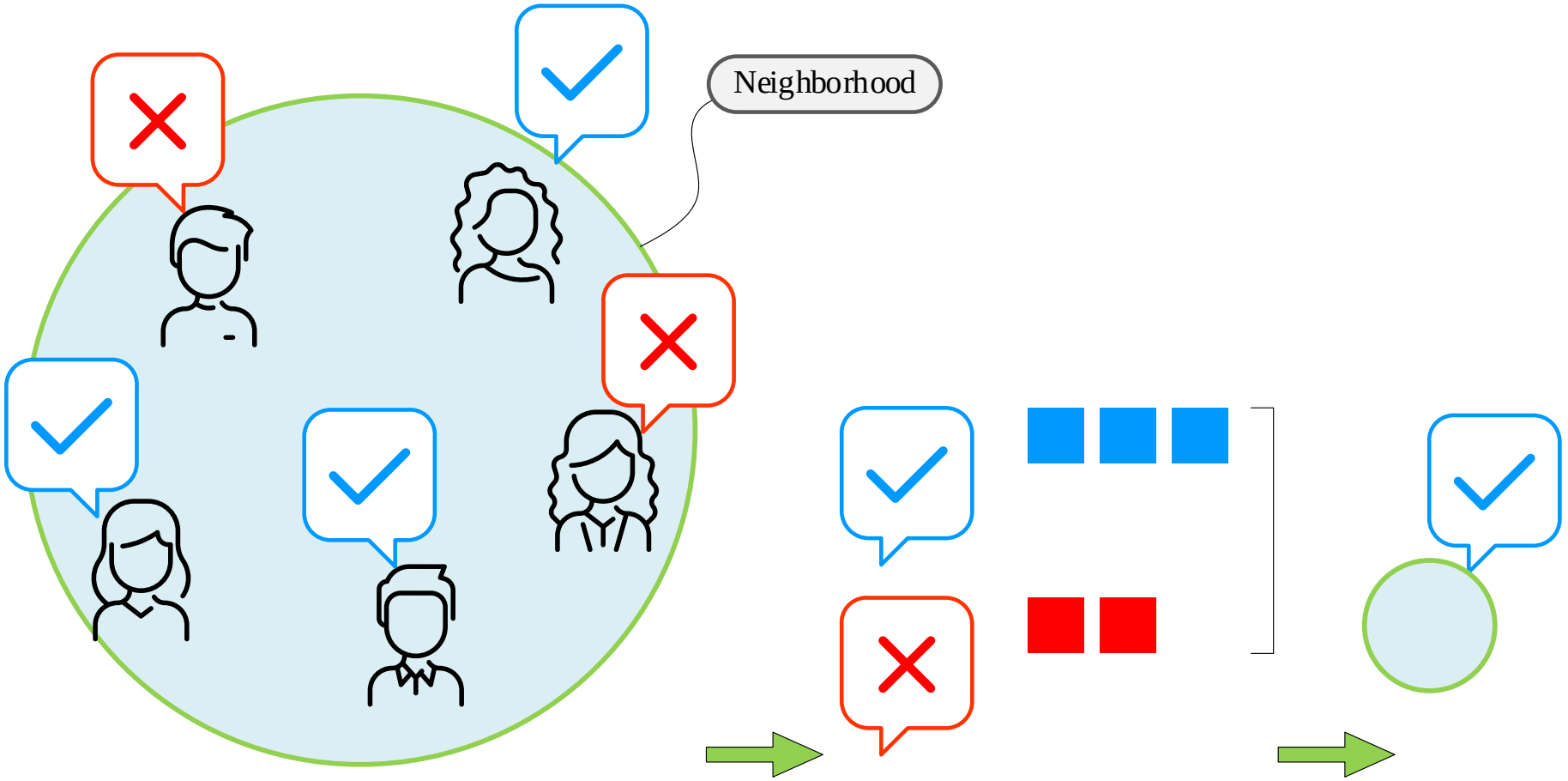
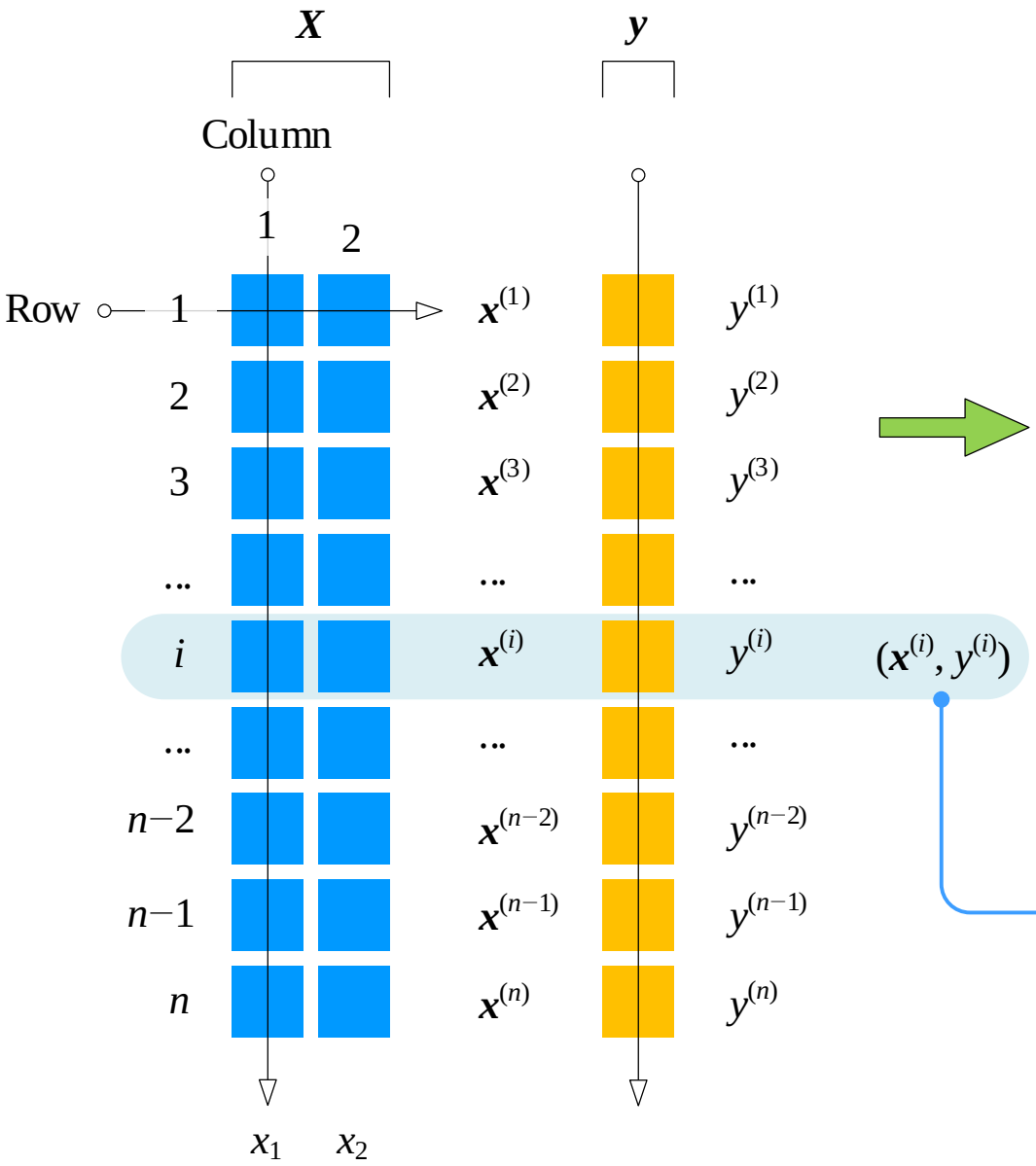
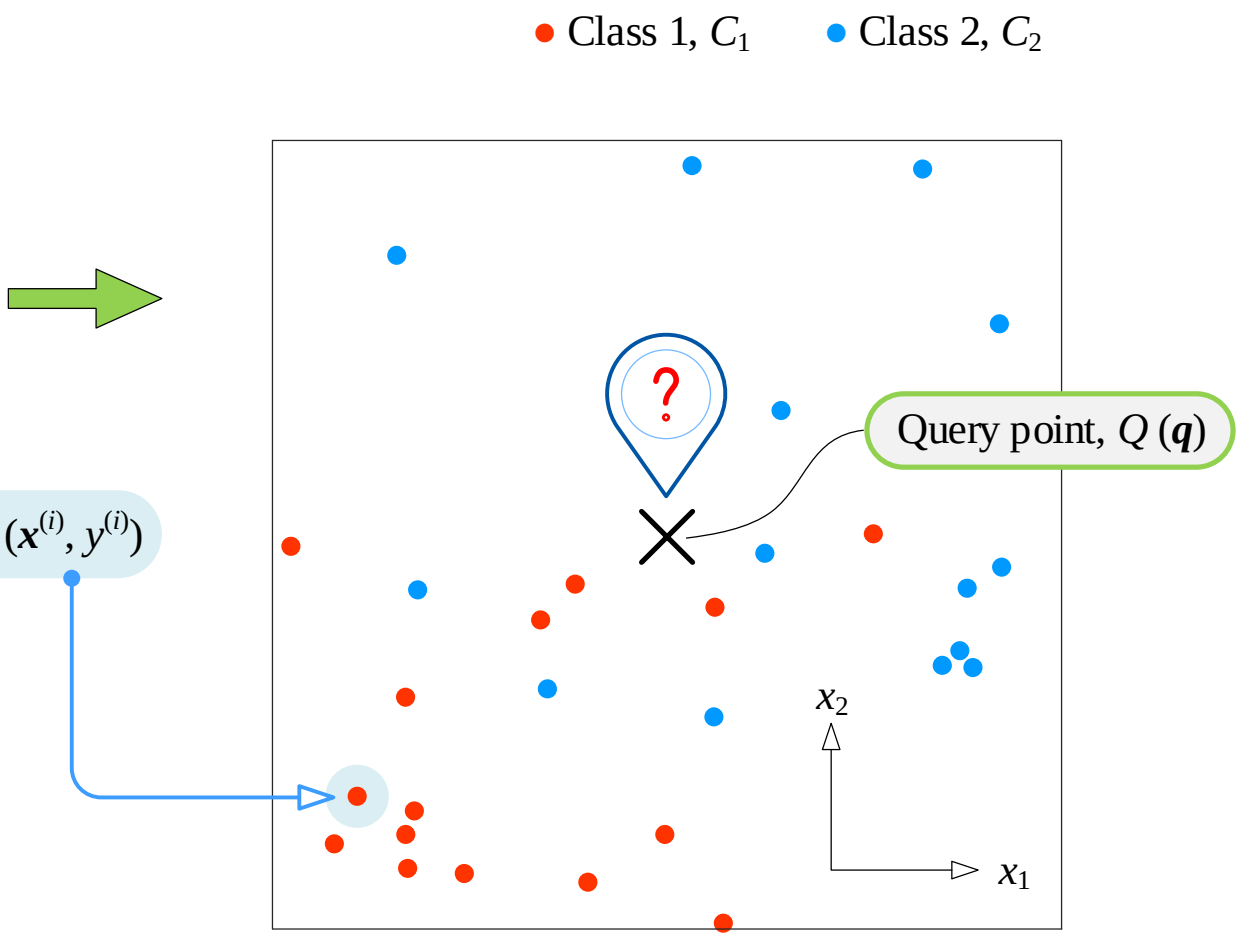
k临近分类kNN——小范围投票，少数服从多数



