第四次作业

Exercise 4.1

下列关于线性回归分析中的残差（Residuals）说法正确的是?

A. 残差均值总是为零
B. 残差均值总是小于零
C. 残差均值总是大于零
D. 以上说法都不对

解析：线性回归分析中，目标是残差最小化。残差平方和是关于参数的函数，为了求残差极小值，令残差平方和关于参数w_0的偏导数为零，会得到残差和为零，即残差均值为零

令$\epsilon_i = y_i - w_0 - \mathbf{w}\mathbf{x}_i$

则$\sum_{i=1}^{N} \epsilon_i^2 = \sum_{i=1}^{N}(y_i - w_0 - \mathbf{w}\mathbf{x}_i)^2$

根据最小值点为极值点，得到：

$$0 = \frac{\partial \sum_{i=1}^{N} \epsilon_i^2}{\partial w_0} = \sum_{i=1}^{N} -2(y_i - w_0 - \mathbf{w}\mathbf{x}_i) = -2\sum_{i=1}^{N} \epsilon_i = -2N\sum_{i=1}^{N} \epsilon_i$$

即：$\sum_{i=1}^{N} \epsilon_i = 0$

Exercise 4.2

假如我们使用 Lasso 回归(L1正则化)来拟合数据集，该数据集输入特征有 100 个(x_1，x_2，...，x_{100})。现在，我们把其中一个特征值扩大 10 倍（例如是特征 x_1），然后用相同的正则化参数对 Lasso 回归进行修正。

那么，下列说法正确的是?

A. 特征x_1 很可能被排除在模型之外
B. 特征x_1 很可能还包含在模型之中
C. 无法确定特征x_1是否被舍弃
D. 以上说法都不对

解析：正则化的目的是减少参数的大小，$L1$ 正则化能够使不少参数为0, 从而出现稀疏化的现象。如果此时把特征值放大， 可以把对应参数缩小同样的倍数，也能达到同样的目的，此时该特征参数不一定为0. 因此特征 x_1 很可能还包含在模型之中

Exercise 4.3

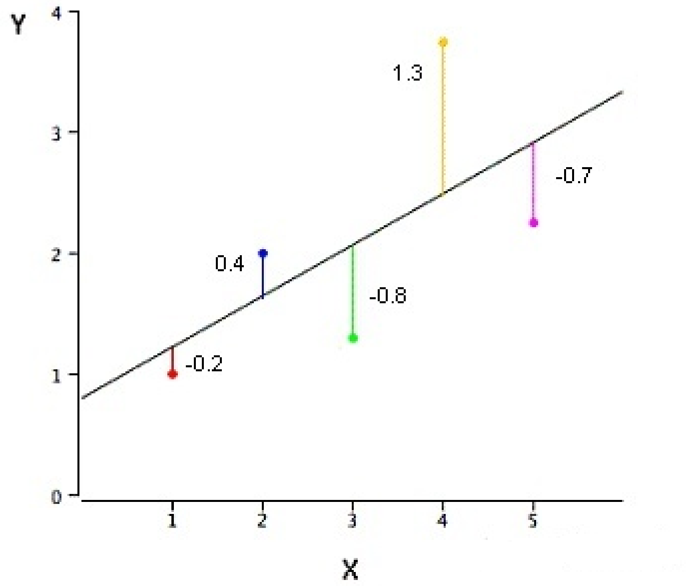
构建一个最简单的线性回归模型需要几个系数?

A. 1 个
B. 2 个
C. 3 个
D. 4 个

解析：最简单的线性回归模型为: $y = w_0 + w_1 x$

Exercise 4.4

下面这张图是一个简单的线性回归模型,图中标注了每个样本点预测值与真实值的残差。计算 SSE 为多少?



