14 逻辑回归

Moving Beyond Linearity

既是回归模型，又是分类模型

毫无争议的是，人类无法毫无错误地判断事物的真伪，我们能做就是遵循更大的可能性。

It is truth very certain that, when it is not in one's power to determine what is true, we ought to follow what is more probable.

—— 勒内·笛卡尔 (René Descartes) | 法国哲学家、数学家、物理学家 | 1596 ~ 1650

- `scipy.special.expit()`
- `sklearn.linear_model.LogisticRegression()` 逻辑回归函数，也可以用来分类
- `seaborn.kdeplot()` 绘制概率密度估计曲线
- `seaborn.scatterplot()` 绘制散点图
- `seaborn.jointplot()` 绘制联合分布/散点图和边际分布
- `matplotlib.pyplot.plot_wireframe()` 绘制线框图
- `matplotlib.pyplot.contour()` 绘制等高线线图
- `matplotlib.pyplot.contourf()` 绘制填充等高线图
- `matplotlib.pyplot.scatter()` 绘制散点图

14.1 逻辑函数

图 1 给出一组数据的散点图，取值为 1 的数据点被标记为蓝色，取值为 0 的数据点被标记为红色。图 2 给出三种可以描述红蓝散点数据的函数。线性函数显然不适合这一问题。阶跃函数虽然可以捕捉函数从 0 到 1 的跳变，但是函数本身不光滑。逻辑函数似乎能够胜任描述红蓝三点数据的任务。线性函数的因变量一般为连读数据；而逻辑函数的因变量为离散数值，即分类数据。

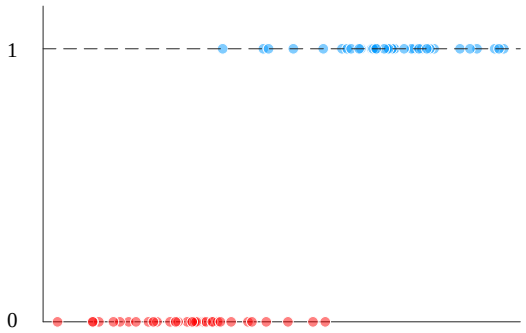


图 1. 红蓝数据的散点图

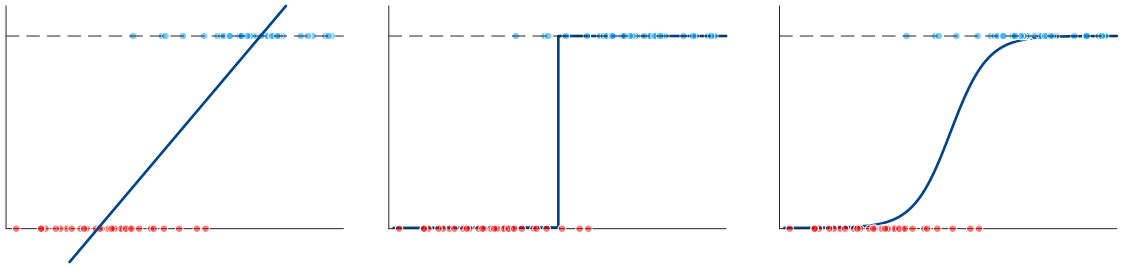


图 2. 可以描述红蓝数据的函数

逻辑函数

回顾《数学要素》12 章讲过的逻辑函数。最简单的逻辑函数：

$$f(x)=\frac{1}{1+e^{-x}}=\frac{e^{x}}{1+e^{x}} \tag{1}$$

更一般的一元逻辑函数：

$$f(x)=\frac{1}{1+\exp\left(-\left(b_0+b_1x\right)\right)} \tag{2}$$

图 3 所示为 b_1 影响一元逻辑函数图像的陡峭程度。图中，$b_0 = 0$。可以发现函数呈现 S 形，取值范围在 [0, 1] 之间；函数在左右两端无限接近 0 或 1。函数的这一性质，方便从概率角度解释，这是下一节要介绍的内容。

