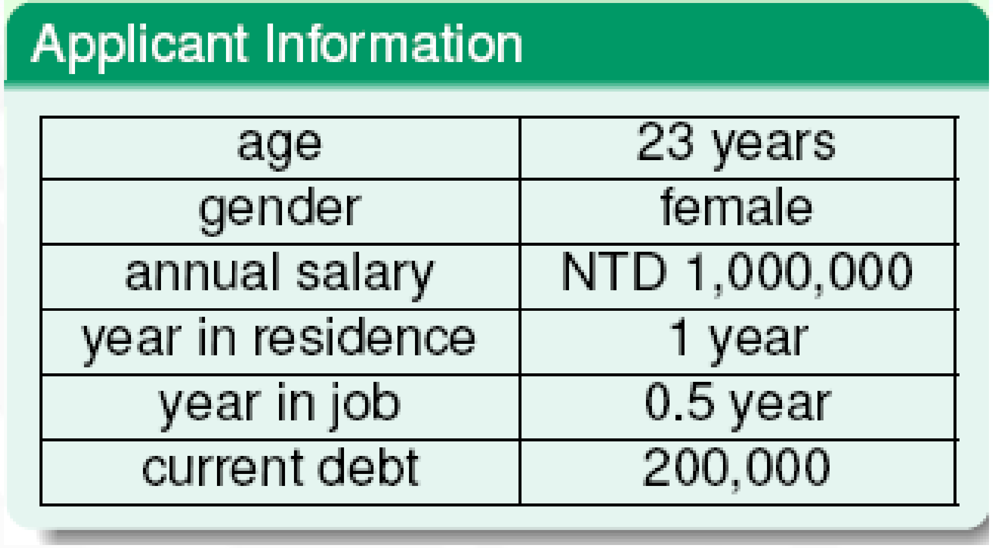
- □ 感知机可以做如下的描述：

$$f(x) = sign(w^T x + b) \quad x \in \mathbf{R}^p, w \in \mathbf{R}^p$$

$$sign(a) = \begin{cases} +1 & a \geq 0 \\ -1 & a < 0 \end{cases}$$



感知机



Applicant Information

age	23 years
gender	female
annual salary	NTD 1,000,000
year in residence	1 year
year in job	0.5 year
current debt	200,000

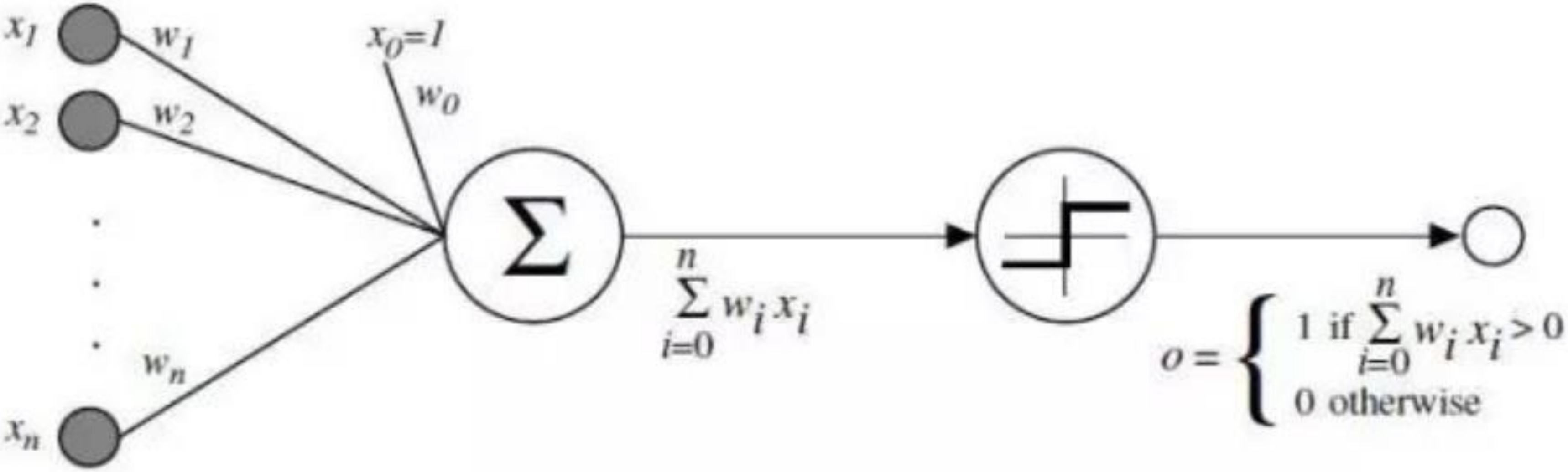
□ 输入数据： $x = (x^1, x^2, \cdots, x^p)^T \in X$

□ 加权分数： if $\omega^T x = \sum_{i=1}^{p} \omega_i x^i \geq \theta$, 通过($y = 1$)

