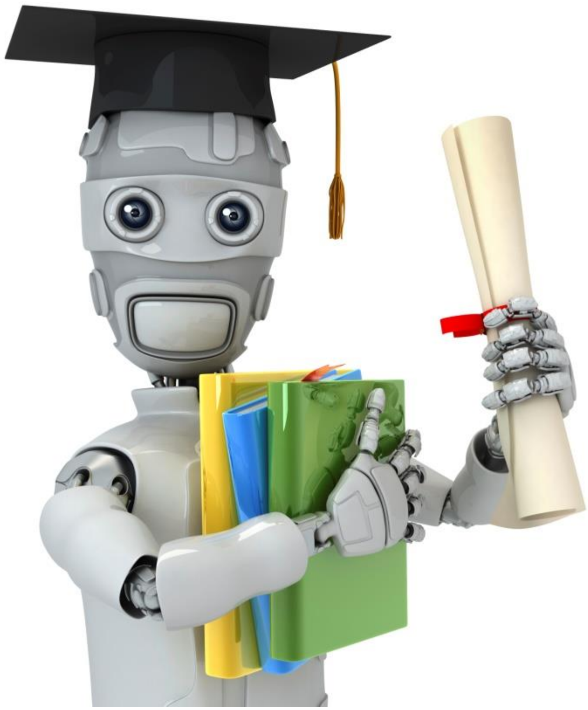


Classification: Logistic Regression

梁毅雄

yxliang@csu.edu.cn

Machine Learning

Some materials from Andrew Ng, Zico Kolter, Hung-yi Lee and others

分类

邮件: 垃圾邮件 / 非垃圾邮件?

网上交易: 欺骗(Yes / No)?

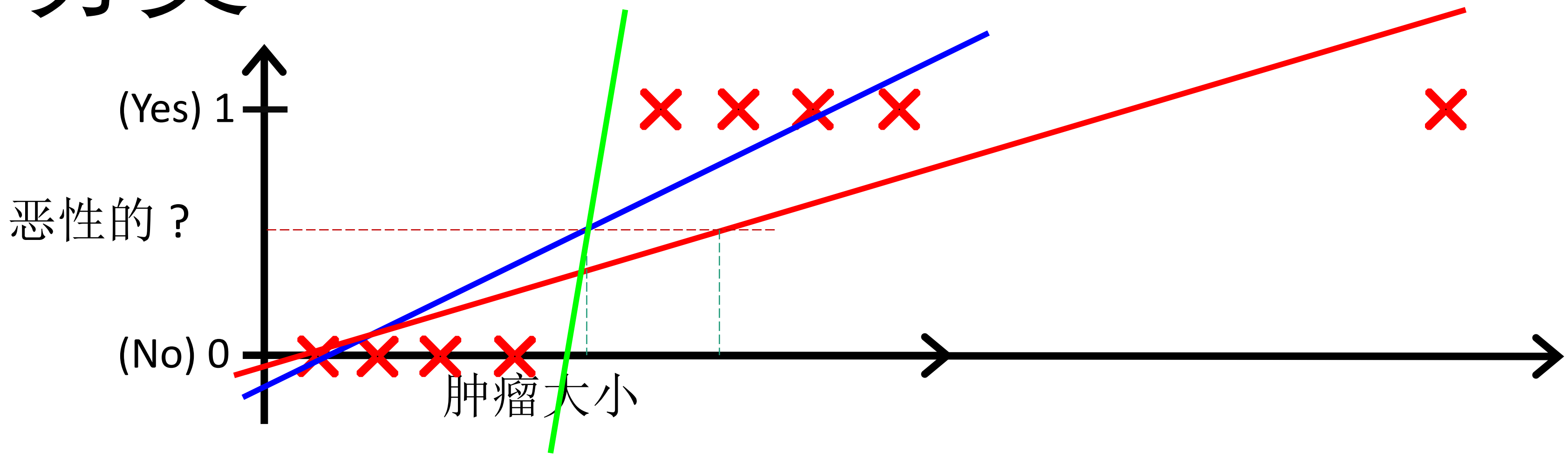
肿瘤: 良性的/ 恶性的 ?

$$y \in \{0, 1\}$$

0: “负类” (e.g., 良性的)
1: “正类” (e.g., 恶性的)

可以直接用上章介绍的回归方法直接进行分类?

分类



分类器输出阈值 $h_\theta(x)$为0.5

If $h_\theta(x) \geq 0.5$, 预测 “y = 1”

If $h_\theta(x) < 0.5$, 预测 “y = 0”

