上海交通大學 SHANGHAI JIAO TONG UNIVERSITY

学生创新中心 Student Innovation Center

分类任务——逻辑回归 Logistic Regression

学生创新中心：肖雄子彦

SJTU

02

逻辑回归
Logistic Regression

学习目标:

•**掌握**逻辑回归**概念、作用、建模方法**

•**能够运用最大似然估计，推导**交叉熵公式、**熟悉**原理

•**运用Numpy基础知识**完成逻辑回归案例**实践**

Logistic Regression

逻辑回归用来解决二分类问题 “是或不是”，“有或无”，“通过或拒绝”…

线性回归?

$$y = w_i x_i + b$$

x_1

x_2

⋮

x_n

分类器

f( )

该不该借钱给他?

是不是垃圾邮件?

要不要发offer?

input依然是样本特征

output（target）是0或1

Logistic Regression

线性回归可以做分类任务吗?

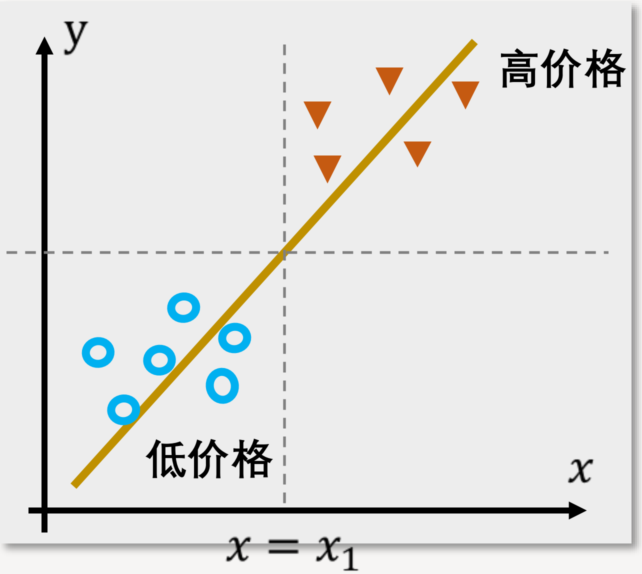
线性回归解决分类问题,效果非常不稳定

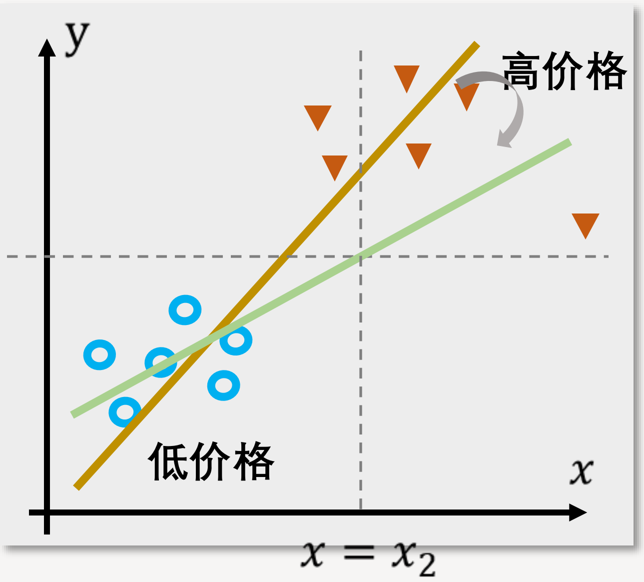
高价格 ▼

低价格 O

Linear Regression

$$y = wx + b$$





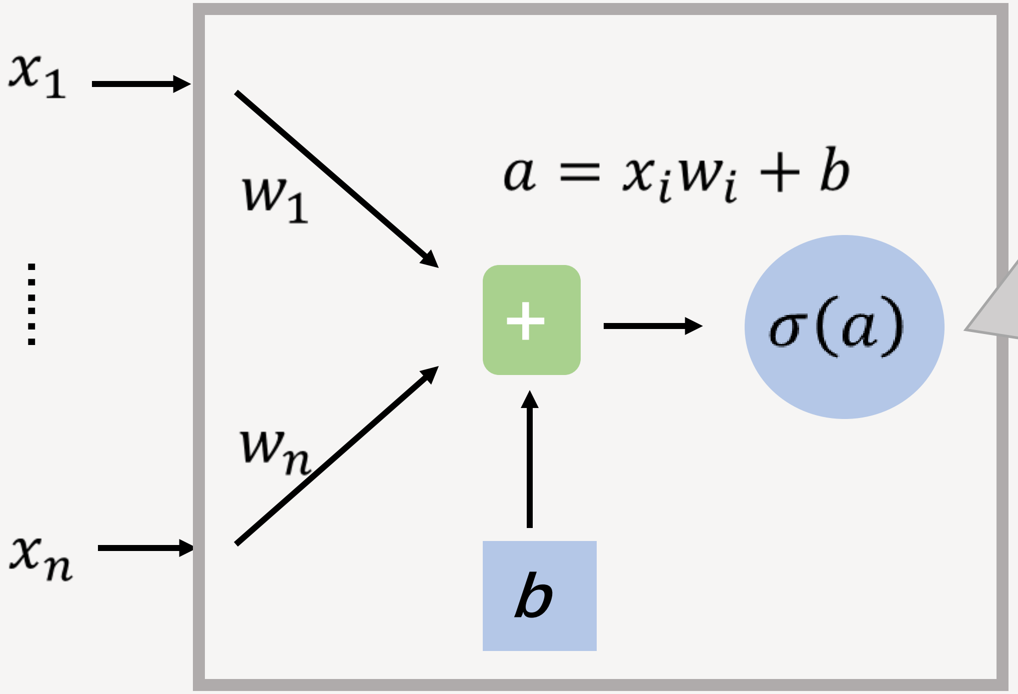
用线性回归模型做，为了减少误差，会得到绿色线

反思 线性回归 $\boldsymbol{y = wx + b}$ 值域为$(-\infty, +\infty)$ 如果是二分类问题，希望得到（0，1）概率

Logistic Regression

把线性回归的值映射为概率, 即把实数空间[-∞,+∞]的输出映射到(0,1)

逻辑回归实质上是求概率，p>0.5 即正样本



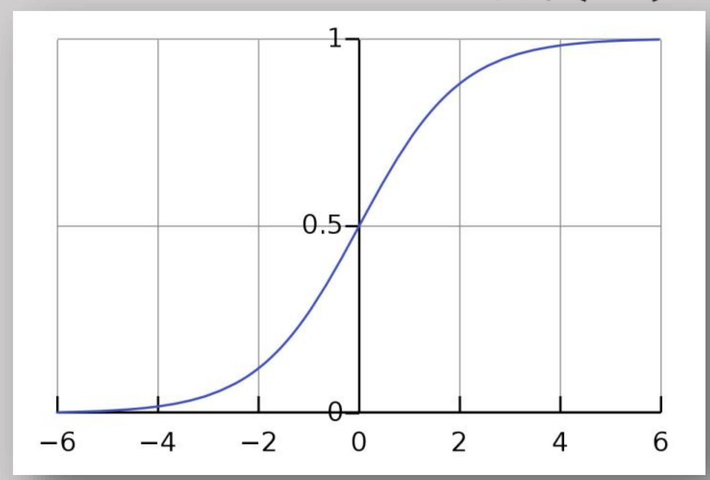
$$\boldsymbol{p(y|x) = f_{w,b}(x) = \sigma(wx + b)}$$