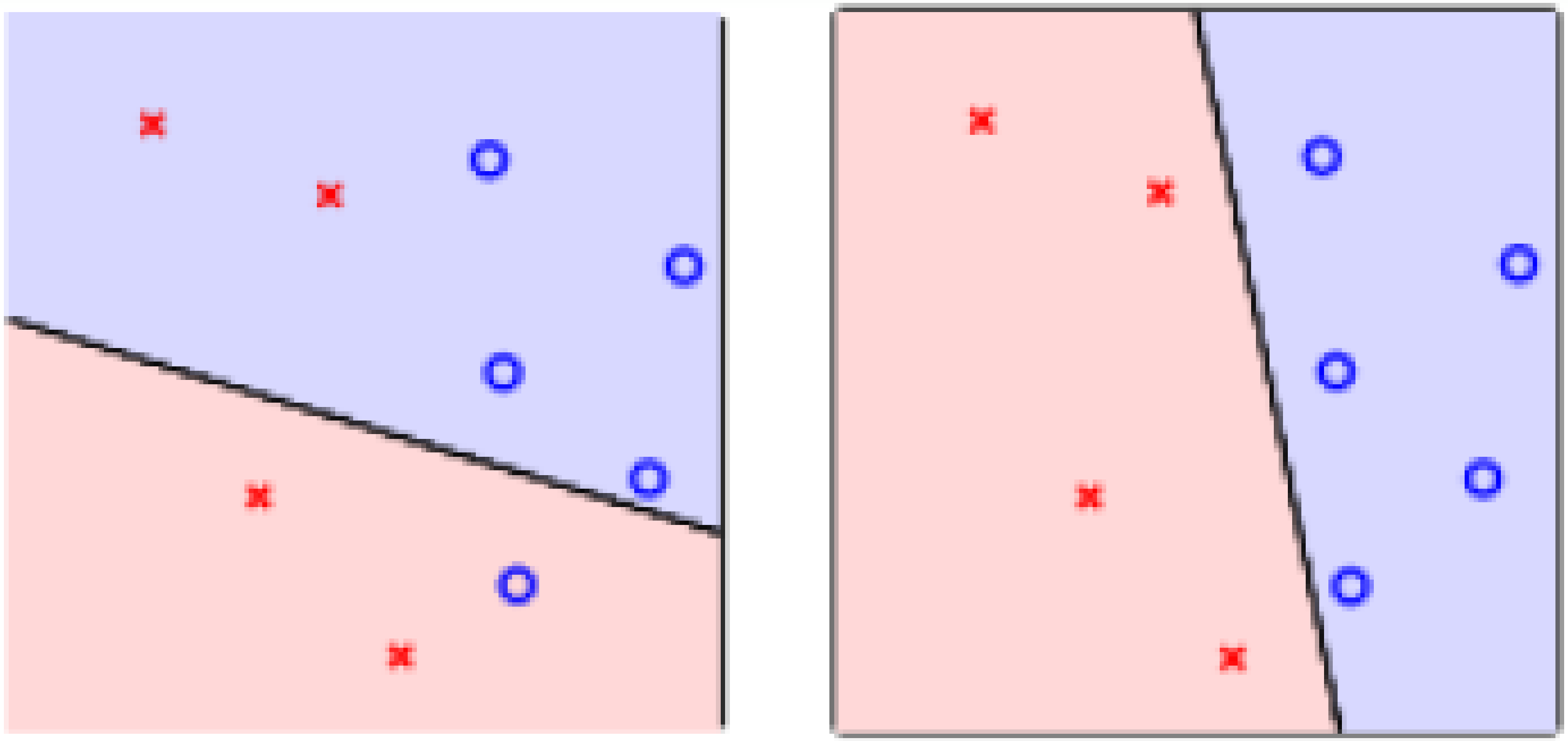
if $\omega^T x = \sum_{i=1}^{p} \omega_i x^i < \theta$, 拒绝($y = -1$)

□ 判别函数： $h(x) = \text{sign} \sum_{i=0}^{p} \omega_i x^i = \text{sign}\, \hat{\omega}^T x$

感知机


x_1
w_1
x_2
w_2
$x_0=1$
w_0
w_n
x_n
Σ
$\sum_{i=0}^{n} w_i x_i$
$o = \begin{cases} 1 \text{ if } \sum_{i=0}^{n} w_i x_i > 0 \\ 0 \text{ otherwise} \end{cases}$

感知机



$$h(x) = \mathrm{sign}\,\widehat{\omega}^T x = \mathrm{sign}(\omega_0 + \omega_1 x^1 + \omega_2 x^2)$$

感知机

- □ 数据集的线性可分性
  - ■ 给定一个数据集$T=\{(\mathbf{x}_1,y_1),(\mathbf{x}_2,y_2),\ldots,(\mathbf{x}_N,y_N)\}$，其中 $\mathbf{x}_i\in\mathcal{X}=\mathbf{R}^n$，$y_i\in\mathcal{Y}=\{+1,-1\}$，$i=\{1,2,\ldots,N\}$

    如果存在某个超平面S：$w\cdot\mathbf{x}+b=0$，能够将数据集的正实例点和负实例点完全正确地划分到超平面的两侧，即对所有的 $y_i=+1$的实例i，有 $w\cdot\mathbf{x}+b>0$，对所有 $y_i=-1$的实例i，有 $w\cdot\mathbf{x}+b<0$，则称数据集T为线性可分数据集（linearly separable data set），否则，称数据集T线性不可分

感知机

□ 损失函数: (M是误分类集合)

$L(\widehat{\omega}) = \sum_i [\![y_i \neq \hat{y}_i]\!]$ 该函数不可导且不是凸函数。

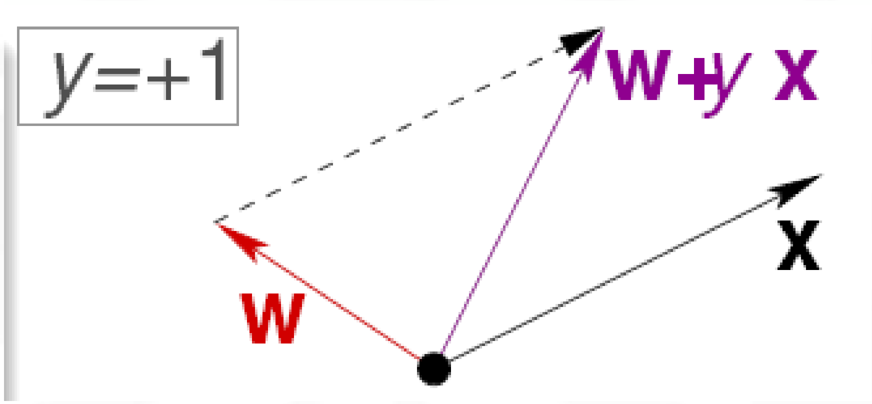
$$L(\hat{w}) \quad = \max(0, -y_i \hat{y}_i) = -\sum_{i \in M} y_i \hat{w}^T x_i$$