分类: $y = 0$ or 1

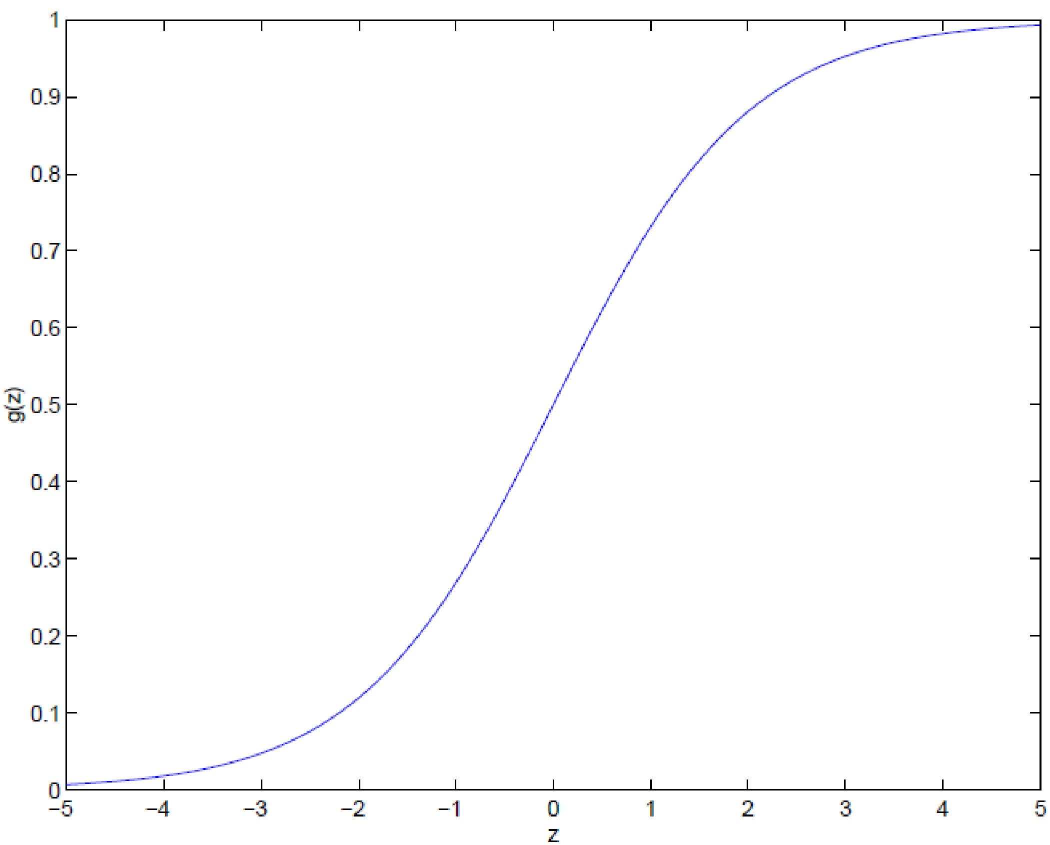
$h_\theta(x)$ 可能 > 1 或 < 0

Logistic回归: $0 \leq h_\theta(x) \leq 1$

Logistic Regression

目标 $0 \leq h_\theta(x) \leq 1$

$$h_\theta(x) \cancel{=} \theta^T x \qquad h_\theta(x) = g(\theta^T x) = \frac{1}{1+e^{-\theta^T x}}$$

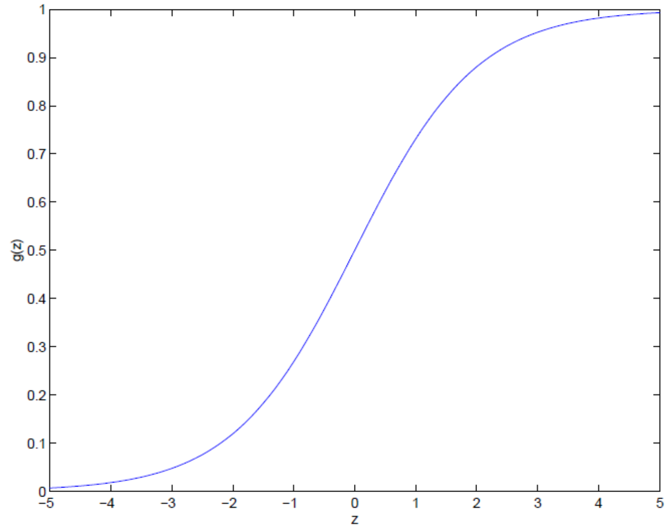


$$g(z) = \frac{1}{1+e^{-z}}$$

logistic 函数 or sigmoid函数

Sigmoid函数的性质

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



$$\begin{aligned} g'(z) &= \frac{d}{dz}\frac{1}{1+e^{-z}} \\ &= \frac{1}{(1+e^{-z})^2}\left(e^{-z}\right) \\ &= \frac{1}{(1+e^{-z})}\cdot\left(1-\frac{1}{(1+e^{-z})}\right) \\ &= g(z)(1-g(z)). \end{aligned}$$

概率解释

$$h_\theta(x) = g(\theta^T x) = \frac{1}{1 + e^{-\theta^T x}}$$

$h_\theta(x)$：对于输入*x*，输出*y* = 1的可能性

例子：如果 $x = \begin{bmatrix} x_0 \\ x_1 \end{bmatrix} = \begin{bmatrix} 1 \\ \text{tumorSize} \end{bmatrix}$

$$h_\theta(x) = 0.7$$

表示肿瘤是恶性的可能性为70%

"给出x，估计y=1的可能性,

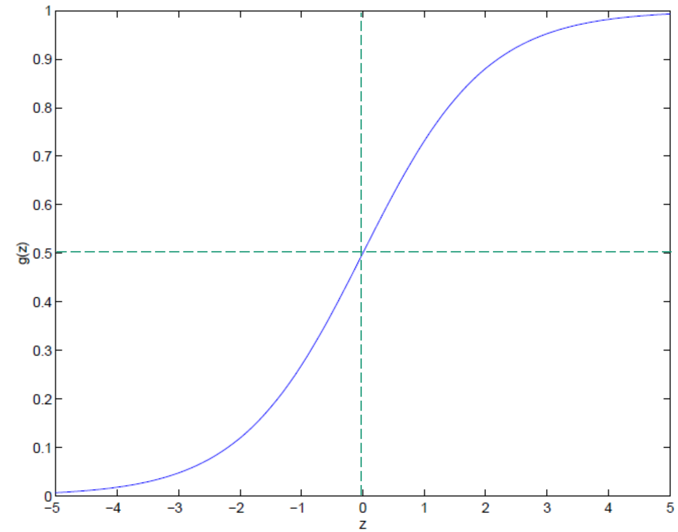
θ 为参数

$$P(y=0|x;\theta) + P(y=1|x;\theta) = 1$$
$$P(y=0|x;\theta) = 1 - P(y=1|x;\theta)$$

