机器学习

逻辑回归的概念

课前线上预习

逻辑回归：

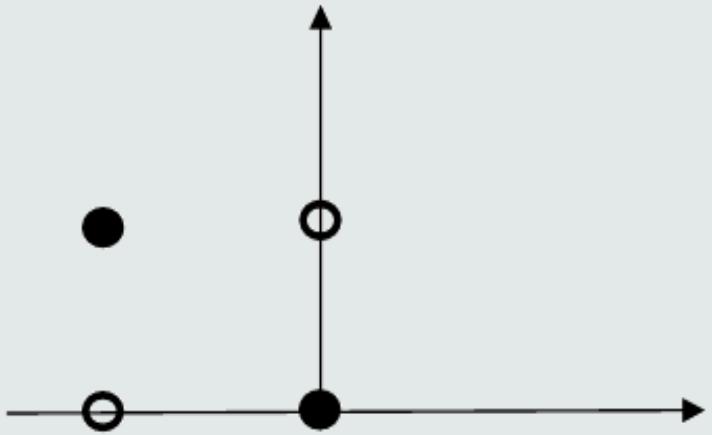
线性和非线性模型之间的桥梁

课堂练习

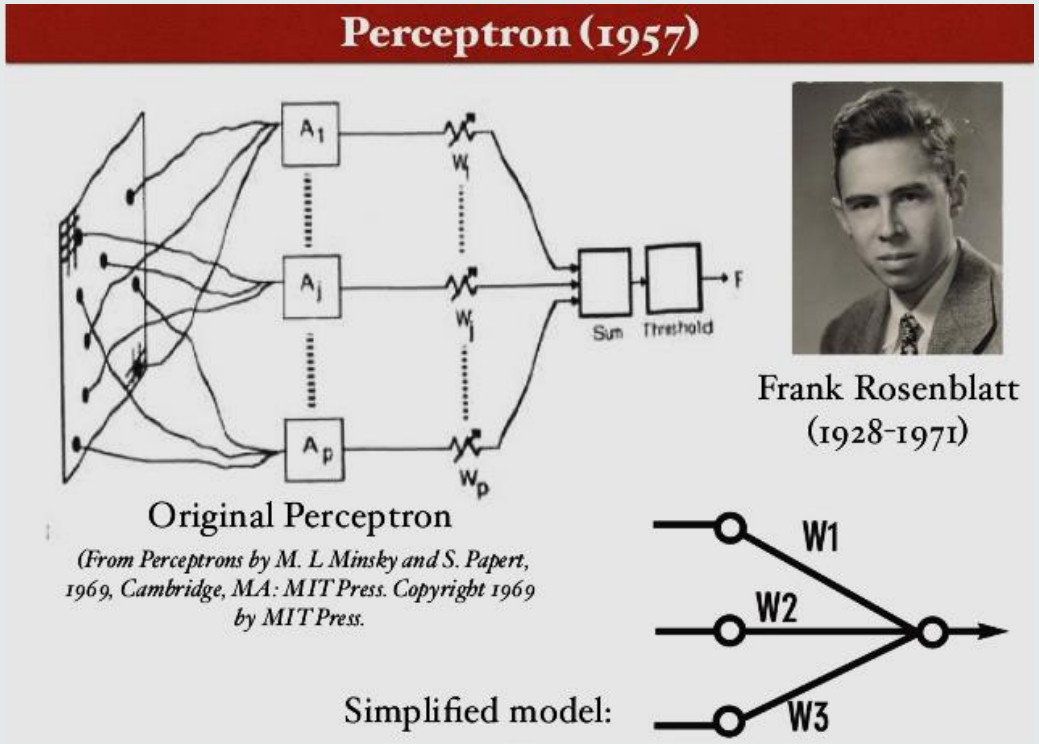
任务：异或问题

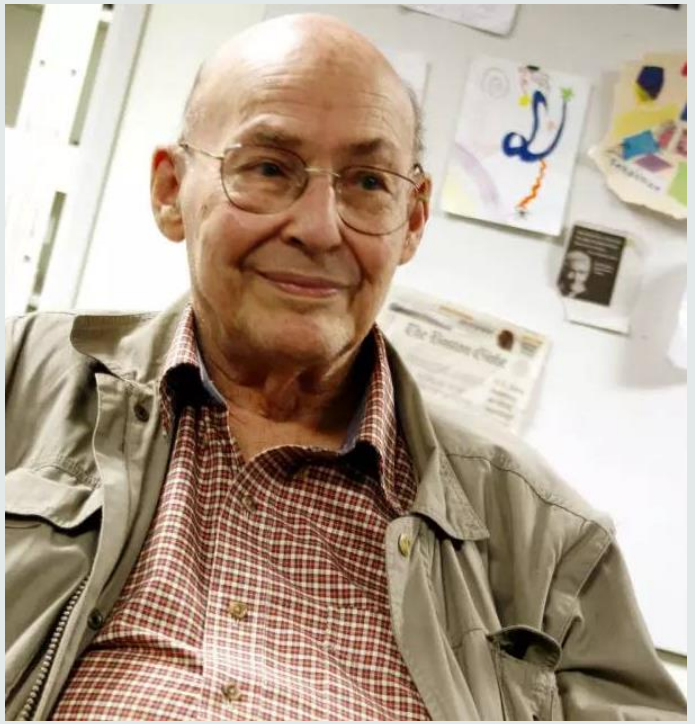
给定如下数据，用课前编好的程序，训练模型并画出分类决策边界。

✓ 4个样本数据：实心是属于类1的样本，空心是属于类2的样本。



感知机小故事


Perceptron (1957)
A1
Aj
Ap
Wi
Wj
Wp
Sum
Threshold
F
Frank Rosenblatt
(1928-1971)
Original Perceptron
(From Perceptrons by M. L Minsky and S. Papert, 1969, Cambridge, MA: MIT Press. Copyright 1969 by MIT Press.
Simplified model:
W1
W2
W3



非线性模型设计思路

假设你是算法设计者，你的思路?