值域为(0,1)

Sigmoid

函数

$$\frac{1}{1+e^{-a}}$$



定义域为(−∞, +∞)

逻辑函数

逻辑回归的sigmoid函数（logit函数）是怎么来的呢?

我们先来复习下这几个概念

- 概率probability：指的是发生的次数/总次数

eg. 抛硬币 $\quad \boldsymbol{p} = \dfrac{\text{正面向上次数}}{\text{总次数}} \quad \boldsymbol{p} \in (\mathbf{0}, \mathbf{1})$

- odds：发生的次数（概率）/没有发生的次数（概率）

$$\boldsymbol{odds} = \frac{\text{正面向上次数}}{\text{反面向上次数}}$$

- 伯努利分布：如果X是伯努利分布中的随机变量, X取值为{1,0}, 如抛硬币的正反面

$$\left.\begin{aligned} P(X=1) &= p \\ P(X=0) &= 1-p \end{aligned}\right\} \quad \boldsymbol{odds} = \frac{\boldsymbol{p}}{\mathbf{1}-\boldsymbol{p}} \qquad \boldsymbol{odds} \in [\mathbf{0}, +\infty)$$

逻辑函数

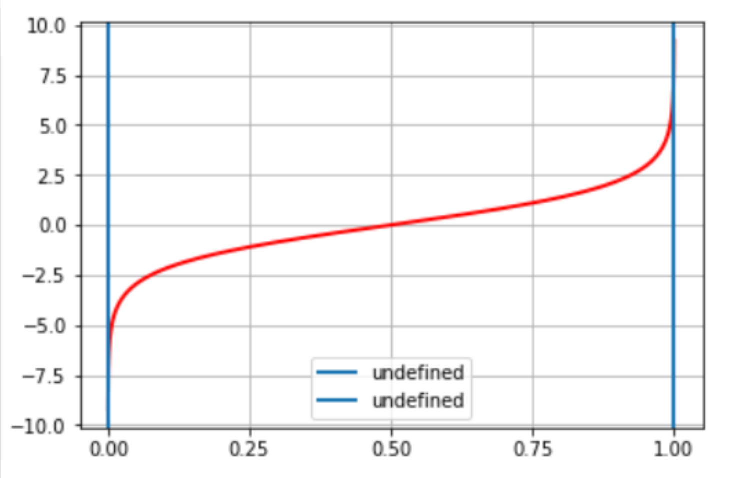
$$\boldsymbol{odds} \in [\mathbf{0}, +\infty)$$

- 我们对$odds$取log, 将它的取值范围扩展到实数空间[-∞,+∞]。这就是$logit$函数:

$$logit(p) = \log_e(odds) = \log_e(\frac{p}{1-p}) \qquad \boldsymbol{p} \in (\mathbf{0}, \mathbf{1}) \qquad \boldsymbol{logit}(\boldsymbol{p}) \in [-\infty, +\infty]$$