A. 3.02
B. 0.75
C. 1.01
D. 0.604

解析：SSE 是平方误差之和（Sum of Squared Error），

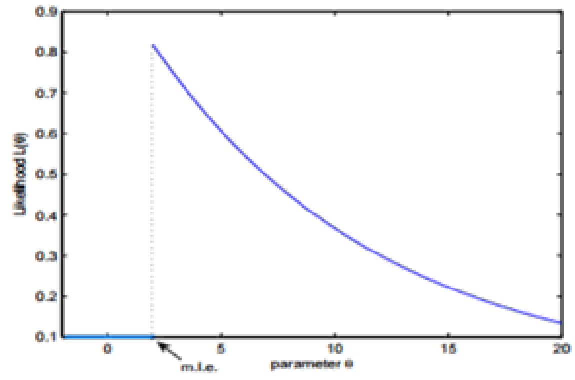
SSE= $(-0.2)^2+(0.4)^2+(-0.8)^2+(1.3)^2+(-0.7)^2=3.02$

Exercise 4.5

下列关于极大似然估计（Maximum Likelihood Estimate，MLE），说法正确的是（多选）？

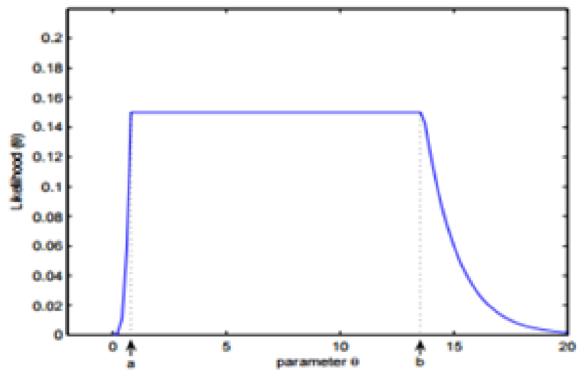
A. MLE 可能并不存在
B. MLE 总是存在
C. 如果 MLE 存在，那么它的解可能不是唯一的
D. 如果 MLE 存在，那么它的解一定是唯一的

解析：如果极大似然函数 L(θ) 在极大值处不连续，一阶导数不存在，则 MLE 不存在，如下图所示



The m.l.e. is a boundary point

另一种情况是 MLE 并不唯一，极大值对应两个 θ。如下图所示：



Any point between a and b is a m.l.e

Exercise 4.6

给出噪声分布符合0均值拉普拉斯分布，模型先验服从均值为0的高斯分布所对应的损失函数的形式？

$$L(\theta)=\frac{1}{m}\sum_{i=1}^{m}|\theta^T x_i-y_i|+\frac{\lambda}{2}||\theta||^2$$

Exercise 4.7

The weight update rule in formula $w(t+1)=w(t)+y(t)x(t)$ has the nice interpretation that it moves in the direction of classifying $x(t)$ correctly.

(a) Show that $y(t)w^T(t)x(t)<0$. [Hint: $x(t)$ is misclassifed by $w(t)$.]
(b) Show that $y(t)w^T(t+1)x(t)>y(t)w^T(t)x(t)$.
(c) As far as classifying $x(t)$ is concerned, argue that the move from $w(t)$ to $w(t+1)$ is a move 'in the right direction ' .

解析:

(a)

Since,

$h(x)=sign(w^Tx)=\begin{cases}+1\ ,\ w^Tx>0\\-1\ ,\ w^Tx<0\end{cases}$ if $x(t)$ is misclassified by $w(t)$, $h(x)$ will obtain opposite value from the label $y(t)$.

It means that when $h(x)=1, y(t)=-1$, or $h(x)=-1, y(t)=1$.

Thus,

$y(t)h(x)=y(t)w^T(t)x(t)=-1\rightarrow y(t)w^T(t)x(t)<0$.

(b)

$$\begin{aligned} y(t)w^T(t+1)x(t) &= y(t)(w(t)+y(t)x(t))^T x(t) \\ &= y(t)w^T(t)x(t) + y^2(t)x^T(t)x(t) \\ &= y(t)w^T(t)x(t) + x^T(t)x(t) \end{aligned}$$

由于$x(t)$的第一个分量为1，因此$x^T(t)x(t) \geqslant 1$，所以$x^T(t)x(t) \geqslant 1$
故$y(t)w^T(t+1)x(t) > y(t)w^T(t)x(t)$