非线性模型设计思路

线性判据

线性

贝叶斯分类器

线性

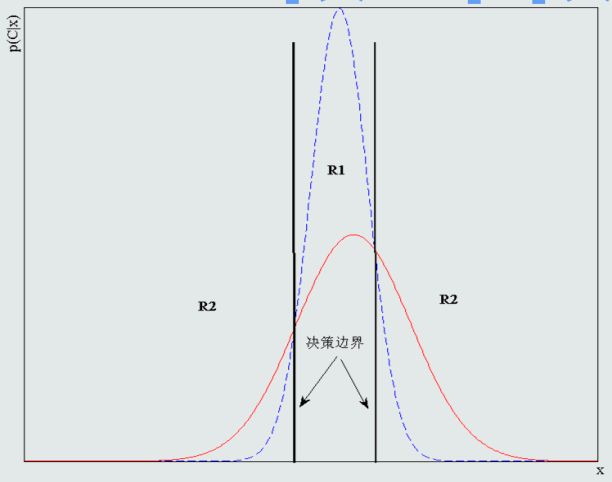
非线性

填空题 4分

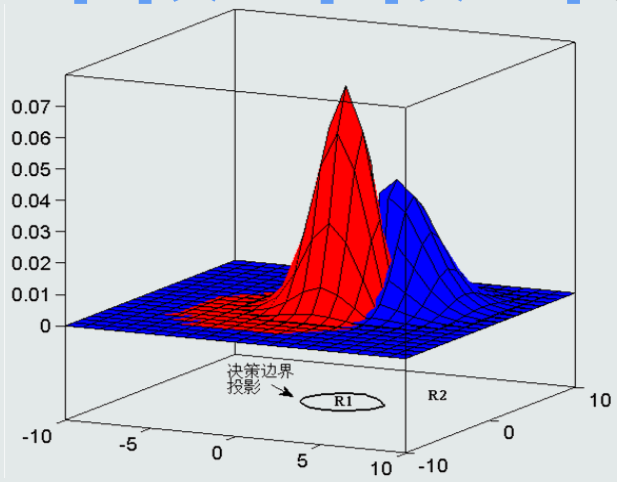
设置

课堂练习：按顺序给出图片对应的均值和方差取值。

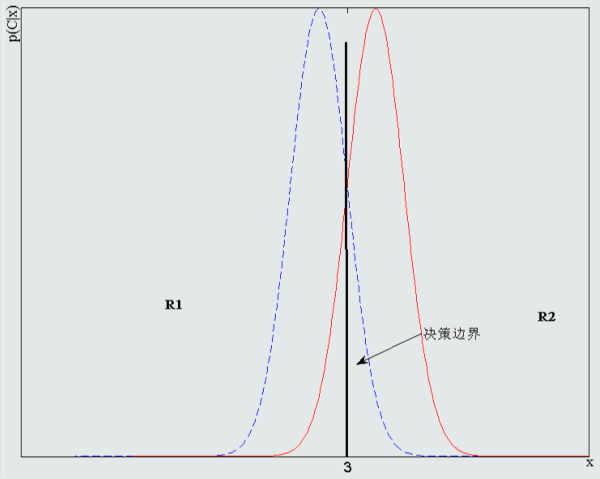
[填空1] [填空2] [填空3] [填空4] 。



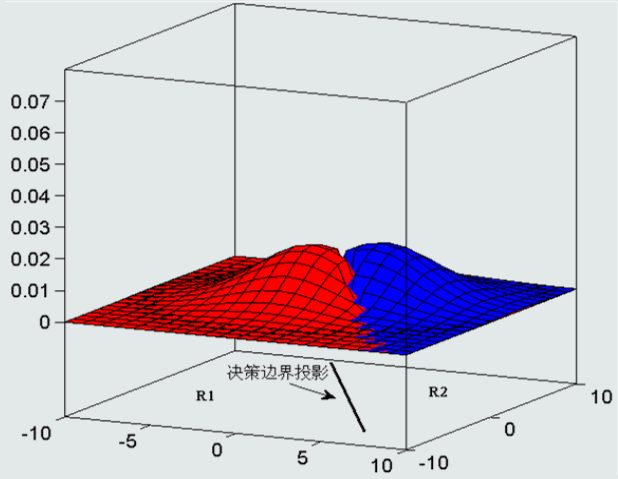
(1)



(2)



(3)



(4)

A $\mu_1 = (3,3), \quad \mu_2 = (1,1)$

$$\Sigma_1 = \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix}, \quad \Sigma_2 = \begin{bmatrix} 1.5 & 0 \\ 0 & 1.5 \end{bmatrix}$$

B $\mu_1 = 1, \ \mu_2 = 5$

$\sigma_1 = 2, \ \sigma_2 = 2$

C $\mu_1 = 2, \ \mu_2 = 3$

$\sigma_1 = 2, \ \sigma_2 = 4$

D $\mu_1 = (2,3), \ \mu_2 = (-1,1)$

$$\Sigma_1 = \begin{bmatrix} 3 & 0 \\ 0 & 3 \end{bmatrix}, \quad \Sigma_2 = \begin{bmatrix} 3 & 0 \\ 0 & 3 \end{bmatrix}$$

正常使用填空题需3.0以上版本雨课堂

作答

贝叶斯分类器

非线性的情况

- 如果观测似然（即每个类的数据分布$p(\boldsymbol{x}|C_i)$）是高斯分布$\mathcal{N}(\boldsymbol{x};\boldsymbol{\mu}_i,\boldsymbol{\Sigma}_i)$，MAP分类器决策方程为：

$$(\boldsymbol{x}-\boldsymbol{\mu}_j)^T\Sigma_j^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_j)-(\boldsymbol{x}-\boldsymbol{\mu}_i)^T\Sigma_i^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_i)-2\ln\left(\frac{P(C_j)|\Sigma_i|}{P(C_i)|\Sigma_j|}\right)\underset{C_i}{\overset{C_j}{\lessgtr}}0$$

- 因此，如果两个类别数据分布的协方差矩阵不同（即$\Sigma_i \neq \Sigma_j$），则MAP分类器的决策边界是一个超二次型曲面，即非线性。

贝叶斯分类器

线性的情况

$$(\boldsymbol{x}-\boldsymbol{\mu}_j)^T\Sigma_j^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_j)-(\boldsymbol{x}-\boldsymbol{\mu}_i)^T\Sigma_i^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_i)-2\ln\left(\frac{P(C_j)|\Sigma_i|}{P(C_i)|\Sigma_j|}\right)\underset{C_i}{\overset{C_j}{\lessgtr}}0$$

- 如果两个类别数据分布的协方差矩阵相同（即$\Sigma_i=\Sigma_j$），则MAP分类器的决策边界是一个超平面，即线性。

非线性模型设计思路

贝叶斯分类器可以在线性和非线性之间切换

MAP分类器输出：后验概率

线性判据输出：样本到决策边界的距离

两者之间有什么联系呢?