- 接着, 我们用线性回归模型来表示$logit(p)$, 因为线性回归模型和$logit$函数的输出有着同样的取值范围，都是 [-∞,+∞]

$$\boldsymbol{logit}(p) = \theta_1 x_1 + \theta_2 x_2 + bias$$

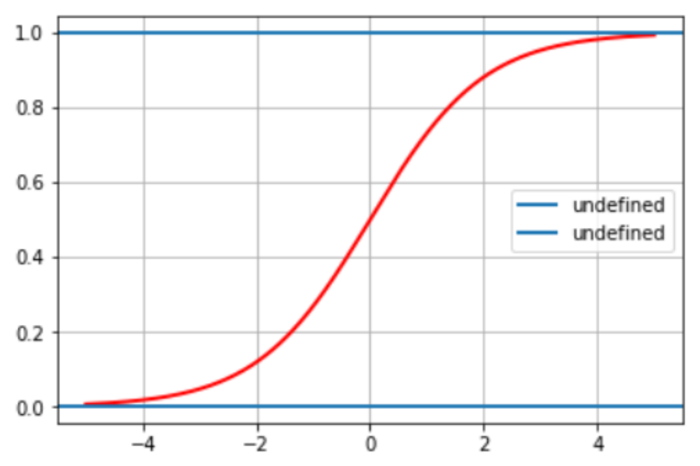


$$\log(\frac{p}{1-p}) = \boxed{\theta_1 x_1 + \theta_2 x_2 + bias} \quad \boldsymbol{z}$$

$$\frac{p}{1-p} = e^z$$

$$p = e^z - e^z p$$

$$\boxed{p = \frac{1}{1+e^{-z}}}$$



Logistic Regression

逻辑回归解决的是一个二分类问题 $\boldsymbol{p(y|x) = f_{w,b}(x) = \sigma(wx+b)}$ $\boldsymbol{y \in \{1,0\}}$

$$\begin{cases} p(y=1|x) = \dfrac{1}{1+e^{-w^Tx+b}} \\ p(y=0|x) = 1 - p(y=1|x) = \dfrac{e^{-w^Tx+b}}{1+e^{-w^Tx+b}} \end{cases}$$

Max

$$p(y|x) = p(y=1|x)^y [1-p(y=1|x)]^{1-y} \longrightarrow \boxed{\text{目标函数}}$$

If $y=1$： $p(y|x) = p(y=1|x)$

If $y=0$： $p(y|x) = 1-p(y=1|x)$

Maximum likelihood

极大似然估计

要理解极大似然估计，首先要明白什么是概率密度函数

（某个随机变量取某个值的时候，所对应的的概率）

现在有一个概率分布，属于正态分布

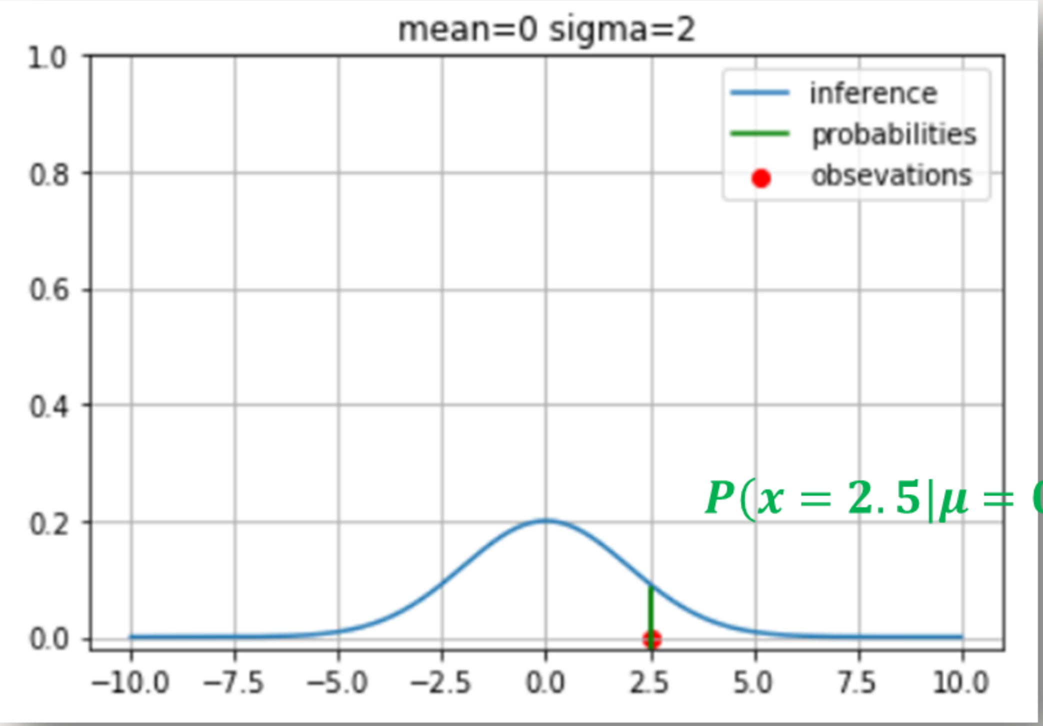
$$X \sim N(\boxed{\mu}, \boxed{\sigma}^2)$$