分类边界

$$h_\theta(x) = g(\theta^T x) = \frac{1}{1+e^{-\theta^T x}}$$

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



预测 “$y=1$“ 如果$h_\theta(x) \geq 0.5$

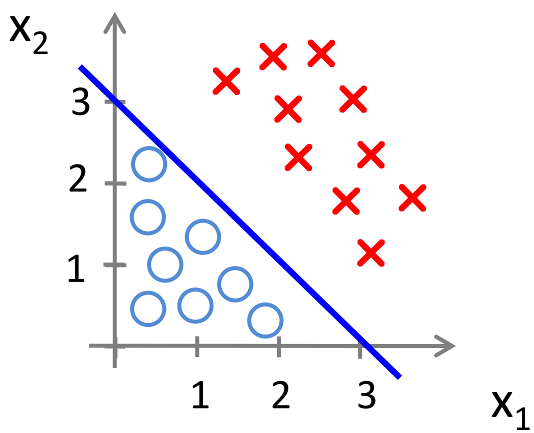
$$\theta^T x \geq 0$$

预测 “$y=0$“ 如果$h_\theta(x) < 0.5$

$$\theta^T x < 0$$

分类边界

线性分类边界

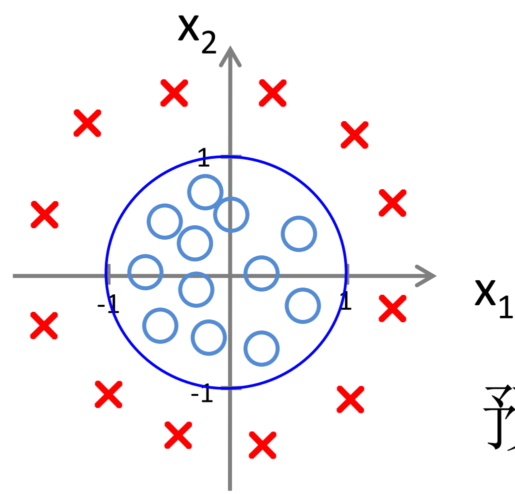


$$h_\theta(x) = g(\theta_0 + \theta_1 x_1 + \theta_2 x_2)$$

预测“$y = 1$“ 如果$-3 + x_1 + x_2 \geq 0$

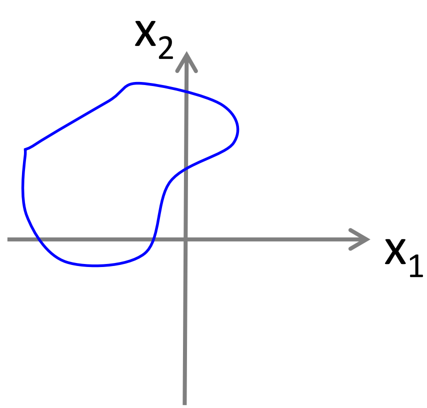
分类边界

非线性分类边界



$$h_\theta(x) = g(\theta_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_1^2 + \theta_4 x_2^2)$$

预测 “$y = 1$“ 如果$-1 + x_1^2 + x_2^2 \geq 0$



$$h_\theta(x) = g(\theta_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_1^2 + \theta_4 x_1^2 x_2 + \theta_5 x_1^2 x_2^2 + \theta_6 x_1^3 x_2 + \dots)$$

训练集: $\{(x^{(1)}, y^{(1)}), (x^{(2)}, y^{(2)}), \cdots, (x^{(m)}, y^{(m)})\}$

m 个样本, n维特征 $x \in \begin{bmatrix} x_0 \\ x_1 \\ \cdots \\ x_n \end{bmatrix}$ $x_0 = 1, y \in \{0, 1\}$

$$h_\theta(x) = \frac{1}{1 + e^{-\theta^T x}}$$

怎样选择参数 θ ?

需要选择损失函数!!!

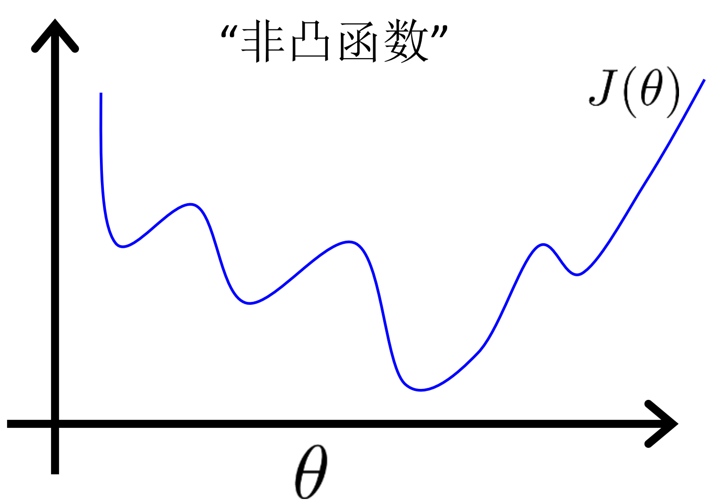
损失函数

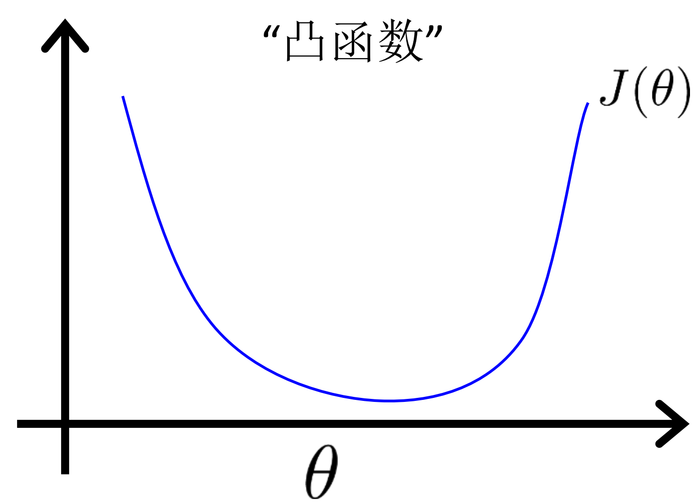
是否可以直接使用线性回归中的平方损失函数？

$$J(\theta) = \frac{1}{m}\sum_{i=1}^{m}\frac{1}{2}\left(h_\theta(x^{(i)}) - y^{(i)}\right)^2$$

$$\mathrm{Cost}(h_\theta(x^{(i)}), y^{(i)}) = \frac{1}{2}\left(h_\theta(x^{(i)}) - y^{(i)}\right)^2$$

$$h_\theta(x) = g(\theta^T x) = \frac{1}{1+e^{-\theta^T x}}$$





Aside: Convex Function

- 凸函数：若函数$f(x)$对任意的$t \in [0,1]$有

$$f(tx_1 + (1-t)x_2) \le tf(x_1) + (1-t)f(x_2),$$

  - 等价于$f''(x) \ge 0, \ \forall x.$
  - 若x为矢量，则对应的条件变为Hessian矩阵H 为半正定矩阵($H \ge 0$)

- Strictly convex: 若$\forall t \in (0,1), \forall x_1 \ne x_2$有

$$f(tx_1 + (1-t)x_2) < tf(x_1) + (1-t)f(x_2)$$

对于矢量，则对应的条件变为Hessian矩阵H正定。

Aside: Hessian matrix

Suppose $f(x) : \mathbb{R}^n \to \mathbb{R}$ is a function taking as input a vector $x \in \mathbb{R}^n$ and outputting a scalar $f(x) \in \mathbb{R}$, if all second partial derivatives of f exist and are continuous over the domain of the function, then the Hessian matrix H of f is a square $n \times n$ matrix, usually defined and arranged as follows:

$$H = \begin{bmatrix} \dfrac{\partial^2 f}{\partial x_1^2} & \dfrac{\partial^2 f}{\partial x_1\,\partial x_2} & \cdots & \dfrac{\partial^2 f}{\partial x_1\,\partial x_n} \\ \dfrac{\partial^2 f}{\partial x_2\,\partial x_1} & \dfrac{\partial^2 f}{\partial x_2^2} & \cdots & \dfrac{\partial^2 f}{\partial x_2\,\partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \dfrac{\partial^2 f}{\partial x_n\,\partial x_1} & \dfrac{\partial^2 f}{\partial x_n\,\partial x_2} & \cdots & \dfrac{\partial^2 f}{\partial x_n^2} \end{bmatrix}.$$