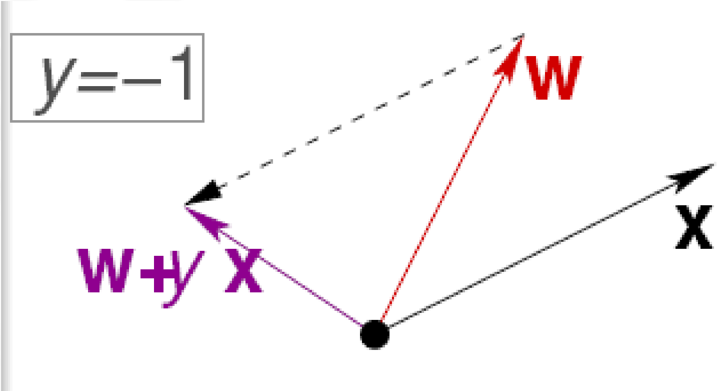
那么，梯度可以表示为：$\nabla_w L = -\sum_{i \in M} y_i \mathbf{x}_i$

权值参数w的更新过程为：$\eta > 0$

$$w_{t+1} = w_t + \eta \sum_{i \in M} y_i \mathbf{x}_i$$

感知机





$$L(\widehat{\omega}) = -\sum_{x_i\in M} y_i\hat{y}_i = -\sum_{x_i\in M} y_i\widehat{\omega}^T x_i$$

感知机: 知错能改，逐点修正

（1）随机选择一条直线（$w = w_0$）进行分类

（2） For $t \; = 0, 1 \cdots$

（3）对于直线w_t, 找一个分类错误的点(x_t, y_t)，

$$sign(w_t^T x_t) \neq y_t$$

（4）试图修改对于该点的分类错误

$$w_{t+1} = w_t + \eta y_t x_t$$

（5）更新$w = w_{t+1}$，循环直到所有点的都分类正确

（6）输出最后的w作为$\hat{\omega}$

感知机

- □ 算法收敛性
  - ■ 数据集线性可分时终止，否则不会停止

（1）终止时：对任意点x_i，均有

$$y_i \omega_f^T x_i \geq \min_n y_n \omega_f^T x_n > 0$$

（2）$\omega_f^T w_{t+1} > \omega_f^T w_t$

（3）$\|w_{t+1}\|^2 \leq \|w_t\|^2 + \eta^2 \max_n \|x_n\|^2$

感知机

如果从$w_0 = 0$开始，我们有$\frac{\omega_f^T}{\|\omega_f^T\|}\frac{w_T}{\|w_T\|} \geq \sqrt{T}\frac{\rho}{R}$

其中$R = \max_n \|x_n\|$，$\rho = \min_n y_n \frac{\omega_f^T}{\|\omega_f^T\|} x_n$

这样，$T \leq \frac{R^2}{\rho^2}$

感知机: 不是线性可分时

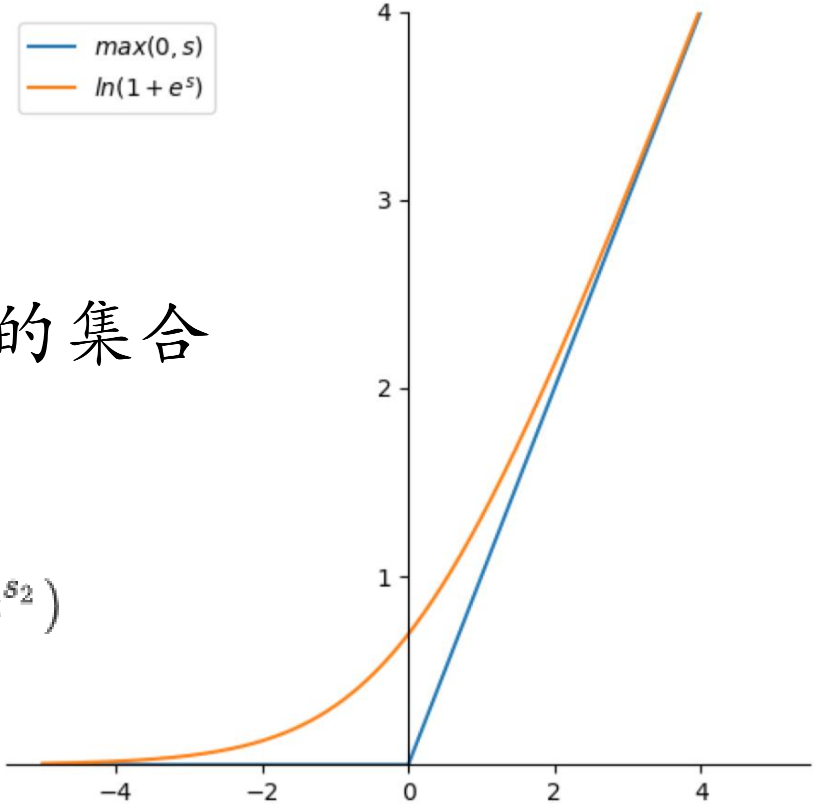
记：$s_i = -y_i \hat{y}_i = -y_i \hat{w}^T \mathbf{x}_i$