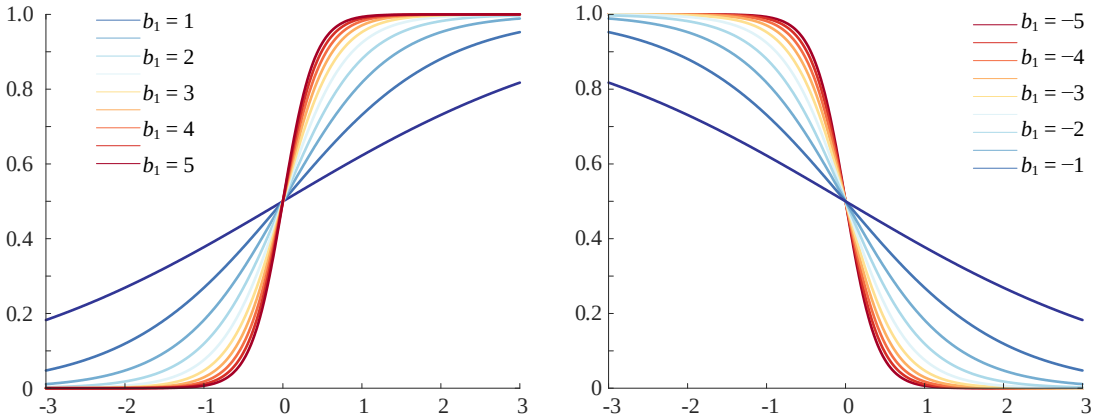


图 3. b_1 影响一元逻辑函数图像的陡峭程度

找到 $f(x) = 1/2$ 位置：

$$f(x) = \frac{1}{1+\exp\left(-\left(b_0 + b_1 x\right)\right)} = \frac{1}{2} \tag{3}$$

整理得到 $f(x) = 1/2$ 对应的 x 值：

$$x = -\frac{b_0}{b_1} \tag{4}$$

也就是当 b_1 确定时，b_0 决定逻辑函数位置。注意，图 4 中，$b_1 = 0$。

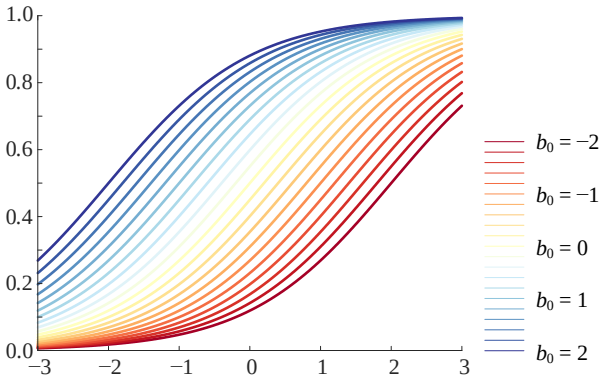


图 4. b_0 决定逻辑函数位置，$b_1 = 0$

图 5 所示为根据数据的分布，选取不同的逻辑函数参数。

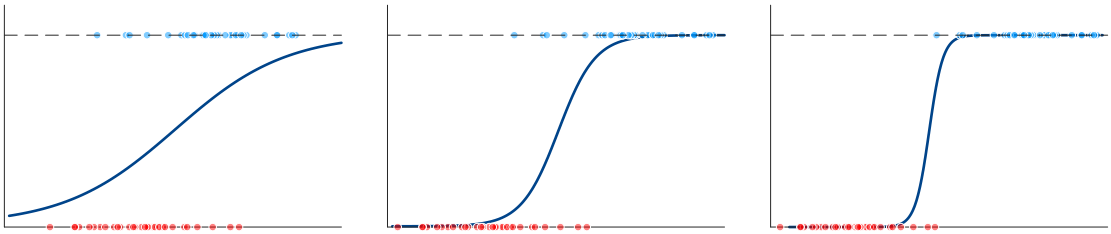


图 5. 根据数据的分布，选取不同的逻辑函数参数

Bk6_Ch14_01.py 绘制逻辑函数图像。

多元

对于多元情况，逻辑函数的一般式如下：

$$f\left(x_1, x_2, ..., x_D\right) = \frac{1}{1+\exp\left(-\left(b_0 + b_1 x_1 + b_2 x_2 + \cdots b_D x_D\right)\right)} \tag{5}$$