标准差
控制着概率分布偏离均值的程度

均值
以均值为中心两边对称

概率密度函数

$$f(x; \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$



$P(x = 2.5 | \mu = 0, \sigma = 2)$

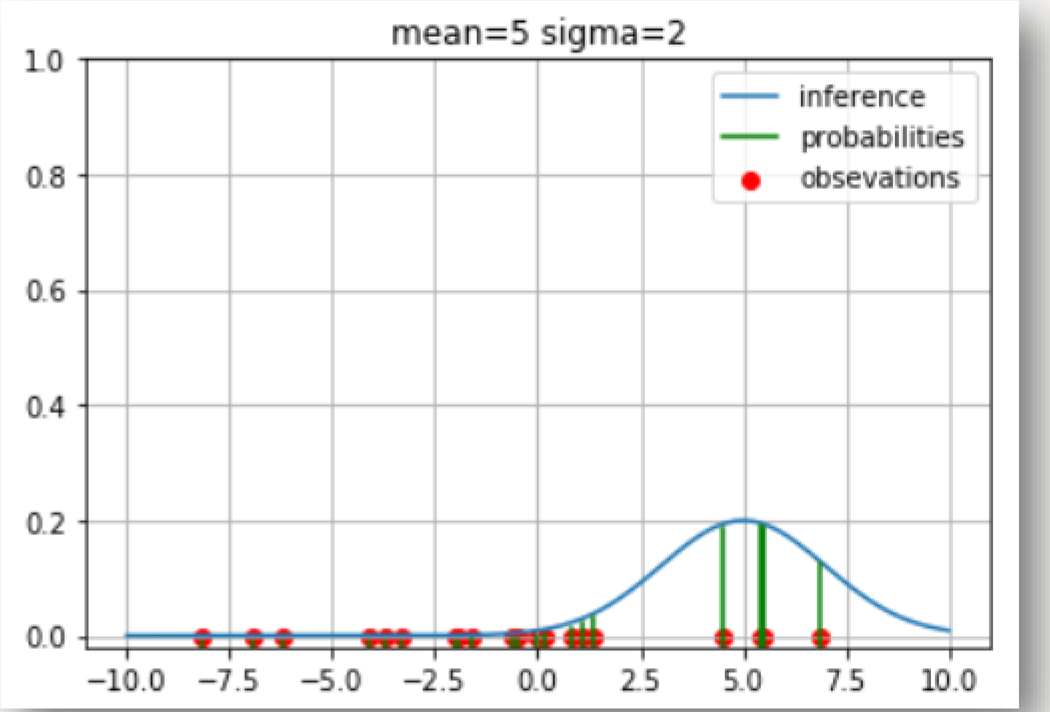
$f(x; \mu = 0, \sigma = 2)$ 随机变量，取值2.5

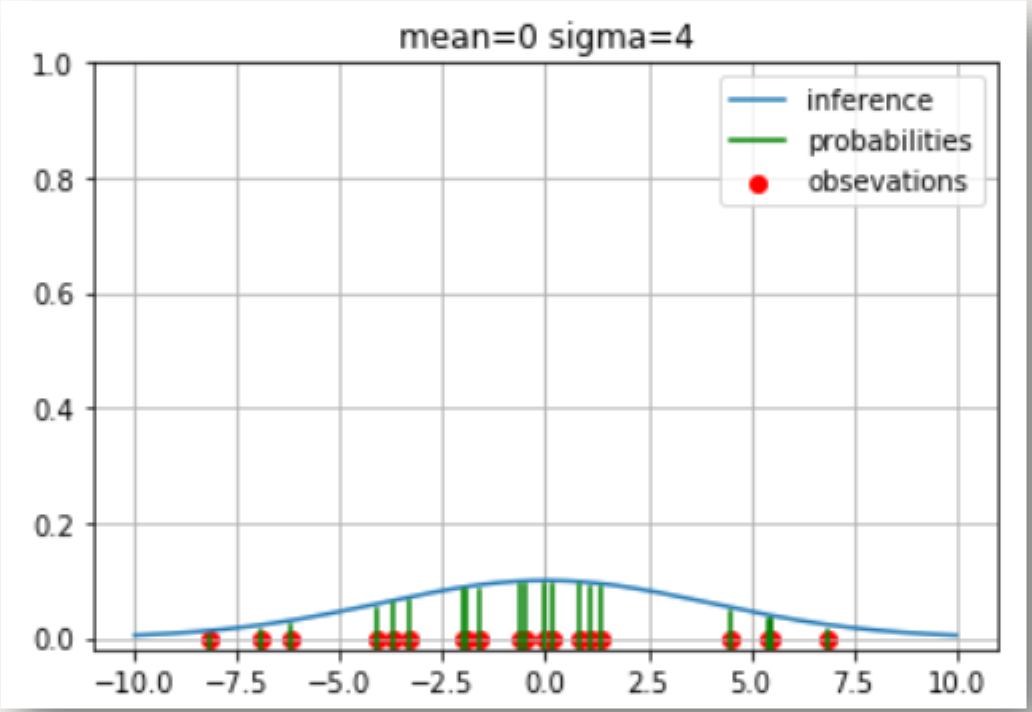
最大似然估计

概率和似然的区别

概率：是在已知概率分布参数的情况下，预测观测的结果
似然：已知观测到的结果，估计观测结果所属于的概率分布

给定观测点的前提下，
我们要求的是
哪一种分布可以得到更高的概率



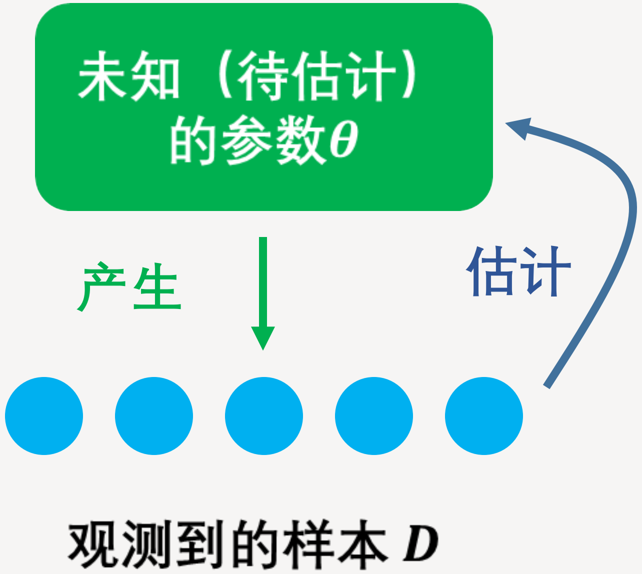


红色圆点是观测到的随机变量，蓝色的线是概率密度函数的图像，绿色直线是观测数据出现的概率

最大似然估计

解决这个问题的方法手段——最大似然估计

- 机器学习领域最为常见的用来构建目标函数的方法
- 核心思想：根据观测到的结果来预测概率分布中的相关参数



Example：扔一枚不均匀硬币，假设出现正面的概率是$\boldsymbol{\theta}$

如Head代表正面，Tail代表反面，投掷5次得到

$D=\{H,T,T,H,H\}$　　$\boldsymbol{\theta}=?$　　通过D反推$\boldsymbol{\theta}$

本质上使用
最大似然估计

$$P(D|\theta)=P(H,T,T,H,H|\theta)=P(H|\theta)P(T|\theta)P(T|\theta)P(H|\theta)P(H|\theta)$$

$$P(D|\theta)=\theta\cdot(1-\theta)\cdot(1-\theta)\cdot\theta\cdot\theta=\theta^3\cdot(1-\theta)^2=f(\theta)$$

$$f'(\theta)=3\theta^2(1-\theta)^2+\theta^3 2(1-\theta)(-1)$$

$$=\theta^2(1-\theta)(3-5\theta)$$

$$\theta=\frac{3}{5}$$

Logistic Regression

回到逻辑回归的目标函数：