代价函数：$L(s) = \sum_i \max(0, s_i)$

可分时：$L(s) = \sum_{i \in M} s_i$，$M$为误分类点的集合

不可分时：用softmax函数近似

$$soft(s_1, s_2) = \log(e^{s_1} + e^{s_2})$$



则损失函数近似为：

$$L(\hat{w}) = \sum_i log(1 + e^{-y_i \hat{w}^T \mathbf{x}_i})$$

感知机: (Pocket Algorithm)不是线性可分时

（1）随机选择一条直线（$w = w_0$）进行分类

（2） For $t = 0, 1 \cdots$

（3）对于直线w_t, 找一个分类错误的点(x_t, y_t),

（4）试图修改对于该点的分类错误

$$w_{t+1} = w_t + \eta y_t x_t$$

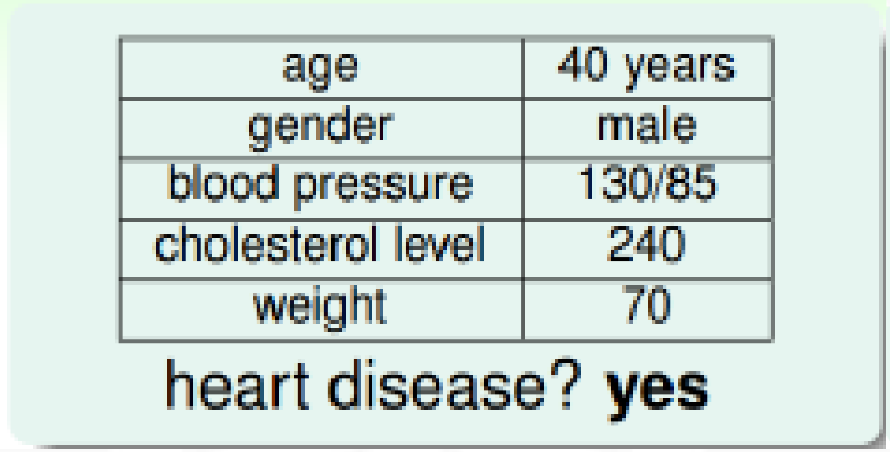
（5）if w_{t+1}分错的点比w_t分错的点少，更新$w = w_{t+1}$

（6）循环直到足够的迭代次数

（7）输出最后的w作为$\hat{\omega}$

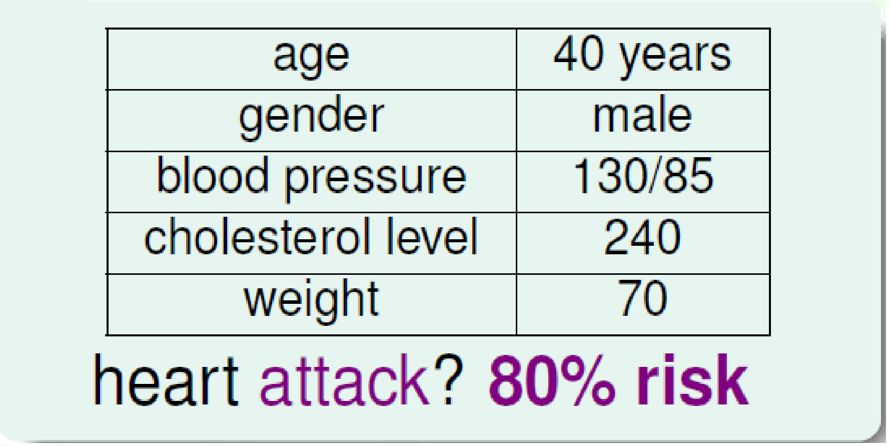
逻辑回归

逻辑回归



age	40 years
gender	male
blood pressure	130/85
cholesterol level	240
weight	70

heart disease? **yes**



age	40 years
gender	male
blood pressure	130/85
cholesterol level	240
weight	70

heart attack? **80% risk**

□ 输入数据：$x = (x^0, x^1, x^2, \cdots, x^p)^T \in X$

□ 加权分数：$s = \omega^T x = \sum_{i=0}^{p} \omega_i \, x^i$

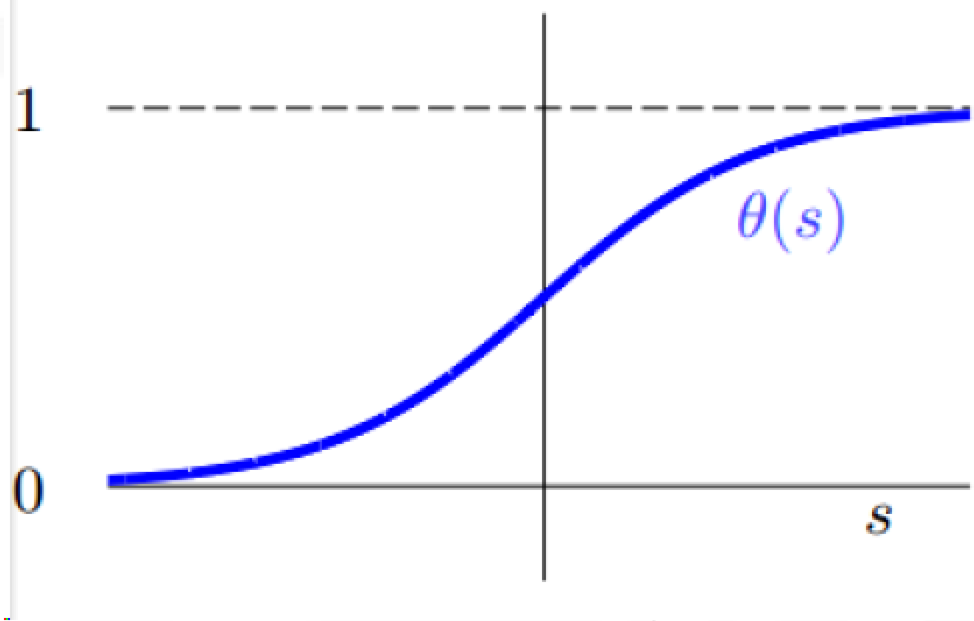
□ 判别函数：$f(x) > 0.5$, 则判断为正类

$f(x) < 0.5$, 则判断为负类

□ 软间隔分类：$f(x) = P(+1|x) \in [0,1]$

Sigmoid Function

- □ Sigmoid函数： $\theta(s) = \frac{1}{1+e^{-s}}$



$$\theta(-s) = \frac{1}{1+e^{s}} = 1 - \frac{e^{s}}{1+e^{s}} = 1 - \theta(s)$$

$$\theta'(s) = \frac{e^{-s}}{(1+e^{-s})^2} = \theta(s)\big(1-\theta(s)\big)$$

逻辑回归

□ 预测函数：$f(x) = \theta(\omega^T x) = \frac{1}{1+e^{-\omega^T x}}$

□ 后验概率：

$$p(y|\mathbf{x}) = \begin{cases} p_1 = \sigma(w^T\mathbf{x}) = \frac{1}{1+exp(-w^T\mathbf{x})} & y = 1 \\ p_0 = 1 - \sigma(w^T\mathbf{x}) = \frac{exp(-w^T\mathbf{x})}{1+exp(-w^T\mathbf{x})} & y = 0 \end{cases}$$