利用矩阵运算表达多元逻辑函数：

$$f\left(\boldsymbol{x}\right) = \frac{1}{1+\exp\left(-\boldsymbol{b}^{\mathrm{T}}\boldsymbol{x}\right)} \tag{6}$$

其中

$$\begin{aligned} \boldsymbol{x} &= \begin{bmatrix} 1 & x_1 & x_2 & \cdots & x_D \end{bmatrix}^{\mathrm{T}} \\ \boldsymbol{b} &= \begin{bmatrix} b_0 & b_1 & b_2 & \cdots & b_D \end{bmatrix}^{\mathrm{T}} \end{aligned} \tag{7}$$

令

$$s\left(\boldsymbol{x}\right) = \boldsymbol{b}^{\mathrm{T}}\boldsymbol{x} = b_0 + b_1 x_1 + b_2 x_2 + \cdots b_D x_D \tag{8}$$

(6) 可以记做：

$$f\left(s\right) = \frac{1}{1+\exp\left(-s\right)} \tag{9}$$

(8) 相当于是线性回归，经过如 (9) 逻辑函数映射，得到逻辑回归。图 6 所示为逻辑回归和线性回归之间关系。图 6 这幅图已经让我们看到神经网络的一点影子，逻辑函数 $f(s)$ 类似激活函数 (activation function)。

特别地，对于二元逻辑函数：

$$f\left(x_1,x_2\right)=\frac{1}{1+\exp\left(-\left(b_0+b_1x_1+b_2x_2\right)\right)} \tag{10}$$

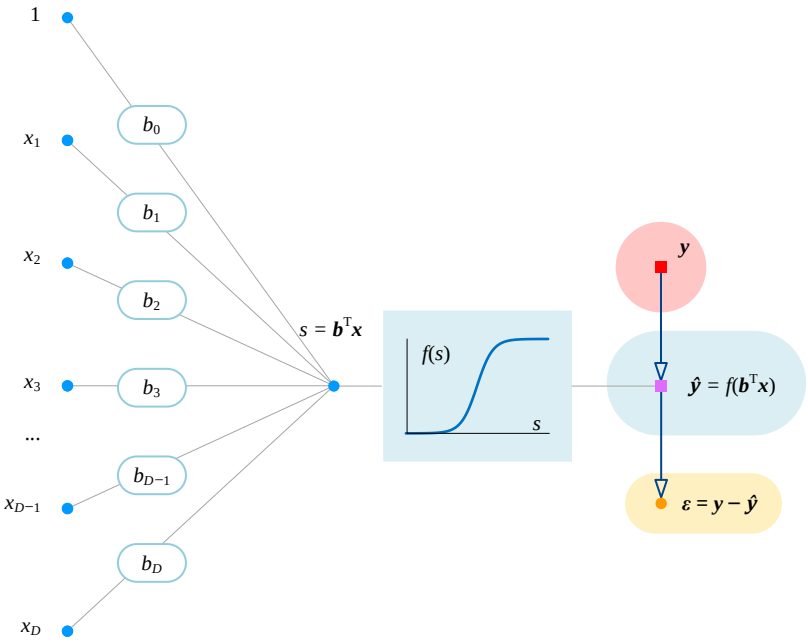


图 6. 逻辑回归和线性回归之间关系

14.2 概率视角

形似 (2) 是逻辑分布的 CDF 曲线，对应的表达式：

$$F\left(x|\mu,s\right)=\frac{1}{1+\exp\left(\frac{-\left(x-\mu\right)}{s}\right)}=\frac{1}{2}+\frac{1}{2}\tanh\left(\frac{x-\mu}{2s}\right) \tag{11}$$