$$f_{w,b}(x) = p(y|x) = p(y=1|x)^y[1-p(y=1|x)]^{1-y}$$

- 现有数据集 $D = \{(x^i, y^i)\}_{i=1}^n$ $x^i \in R^d$ $y^i \in \{0,1\}$

（假设样本是独立同分布的）

- 我们需要最大化目标函数（似然函数）

$$w^*, b^* = argmax \prod_{i=1}^{n} p(y^i|x^i, w, b)$$

$$= argmax\ ln \prod_{i=1}^{n} p(y^i|x^i, w, b)$$

x^i	x^1	x^2	x^3	……
y^i	1	0	1	

$$\prod_{i=1}^{n} p(y^i|x^i, w, b)$$

$$= p(y=1|x^1)p(y=0|x^2)p(y=1|x^3) \cdots\cdots$$

浮点数、累乘很难计算

Logistic Regression

$$argmax\, ln \prod_{i=1}^{n} p(y^i|x^i, w, b)$$

ln里的累乘转换为ln外的累加

$$= \underset{w,b}{\operatorname{argmax}} \sum_{i=1}^{n} ln\, \underline{p(y^i|x^i)}$$

$$p(y=1|x)^y[1-p(y=1|x)]^{1-y}$$

$$= \underset{w,b}{\operatorname{argmax}} \sum_{i=1}^{n} ln\, p(y^i=1|x^i)^{y^i} + ln\ [1-p(y^i=1|x^i)]^{1-y^i}$$

$$= \underset{w,b}{\operatorname{argmax}} \sum_{i=1}^{n} y^i\, ln \boxed{p(y^i=1|x^i)} + (1-y^i)\, ln[1-p(y^i=1|x^i)]$$

$$= -\underset{w,b}{\operatorname{argmin}} \sum_{i=1}^{n} y^i\, ln \boxed{f(x^i)} + (1-y^i)\, ln[1-f(x^i)]$$

Logistic Regression

$$-\sum_{i=1}^{n} y^i \ln f\left(x^i\right) + (1 - y^i) \ln[1 - f\left(x^i\right)]$$