或者

$$H_{i,j} = \frac{\partial^2 f}{\partial x_i \partial x_j}$$

https://en.wikipedia.org/wiki/Hessian_matrix

损失函数

0-1损失函数？

$$\ell\left(h_\theta(x), y\right) = \mathbf{1}(h_\theta(x) \neq y) = \begin{cases} 1, & \text{if } h_\theta(x) \neq y \\ 0, & \text{otherwise} \end{cases}$$

如何找到合适的损失函数以替代0-1损失？

$$\begin{aligned} P(y = 1 \mid x; \theta) &= h_\theta(x) \\ P(y = 0 \mid x; \theta) &= 1 - h_\theta(x) \end{aligned}$$

$$p(y \mid x; \theta) = (h_\theta(x))^y \left(1 - h_\theta(x)\right)^{1-y}$$

$$\begin{aligned} L(\theta) &= p(y \mid X; \theta) \\ &= \prod_{i=1}^{m} p(y^{(i)} \mid x^{(i)}; \theta) \\ &= \prod_{i=1}^{m} \left(h_\theta(x^{(i)})\right)^{y^{(i)}} \left(1 - h_\theta(x^{(i)})\right)^{1-y^{(i)}} \end{aligned}$$

估计参数：最大似然估计

损失函数

$$\begin{aligned} P(y=1 \mid x;\theta) &= h_\theta(x) \\ P(y=0 \mid x;\theta) &= 1-h_\theta(x) \end{aligned}$$

$$p(y \mid x;\theta) = (h_\theta(x))^y \, (1-h_\theta(x))^{1-y}$$

$$\begin{aligned} L(\theta) &= p(y \mid X;\theta) \\ &= \prod_{i=1}^{m} p(y^{(i)} \mid x^{(i)};\theta) \\ &= \prod_{i=1}^{m} \left(h_\theta(x^{(i)})\right)^{y^{(i)}} \left(1-h_\theta(x^{(i)})\right)^{1-y^{(i)}} \end{aligned}$$

Logistic损失函数

$$\begin{aligned} \ell(\theta) &= -\log L(\theta) \\ &= -\left[\sum_{i=1}^{m} y^{(i)} \log h_\theta(x^{(i)}) + (1-y^{(i)}) \log(1-h_\theta(x^{(i)}))\right] \end{aligned}$$

Cross entropy

https://en.wikipedia.org/wiki/Cross_entropy

In information theory, the cross entropy between two probability distributions p and q over the same underlying set of events measures the average number of bits needed to identify an event drawn from the set, if a coding scheme is used that is optimized for an "unnatural" probability distribution q, rather than the "true" distribution p.

The cross entropy for the distributions p and q over a given set is defined as follows:

$$H(p,q) = E_p[-\log q] = H(p) + D_{\mathrm{KL}}(p\|q),$$

where $H(p)$ is the entropy of p, and $D_{\mathrm{KL}}(p\|q)$ is the Kullback–Leibler divergence of q from p (also known as the relative entropy of p with respect to q — note the reversal of emphasis).

For discrete p and q this means

$$H(p,q) = -\sum_x p(x)\,\log q(x).$$

损失函数

cross entropy

$$H(p,q) = -\sum_x p(x) ln\big(q(x)\big)$$

Given two Bernoulli distribution

Distribution p :
$p(x=1) = y$
$p(x=0) = 1-y$

cross entropy

Distribution q :
$q(x=1) = h(x)$
$q(x=0) = 1-h(x)$

Logistic损失函数(Cross entropy**)**

$$\begin{aligned}\ell(\theta) &= -\log L(\theta) \\ &= -\left[\sum_{i=1}^{m} y^{(i)} \log h_\theta(x^{(i)}) + (1-y^{(i)}) \log(1-h_\theta(x^{(i)}))\right]\end{aligned}$$

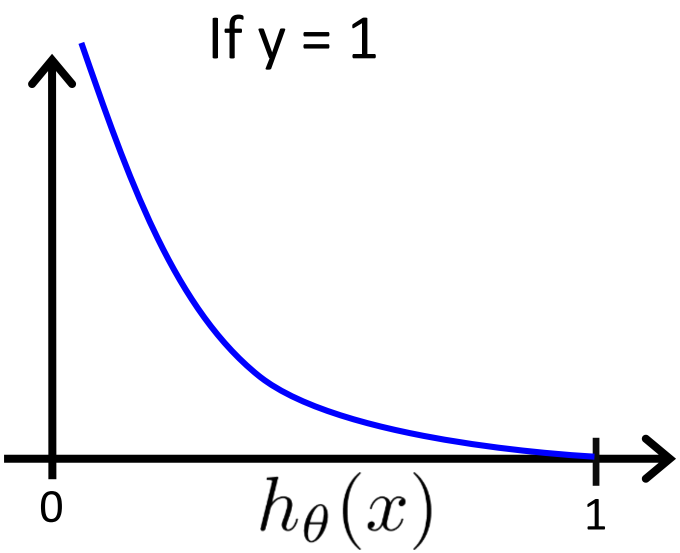
Cross entropy between two Bernoulli distribution

损失函数

Logistic损失函数(Cross entropy**)**

$$
\begin{aligned}
\ell(\theta) &= -\log L(\theta) \\
&= -\left[\sum_{i=1}^{m} y^{(i)} \log h_\theta(x^{(i)}) + (1-y^{(i)}) \log(1-h_\theta(x^{(i)}))\right]
\end{aligned}
$$

$$
\mathrm{Cost}(h_\theta(x), y) = \begin{cases} -\log(h_\theta(x)) & \text{if } y = 1 \\ -\log(1-h_\theta(x)) & \text{if } y = 0 \end{cases}
$$



$\mathrm{Cost} = 0$ if $y = 1, h_\theta(x) = 1$

But as $h_\theta(x) \to 0$

$Cost \to \infty$

Captures intuition that if $h_\theta(x) = 0$, (predict $P(y=1|x;\theta) = 0$), but $y = 1$, we'll penalize learning algorithm by a very large cost.