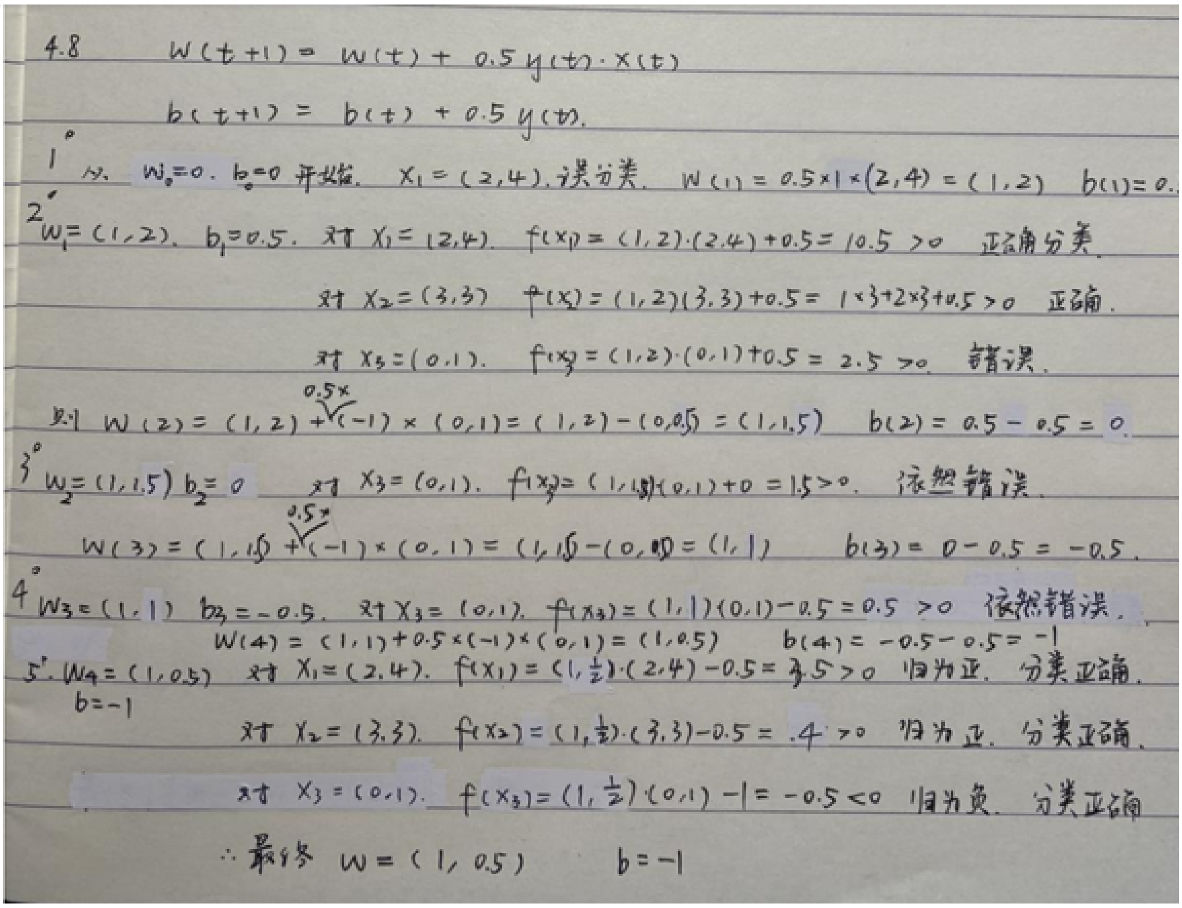
（c）由（a）可知，当分类错误时，$y(t)w^T(t)x(t) < 0$，而由（b）可知每次更新后$y(t)w^T(t+1)x(t) > y(t)w^T(t)x(t)$，
yw^Tx朝着正方向前进，因此若数据集是线性可分的，必然经过有限次更新后，使得$yw^Tx > 0$，因此前进的方向是正确的。

Exercise 4.8

已知一个训练数据集，其正实例点$x_1 = (2,4)$，$x_2 = (3,3)$；负实例点是 $x_3 = (0,1)$，试用感知机学习算法，求感知机模型$f(x) = sign(w \cdot x + b)$ (注每次的学习率为0.5)，其中损失函数为均方差。