样本数据及可视化



$(x^{(i)}, y^{(i)})$



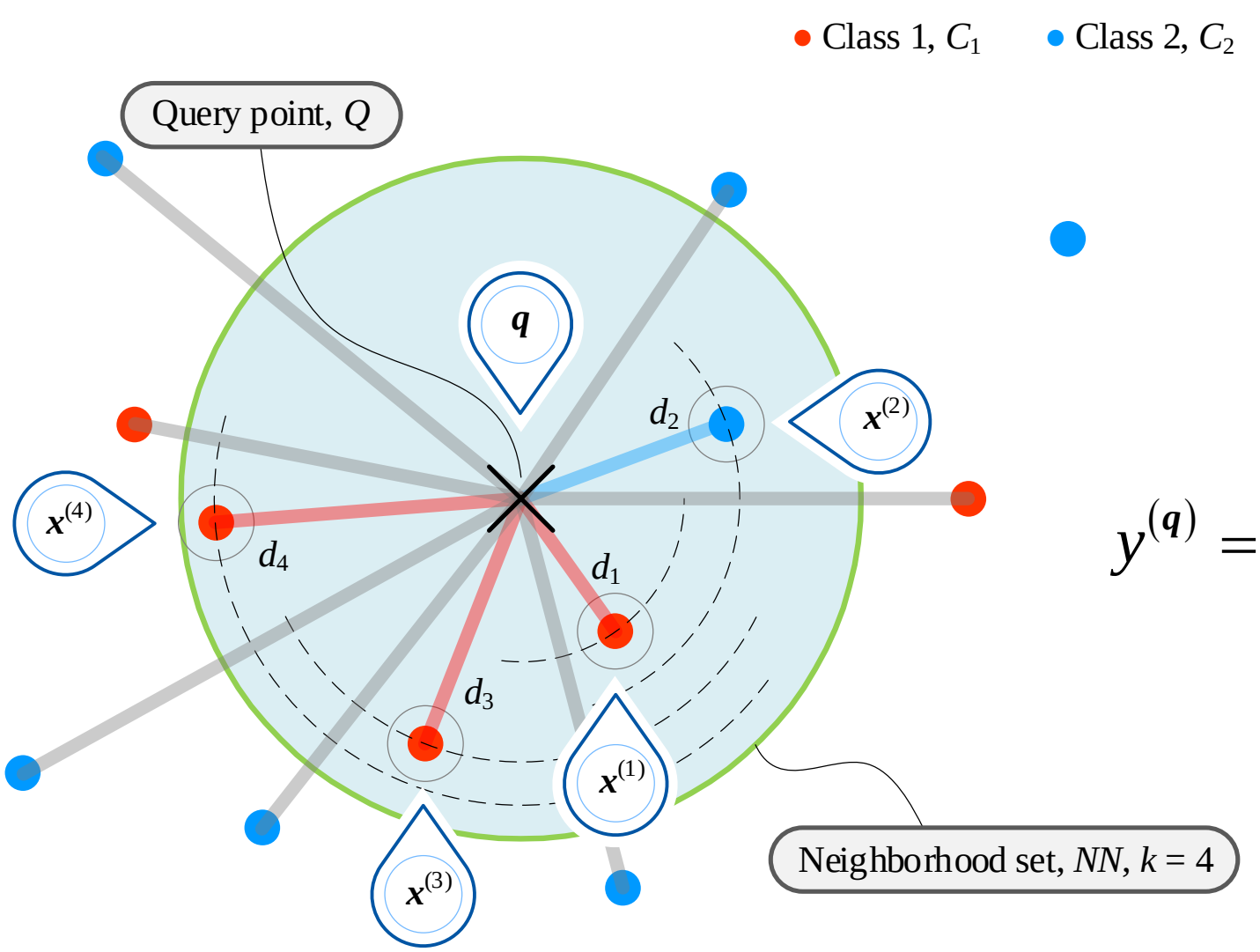
分类问题

$$y^{(q)} = \arg\max_{C_l} \sum_{i \in kNN(q)} I\left(y^{(i)} = C_l\right)$$

$$K = 2$$

$$y^{(q)} = \max_{C_1, C_2} \left\{ \sum_{i \in kNN(q)} I\left(y^{(i)} = C_1\right), \sum_{i \in kNN(q)} I\left(y^{(i)} = C_2\right) \right\}$$

k临近原理



$$\sum_{i \in kNN(q)} I\left(y^{(i)} = C_1\right) = 3_{(C_1)}$$

$$\sum_{i \in kNN(q)} I\left(y^{(i)} = C_2\right) = 1_{(C_2)}$$

$$y^{(q)} = \max_{C_1, C_2} \left\{3_{(C_1)}, 1_{(C_2)}\right\} = C_1$$

sklearn

- `sklearn.neighbors.KNeighborsClassifier`
- 函数默认的近邻数量`n_neighbors`为5，默认距离度量`metric`为欧氏距离 (Euclidean distance)。
- 这个函数常用的mehods为`fit(X, y)` 和 `predit(q)`;
- `fit(X, y)`用来加载样本数据，`predit(q)`用来预测查询点***q***的分类

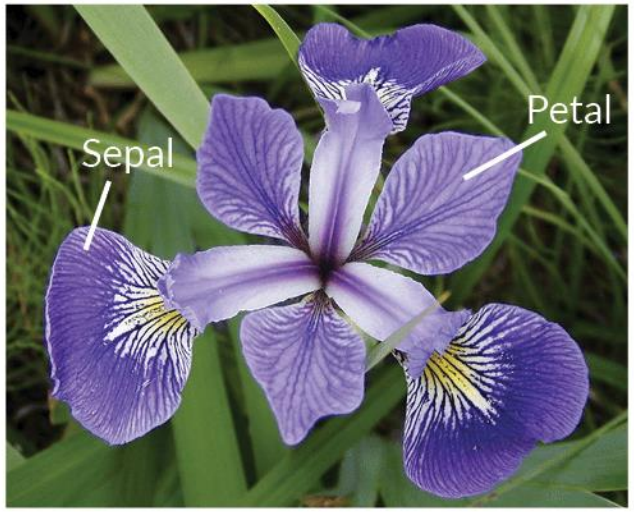
Scikit-learn（曾叫做scikits.learn还叫做sklearn）是用于Python编程语言的自由软件机器学习库。它的特征是具有各种分类、回归和聚类算法