统一表示为：$p(y|\mathbf{x}) = p_1^y \cdot p_0^{1-y}$

□ 事件的几率(odds)：事件发生的概率与该事件不发生的概率的比值：$\frac{p}{1-p}$

□ 对数几率(logit函数)：$logit(p) = log\frac{p_1}{1-p_1} = w \cdot \mathbf{x}$

逻辑回归

- □ 极大似然法：

$$
\begin{aligned}
\hat{w} &= argmax_w \log p(y|\mathbf{x}) \\
&= argmax_w \log \Pi_i p(y_i|\mathbf{x}_i) \\
&= argmax_w \sum_i \log p(y_i|\mathbf{x}_i) \\
&= argmax_w \sum_i (y_i \log p_1 + (1 - y_i) \log p_0) \\
&= argmin_w - \frac{1}{N} \sum_i (y_i \log p_1 + (1 - y_i) \log p_0) \\
&= argmin_w J(w)
\end{aligned}
$$

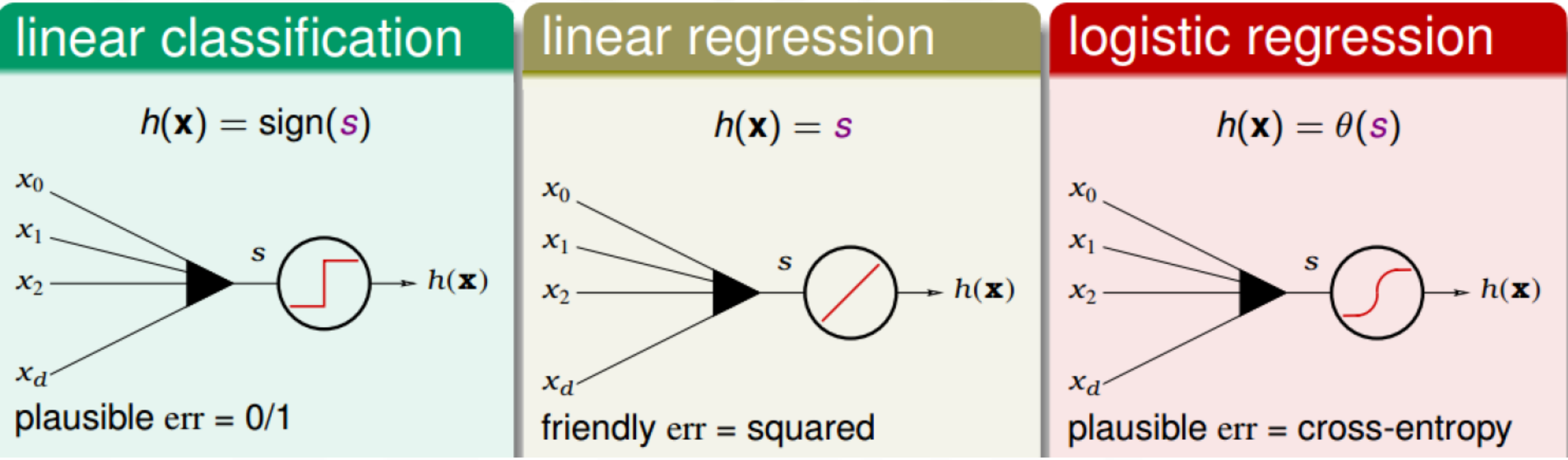
- □ 参数估计值:

$$
\begin{aligned}
\hat{w}_j = \frac{\partial J(w)}{\partial w_j} &= -\frac{1}{N} \sum_i [\frac{y_i}{\sigma(w^T \mathbf{x}_i)} \sigma(w^T \mathbf{x}_i) \sigma(-w^T \mathbf{x}_i) x_j - \frac{(1-y_i)}{\sigma(-w^T \mathbf{x}_i)} \sigma(-w^T \mathbf{x}_i) \sigma(w^T \mathbf{x}_i) x_j] \\
&= -\frac{1}{N} \sum_i [y_i \sigma(-w^T \mathbf{x}_i) - (1 - y_i) \sigma(w^T \mathbf{x}_i)] x_j \\
&= \frac{1}{N} (\sigma(w^T \mathbf{x}_i) - y_i) x_i \\
&= \frac{1}{N} (\hat{y} - y_i) x_i
\end{aligned}
$$