Logit变换

通过定义Logit变换来建立起

线性判据和贝叶斯分类器之间的联系

Logit变换

后验概率的比率

- 对于二类分类，MAP分类器通过比较后验概率的大小来决策。
- 也可以通过比较两个后验概率的比率来做决策。

$$\frac{p(C_1|\boldsymbol{x})}{p(C_2|\boldsymbol{x})}, \text{ given that } p(C_1|\boldsymbol{x}) + p(C_2|\boldsymbol{x}) = 1$$

$$\text{So } \frac{p(C_1|\boldsymbol{x})}{p(C_2|\boldsymbol{x})} = \frac{p(C_1|\boldsymbol{x})}{1-p(C_1|\boldsymbol{x})}$$

Logit变换

后验概率比率取log

- ✓ 观测是高斯分布
- ✓ $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \boldsymbol{\Sigma}$

$$\log \frac{p(C_1|\boldsymbol{x})}{p(C_2|\boldsymbol{x})} = \log \frac{p(\boldsymbol{x}|C_1)\,p(C_1)}{p(\boldsymbol{x}|C_2)\,p(C_2)} = \log \frac{p(\boldsymbol{x}|C_1)}{p(\boldsymbol{x}|C_2)} + \log \frac{p(C_1)}{p(C_2)}$$

$$= \log \frac{(2\pi)^{-p/2}|\boldsymbol{\Sigma}|^{-1/2}\exp\left[-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_1)^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_1)\right]}{(2\pi)^{-p/2}|\boldsymbol{\Sigma}|^{-1/2}\exp\left[-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_2)^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_2)\right]} + \log \frac{p(C_1)}{p(C_2)}$$

$$= -\frac{1}{2}\left[(\boldsymbol{x}-\boldsymbol{\mu}_1)^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_1) - (\boldsymbol{x}-\boldsymbol{\mu}_2)^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_2)\right] + \log \frac{p(C_1)}{p(C_2)}$$

$$= (\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)^T\boldsymbol{\Sigma}^{-1}\boldsymbol{x} - \frac{1}{2}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1+\boldsymbol{\mu}_2) + \log \frac{p(C_1)}{p(C_2)}$$

Logit变换

后验概率对数比率 = 线性判据输出

$$\log\frac{p(C_1|\boldsymbol{x})}{p(C_2|\boldsymbol{x})}=(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)^T\boldsymbol{\Sigma}^{-1}\boldsymbol{x}-\frac{1}{2}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1+\boldsymbol{\mu}_2)+\log\frac{p(C_1)}{p(C_2)}$$

令 $\quad \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)=\boldsymbol{w}$

$$-\frac{1}{2}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1+\boldsymbol{\mu}_2)+\ln\frac{p(C_1)}{p(C_2)}=w_0$$

得到 $\log\frac{p(C_1|\boldsymbol{x})}{p(C_2|\boldsymbol{x})}=\boldsymbol{w}^T\boldsymbol{x}+w_0=f(\boldsymbol{x})$

后验概率是否可以直接由线性判据表达?

Sigmoid函数

Sigmoid函数

$$\log \frac{p(C_1|\boldsymbol{x})}{1-p(C_1|\boldsymbol{x})} = \boldsymbol{w}^T\boldsymbol{x} + w_0 \quad \Rightarrow p(C_1|\boldsymbol{x}) = \frac{1}{1+\exp\left[-\left(\boldsymbol{w}^T\boldsymbol{x}+w_0\right)\right]}$$

- 设$y = f(\boldsymbol{x}) = \boldsymbol{w}^T\boldsymbol{x} + w_0$，根据上式可以定义Sigmoid函数：

$$\text{sigmoid}\ (y) = \frac{1}{1+\exp(-y)} = \frac{1}{1+\exp[-(\boldsymbol{w}^T\boldsymbol{x}+w_0)]}$$

- 线性判据$f(\boldsymbol{x})$放入Sigmoid函数，可得到$\boldsymbol{x}$属于C_1类的后验概率：

$$\text{sigmoid}(y) = p(C_1|\boldsymbol{x})$$

单选题 10分

设置

关于Sigmoid函数，如下哪个描述是正确的：

A 是一个以x为自变量、输出范围为[0,1]的函数。

B 是一个以x为自变量、输出范围为[-1,1]的函数。

C 是一个以y为自变量、输出范围为[0,1]的函数。

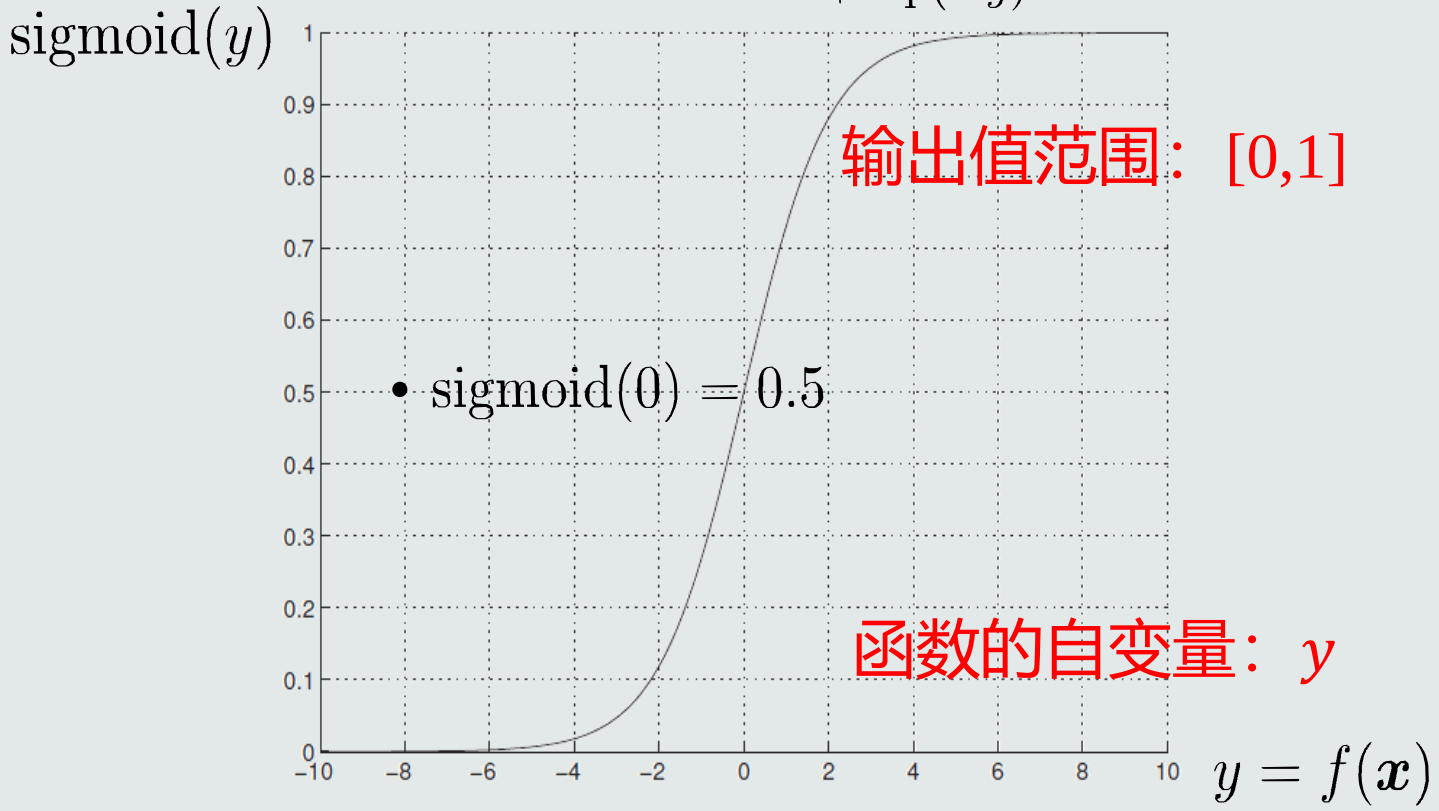
D 是一个以y为自变量、输出范围为[-1,1]的函数。

提交

Sigmoid函数

Sigmoid函数：波形

$$\text{sigmoid}\ (y) = \frac{1}{1+\exp(-y)}$$



输出值范围：[0,1]