其中，μ 为位置参数，s 为形状参数。注意，对于逻辑分布，$s>0$。

逻辑回归可以用来解决二分类，标签为 0 或 1；这是因为逻辑回归可以用来估计事件发生的可能性。

标签为 1 对应的概率为：

$$\Pr\left(y=1|x\right)=\frac{1}{1+\exp\left(-\left(b_0+b_1x\right)\right)} \tag{12}$$

标签为 0 对应的概率为：

$$\Pr\left(y=0|x\right)=1-\Pr\left(y=1|x\right)=\frac{\exp\left(-\left(b_0+b_1x\right)\right)}{1+\exp\left(-\left(b_0+b_1x\right)\right)} \tag{13}$$

图 7 所示为标签为 1 和为 0 的概率关系。

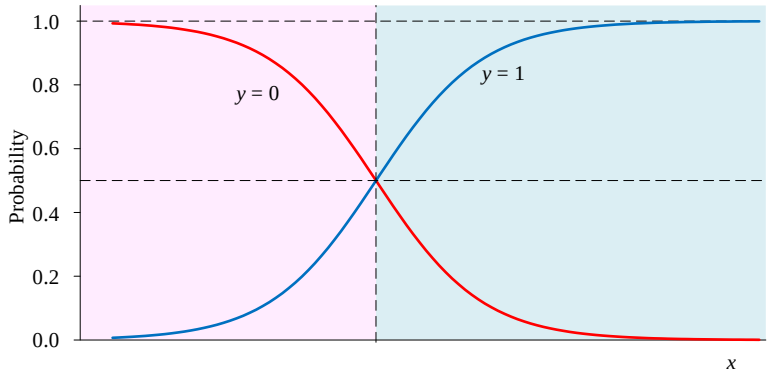


图 7. 标签为 1 和为 0 的概率关系

显然，对于二分类问题，对于任意一点 x，标签为 1 的概率和标签为 0 的概率相加为 1：

$$\mathrm{P}\left(y=0|x\right)+\mathrm{P}\left(y=1|x\right)=1 \tag{14}$$

白话说，某一点要么标签为 1，要么标签为 0，如图 8 所示。

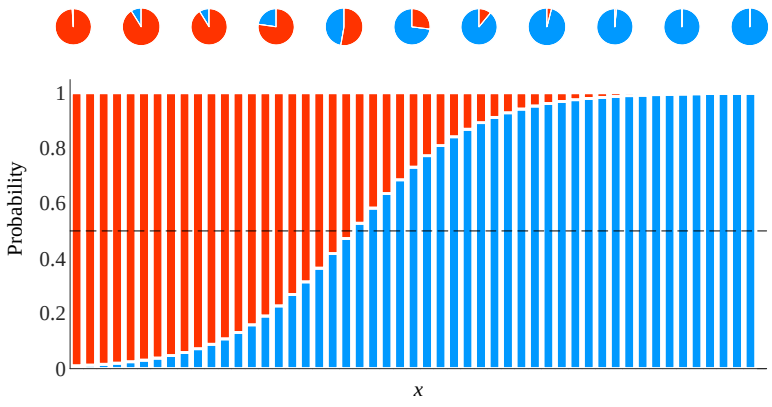


图 8. 逻辑回归模型用于二分类问题

优势率 (odds ratio, OR)，比值比；缩写词为 OR 的对数值：

$$\mathrm{OR}=\text{odds ratio}=\frac{\Pr\left(y=1|x\right)}{\Pr\left(y=0|x\right)}=\frac{1}{\exp\left(-\left(b_0+b_1x\right)\right)} \tag{15}$$