线性模型

线性模型

- □ 线性评分函数：$s = \omega^T x$



linear classification

$h(\mathbf{x}) = \mathrm{sign}(s)$

plausible err = 0/1

linear regression

$h(\mathbf{x}) = s$

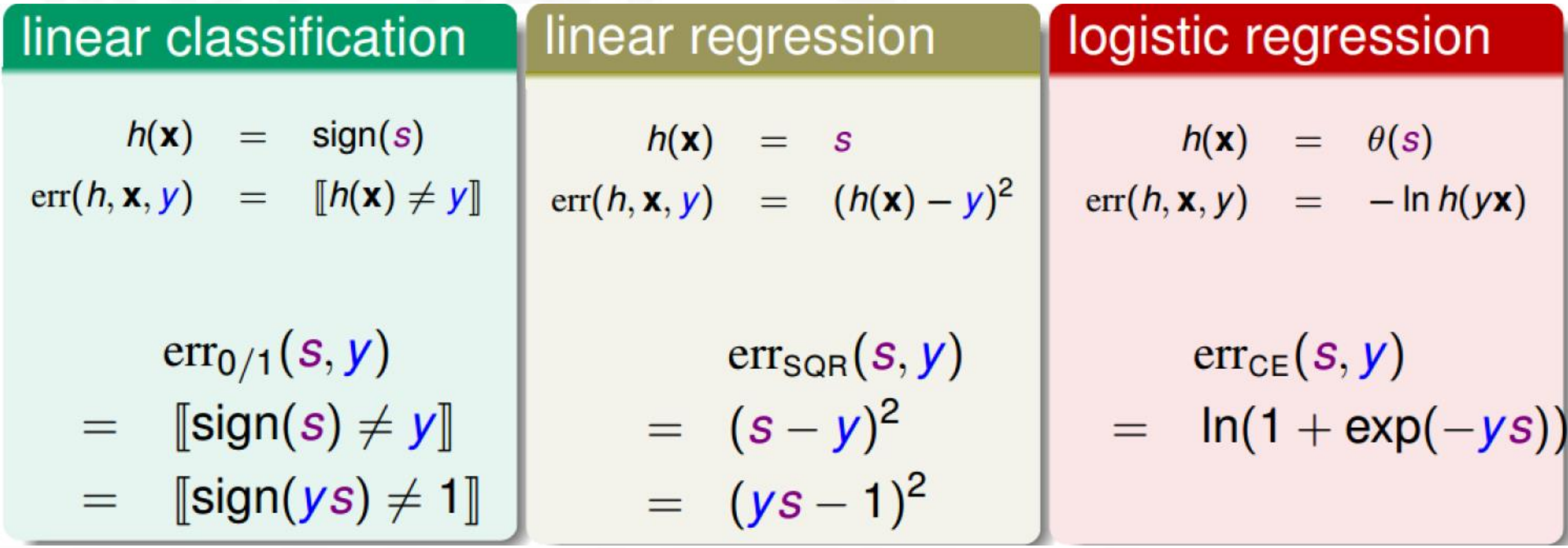
friendly err = squared

logistic regression

$h(\mathbf{x}) = \theta(s)$

plausible err = cross-entropy

线性模型



linear classification

$$h(\mathbf{x}) = \text{sign}(s)$$
$$\text{err}(h, \mathbf{x}, y) = [\![h(\mathbf{x}) \neq y]\!]$$

$$\begin{aligned} &\text{err}_{0/1}(s, y) \\ = \;& [\![\text{sign}(s) \neq y]\!] \\ = \;& [\![\text{sign}(ys) \neq 1]\!] \end{aligned}$$

linear regression

$$h(\mathbf{x}) = s$$
$$\text{err}(h, \mathbf{x}, y) = (h(\mathbf{x}) - y)^2$$

$$\begin{aligned} &\text{err}_{\text{SQR}}(s, y) \\ = \;& (s - y)^2 \\ = \;& (ys - 1)^2 \end{aligned}$$

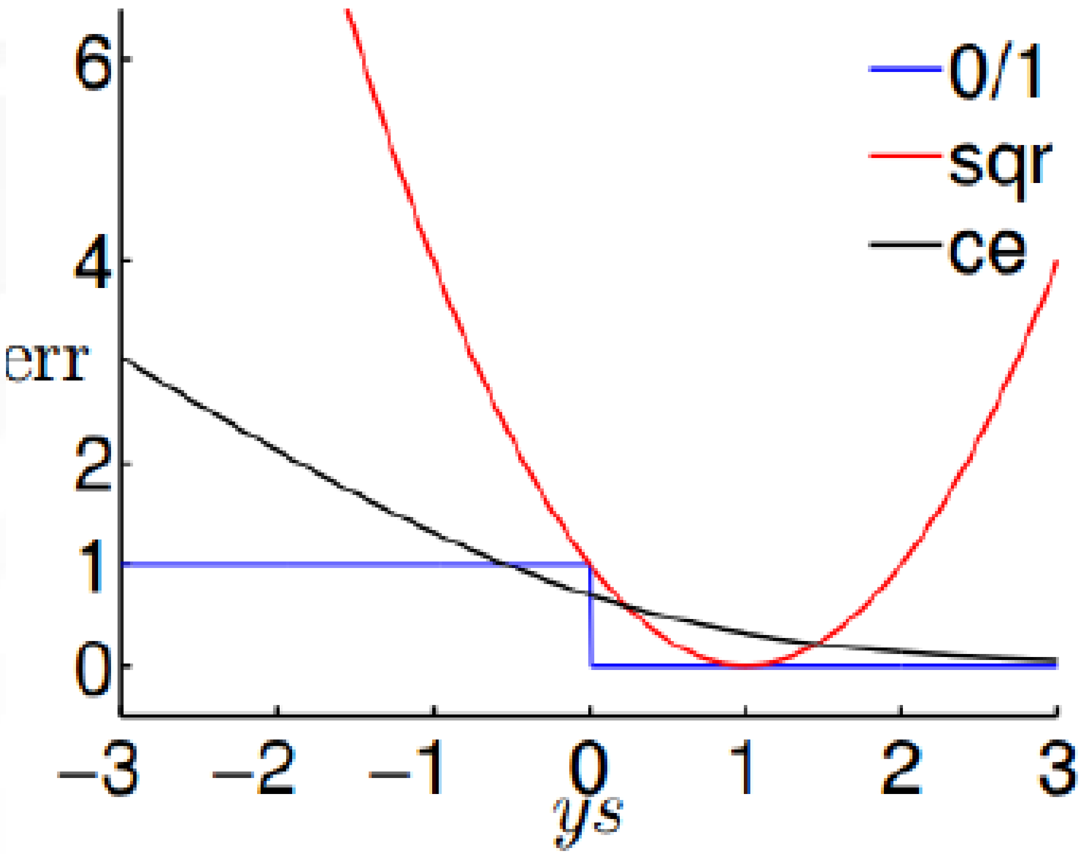
logistic regression

$$h(\mathbf{x}) = \theta(s)$$
$$\text{err}(h, \mathbf{x}, y) = -\ln h(y\mathbf{x})$$

$$\begin{aligned} &\text{err}_{\text{CE}}(s, y) \\ = \;& \ln(1 + \exp(-ys)) \end{aligned}$$

□ 分类的正确率得分：ys

线性模型



多分类问题

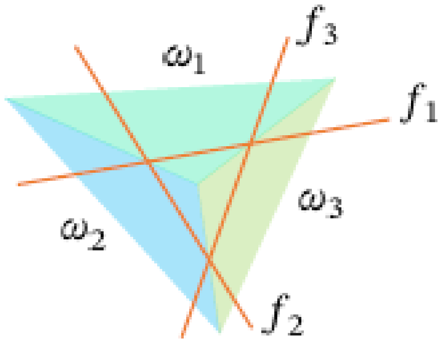
□ 一对多（One-Versus-All, OVA/One-Vs-Rest, OVR）