函数的自变量：y

Sigmoid函数：小结

Sigmoid函数：连接线性模型和后验概率的桥梁

线性模型$f(\boldsymbol{x})$ + Sigmoid函数 = 后验概率

参考文献

[1] A. Krizhevsky, I. Sutskever, G. E. Hinton, “ImageNet classification with deep convolutional neural networks”, in *Proceedings of Advances in Neural Information Processing Systems*, vol. 60, no.2, pp. 1097-1105, 2012.

[2] K. He, X. Zhang, S. Ren, and J. Sun, “Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification,” in *Proceedings of International Conference on Computer Vision*, pp.1026-1034, 2015.

逻辑回归

定义

- 逻辑回归（Logistic Regression）：线性模型$f(\boldsymbol{x})$ + sigmoid函数。

$$\text{Logistic:}\quad \text{sigmoid}(\boldsymbol{w}^T\boldsymbol{x}+w_0)$$

  - 给定测试样本$\boldsymbol{x}$，Logistic回归输出其属于C_1类的后验概率。

逻辑回归：分类

决策边界

- 单个逻辑回归可以用于二类分类，其决策过程如下：

$$l(\boldsymbol{x}) = \begin{cases} 1\ (C_1) & \text{if}\ \ \text{sigmoid}(\boldsymbol{w}^T\boldsymbol{x} + w_0) > 0.5 \\ 0\ (C_2) & \text{if}\ \ \text{sigmoid}(\boldsymbol{w}^T\boldsymbol{x} + w_0) \leq 0.5 \end{cases}$$

- 给定两个类，逻辑回归的决策边界仍然是线性的超平面。

$$\text{sigmoid}(\boldsymbol{w}^T\boldsymbol{x} + w_0) = 0.5$$

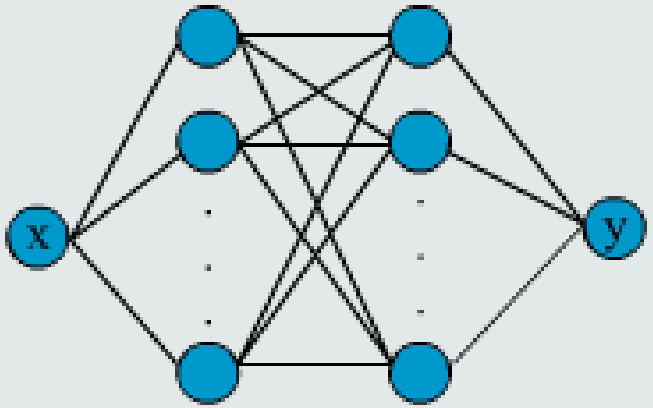
$$\Rightarrow \boldsymbol{w}^T\boldsymbol{x} + w_0 = 0$$

分组讨论

逻辑回归分类决策边界是线性，

逻辑回归如何用于非线性分类?

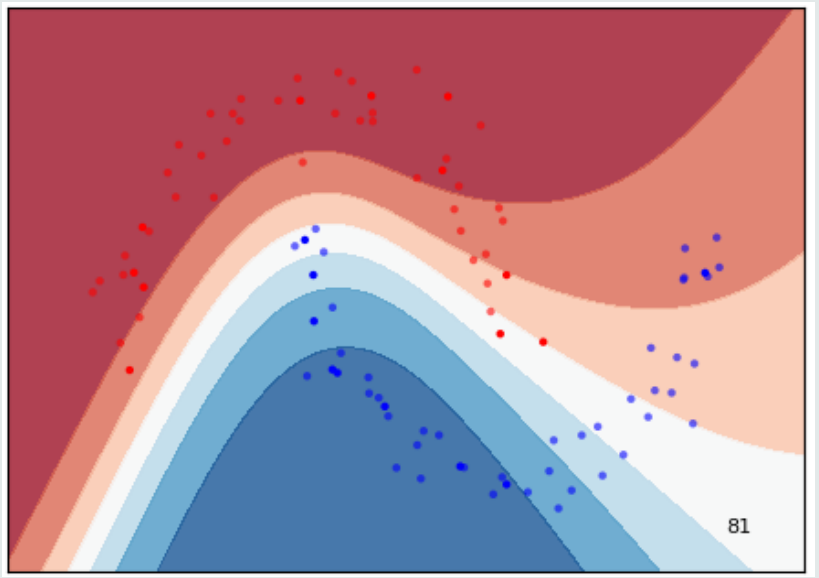
逻辑回归与非线性



$\sigma(w_1x+b_1)$ $\sigma(w_2\sigma(w_1x+b_1)+b_2)$

$\sigma(x) = \text{sigmoid}(x)$

分类边界



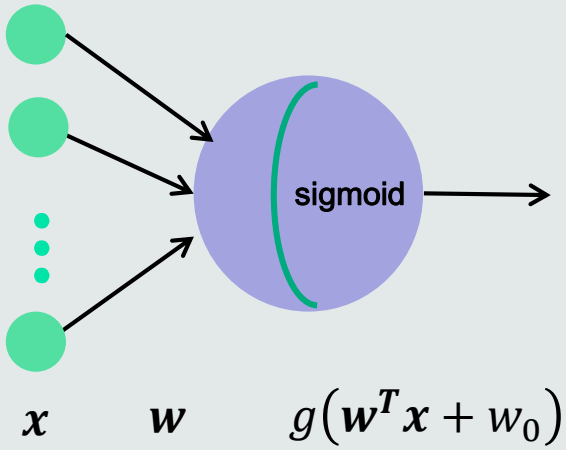
逻辑回归

逻辑回归与神经元

- 单个逻辑回归就是一个神经元模型：

$$g(\boldsymbol{w}^T\boldsymbol{x} + w_0)$$

其中，函数g是sigmoid函数。



- 多层逻辑回归嵌套可应用于非线性分类。

逻辑回归

总结

- 逻辑回归本身是一个非线性模型。
- 逻辑回归用于分类：仍然只能处理两个类别线性可分的情况。但是，sigmoid函数输出了后验概率，使得逻辑回归成为一个非线性模型。因此，逻辑回归比线性模型向前迈进了一步。
- 逻辑回归用于拟合：可以拟合有限的非线性曲线。