https://en.wikipedia.org/wiki/Scikit-learn

鸢尾花数据

数据集包含了150个样本，都属于鸢尾属下的三个亚属，分别是山鸢尾、变色鸢尾和维吉尼亚鸢尾。四个特征被用作样本的定量分析，它们分别是花萼和花瓣的长度和宽度。

Iris Setosa



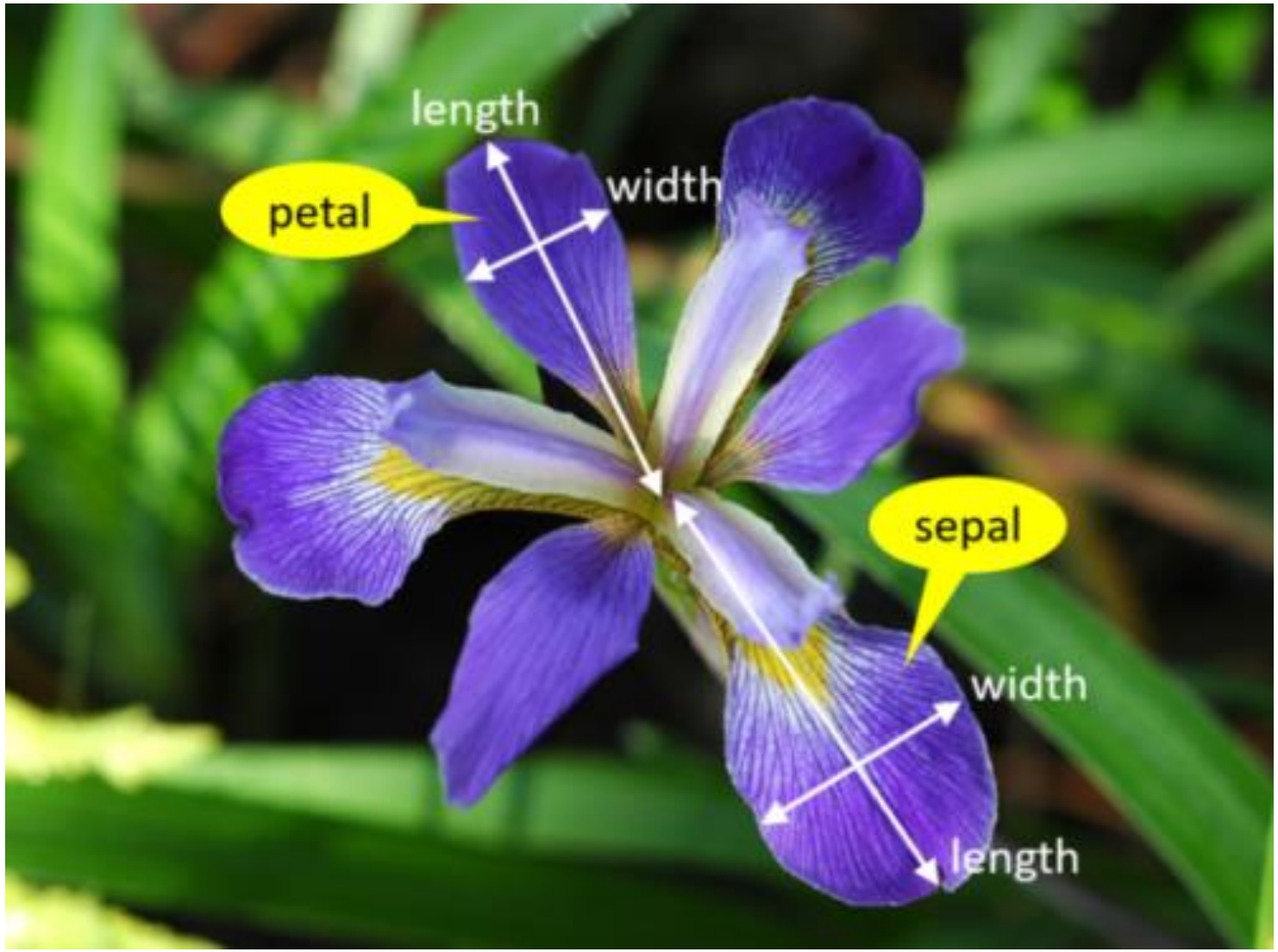
Iris Versicolor

Iris Virginica

- Setosa, C_1
- Versicolor, C_2
- Virginica, C_3

https://en.wikipedia.org/wiki/Iris_flower_data_set

鸢尾花数据



鸢尾花数据

度量单位为厘米：

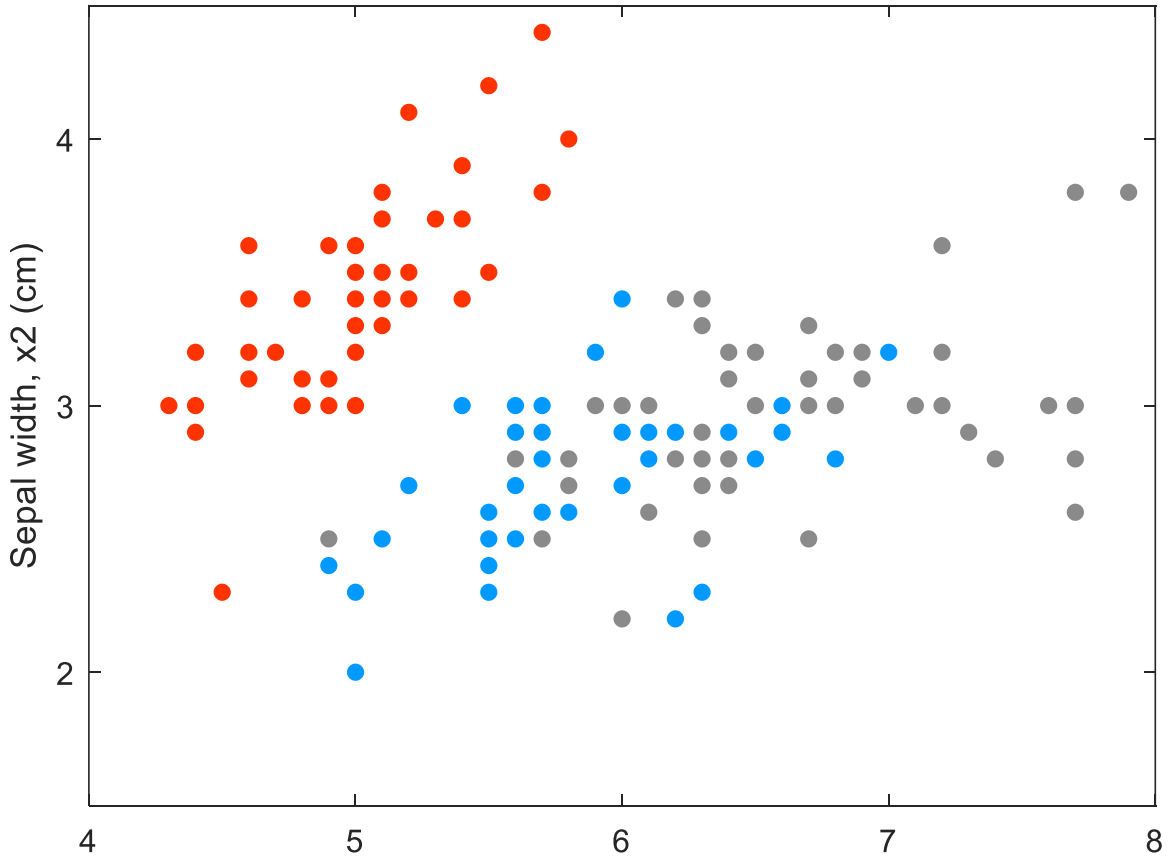
费雪鸢尾花卉数据集

花萼长度	花萼宽度	花瓣长度	花瓣宽度	属种
5.1	3.5	1.4	0.2	*setosa*
4.9	3.0	1.4	0.2	*setosa*
4.7	3.2	1.3	0.2	*setosa*
4.6	3.1	1.5	0.2	*setosa*
5.0	3.6	1.4	0.2	*setosa*
5.4	3.9	1.7	0.4	*setosa*
4.6	3.4	1.4	0.3	*setosa*