1. 对每个label $k \in Y$, 构造数据集

$$D_{[k]} = \{(x_n, y'_n = [\![y_n = k]\!])\}_{n=1}^{N}$$

2. 对每个数据集$D_{[k]}$，执行逻辑回归，得到$\omega_{[k]}$

3. 预测函数$g(x) = \underset{k \in Y}{\mathrm{argmax}}\left(\omega_{[k]}^T x\right)$



□ 数据不平衡的问题

多分类问题

□ 一对一（One-Versus-One，OVO,MVM的特例）

1. 对每个label对 $(k, l) \in Y \times Y$, 构造数据集

$$D_{[k,l]} = \{(x_n, y_n' = [\![y_n = k]\!]): y_n = k \; or \; y_n = l\}$$

2. 对每个数据集$D_{[k,l]}$，执行线性分类，得到$\omega_{[k,l]}$

3. 预测函数$g(x) = tournament \; champion\{\omega_{[k,l]}^T x\}$

□ 更多的训练，更慢的预测

多元逻辑回归

□ 假设是K元分类模型，即$\mathcal{Y}=\{1,2,\ldots,K\}$

根据二元逻辑回归的经验，我们有：

$$
\begin{aligned}
ln\frac{P(y=1|\mathbf{x})}{P(y=K|\mathbf{x})} &= w_1^T\mathbf{x} \\
ln\frac{P(y=2|\mathbf{x})}{P(y=K|\mathbf{x})} &= w_2^T\mathbf{x} \\
&\cdots \\
ln\frac{P(y=K-1|\mathbf{x})}{P(y=K|\mathbf{x})} &= w_{K-1}^T\mathbf{x}
\end{aligned}
$$

和

$$\sum_{i=1}^{K} P(y=i|\mathbf{x})=1$$

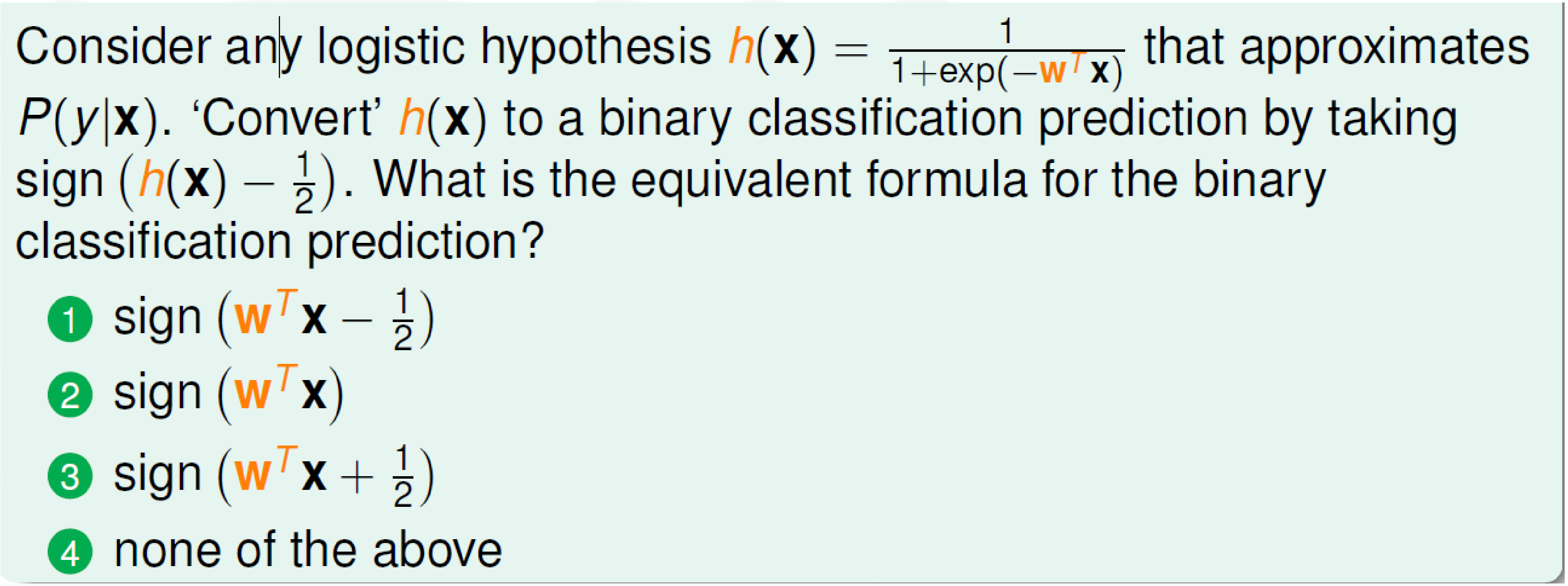
求解得到：

$$P(y=k|\mathbf{x})=\frac{e^{w_k^T\mathbf{x}}}{1+\sum_{t=1}^{K-1}e^{w_t^T\mathbf{x}}},\quad k=1,2,\ldots,K-1$$

$$P(y=K|\mathbf{x})=\frac{1}{1+\sum_{t=1}^{K-1}e^{w_t^T\mathbf{x}}}$$

练习题

题1



Consider any logistic hypothesis $h(\mathbf{x}) = \frac{1}{1+\exp(-\mathbf{w}^T\mathbf{x})}$ that approximates $P(y|\mathbf{x})$. 'Convert' $h(\mathbf{x})$ to a binary classification prediction by taking $\text{sign}\left(h(\mathbf{x}) - \frac{1}{2}\right)$. What is the equivalent formula for the binary classification prediction?

1. $\text{sign}\left(\mathbf{w}^T\mathbf{x} - \frac{1}{2}\right)$
2. $\text{sign}\left(\mathbf{w}^T\mathbf{x}\right)$
3. $\text{sign}\left(\mathbf{w}^T\mathbf{x} + \frac{1}{2}\right)$
4. none of the above

练习题

题2

The four statements below help us understand more about the cross-entropy error $\mathrm{err}(\mathbf{w}, \mathbf{x}, y) = \ln\left(1 + \exp(-y\mathbf{w}^T\mathbf{x})\right)$. Consider $\mathbf{w}^T\mathbf{x} \neq 0$. Which statement is not true?