注：按照感知机算法给出每次过程

解析：


4.8 $W(t+1) = W(t) + 0.5\, y(t) \cdot X(t)$
$b(t+1) = b(t) + 0.5\, y(t)$.
1° 以 $W_0=0$. $b_0=0$ 开始. $X_1=(2,4)$. 误分类. $W(1) = 0.5\times1\times(2,4) = (1,2)$ $b(1)=0.5$
2° $W_1=(1,2)$. $b_1=0.5$. 对 $X_1=(2,4)$. $f(X_1) = (1,2)\cdot(2,4)+0.5 = 10.5 > 0$ 正确分类.
对 $X_2=(3,3)$ $f(X_2) = (1,2)(3,3)+0.5 = 1\times3+2\times3+0.5 > 0$ 正确.
对 $X_3=(0,1)$. $f(X_3) = (1,2)\cdot(0,1)+0.5 = 2.5 > 0$. 错误.
则 $W(2) = (1,2) + 0.5\times(-1)\times(0,1) = (1,2)-(0,0.5) = (1,1.5)$ $b(2) = 0.5-0.5 = 0$.
3° $W_2=(1,1.5)$ $b_2=0$ 对 $X_3=(0,1)$. $f(X_3) = (1,1.5)(0,1)+0 = 1.5>0$. 依然错误.
$W(3) = (1,1.5) + 0.5\times(-1)\times(0,1) = (1,1.5)-(0,0.5) = (1,1)$ $b(3) = 0-0.5 = -0.5$.
4° $W_3=(1,1)$ $b_3=-0.5$. 对 $X_3=(0,1)$. $f(X_3) = (1,1)(0,1)-0.5 = 0.5 > 0$ 依然错误.
$W(4) = (1,1)+0.5\times(-1)\times(0,1) = (1,0.5)$ $b(4) = -0.5-0.5 = -1$
5° $W_4=(1,0.5)$ $b=-1$ 对 $X_1=(2,4)$. $f(X_1) = (1,\frac{1}{2})\cdot(2,4)-0.5 = 3.5>0$ 归为正. 分类正确.
对 $X_2=(3,3)$. $f(X_2) = (1,\frac{1}{2})\cdot(3,3)-0.5 = 4>0$ 归为正. 分类正确.
对 $X_3=(0,1)$. $f(X_3) = (1,\frac{1}{2})\cdot(0,1)-1 = -0.5<0$ 归为负. 分类正确
∴ 最终 $W=(1,0.5)$ $b=-1$

Exercise 4.9

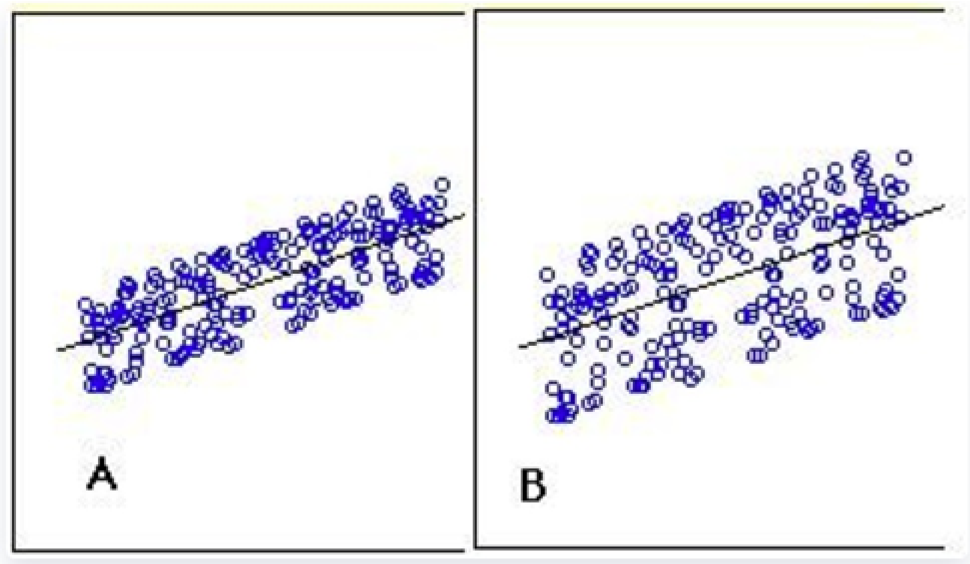
以下关于sigmoid函数的优点说法错误的是？

A. 可以压缩数据值到(0,1)之间，便于后续处理
B. 可以用于处理二分类问题
C. 函数处处连续，便于求导
D. 在深层次神经网络反馈传输中，不易出现梯度消失

解析: Sigmod求导值较小，容易出现梯度消失

Exercise 4.10

下面两张图展示两个拟合回归线（A 和 B），原始数据是随机产生的。现在，我想要计算 A 和 B 各自的残差之和。注意：两种图中的坐标尺度一样。



关于 A 和 B 各自的残差之和，下列说法正确的是?

A. A 比 B 高
B. A 比 B 小
C. A 与 B 相同
D. 以上说法都不对

解析：根据**Exercise 4.1**可知，他们的残差和均为0.