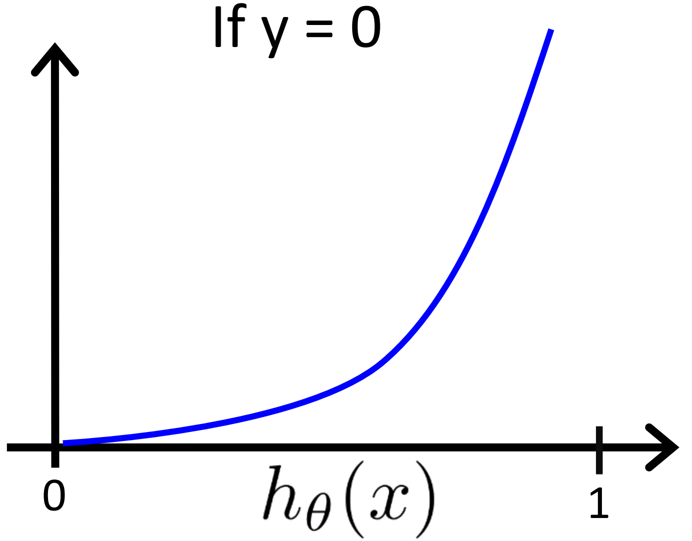
损失函数

Logistic损失函数(Cross entropy**)**

$$
\begin{aligned}
\ell(\theta) &= -\log L(\theta) \\
&= -\left[\sum_{i=1}^{m} y^{(i)} \log h_\theta(x^{(i)}) + (1 - y^{(i)}) \log(1 - h_\theta(x^{(i)}))\right]
\end{aligned}
$$

$$
\mathrm{Cost}(h_\theta(x), y) = \begin{cases} -\log(h_\theta(x)) & \text{if } y = 1 \\ -\log(1 - h_\theta(x)) & \text{if } y = 0 \end{cases}
$$

If y = 0



梯度下降

$$J(\theta) = -\left[\sum_{i=1}^{m} y^{(i)} \log h_\theta(x^{(i)}) + (1 - y^{(i)}) \log(1 - h_\theta(x^{(i)}))\right]$$

为了找到合适的参数θ:

$$\min_\theta J(\theta)$$

梯度下降

repeat $\{$

$$\theta_j := \theta_j - \alpha \frac{\partial}{\partial \theta_j} J(\theta)$$

(simultaneously update all θ_j)

$\}$

对于新给的x 给出一个预测 :

输出 $h_\theta(x) = \frac{1}{1+e^{-\theta^T x}}$

$$J(\theta) = -\left[\sum_{i=1}^{m} y^{(i)} \log h_\theta(x^{(i)}) + (1-y^{(i)}) \log(1-h_\theta(x^{(i)}))\right]$$

关键在于如何求梯度，推导如下（为了方便省略求和与上标i）

$$\begin{aligned}
\frac{\partial}{\partial\theta_j}J(\theta) &= -\left(y\frac{1}{g(\theta^T x)} - (1-y)\frac{1}{1-g(\theta^T x)}\right)\frac{\partial}{\partial\theta_j}g(\theta^T x) \\
&= -\left(y\frac{1}{g(\theta^T x)} - (1-y)\frac{1}{1-g(\theta^T x)}\right)g(\theta^T x)(1-g(\theta^T x)\frac{\partial}{\partial\theta_j}\theta^T x \\
&= -\left(y(1-g(\theta^T x)) - (1-y)g(\theta^T x)\right)x_j \\
&= (h_\theta(x) - y)\,x_j
\end{aligned}$$

repeat $\{$

$$\theta_j := \theta_j - \alpha \sum_{i=1}^{m} (h_\theta(x^{(i)}) - y^{(i)})x_j^{(i)}$$

(同步更新所有的 θ_j)

$\}$

看起来与线性回归算法是相同的，是否真正相同？？？

是否可以直接使用线性回归中的平方损失函数？

$$J(\theta) = \frac{1}{m}\sum_{i=1}^{m}\frac{1}{2}\left(h_\theta(x^{(i)}) - y^{(i)}\right)^2$$

$$h_\theta(x) = g(\theta^T x) = \frac{1}{1+e^{-\theta^T x}} \qquad g'(z) = g(z)(1-g(z)).$$