逻辑回归如何学习？

课后预习与作业

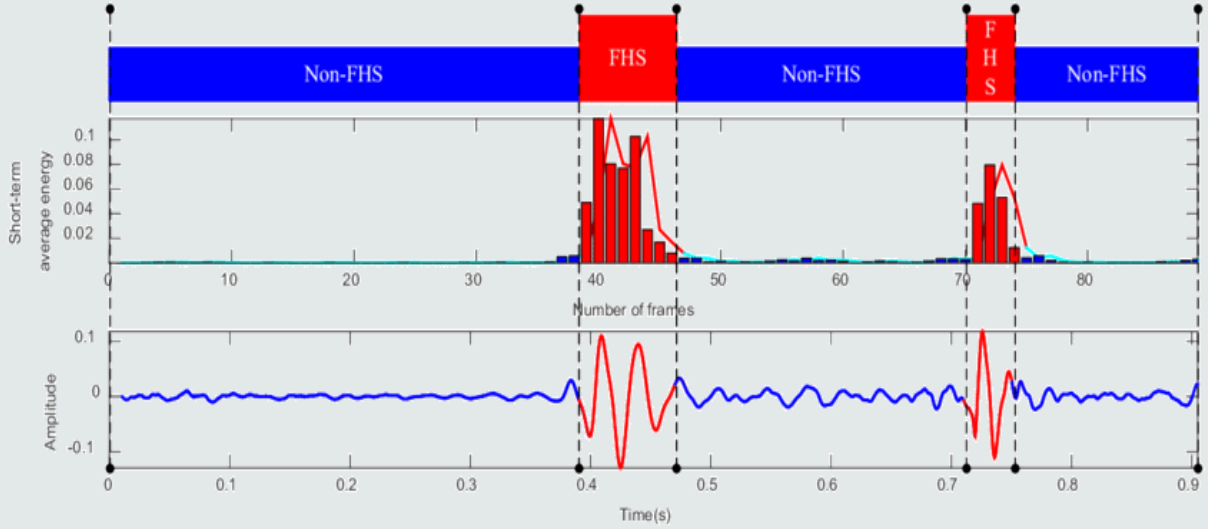
- ✓ 预习MOOC第4.14节课：Logistic学习
- ✓ 编程：基于梯度下降法的Logistic学习算法

科研实训作业

编程

- ✓ 设计逻辑回归模型，实现心脏的心音分段（正常/异常）。
- ✓ 数据集：

  https://www.physionet.org/challenge/2016



参考文献

[1] D. B. Springer, L. Tarassenko, and G. D. Clifford, "Logistic Regression-HSMM-Based Heart Sound Segmentation", *IEEE Transactions on Biomedical Engineering*, vol. 33, no. 4, pp. 822-832, 2016.

[2] E. Adeli, X. Li, D. Kwon, Y. Zhang, and K. M. Pohl, "Logistic Regression Confined by Cardinality-Constrained Sample and Feature Selection", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, no. 7, pp. 1713-1728, 2020.

机器学习

逻辑回归的学习

待学习参数

学什么

$$\text{Logistic:}\quad \text{sigmoid}(\boldsymbol{w}^T\boldsymbol{x} + w_0)$$

- 给定训练样本，学习参数$\boldsymbol{w}$和w_0。

单选题 1分

设置

训练样本真值设置

针对Logistic学习算法，有关真值设置的描述，如下哪个是正确的：

A 正类的输出真值是1，负类的输出真值是0。

B 正类的输出真值是1，负类的输出真值是-1。

C 正类的输出真值可以是任意正数，负类的输出真值可以是任意负数。

D 正类的输出真值可以是任意正数，负类的输出真值是0。

提交

目标函数设计

由于输出是概率，如何表达真值的分布?

输出真值的概率表达

输出真值：伯努利分布

- 给定单个输入样本$\boldsymbol{x}$，模型输出的真值标签l可以看做一个随机变量。
  - ✓ 单个训练样本$\boldsymbol{x}_n$放入模型，相当于对随机变量l的一次采样试验（trial），输出真值标签t_n相当于指定此次试验的结果。
  - ✓ 该随机变量只有两个取值：1（正类）或0（负类），符合伯努利分布。

伯努利分布（Bernoulli）

- 随机变量x只有两个取值：成功（1）或失败（0），成功的概率为p，失败的概率则为$1-p$。
- 该随机变量x的概率分布即为伯努利分布，可以表达为：

$$\mathrm{Bernoulli}(x;p) = p^x(1-p)^{1-x},\ x \in \{0,1\}$$

输出真值的概率表达

在逻辑回归中，伯努利分布的参数p怎么设置？

请问下老师，在逻辑回归中伯努利分布的参数p 是如何设置的？

mooc7418309...　　顶 0　　回复 0

输出真值的概率表达

伯努利的参数p的设置

- 真值l可以看做是伯努利分布。

$$\text{Bernoulli}(x;p) = p^x(1-p)^{1-x},\ x \in \{0,1\}$$

$$p(l|\boldsymbol{x}) = \text{Bernoulli}(l;z) = z^l(1-z)^{1-l}$$

为什么使用模型实际输出的后验概率z来设置参数p？

输出真值的概率表达

例子

- 例1: for a sample x_n, given $t_n = 1$ and $z_n = 0.3$

$$p(t_n|x_n) = 0.3^1(1-0.3)^{1-1} = 0.3$$

for a sample x_m, given $t_m = 0$ and $z_m = 0.7$

$$p(t_m|x_m) = 0.7^0(1-0.7)^{1-0} = 0.3$$

针对样本输入，如果模型输出概率较低，说明模型参数不是最优的。

输出真值的概率表达

输出真值的概率分布

- 给定单个输入样本$\boldsymbol{x}$，模型输出真值l符合伯努利分布。
  - ✓ 分布参数p：模型输出的属于正类（C_1类）的后验概率z。

$$p(l|\boldsymbol{x}) = \text{Bernoulli}(l; z) = z^l(1-z)^{1-l}$$

$$\text{where } l \in \{0,1\}, \quad z = p(C_1|\boldsymbol{x}) = \frac{1}{1+\exp[-(\boldsymbol{w}^T\boldsymbol{x}+w_0)]}$$

概率z是由模型参数决定的，是待学习的。

目标函数设计

如何设计目标函数?