https://en.wikipedia.org/wiki/Iris_flower_data_set

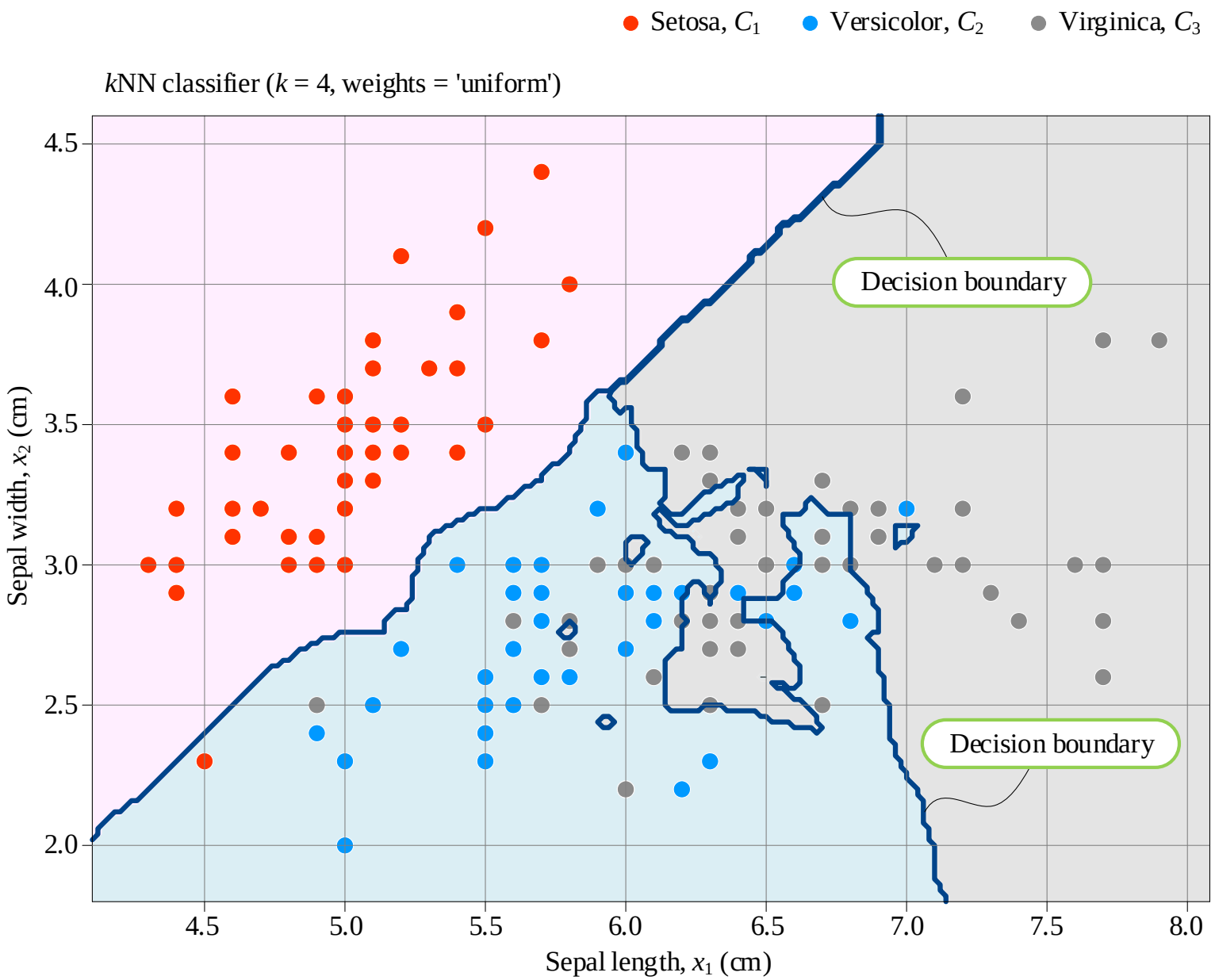
鸢尾花数据


```
# import the iris data
iris = datasets.load_iris()

# Only use the first two features: sepal length, sepal width
X = iris.data[:, :2]

# Vector of labels
y = iris.target
```



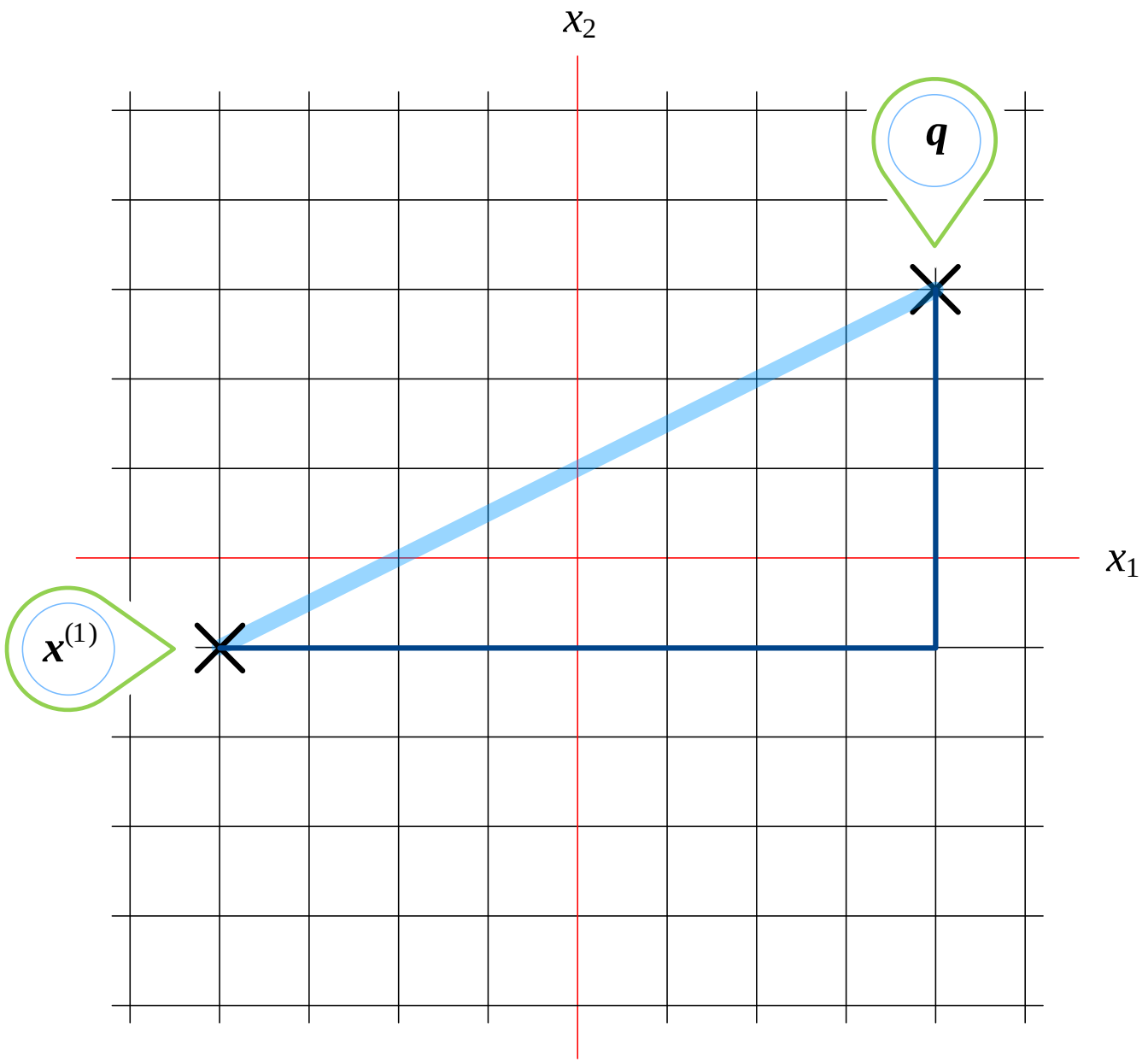
● Setosa, C_1 ● Versicolor, C_2 ● Virginica, C_3

