

Python 3 玩儿转机器学习

讲师：liuyubobobo

版权所有 侵权必究
liuyubobobo

kNN - k近邻算法

k-Nearest Neighbors

讲师：liuyuboboo

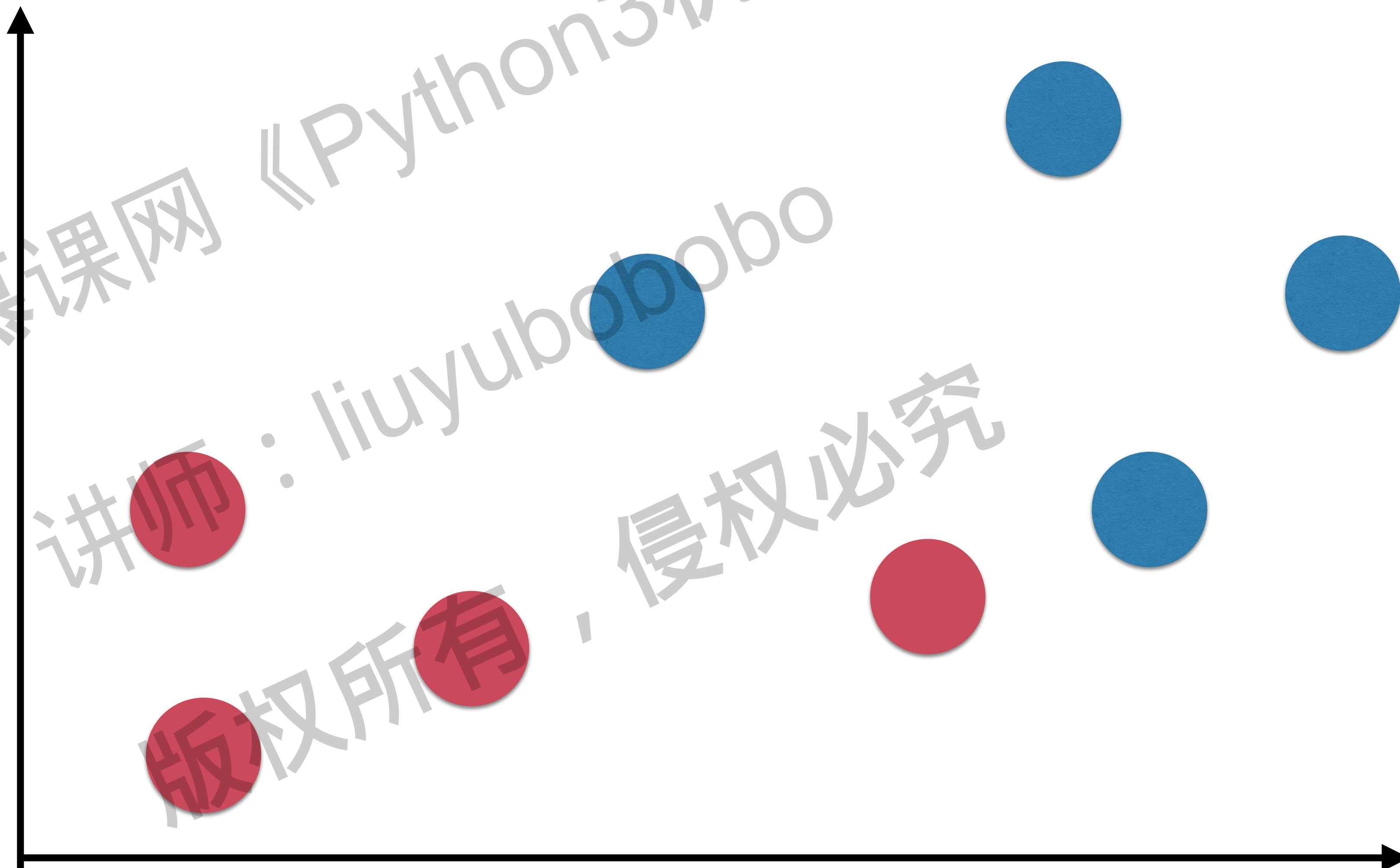
版权所有，侵权必究

k近邻算法

- 思想极度简单
- 应用数学知识少（近乎为零）