为什么不能用均方误差来表达目标函数？

mooc7418309...　　顶 0　　回复 0

最大似然估计法

似然函数

- 如果参数$\boldsymbol{w}$和w_0是最优的，意味着对大部分样本$(\boldsymbol{x}_n, t_n)$而言，$p(t_n|\boldsymbol{x}_n)$ 应该是较大的（无论t_n取值是1还是0）。
- 因此，使用最大似然估计：针对所有训练样本$\mathcal{X}$，最大化输出标签分布的似然函数，以此求得参数$\boldsymbol{w}$和w_0的最优值。
- 似然函数为所有训练样本输出概率的乘积，表达为：

$$\max L(\boldsymbol{w}, w_0|\mathcal{X}) = \max \prod_{n=1}^{N} p(t_n|\boldsymbol{x}_n)$$

最大似然估计法

目标函数

- 对似然函数求取log：

$$\log L(\boldsymbol{w}, w_0|\mathcal{X}) = \log \prod_{n=1}^{N} p(t_n|\boldsymbol{x}_n)$$

$$= \log \prod_{n=1}^{N} z_n^{t_n}(1-z_n)^{1-t_n}$$

$$= \sum_{n=1}^{N} t_n \log z_n + (1-t_n)\log(1-z_n)$$

最大似然估计法

目标函数

- 由于log是凹函数，所以对目标函数取反。相应的，最大化变为最小化。
- 因此，基于最大似然估计策略，最终得到的目标函数为：

$$\min\ J = \min -\sum_{n=1}^{N} t_n \ln z_n + (1-t_n)\ln(1-z_n)$$

?

既然知道了真值和模型实际输出的概率，

是否可以使用概率分布相似度衡量方法?

交叉熵表达

训练目标

- 给定单个样本$\boldsymbol{x}_n$，希望模型预测输出的概率分布$p\ (l_n|\boldsymbol{x}_n)$符合输出真值的概率分布$q(l_n|\boldsymbol{x}_n)$，即两种分布的差异程度最小。
- $p\ (l_n|\boldsymbol{x}_n)$和$q(l_n|\boldsymbol{x}_n)$的分布情况如下表：

$$p^{t_n}(1-p)^{1-t_n}$$

?

给定样本$\boldsymbol{x}_n$	模型预测输出的概率值 $p\ (l_n\|\boldsymbol{x}_n)$	输出真值的概率值 $q(l_n\|\boldsymbol{x}_n)$
属于正类（1）的概率	z_n	t_n
属于负类（0）的概率	$1-z_n$	$1-t_n$

课堂练习

题目

- 给定单个样本$\boldsymbol{x}_n$，模型预测输出的概率分布$p\ (l_n|\boldsymbol{x}_n)$与输出真值的概率分布$q(l_n|\boldsymbol{x}_n)$之间的交叉熵为：

$$(a)\ H(p,q)=\sum_{n=1}^{N} t_n \ln z_n+(1-t_n)\ln(1-z_n),\quad (b)\ H(p,q)=t_n \ln z_n+(1-t_n)\ln(1-z_n)$$

$$(c)\ H(p,q)=t_n \ln z_n-(1-t_n)\ln(1-z_n),\qquad (d)\ H(p,q)=t_n \ln(1-z_n)-(1-t_n)\ln z_n$$

交叉熵（Cross Entropy）

- 给定两个概率分布$p(x)$和$q(x)$，两个分布之间的交叉熵计算如下：

离散随机变量x：有M个取值状态

$$H(p,q)=-\sum_{m=1}^{M} p(x_m)\log q(x_m)$$

连续随机变量x：

$$H(p,q)=-\int_x p(x)\log q(x)dx$$

单选题 1分

设置

给定单个样本$\boldsymbol{x}_n$，模型预测输出的概率分布$p\ (l_n|\boldsymbol{x}_n)$与输出真值的概率分布$q(l_n|\boldsymbol{x}_n)$之间的交叉熵为：

A $H(p,q)=\sum_{n=1}^{N} t_n \ln z_n+(1-t_n)\ln(1-z_n)$

B $H(p,q)=t_n \ln z_n+(1-t_n)\ln(1-z_n)$

C $H(p,q)=t_n \ln z_n-(1-t_n)\ln(1-z_n)$

D $H(p,q)=t_n\ln(1-z_n)-(1-t_n)\ln z_n$

提交

交叉熵表达

针对N个训练样本：目标函数

- 给定N个训练样本，把每个训练样本的交叉熵求和，得到最终的目标函数：

$$\min_{\boldsymbol{w},w_0} -\Big[\sum_{n=1}^{N} t_n \log z_n + (1-t_n)\log(1-z_n)\Big]$$

为什么不用乘积？

如何优化目标函数?

目标函数优化：梯度下降法

对参数$\boldsymbol{w}$求偏导

$$J(\boldsymbol{w}, w_0) = -\Big[\sum_{n=1}^{N} t_n \log z_n + (1-t_n)\log(1-z_n)\Big]$$

$$\text{where} \quad z_n = \text{sigmoid}(y_n), \quad y_n = \boldsymbol{w}^T \boldsymbol{x}_n + w_0$$

- 目标函数对$\boldsymbol{w}$求偏导：

$$\frac{\partial J(\boldsymbol{w}, w_0)}{\partial \boldsymbol{w}} = \sum_{n=1}^{N} \frac{\partial J}{(?)} \frac{(?)}{\partial y_n} \frac{\partial y_n}{(?)}$$

$$\frac{\partial J}{\partial z_n} = -\left(\frac{t_n}{z_n} - \frac{1-t_n}{1-z_n}\right)$$

$$\frac{\partial z_n}{\partial y_n} = z_n(1-z_n)$$

$$\frac{\partial J(\boldsymbol{w}, w_0)}{\partial \boldsymbol{w}} = \sum_{n=1}^{N} (z_n - t_n)\boldsymbol{x}_n$$

$$\frac{\partial y_n}{\partial \boldsymbol{w}} = \boldsymbol{x}_n$$

小贴士

$$\frac{\partial \log \boldsymbol{x}}{\partial \boldsymbol{x}} = \frac{1}{\boldsymbol{x}}$$

$$\frac{\partial f(\boldsymbol{x})}{\partial \boldsymbol{x}} = f(\boldsymbol{x})\left[1 - f(\boldsymbol{x})\right]$$

$f(x)$ is sigmoid

目标函数优化：梯度下降法

对参数w_0求偏导

$$J(\boldsymbol{w}, w_0) = -\Big[\sum_{n=1}^{N} t_n \log z_n + (1-t_n)\log(1-z_n)\Big]$$

$$\text{where} \quad z_n = \text{sigmoid}(y_n), \quad y_n = \boldsymbol{w}^T \boldsymbol{x}_n + w_0$$