分界 OR = 1，两者概率相同：

$$\frac{1}{\exp\left(-\left(b_0+b_1x\right)\right)}=1 \tag{16}$$

整理得到：

$$b_0+b_1x=0 \tag{17}$$

即

$$x=-\frac{b_0}{b_1} \tag{18}$$

本章后文介绍如何用 sklearn 中逻辑回归函数解决三分类问题。

14.3 单特征分类

本节介绍用 sklearn.linear_model.LogisticRegression() 逻辑回归模型，根据鸢尾花花萼长度这一单一特征数据进行分类。

图 9 所示为鸢尾花花萼长度数据和真实三分类 y 之间关系。

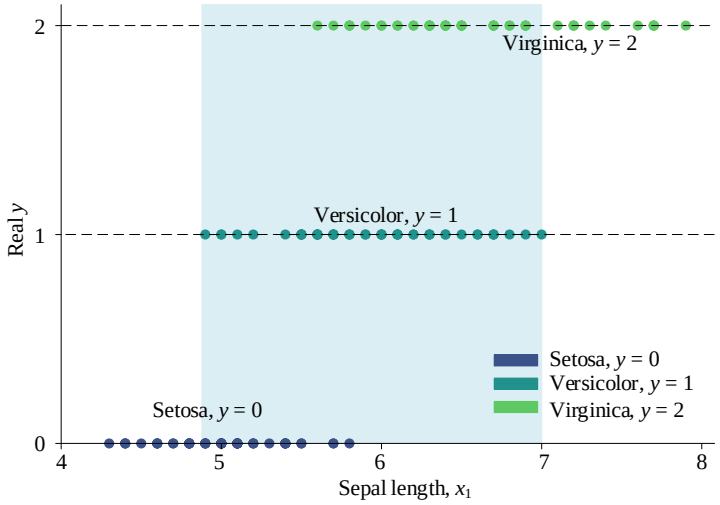


图 9. 鸢尾花花萼长度和真实分类之间关系

图 10 所示为鸢尾花花萼长度数据分类概率密度估计。这幅图实际上已经能够透露出比较合适的分类区间。

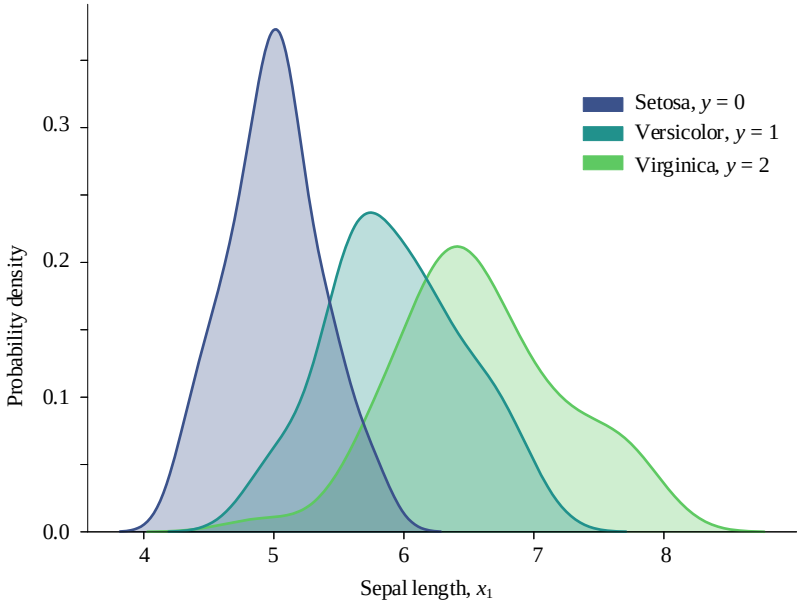


图 10. 鸢尾花花萼长度数据分类概率密度估计

sklearn.linear_model.LogisticRegression() 模型结果可以输出各个分类的概率，得到的图像如图 11 所示。比较三个类别的概率，可以进行分类预测。