- 效果好（缺点？）
- 可以解释机器学习算法使用过程中的很多细节问题
- 更完整的刻画机器学习应用的流程

k近邻算法

时间



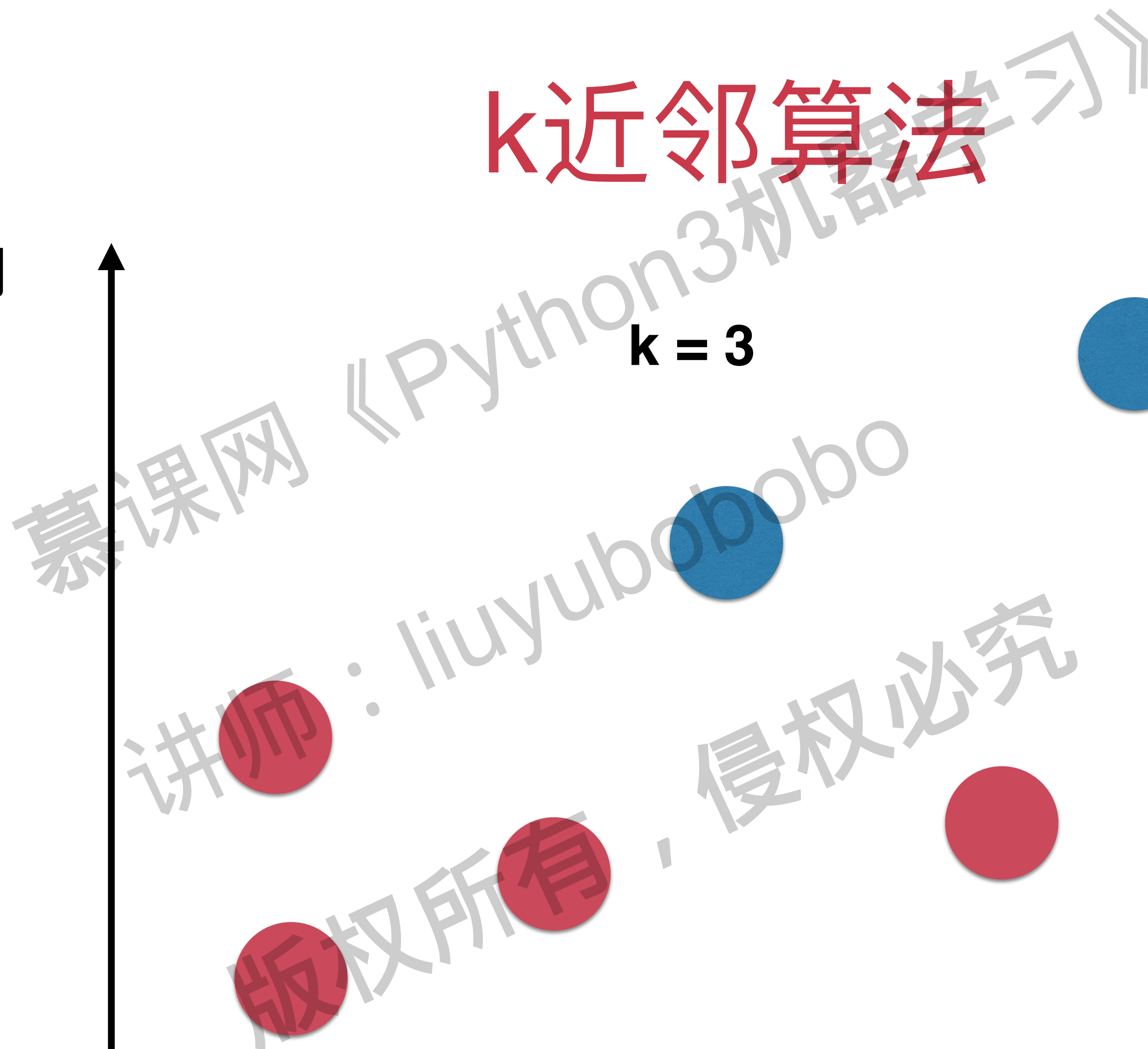
肿瘤大小

k近邻算法

时间

$k = 3$

肿瘤大小

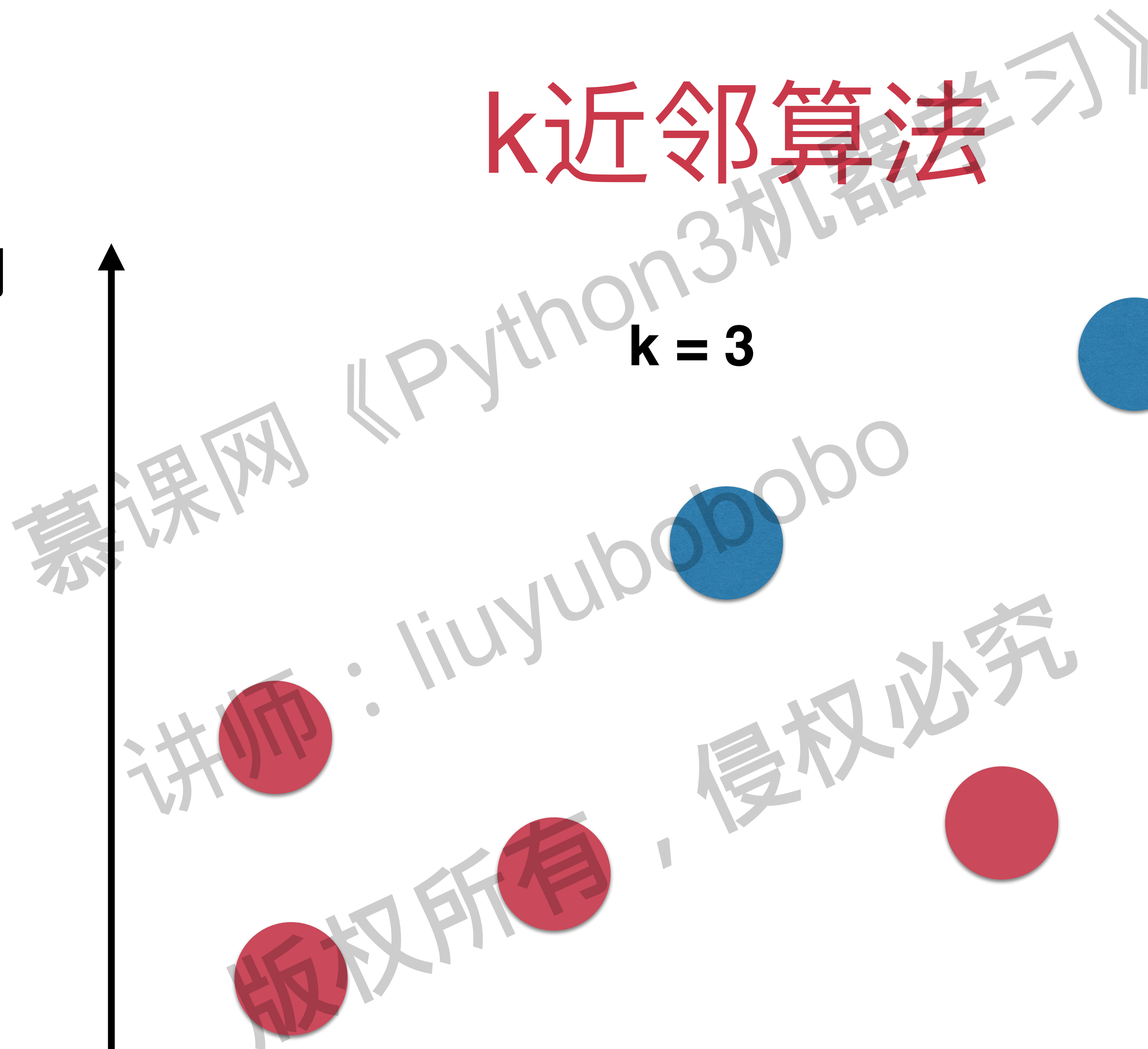