最小化

Cross Entropy

交叉熵
损失函数

衡量两个伯努利分布之间的分散程度

Distribution p

$$p(x = 1) = y^i$$
$$p(x = 0) = 1 - y^i$$

交叉熵

Distribution q

$$q(x = 1) = f(x^i)$$
$$q(x = 0) = 1 - f(x^i)$$

Target（数据本身）

Model（逻辑回归预测的）

$$H_{(p,q)} = -\sum_{x} p(x)\ln(q(x))$$

最小化

这个值越小说明两个分布越接近
我们要找的就是最接近于Target的Model

多选题 5分

设置

以下表述正确的是

A 概率是在已知概率分布参数的情况下，预测观测的结果

B 似然是已知观测到的结果，估计观测结果所属于的概率分布

C 我们希望目标函数最大化，因此交叉熵的值越大越好

D 机器学习中常常假设样本服从独立同分布（IID）

提交

Logistic Regression

接下来，要做的事就是——梯度下降（求导微分）

1

$$-\sum_{i=1}^{n} y^i \boxed{\frac{\ln f\left(x^i\right)}{\partial w_i}} + (1-y^i)\ln[1-f\left(x^i\right)]$$

尽可能地小

1 $$\frac{\partial \ln f\left(x^i\right)}{\partial w_i} = \frac{\partial \ln f\left(x^i\right)}{\partial a}\frac{\partial a}{\partial w_i}$$

$$\boldsymbol{f(x) = \sigma(a) = \sigma(wx+b)}$$

$$= \frac{\partial \ln \sigma(a)}{\partial a}\frac{\partial a}{\partial w_i} = \frac{1}{\sigma(a)}\frac{\partial \sigma(a)}{\partial a}\frac{\partial a}{\partial w_i}$$

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

$$\boldsymbol{\sigma'(a) = \sigma(a)(1-\sigma(a))}$$

$$= \frac{1}{\cancel{\sigma(a)}}\cancel{\sigma(a)}(1-\sigma(a))\frac{\partial a}{\partial w_i} = (1-\sigma(a))\frac{\partial a}{\partial w_i} = (1-\sigma(a))x_i$$

Logistic Regression

接下来，要做的事就是：

1 $(1-\sigma(a))x_i$　　2 $-\sigma(a)\,x_i$

$$-\sum_{i=1}^{n} y^i \frac{\ln f(x^i)}{\partial w_i} + (1-y^i)\boxed{\frac{\ln[1-f(x^i)]}{\partial w_i}}$$

→ 尽可能地小

2
$$\frac{\partial \ln[1-f(x^i)]}{\partial w_i} = \frac{\partial \ln[1-f(x^i)]}{\partial a}\frac{\partial a}{\partial w_i}$$

$$\boldsymbol{f(x)=\sigma(a)=\sigma(wx+b)}$$

$$= \frac{\partial \ln[1-\sigma(a)]}{\partial a}\frac{\partial a}{\partial w_i} = -\frac{1}{1-\sigma(a)}\frac{\partial \sigma(a)}{\partial a}\frac{\partial a}{\partial w_i}$$

$$\boldsymbol{\sigma'(a)=\sigma(a)(1-\sigma(a))}$$

$$= -\frac{1}{1-\sigma(a)}\sigma(a)(1-\sigma(a))\frac{\partial a}{\partial w_i} = -\sigma(a)\frac{\partial a}{\partial w_i} = -\sigma(a)x_i$$

Logistic Regression

1 $(1-\sigma(a))x_i$ 2 $-\sigma(a)\ x_i$

$$-\sum_{i=1}^{n} y^i \frac{ln\, f\left(x^i\right)}{\partial w_i} + (1-y^i)\frac{ln\left[1-f\left(x^i\right)\right]}{\partial w_i} = \frac{\partial L(w,b)}{\partial w}$$

$$= \sum_{i=1}^{n} -[y^i(1-\sigma(a))\ x_i - \left(1-y^i\right)\sigma(a)\ x_i]$$

$$= \sum_{i=1}^{n} -[y^i - \cancel{\sigma(a)y^i} - \sigma(a) + \cancel{y^i\sigma(a)}]x_i$$

$$= \sum_{i=1}^{n} -[y^i - \sigma(a)]x_i = \sum_{i=1}^{n}[f(x_i) - y^i]x_i$$

线性回归 vs 逻辑回归

	Linear Regression	Logistic Regression
损失函数	$L(f) = \frac{1}{2}\sum_{i=1}^{n}\left(f(x^i) - \hat{y}^i\right)^2$ SE	$-\sum_{i=1}^{n} y^i \ln f(x^i) + (1 - y^i)\ln[1 - f(x^i)]$ 交叉熵
$\frac{\partial L(w,b)}{\partial w}$	$\frac{\partial L}{\partial w} = \sum_{i=1}^{n}[f(x^i) - \hat{y}^i]\,x^i$	$\frac{\partial L}{\partial w} = \sum_{i=1}^{n}[f(x^i) - \hat{y}^i]\,x^i$

Logistic Regression可以用SE作为损失函数吗?