- 目标函数对w_0求偏导：

$$\frac{\partial J(\boldsymbol{w}, w_0)}{\partial w_0} = \sum_{n=1}^{N} \frac{\partial J}{\partial z_n} \frac{\partial z_n}{\partial y_n} \frac{\partial y_n}{\partial w_0}$$

$$\frac{\partial J}{\partial z_n} = -\left(\frac{t_n}{z_n} - \frac{1-t_n}{1-z_n}\right)$$

$$\frac{\partial z_n}{\partial y_n} = z_n(1-z_n)$$

$$\frac{\partial J(\boldsymbol{w}, w_0)}{\partial w_0} = \sum_{n=1}^{N} (z_n - t_n)$$

$$\frac{\partial y_n}{\partial w_0} = 1$$

目标函数优化：梯度下降法

参数更新

- 采用梯度下降法更新$\boldsymbol{w}$和w_0：
  - ✓ 设当前时刻为k，下一个时刻为$k+1$
  - ✓ η为更新步长。

$$\boldsymbol{w}^{(k+1)} = \boldsymbol{w}^{(k)} - \eta^{(k)} \sum_{n=1}^{N} \left(z_n \big|_{\boldsymbol{w}^{(k)}, w_0^{(k)}} - t_n \right) \boldsymbol{x}_n$$

$$w_0^{(k+1)} = w_0^{(k)} - \eta^{(k)} \sum_{n=1}^{N} \left(z_n \big|_{\boldsymbol{w}^{(k)}, w_0^{(k)}} - t_n \right)$$

前沿文献阅读

参考文献

[1] X. Shen and Y. Gu, “Nonconvex Sparse Logistic Regression with Weakly Convex Regularization”, *IEEE Transactions on Signal Processing*, vol. 66, no. 12, pp. 3199-3211, 2018.

[2] R. Wang, N. Xiu, and C. Zhang, “Greedy Projected Gradient-Newton Method for Sparse Logistic Regression”, *IEEE Transactions on Neural Networks and Learning Systems*, vol. 31, no.2, pp. 527-538, 2020.

课堂练习

任务

给定数据集，用课前编好的逻辑回归学习算法程序，实现如下功能：

- ✓ 训练逻辑回归模型
- ✓ 观察每个迭代周期，梯度的变化情况

主观题 10分

设置

上传程序运行结果图。

正常使用主观题需2.0以上版本雨课堂

作答

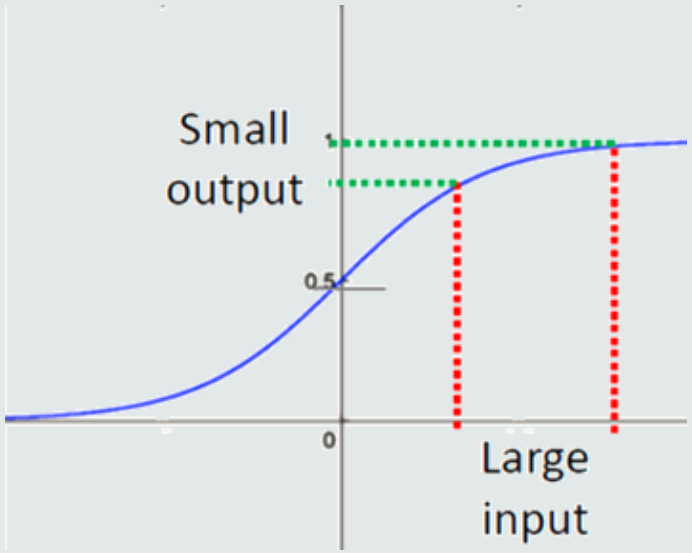
梯度消失问题

梯度消失问题

$$z = f(y) = \text{sigmoid}(\boldsymbol{w}^T\boldsymbol{x} + w_0)$$

$$\frac{\partial f(y)}{\partial y} = f(y)[1 - f(y)]$$

- 当$y = \boldsymbol{w}^T\boldsymbol{x} + w_0$较大时，sigmoid函数输出$z$会出现饱和：输入变化量$\Delta y$很大时，输出变化量$\Delta z$很小。
- 在饱和区，输出量z接近于1，导致sigmoid函数梯度值接近于0，出现梯度消失问题。



参数的初始化

参数初始化

$$\frac{\partial J(\boldsymbol{w}, w_0)}{\partial \boldsymbol{w}} = \sum_{n=1}^{N} \frac{\partial J}{\partial z_n} \frac{\partial z_n}{\partial y_n} \frac{\partial y_n}{\partial \boldsymbol{w}}$$