k临近分类结果

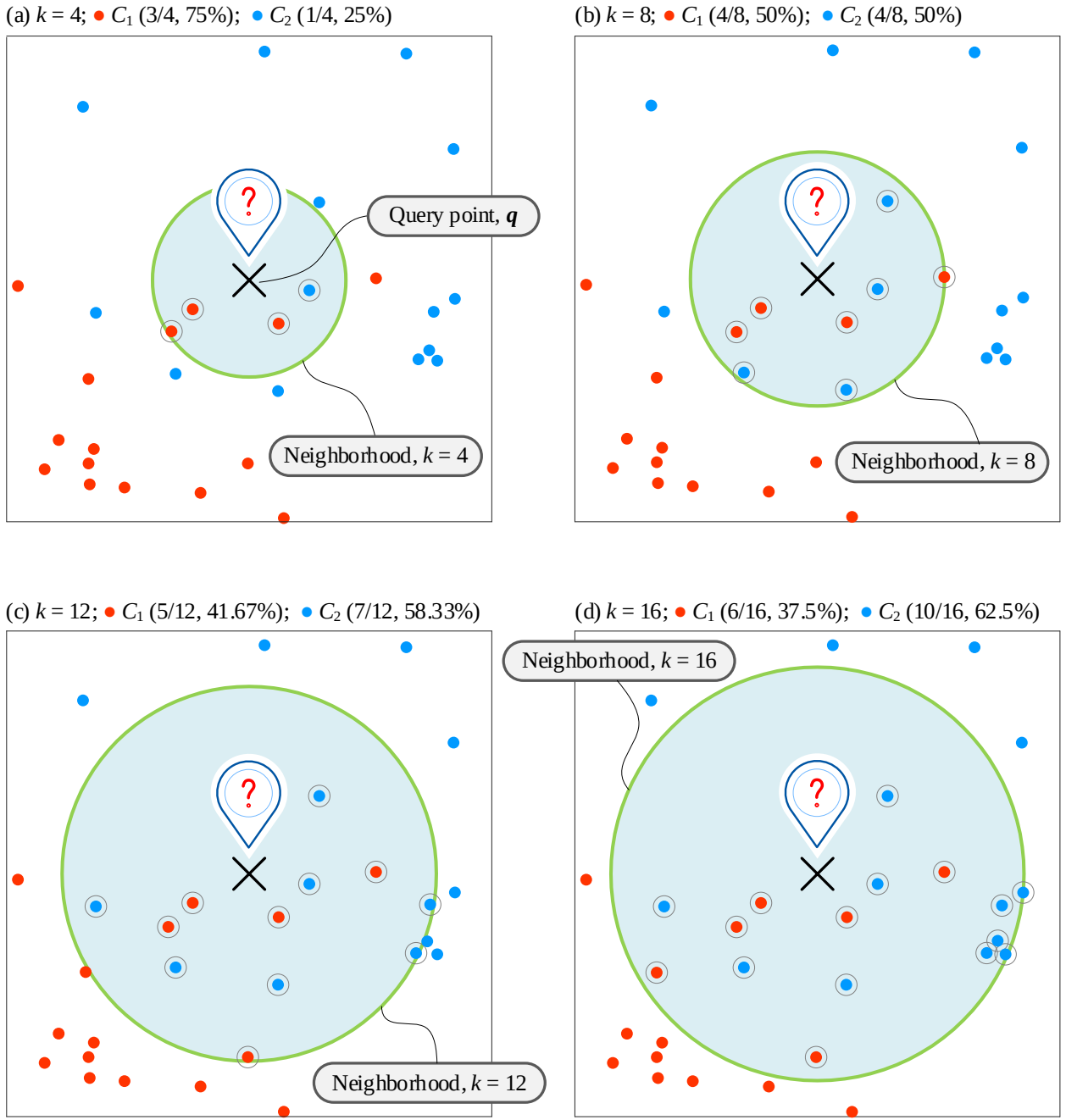


采用2个特征 (花萼长度 和 花萼宽度) 分类三种鸢尾花

欧氏距离



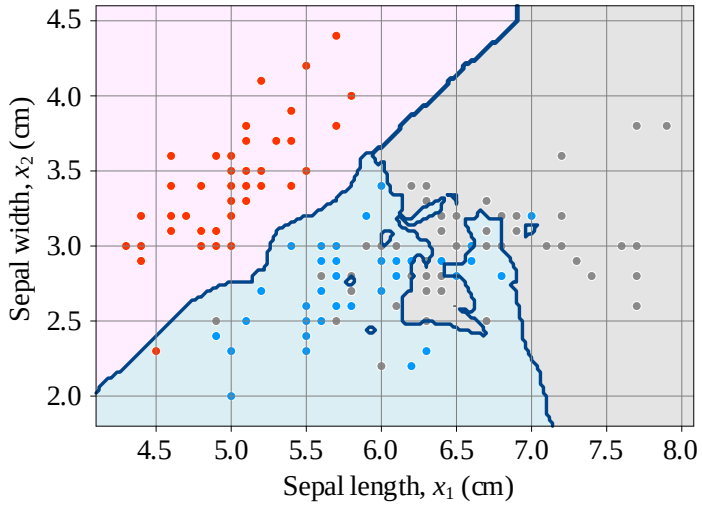
k近邻数量



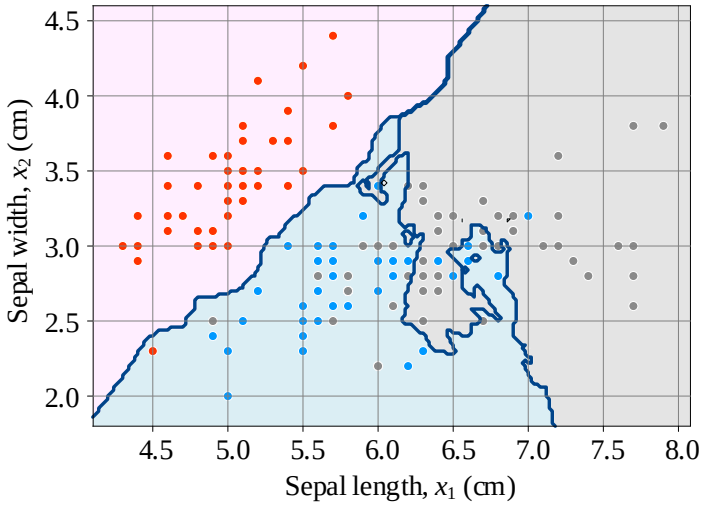
k选取不同值时对鸢尾花分类影响

Setosa, C_1 Versicolor, C_2 Virginica, C_3

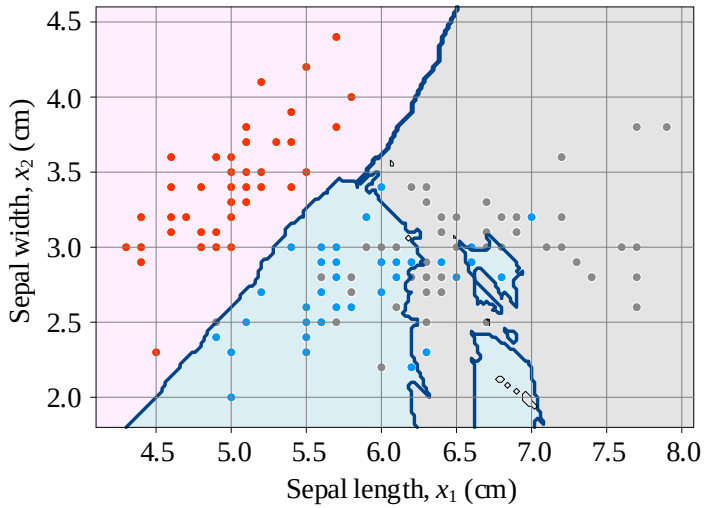
(a) kNN classifier (k = 4, weights = 'uniform')



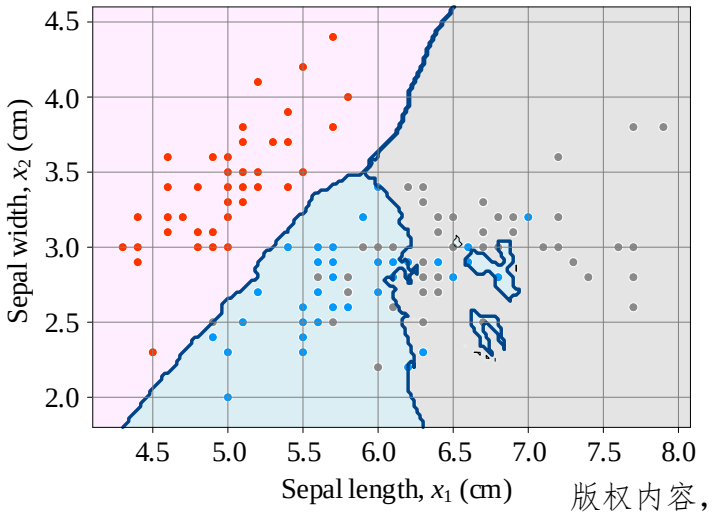
(b) kNN classifier (k = 8, weights = 'uniform')



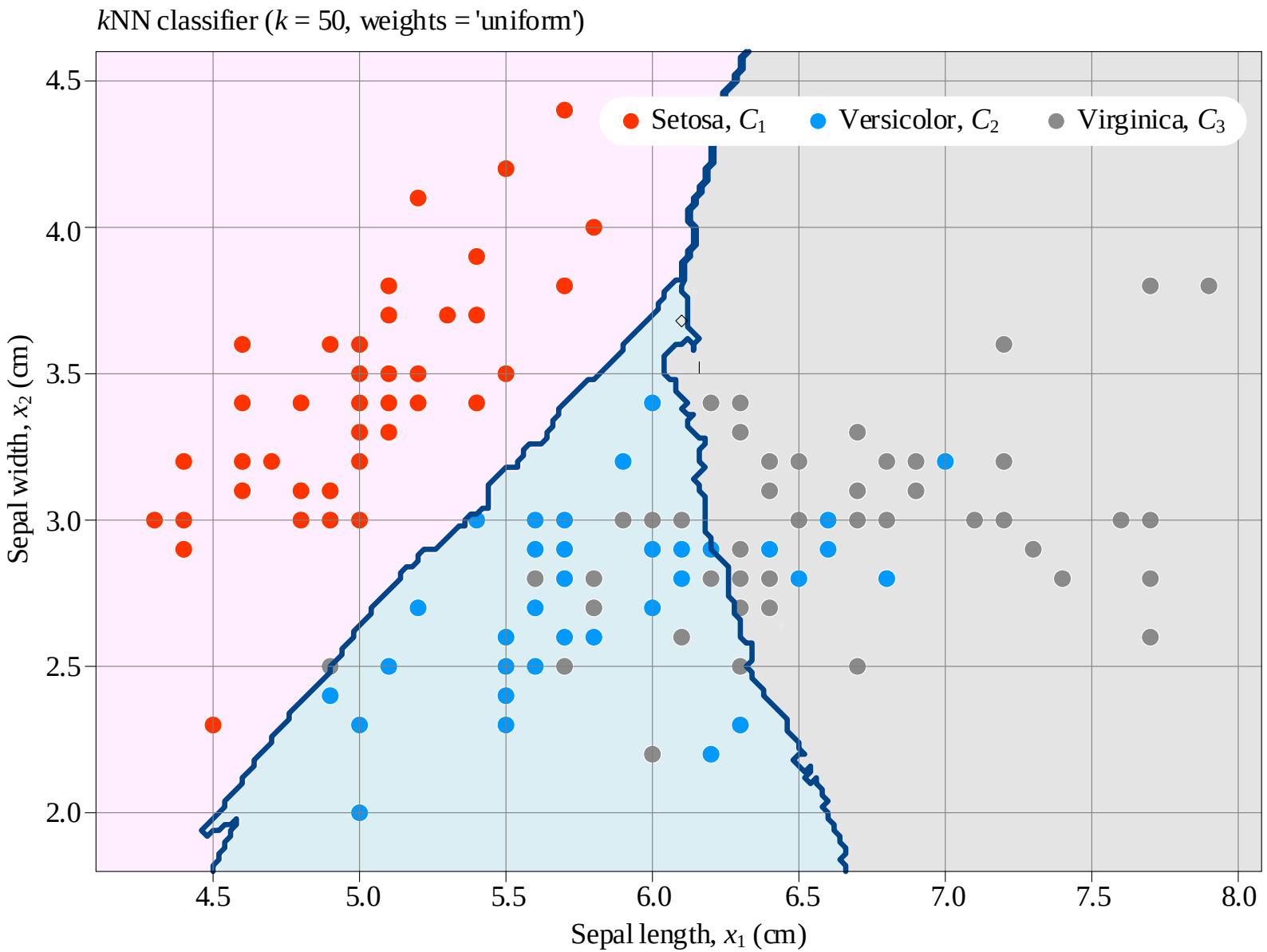
(c) kNN classifier (k = 12, weights = 'uniform')