为什么逻辑回归不能用SE

假如用SE作为Logistic Regression的损失函数

逻辑回归模型 $f(x)=\sigma\left(\boxed{\sum_{i=1}^{n} w_i x_i + b}\right)$

损失函数 $L(f)=\frac{1}{2}\sum_{n}\left(f(x^n)-\hat{y}^n\right)^2$

1 0

Gradient Descent

$$\frac{\partial L(f)}{\partial w_i}=\left(f(x)-\hat{y}\right)\underline{\frac{\partial f(x)}{\partial a}}\frac{\partial a}{\partial w_i}$$

$$=\left(f(x)-\hat{y}\right)f(x)\left[1-f(x)\right]x_i$$

如果 $\hat{y}$ 为class A

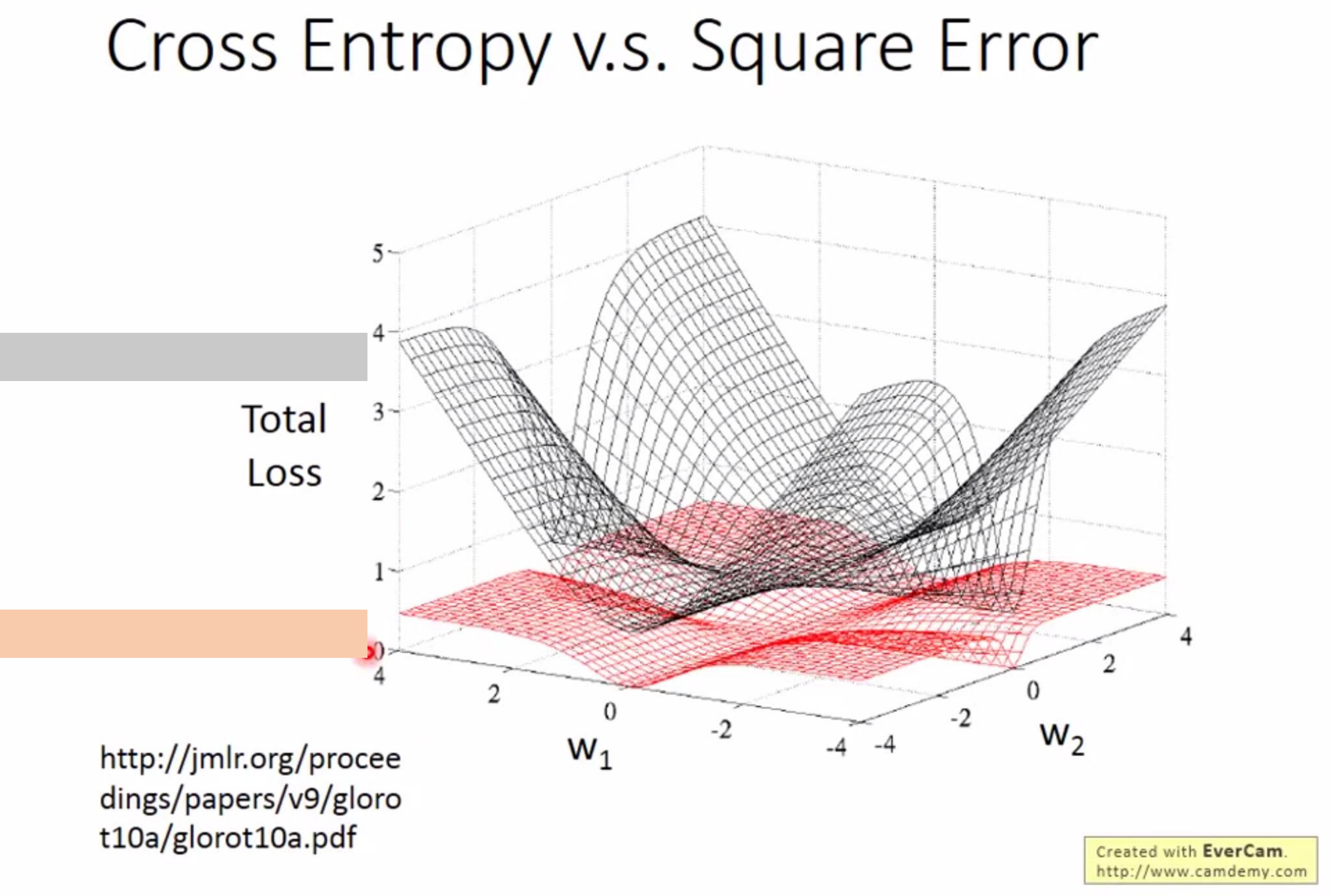
$$if\quad f(x)=1 \xrightarrow{\text{是我想要的}} \frac{\partial L(f)}{\partial w_i}=0$$

$$if\quad f(x)=0 \xrightarrow{\text{不是我想要的}} \frac{\partial L(f)}{\partial w_i}=0$$

为什么逻辑回归不能用SE

Cross Entropy

Square Error



投票 最多可选1项

设置

你认为逻辑回归是线性分类器还是非线性分类器？

- A 线性分类器
- B 非线性分类器

提交

Logistic Regression

Final Ques：逻辑回归是线性分类器还是非线性分类器?