Exercise 4.11

一监狱人脸识别准入系统用来识别待进入人员的身份，此系统一共包括识别4种不同的人员：狱警，小偷，送餐员，其他。下面哪种学习方法最适合此种应用需求：

A. 回归问题　B. 二分类问题　**C. 多分类问题**　D. 聚类问题

解析：该问题需要识别 **4** 类对象，所以是多分类问题。

Exercise 4.12

以下关于分类问题的说法错误的是?

A. 回归问题在一定条件下可被转化为多分类问题
B. 分类问题输入属性必须是离散的
C. 分类属于监督学习
D. 多分类问题可以被拆分为多个二分类问题

解析：分类问题对输入没有限制，输出必须是离散的。

Exercise 4.13

以下关于逻辑回归与线性回归问题的描述错误的是

A. 逻辑回归一般要求变量服从正态分布，线性回归一般不要求
B. 逻辑回归用于处理分类问题，线性回归用于处理回归问题
C. 线性回归计算方法一般是最小二乘法，逻辑回归的参数计算方法是似然估计法。
D. 线性回归要求输入输出值呈线性关系，逻辑回归不要求

解析：逻辑回归和线性回归都对变量没有限制

Exercise 4.14

假设有三类数据，用OVR方法需要分类几次才能完成?

A. 2 B. 3 C. 1 D. 4

解析：一次逻辑回归判定一个种类，两次判定两种，同时两次都未被判定成功的则是第三种

Exercise 4.15

逻辑回归的损失函数是哪个?

A. MAE B. RMSE C. MSE **D. 交叉熵(Cross-Entropy)损失函数**

解析：按照概率模型的损失函数去推导，可得到交叉熵损失函数。

Exercise 4.16

你正在训练一个分类逻辑回归模型。以下哪项陈述是正确的?

A. 将正则化引入到模型中，总是能在训练集上获得相同或更好的性能
B. 向模型中添加新特征总是会在训练集上获得相同或更好的性能
C. 将正则化引入到模型中，对于训练集中没有的样本，总是可以获得相同或更好的性能
D. 在模型中添加许多新特性有助于防止训练集过度拟合

解析：若新特征没有被使用，则性能相同，若被使用，则增强输入和输出的相关性，性能更好

选项A一般会降低在训练集上的性能。

选项C 引入正则化，能够防止过拟合，降低泛化误差，但不是每个测试样本的性能都能提高。

选项D，添加新特征，反而更容易过拟合。

Exercise 4.17

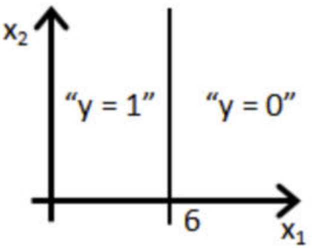
假设您进行了两次逻辑回归，一次是λ=0，一次是λ=1（λ是正则化参数）。其中一次，得到参数w=[81.47，12.69]，另一次，得w=[13.0，10.91]。 但是，您忘记了哪个λ值对应于哪个w值。你认为哪个对应于λ=1?

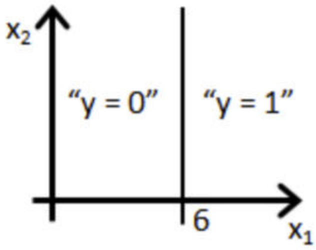
A. w=[13.0，10.91]
B. w=[81.47，12.69]

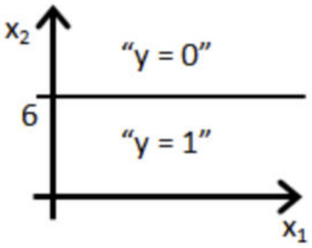
解析：正则化参数 λ 越大，则代表模型本身的参数在损失函数中比重越重，为减小loss值，λ 越大训练出来的模型参数越小

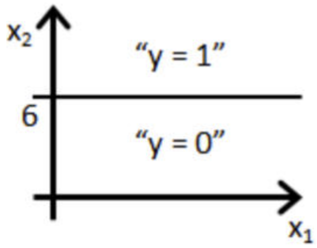
Exercise 4.18

假设训练一个逻辑回归分类器$h_{\mathbf{w}}(\mathbf{x})=\theta(w_0+w_1x_1+w_2x_2)$。假设$w_0=6, w_1=-1, w_2=0$，下列哪个图表示分类器找到的决策边界?

