

2-1 机器学习世界的的数据

• Part-1

数据

萼片长度	萼片宽度	花瓣长度	花瓣宽度	种类
5.1	3.5	1.4	0.2	se (0)
7.0	3.2	4.7	1.4	ve (1)
6.3	3.3	6	2.5	vi (2)

- 数据整体叫数据集 (data set)
- 每一行数据称为一个样本(sample)
- 除最后一列，每一列表达样本的一个特征(feature)
- 最后一列，称为标记(label)

X

y

第i个样本行写作 $X^{(i)}$ 第i个样本第j个特征值 $X_j^{(i)}$ 第i个样本的标记写作 $y^{(i)}$

课程

• Part-2

数据

萼片长度	萼片宽度	花瓣长度	花瓣宽度
5.1	3.5	1.4	0.2
7.0	3.2	4.7	1.4
6.3	3.3	6	2.5

→ 特征

→ 特征向量 $X^{(i)}$

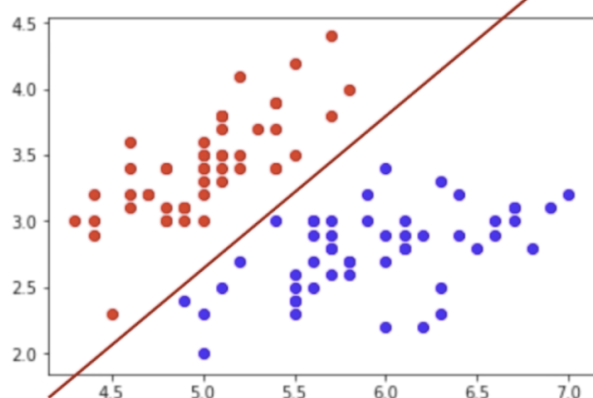
$$\begin{pmatrix} (X^{(1)})^T \\ (X^{(2)})^T \\ (X^{(3)})^T \\ \dots \end{pmatrix}$$

$$\begin{pmatrix} 5.1 \\ 3.5 \\ 1.4 \\ 0.2 \end{pmatrix}$$

课程

• Part-3

数据



- 特征空间 (feature space)
- 分类任务本质就是在特征空间切分
- 在高维空间同理

• Part-4

特征可以很抽象



- 图像，每一个像素点都是特征
- 28×28 的图像有 $28 \times 28 = 784$ 个特征
- 如果是彩色图像特征更多

- 特征工程
- 深度学习完全可以理解成算法来帮助我们完成特征工程