(d) kNN classifier (k = 16, weights = 'uniform')



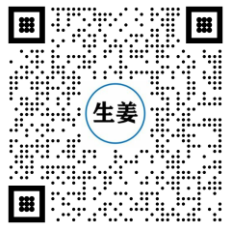
k = 50时，鸢尾花分类决策边界，kNN，等权重投票



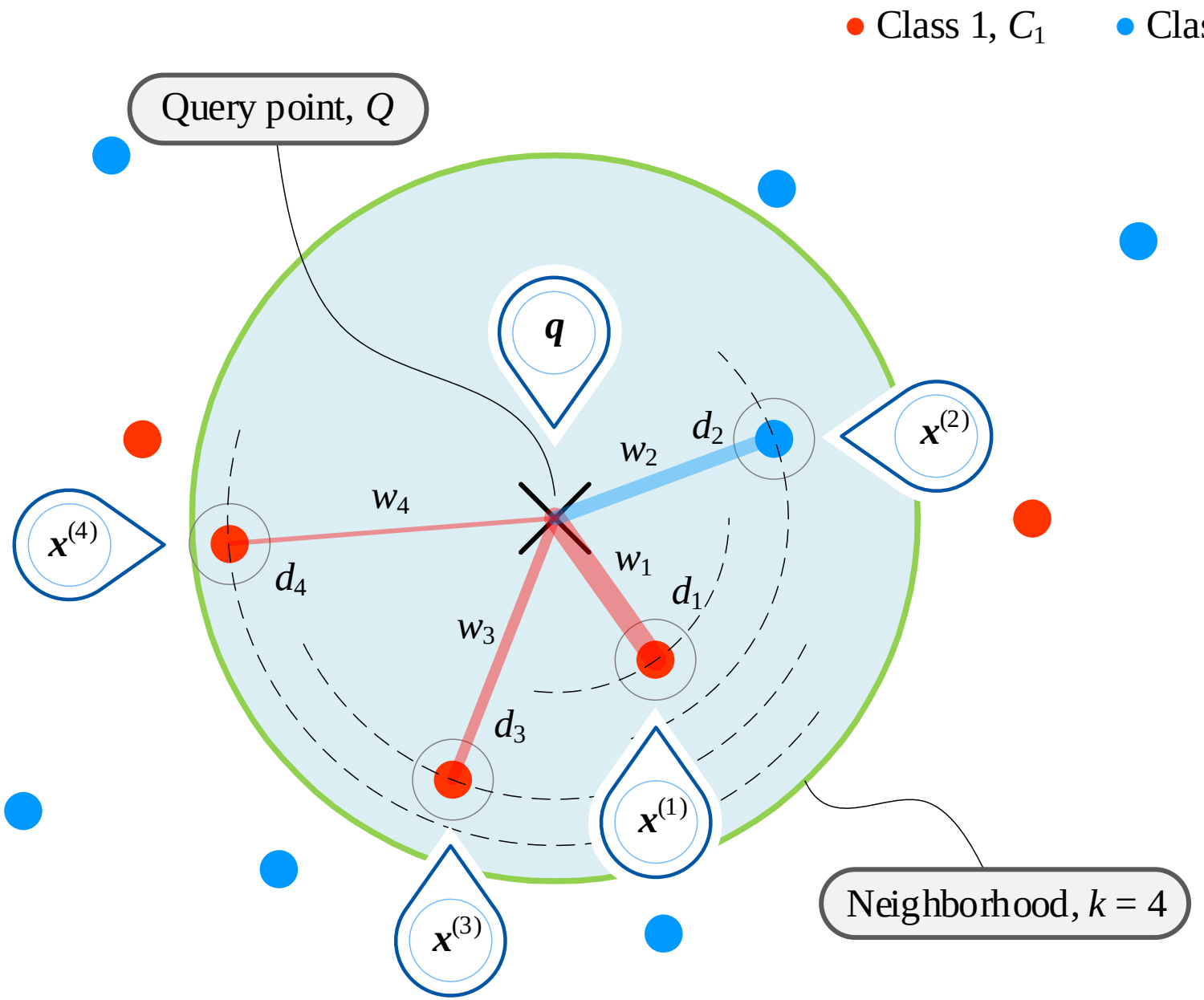
投票权重

$$y^{(q)} = \arg\max_{C_l} \sum_{i \in kNN(q)} w_i \cdot I\left(y^{(i)} = C_l\right)$$

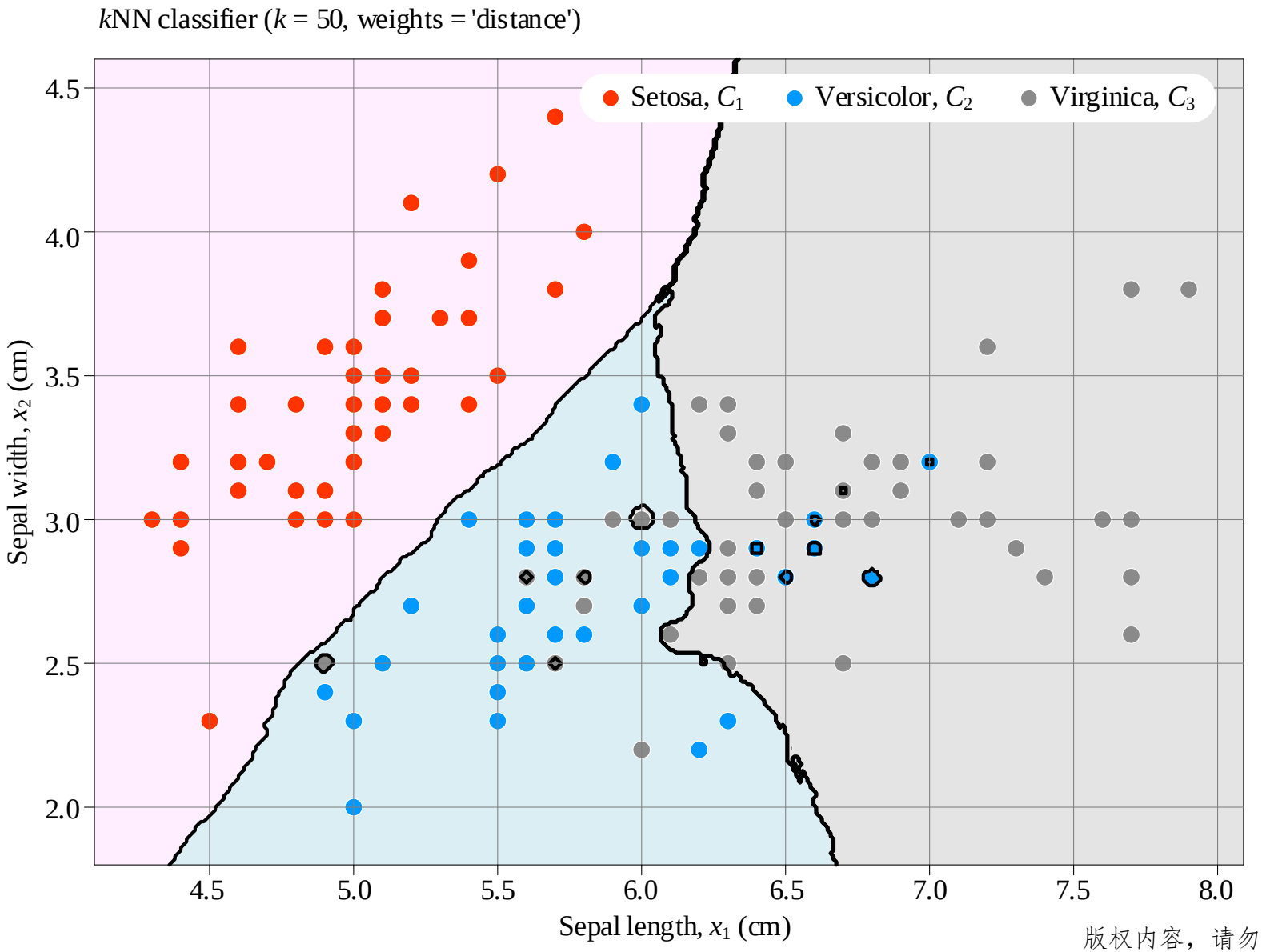
知乎



投票权重



k = 50时，鸢尾花分类决策边界，kNN



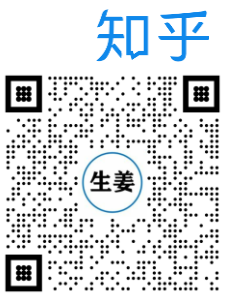
投票权重与查询点距离成反

最近质心分类器 (Nearest Centroid Classifier)