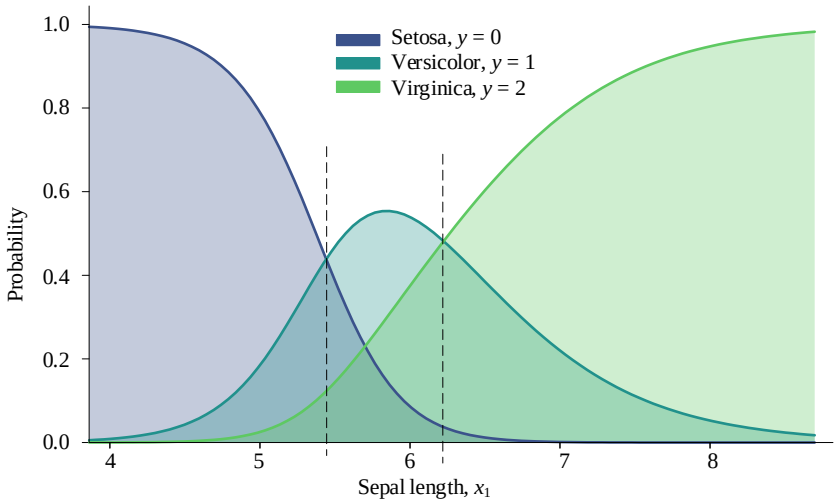


图 11. 逻辑回归估算得到的分类概率

图 12 所示为鸢尾花分类预测结果。

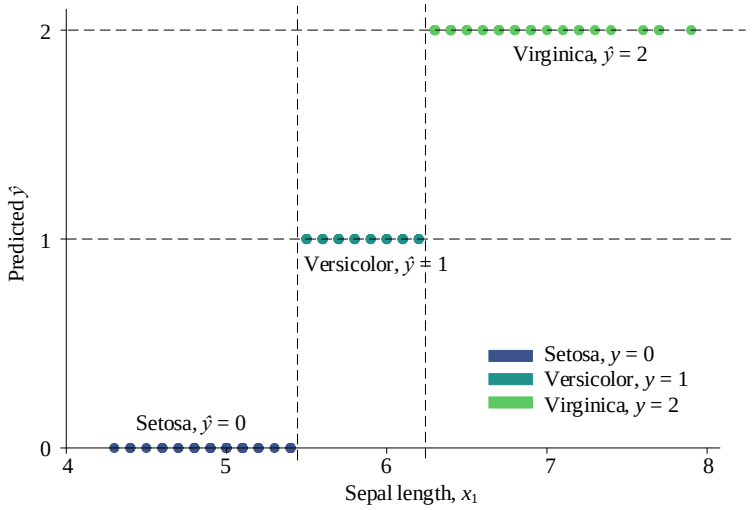


图 12. 鸢尾花花萼长度和预测分类之间关系

Bk6_Ch14_02.py 绘制本节图像。

14.4 双特征分类

本节介绍用 sklearn.linear_model.LogisticRegression() 逻辑回归模型，根据鸢尾花花萼长度和花萼宽度这两个特征数据进行分类。