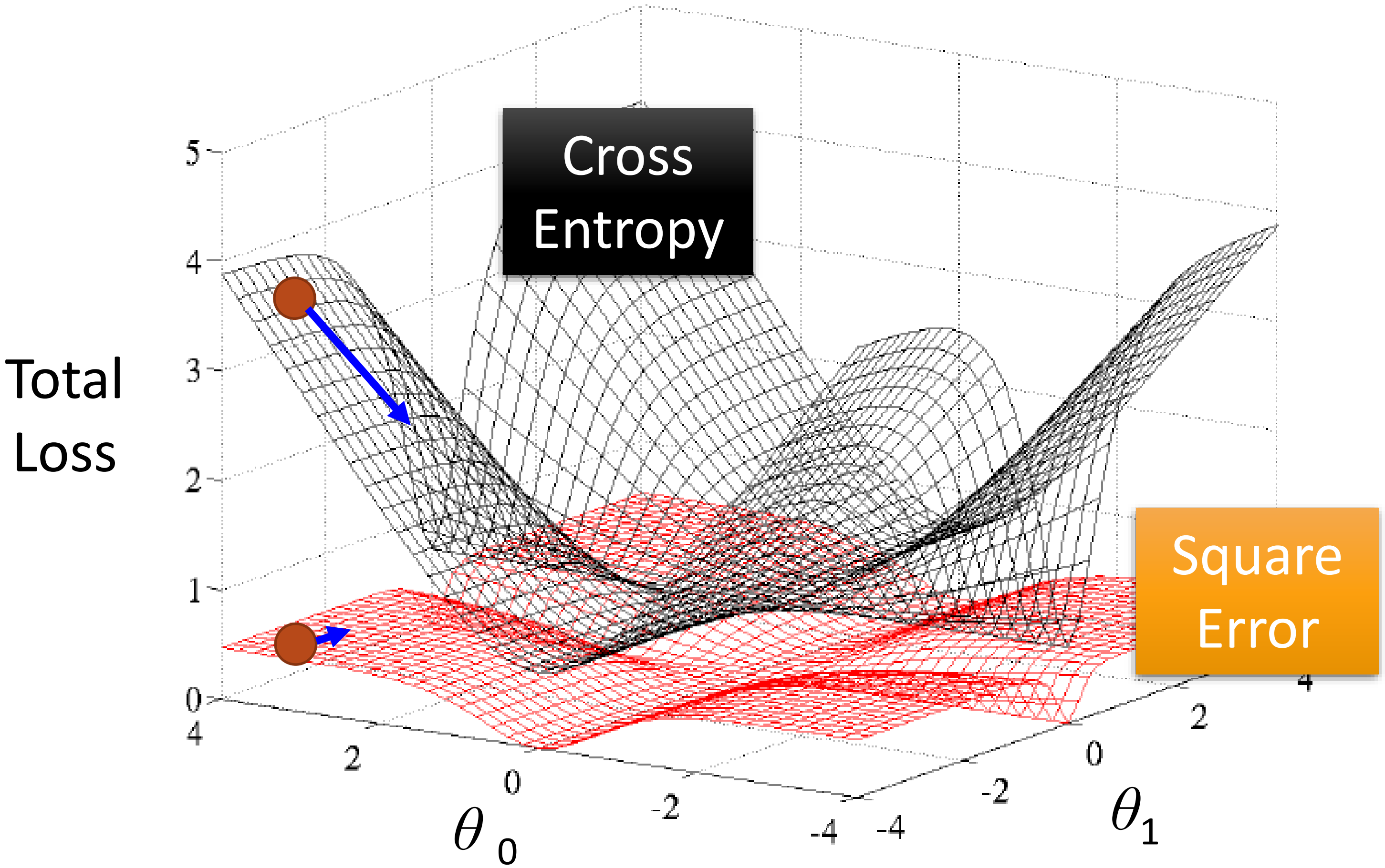
为了方便同样省略求和、系数与上标i

$$\begin{aligned}\frac{\partial}{\partial\theta_j}J(\theta) &= \left(g(\theta^T x) - y\right)\frac{\partial}{\partial\theta_j}g(\theta^T x) \\ &= \left(g(\theta^T x) - y\right)g(\theta^T x)\left(1 - g(\theta^T x)\right)\frac{\partial}{\partial\theta_j}\theta^T x \\ &= \left(g(\theta^T x) - y\right)g(\theta^T x)\left(1 - g(\theta^T x)\right)x_j\end{aligned}$$

假设$y = 0$, If $h_\theta(x) = 1$ (far from target) $\Rightarrow$ $\partial J/\partial\theta_i = 0$

If $h_\theta(x) = 0$ (close to target) $\Rightarrow$ $\partial J/\partial\theta_i = 0$

Cross Entropy vs. Square Error



http://jmlr.org/proceedings/papers/v9/glorot10a/glorot10a.pdf

Multi-class Classification

邮件标签: 工作, 朋友, 家人, 同伴

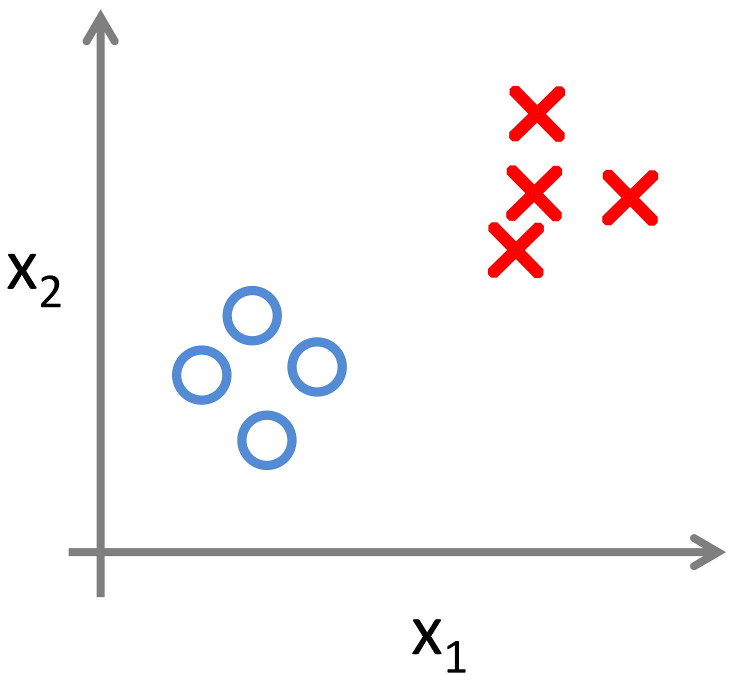
医学诊断: 没有生病, 感冒, 流感

天气: 晴天, 多云, 有雨, 下雪

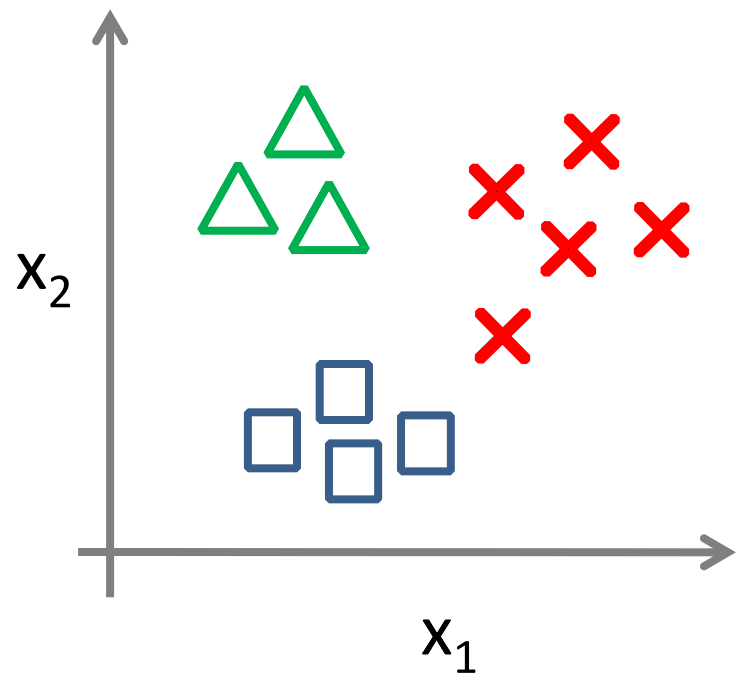
手写数字识别? $y = \{0, 1, \cdots, 9\}$

Multi-class Classification

二分类:

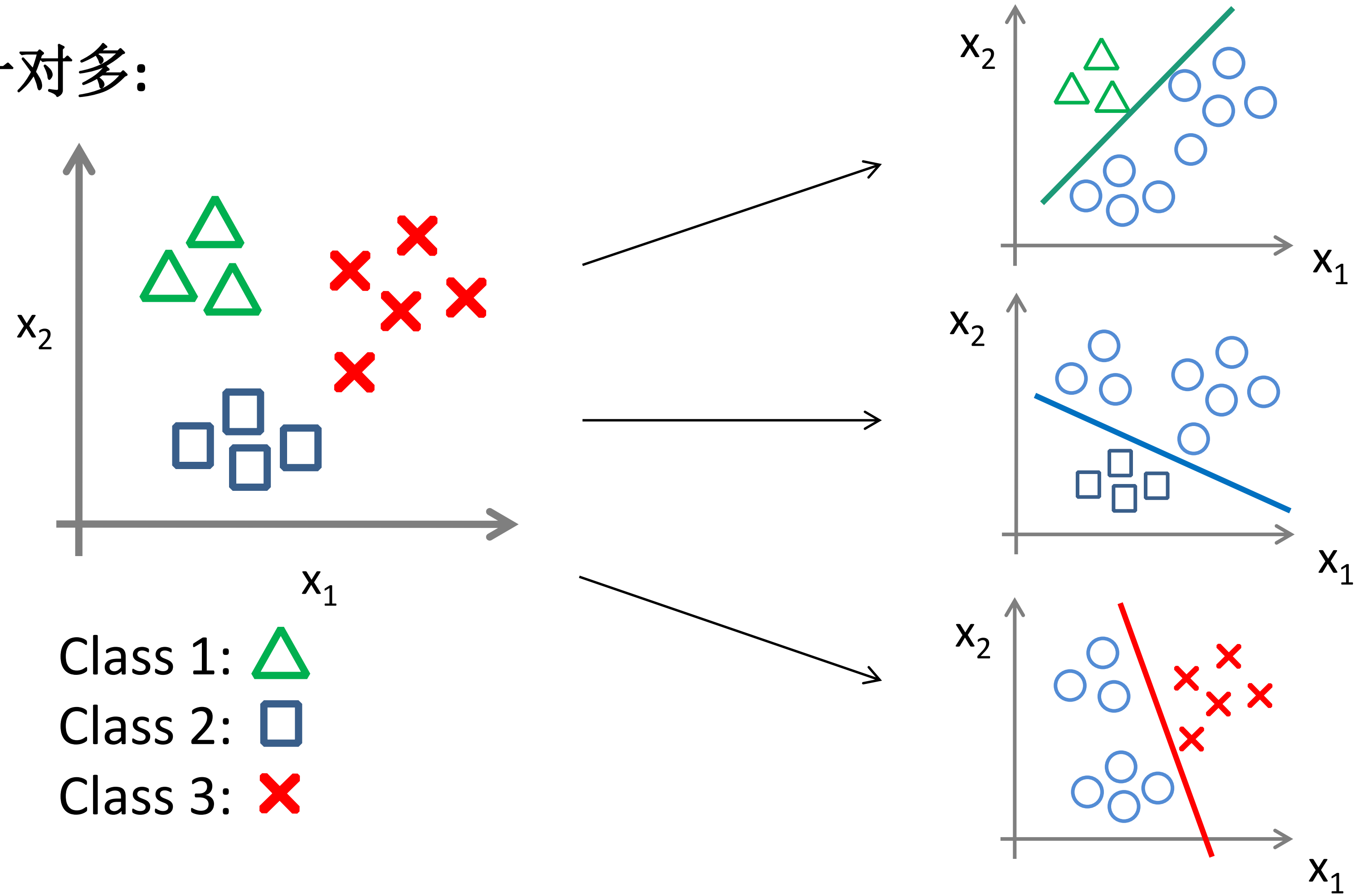


多类分类:



Multi-class Classification

一对多:



$$h_\theta^{(i)}(x) = P(y = i | x; \theta) \qquad (i = 1, 2, 3)$$

Multi-class Classification

一对多

为每类训练一个逻辑回归分类器 $h_{\theta^i}(x)$ 用来预测 $y=i$ 的可能性 .

对于一个新输入x , 做一个预测, 选择一个类别 i 使得:

$$\max_i h_{\theta^i}(x)$$

Softmax Regression

C$_1$: θ^1 $\quad z_1 = \theta^1 \cdot x$

C$_2$: θ^2 $\quad z_2 = \theta^2 \cdot x$

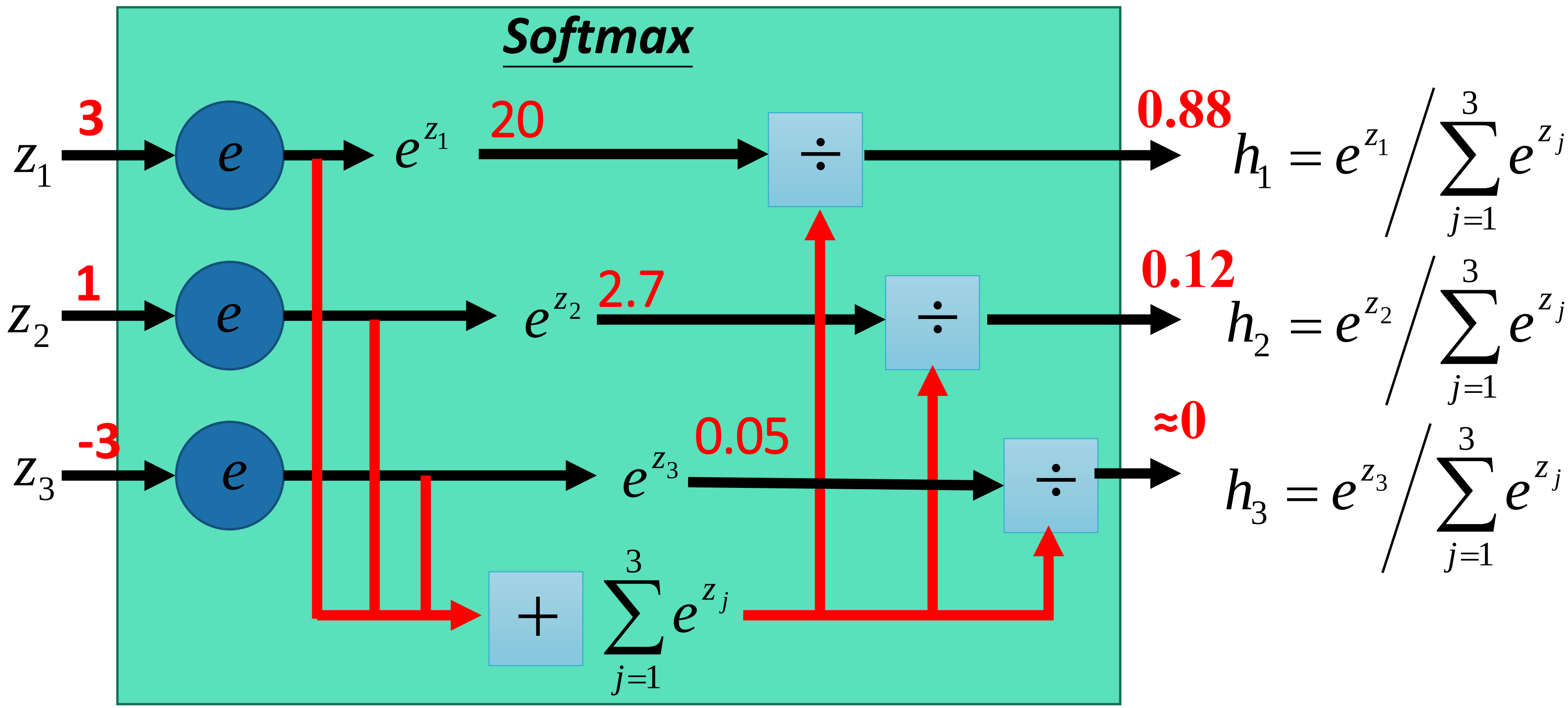
C$_3$: θ^3 $\quad z_3 = \theta^3 \cdot x$