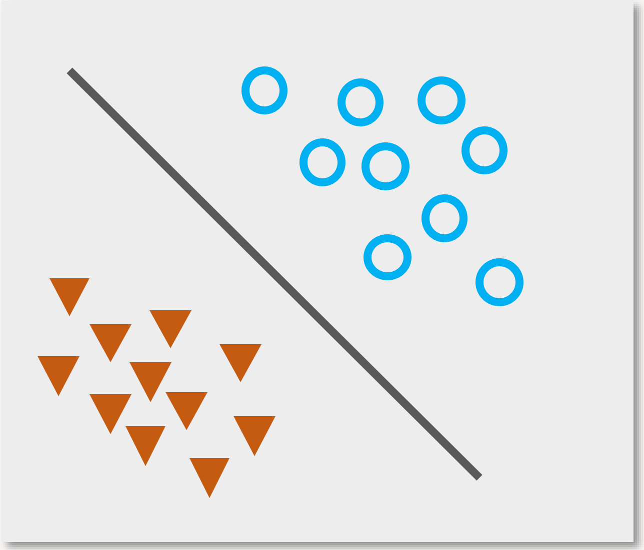
线性分类器

怎么判断？ ——决策边界（decision boundary）

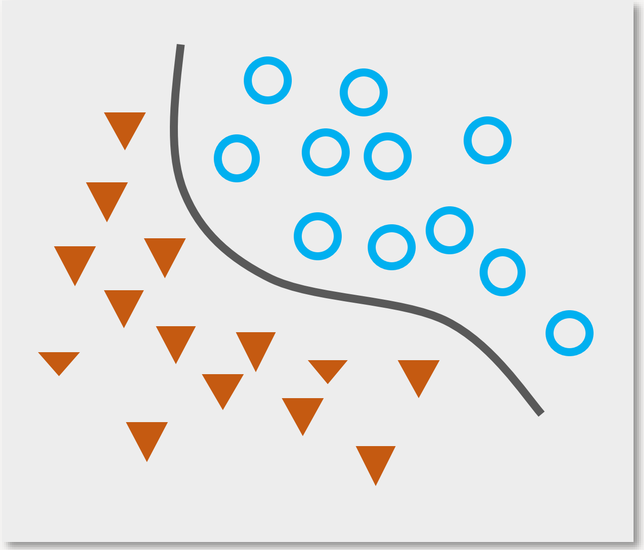
在决策边界上两种类别概率相同

$$\frac{p(y=1|x)}{p(y=0|x)}=1$$

$$\frac{\dfrac{1}{1+e^{-w^Tx+b}}}{\dfrac{e^{-w^Tx+b}}{1+e^{-w^Tx+b}}}=1 \quad \left\{ \begin{array}{l} \dfrac{1}{e^{-w^Tx+b}}=1 \\ \Downarrow \\ w^Tx+b=0 \end{array} \right.$$



线性分类器



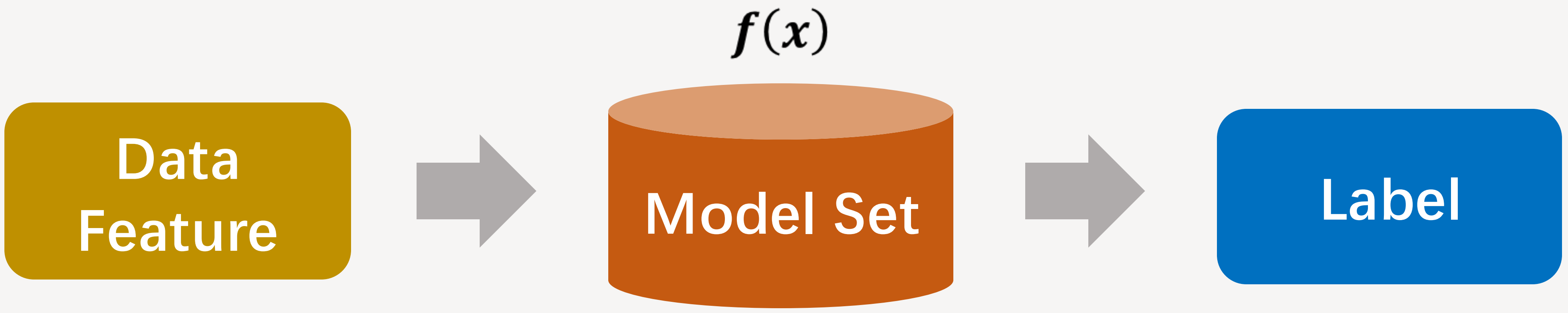
非线性分类器

逻辑回归会学到的是线性决策边界

总结

Linear Regression VS Logistic Regression

本质上都是特征（feature）到结果\标签（Label）之间的映射



- Linear Regression　结果是连续的
- Logistic Regression　结果是离散的

逻辑回归想要
找到一个线性分类边界

多选题 5分

设置

下列对逻辑回归的描述正确的是：

A 逻辑回归比线性回归更擅长解决分类问题

B 逻辑回归的目标函数本质上用了极大似然估计原理

C 逻辑回归是线性分类器

D 逻辑回归一般用交叉熵作为损失函数

提交

Exercise time：

Numpy