k近邻算法

时间

$k = 3$

肿瘤大小

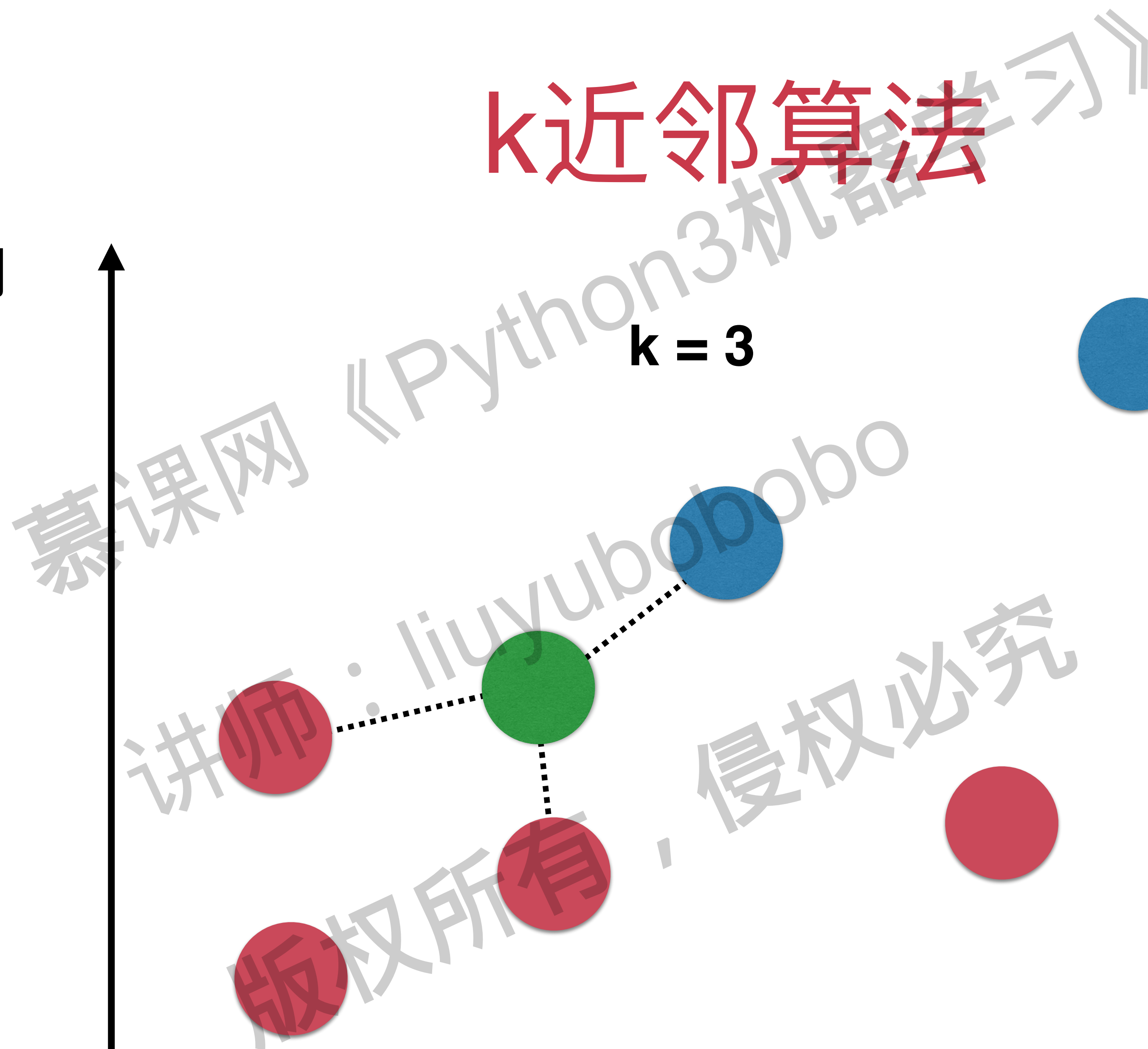


k近邻算法

时间

$k = 3$

肿瘤大小

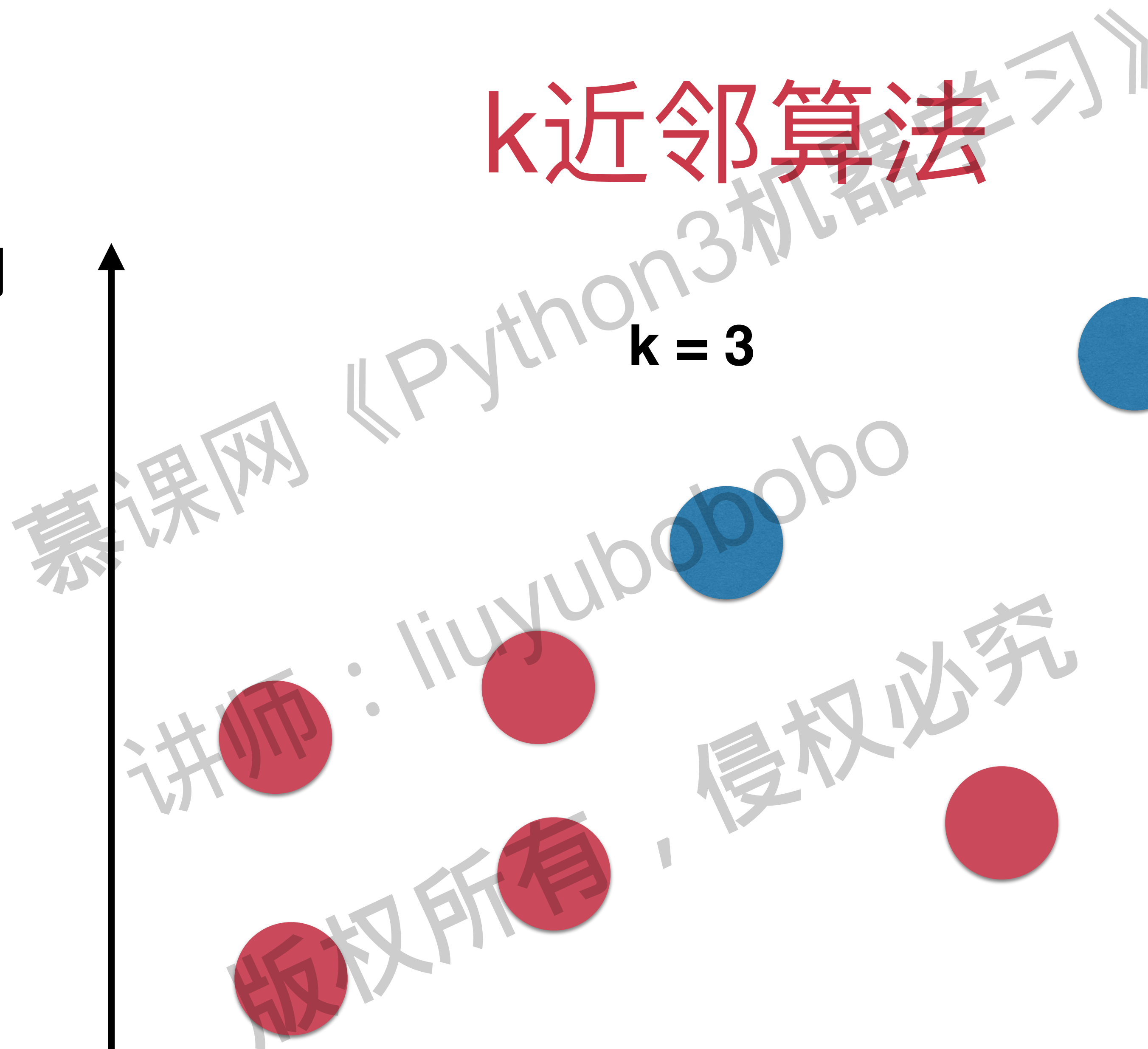


k近邻算法

时间

$k = 3$

肿瘤大小



欧拉距离