图 13 所示为鸢尾花花萼长度和花萼宽度两个特征数据散点图，和分类边际分布概率密度估计曲线。

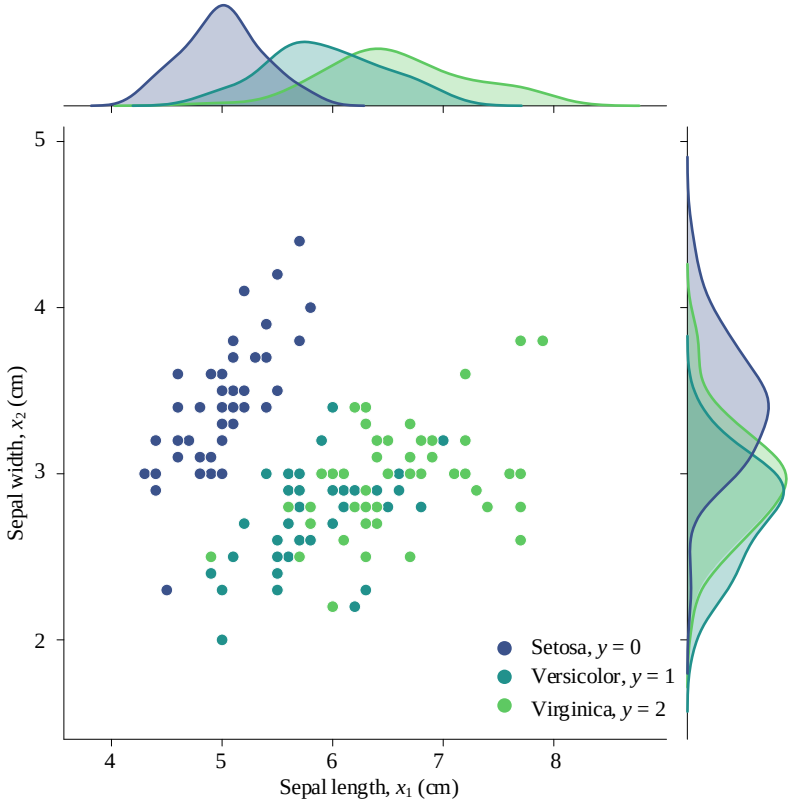


图 13. 鸢尾花双特征数据和分类边际分布

图 14 ~ 图 16 三幅图分别给出鸢尾花双特征分类概率预测曲面。比较三个曲面高度可以得到分类决策边界。在分类问题中，决策边界 (decision boundary) 指的是将不同类别样本分开的平面或曲面。