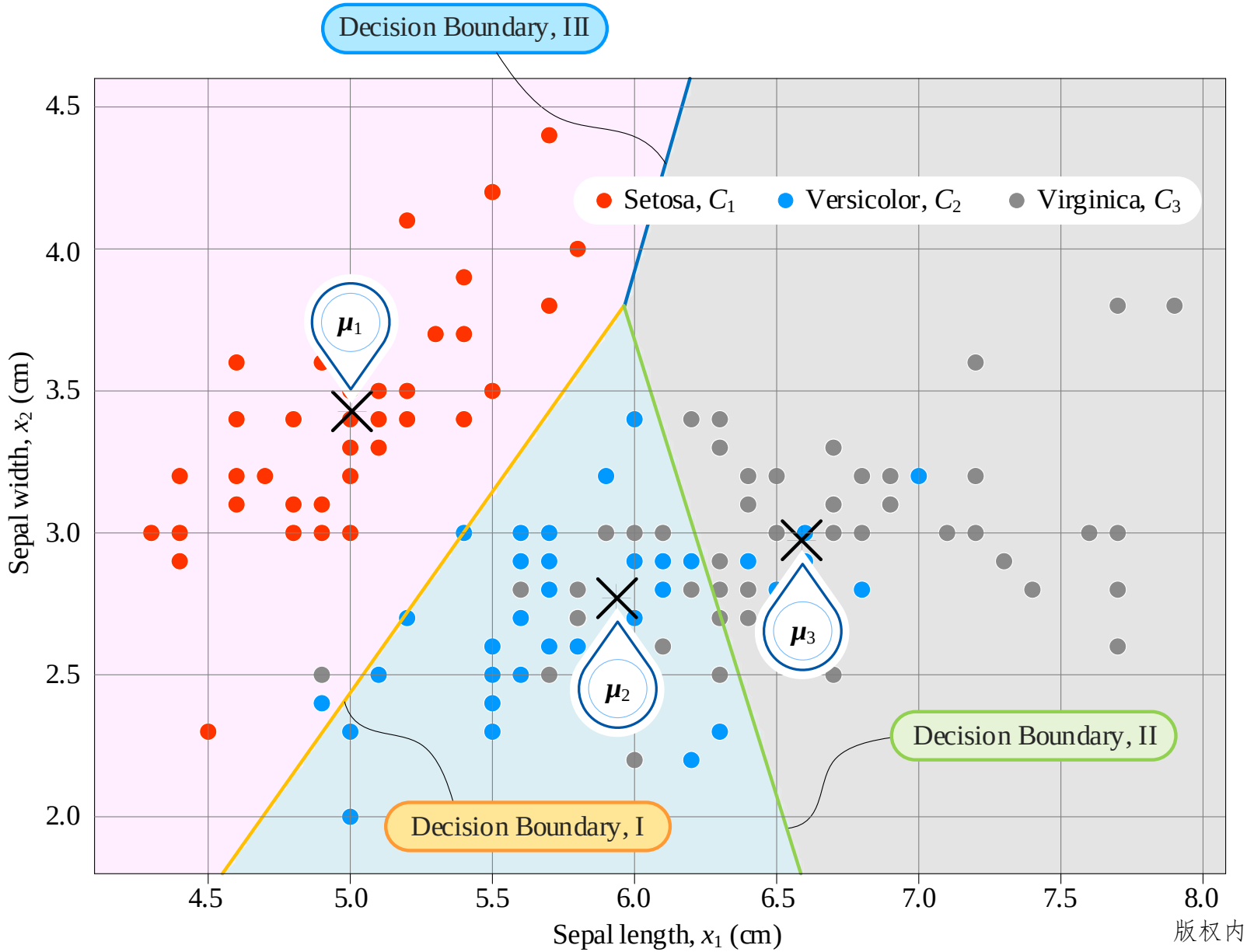
$$\boldsymbol{\mu}_l = \frac{1}{\text{count}\left(C_l\right)} \sum_{i \in C_l} \boldsymbol{x}_i$$

$$\boldsymbol{x}^{(2)} = \begin{bmatrix} 2 & 3 \end{bmatrix}, \quad \boldsymbol{x}^{(5)} = \begin{bmatrix} 3 & 1 \end{bmatrix}, \quad \boldsymbol{x}^{(6)} = \begin{bmatrix} -2 & 2 \end{bmatrix}, \quad \boldsymbol{x}^{(9)} = \begin{bmatrix} 1 & 6 \end{bmatrix}$$

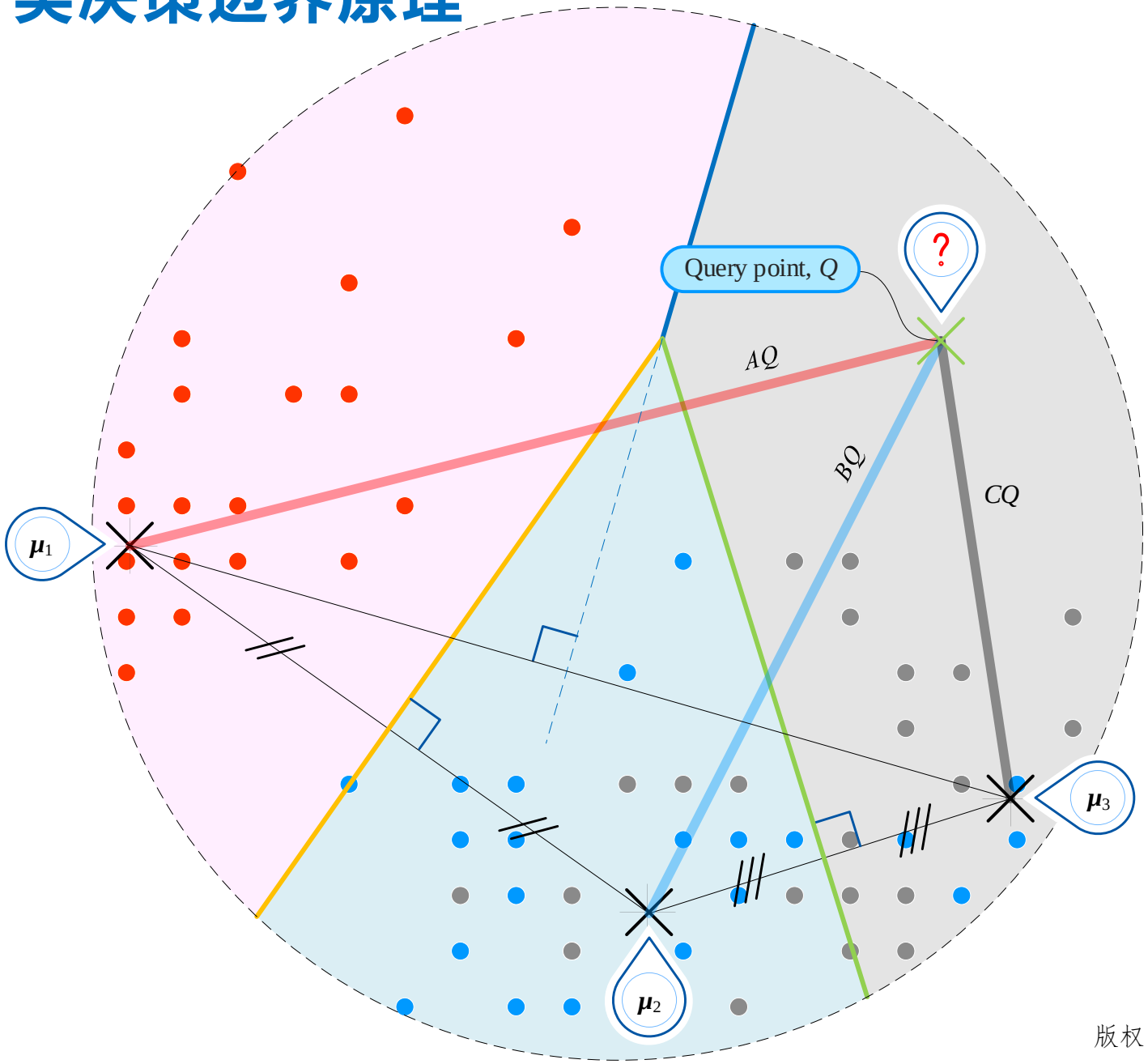
$$\begin{aligned} \boldsymbol{\mu}_1 &= \frac{1}{\text{count}\left(C_1\right)} \sum_{i \in C_l} \boldsymbol{x}_i = \frac{1}{\text{count}\left(C_1\right)} \left(\boldsymbol{x}^{(2)} + \boldsymbol{x}^{(5)} + \boldsymbol{x}^{(6)} + \boldsymbol{x}^{(9)} \right) \\ &= \frac{1}{4} \left(\begin{bmatrix} 2 & 3 \end{bmatrix} + \begin{bmatrix} 3 & 1 \end{bmatrix} + \begin{bmatrix} -2 & 2 \end{bmatrix} + \begin{bmatrix} 1 & 6 \end{bmatrix} \right) \\ &= \begin{bmatrix} 1 & 3 \end{bmatrix} \end{aligned}$$