(3 classes as example)

Probability:

- $1 > y_i > 0$
- $\sum_i y_i = 1$

$$p(y = i | x; \theta) = h_\theta^i(x) = \frac{e^{(\theta^i)^T x}}{\sum_{j=1}^{3} e^{(\theta^j)^T x}}$$



Softmax Loss

$$p(y=i|x;\theta) = h_\theta^i(x) = \frac{e^{z_i}}{\sum_{j=1}^K e^{z_j}} \qquad z_j = (\theta^j)^T x$$

- 对数似然为 $$\begin{aligned} L(\theta) &= \sum_{i=1}^m \log p(y^{(i)}|x^{(i)};\theta) \\ &= \sum_{i=1}^m \log\left(\frac{e^{z_{y^{(i)}}}}{\sum_{j=1}^K e^{z_j}}\right) \end{aligned}$$

- 总损失为 $$\begin{aligned} \ell(\theta) = -L(\theta) &= -\sum_{i=1}^m \log\left(\frac{e^{z_{y^{(i)}}}}{\sum_{j=1}^K e^{z_j}}\right) \\ &= \sum_{i=1}^m \left[\log\left(\sum_{j=1}^K e^{z_j}\right) - z_{y^{(i)}}\right] \end{aligned}$$

Softmax Loss

$$p(y=i|x;\theta) = h_\theta^i(x) = \frac{e^{z_i}}{\sum_{j=1}^K e^{z_j}} \qquad z_j = (\theta^j)^T x$$

$$\ell_i = \log\left(\sum_{j=1}^K e^{z_j}\right) - z_{y(i)}$$

单个训练样本对应的损失

$(x^{(i)}, y^{(i)}),\ i \in [1, \cdots, m]$

$$\ell(\theta) = \sum_{i=1}^m \left[\log\left(\sum_{j=1}^K e^{z_j}\right) - z_{y(i)}\right]$$

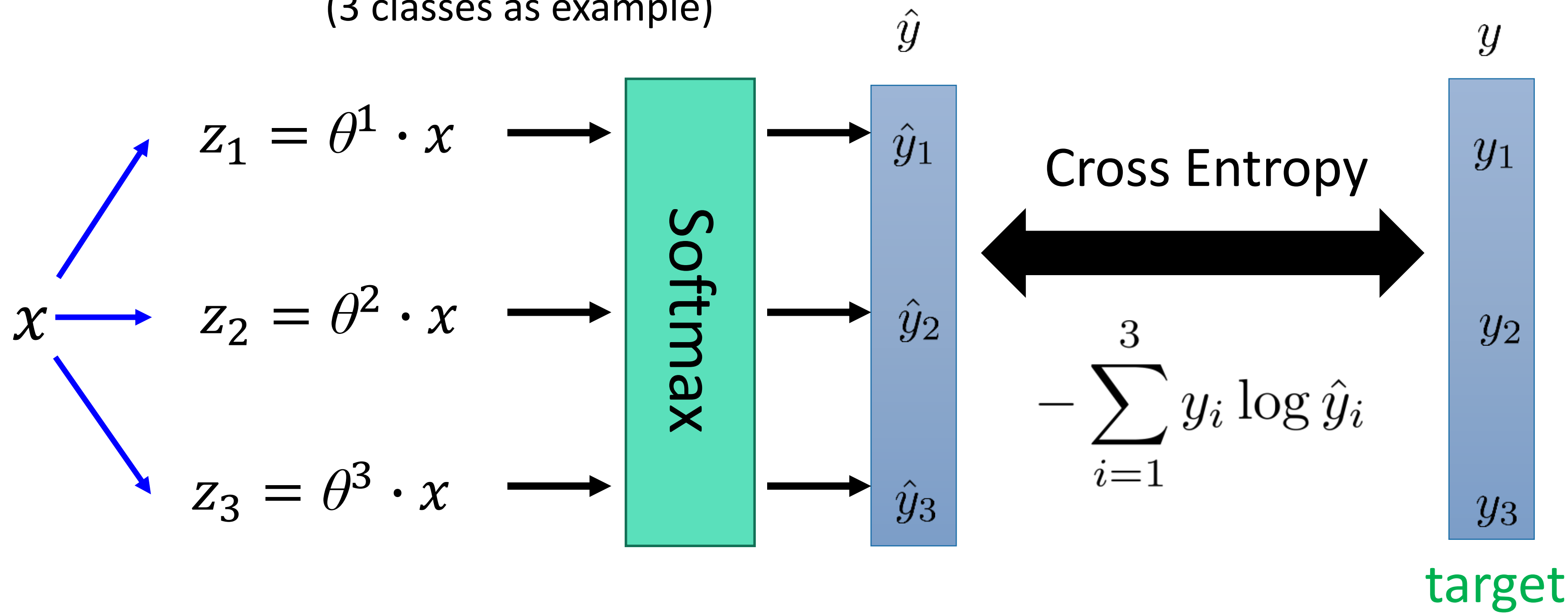
思考:

1. ℓ_i 的取值范围是多少?
2. 初始化每类的参数 $\theta^l \approx 0, \ell_i =$?

Softmax Regression

$$\hat{y}_i = h_\theta^i(x)$$

(3 classes as example)



If $x \in$ class 1

$$y = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}$$

$$-\log \hat{y}_1$$

If $x \in$ class 2

$$y = \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}$$

$$-\log \hat{y}_2$$

If $x \in$ class 3

$$y = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}$$

$$-\log \hat{y}_3$$

Thanks!

Any questions?