A. 

B. 

C. 

D. 

解析：选择特殊点去测试，可选$(0,0),(7,0)$，满足要求，故选 A.

Exercise 4.19

[**不定项选择题**] 假设您有以下训练集，并拟合logistic回归分类器

$$h_{\mathbf{w}}(\mathbf{x})=g(w_0+w_1x_1+w_2x_2)$$

x_1	x_2	y
1	0.5	0
1	1.5	0
2	1	1
3	1	0

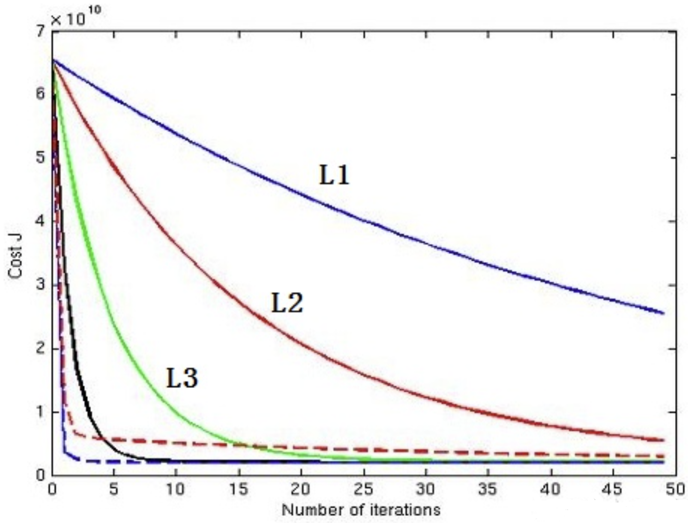
以下哪项是正确的？选出所有正确项

A. 添加多项式特征（例如，使用$h_{\mathbf{w}}(\mathbf{x})=g(w_0+w_1x_1+w_2x_2+w_3x_1^2+w_4x_1x_2+w_5x_2^2)$)可以增加我们拟合训练数据的程度
B. 在$\mathbf{w}$的最佳值处，$J(\mathbf{w})\geq 0$
C. 添加多项式特征（例如，使用$h_{\mathbf{w}}(\mathbf{x})=g(w_0+w_1x_1+w_2x_2+w_3x_1^2+w_4x_1x_2+w_5x_2^2)$)将增加$J(\mathbf{w})$，因为我们现在正在对更多项进行求和
D.如果我们训练梯度下降迭代足够多次，对于训练集中的一些例子$\mathbf{x}_i$，可能得到$h_{\mathbf{w}}(\mathbf{x}_i)>1$

解析：适当增强模型的复杂度可以加强模型的表达能力；

Exercise 4.20

如图显示了逻辑回归中3种不同学习速率值的代价函数和迭代次数之间的关系，L_1、L_2、L_3为对应的学习速率，下面哪一个选项是正确的？



A、L1> L2> L3　　B、L1 = L2 = L3

C、L1 < L2 < L3　　D、都不是

解析：学习速率越大，代价函数下降的越快。从图中可以看出，相同迭代次数下，绿色曲线的代价函数下降的最多，故L3最大，答案选C。

Exercise 4.21

为什么要使用：

a. 岭回归而不是简单的线性回归（即没有任何正则化）？

b. Lasso而不是岭回归?

c. 弹性网络而不是Lasso?

解析： 具有某些正则化的模型通常比没有任何正则化的模型要好，因此，你通常应优先选择岭回归而不是简单的线性回归。

Lasso回归使用 1惩罚，这通常会将权重降低为零。这将导致稀疏模型，其中除了最重要的权重之外，所有权重均为零。这是一种自动进行特征选择的方法，如果你怀疑实际上只有很少的特征很重要，那么这是一种很好的方法。如果你不确定，则应首选岭回归。

与Lasso相比，弹性网络通常更受青睐，因为Lasso在某些情况下可能产生异常（当几个特征强相关或当特征比训练实例更多时）。但是，它确实增加了额外需要进行调整的超参数。如果你希望Lasso没有不稳定的行为，则可以仅使用l1_ratio接近1的弹性网络。

Exercise 4.22

训练逻辑回归模型时，梯度下降会卡在局部最小值中吗?

解析：训练逻辑回归模型时，梯度下降不会陷入局部最小值，因为成本函数是凸函数