1. For any $\mathbf{w}$, $\mathbf{x}$, and y, $\mathrm{err}(\mathbf{w}, \mathbf{x}, y) > 0$.
2. For any $\mathbf{w}$, $\mathbf{x}$, and y, $\mathrm{err}(\mathbf{w}, \mathbf{x}, y) < 1126$.
3. When $y = \mathrm{sign}\left(\mathbf{w}^T\mathbf{x}\right)$, $\mathrm{err}(\mathbf{w}, \mathbf{x}, y) < \ln 2$.
4. When $y \neq \mathrm{sign}\left(\mathbf{w}^T\mathbf{x}\right)$, $\mathrm{err}(\mathbf{w}, \mathbf{x}, y) \geq \ln 2$.

练习题

题3

Consider the gradient $\nabla E_{\text{in}}(\mathbf{w}) = \frac{1}{N}\sum_{n=1}^{N} \theta\left(-y_n\mathbf{w}^T\mathbf{x}_n\right)\left(-y_n\mathbf{x}_n\right)$. That is, each example $(\mathbf{x}_n, y_n)$ contributes to the gradient by an amount of $\theta\left(-y_n\mathbf{w}^T\mathbf{x}_n\right)$. For any given $\mathbf{w}$, which example contributes the most amount to the gradient?

1. the example with the smallest $y_n\mathbf{w}^T\mathbf{x}_n$ value
2. the example with the largest $y_n\mathbf{w}^T\mathbf{x}_n$ value
3. the example with the smallest $\mathbf{w}^T\mathbf{x}_n$ value
4. the example with the largest $\mathbf{w}^T\mathbf{x}_n$ value

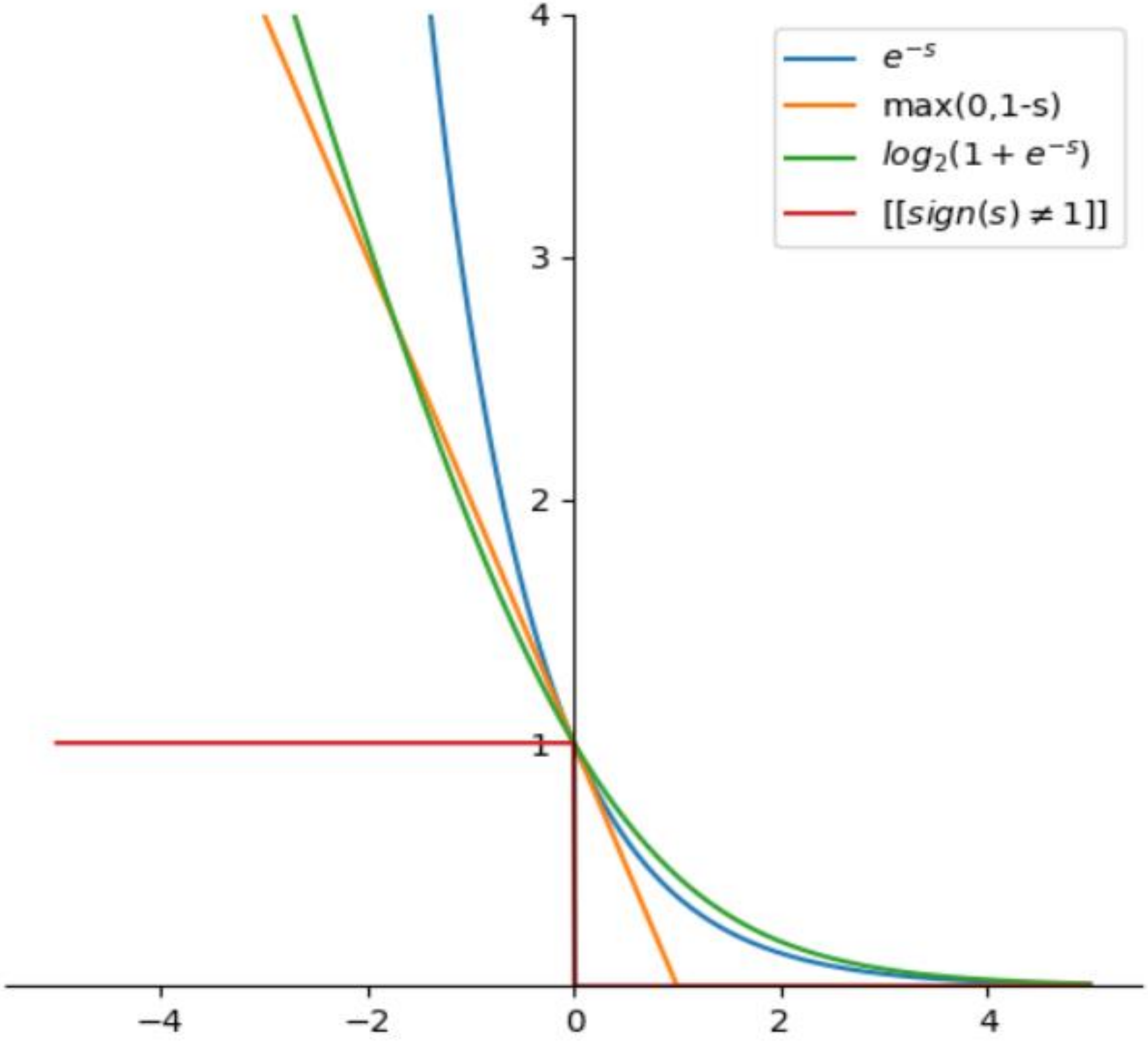
练习题

题4

Which of the following functions are upper bounds of the pointwise 0/1 error $[\![\text{sign}(\mathbf{w}^T\mathbf{x}) \neq y]\!]$ for $y \in \{-1, +1\}$?

1. $\exp(-y\mathbf{w}^T\mathbf{x})$
2. $\max(0, 1 - y\mathbf{w}^T\mathbf{x})$
3. $\log_2(1 + \exp(-y\mathbf{w}^T\mathbf{x}))$
4. all of the above

线性模型



THE END