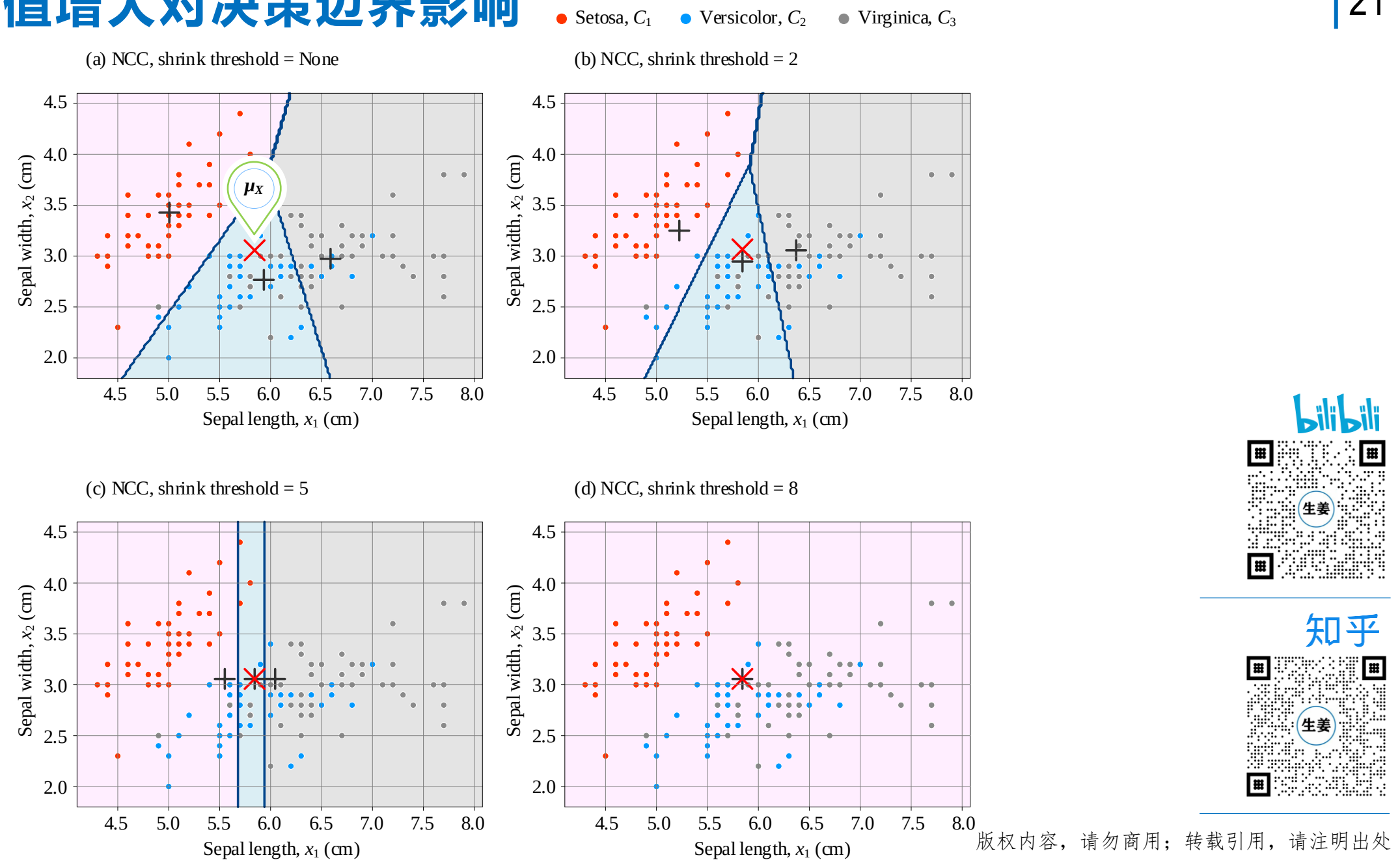
鸢尾花分类决策边界，最近质心分类



最近质心分类决策边界原理

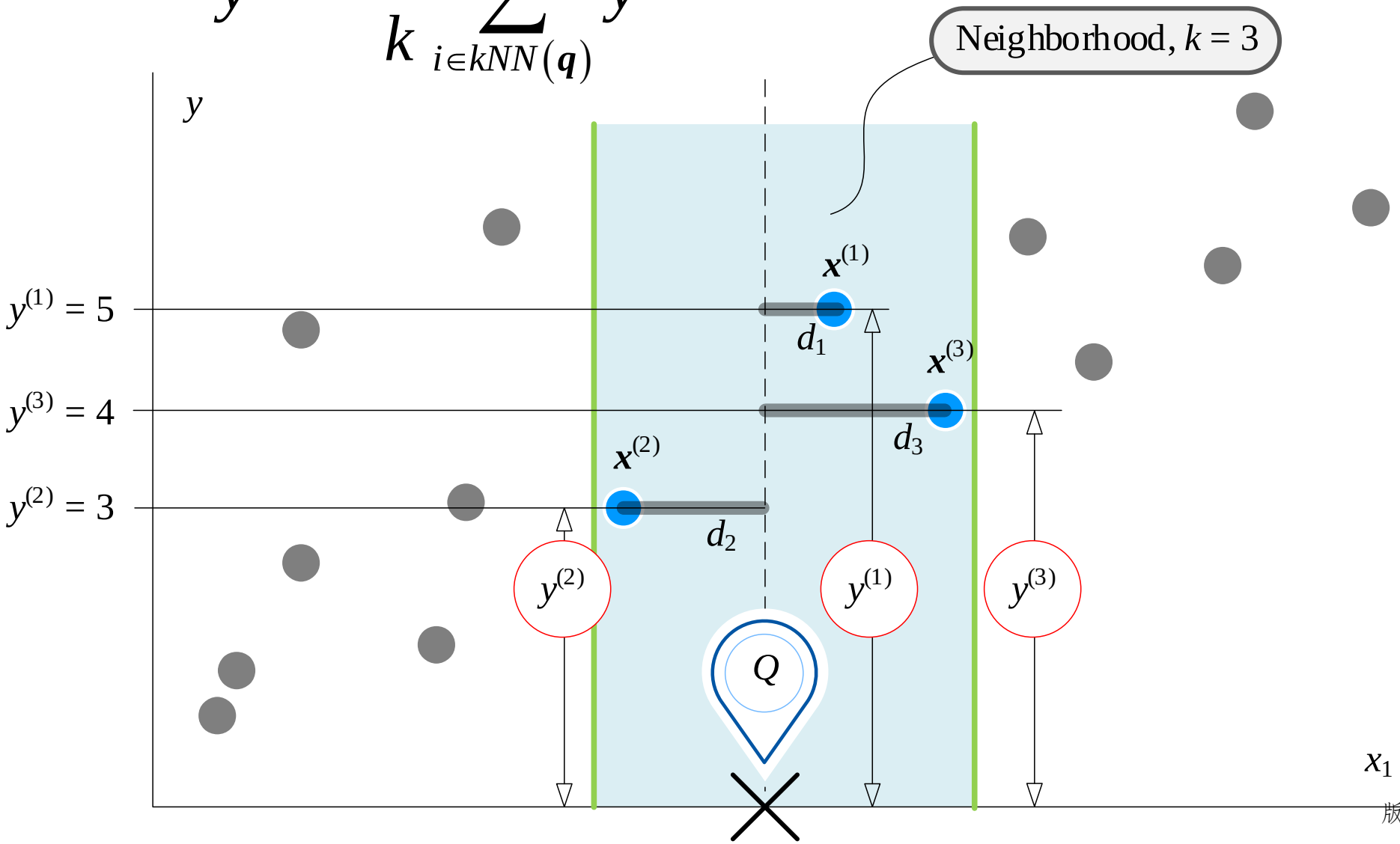


收缩阈值增大对决策边界影响



$$y^{(q)} = \frac{1}{3}\left(y^{(1)} + y^{(2)} + y^{(3)}\right) = \frac{1}{3}(5+3+4) = 4$$

$$y^{(q)} = \frac{1}{k}\sum_{i \in kNN(q)} y^{(i)}$$



鸢尾花花萼数据回归

kNN回归，不同种类鸢尾花花萼长度 x_1和花萼宽度x_2回归关系