$$\frac{\partial z_n}{\partial y_n} = z_n(1 - z_n)$$

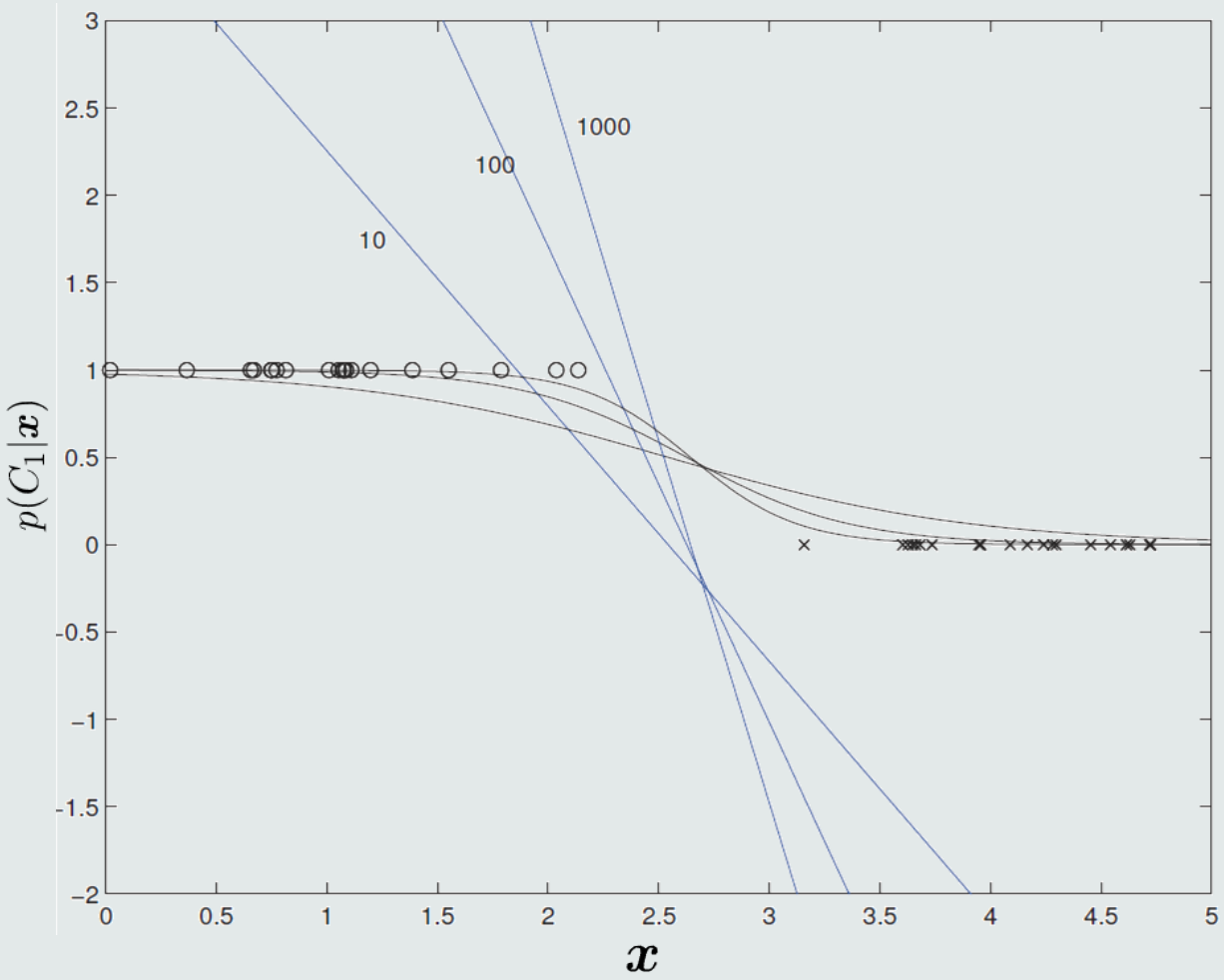
- 在迭代训练过程中，如果参数$\boldsymbol{w}$选择较大的初始值，输出z_n很快会进入sigmoid饱和区（即z_n的值接近于1），梯度$\partial z_n / \partial y_n$接近于0，出现梯度消失。根据链式法则，导致目标函数关于参数的梯度$\partial J / \partial \boldsymbol{w}$接近于0，使得后续迭代更新不起作用。
- 因此，参数$\boldsymbol{w}$尽量选择较小的初始值，避免出现梯度消失问题。

迭代什么时候停止?

Stop Early

- 如果迭代停止条件设为训练误差为0，或者所有训练样本都正确分类的时候才停止，则会出现过拟合问题。
- 所以，在达到一定训练精度后，提前停止迭代，可以避免过拟合。

训练迭代过程示意



◦：正类样本。

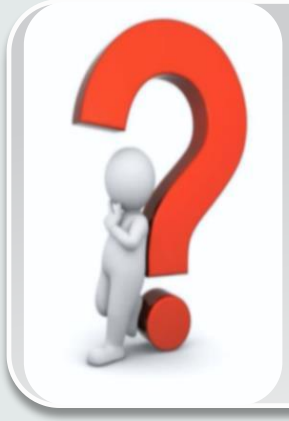
×：负类样本。

分别迭代训练10次、100次和1000次的结果。

直线：决策边界$\boldsymbol{w}^T\boldsymbol{x}+w_0=0$；

曲线：$\text{sigmoid}(\boldsymbol{w}^T\boldsymbol{x}+w_0)$。

- ✓ 迭代次数要达到一定程度，训练性能较好。
- ✓ 逻辑回归输出的非线性形式就是sigmoid函数的非线性形式。



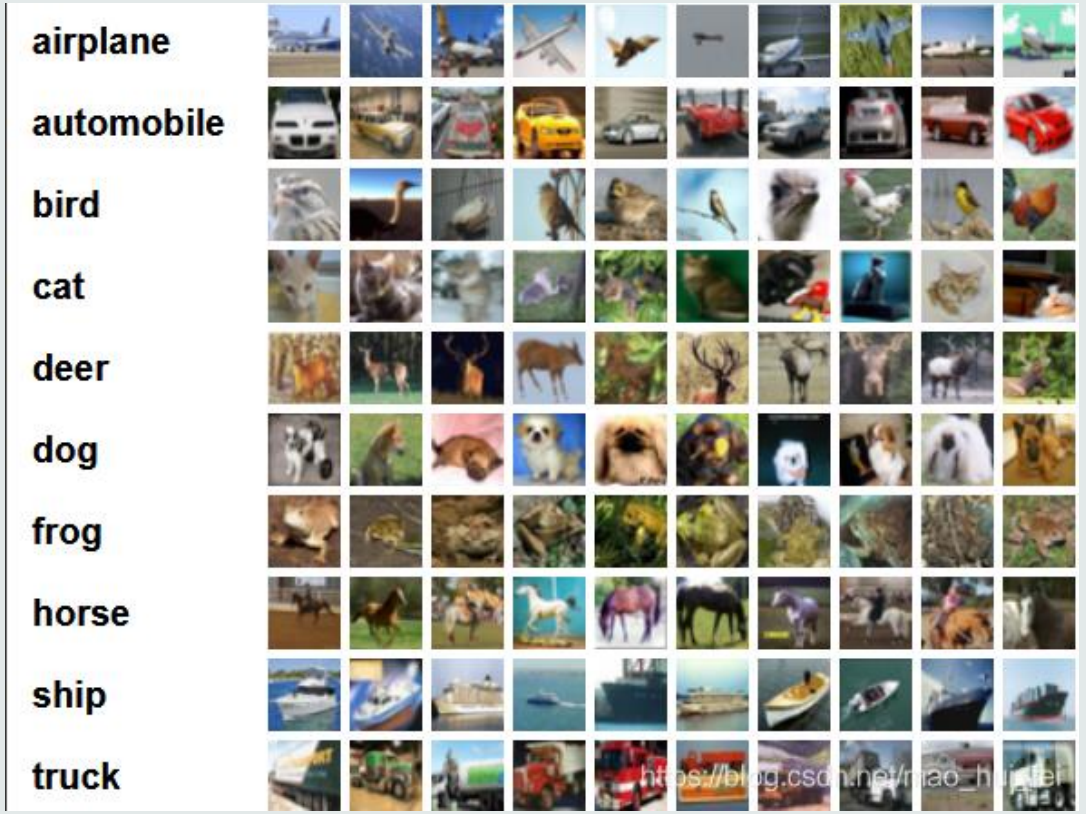
单个逻辑回归只能处理二类分类

如何实现单个模型处理多类分类?

科研实训作业

编程

编写一个多层嵌套的逻辑回归模型，并在Cifar10数据集测试分类性能



参考文献

[1] K. Kayabol, “Approximate Sparse Multinomial Logistic Regression for Classification”, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, no. 2, pp. 490-493, 2020.

课后预习

✓ 预习MOOC第4.15节课：Softmax判据