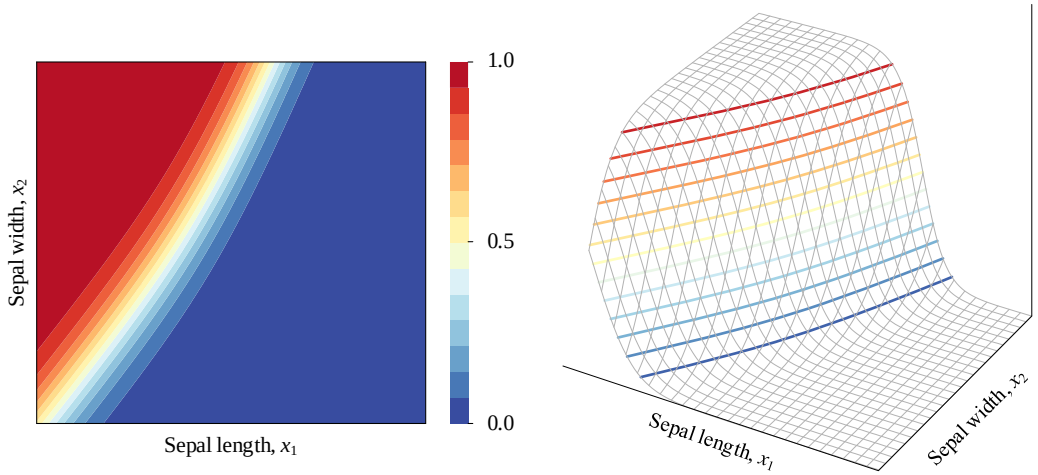


图 14. 鸢尾花双特征分类预测，$\hat{y} = 0$

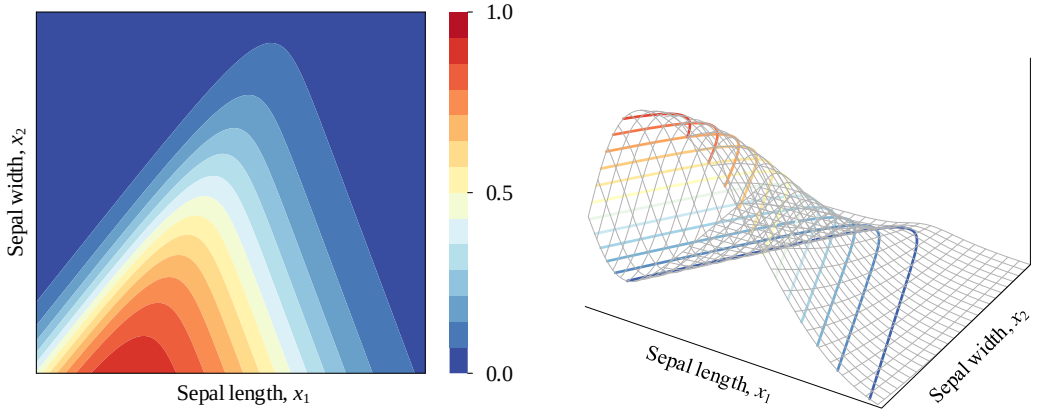


图 15. 鸢尾花双特征分类预测，$\hat{y} = 1$

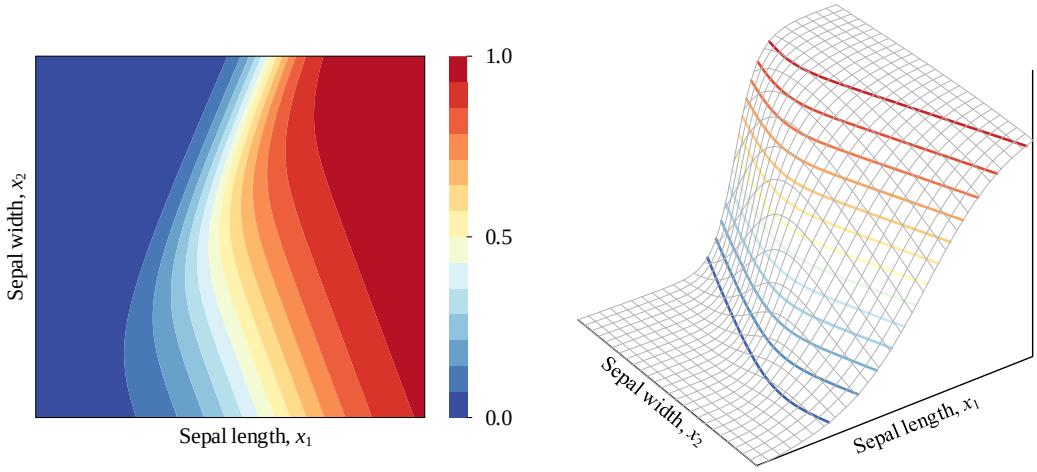


图 16. 鸢尾花双特征分类预测，$\hat{y} = 2$

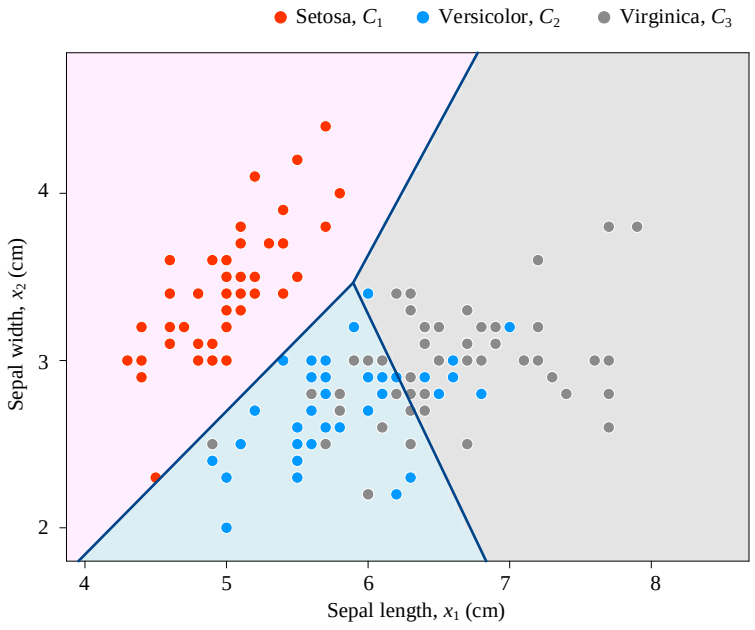
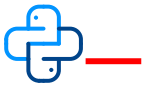


图 17. 利用逻辑回归得到的分类决策边界



Bk6_Ch14_03.py 绘制本节图像。

下例介绍在逻辑回归中引入 L1 正则项，并绘制系数轨迹。

https://scikit-learn.org/stable/auto_examples/linear_model/plot_logistic_path.html