$$\sqrt{(x^{(a)} - x^{(b)})^2 + (y^{(a)} - y^{(b)})^2}$$

$$\sqrt{(x^{(a)} - x^{(b)})^2 + (y^{(a)} - y^{(b)})^2 + (z^{(a)} - z^{(b)})^2}$$

$$\sqrt{(X_1^{(a)} - X_1^{(b)})^2 + (X_2^{(a)} - X_2^{(b)})^2 + \dots + (X_n^{(a)} - X_n^{(b)})^2}$$

欧拉距离

$$\sqrt{(X_1^{(a)} - X_1^{(b)})^2 + (X_2^{(a)} - X_2^{(b)})^2 + \dots + (X_n^{(a)} - X_n^{(b)})^2}$$

$$\sqrt{\sum_{i=1}^n (X_i^{(a)} - X_i^{(b)})^2}$$

慕课网《Python3机器学习》

演示：KNN基础

讲师：liuyubobobo

版权所有，侵权必究

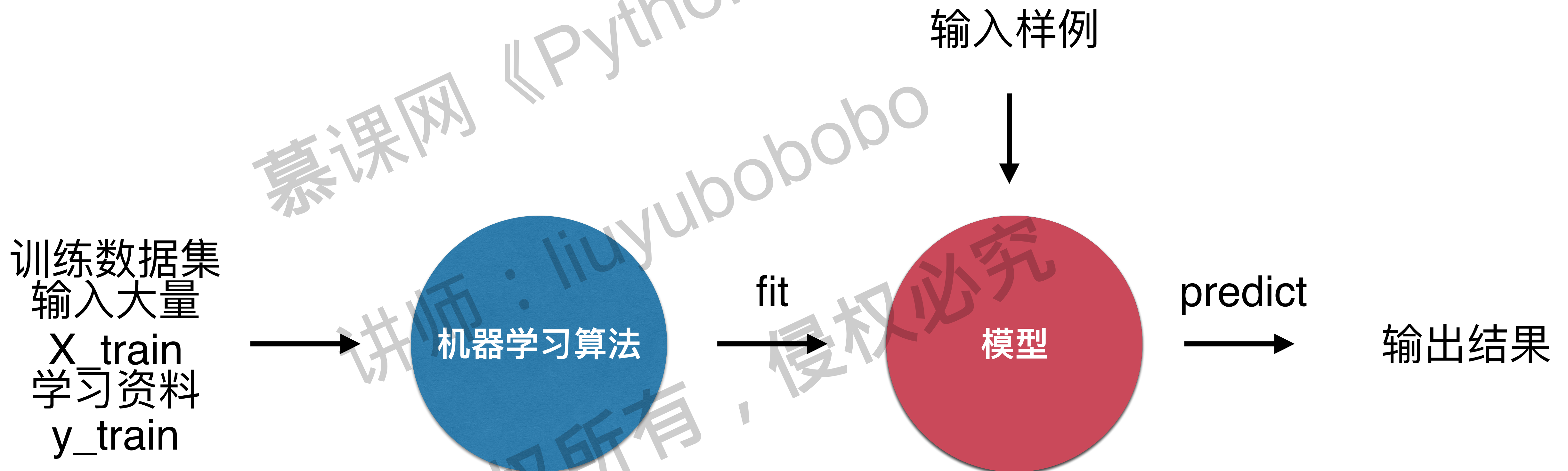
scikit-learn中的 kNN算法

慕课网《Python3机器学习》

讲师：liuyubobobo

版权所有，侵权必究

什么是机器学习



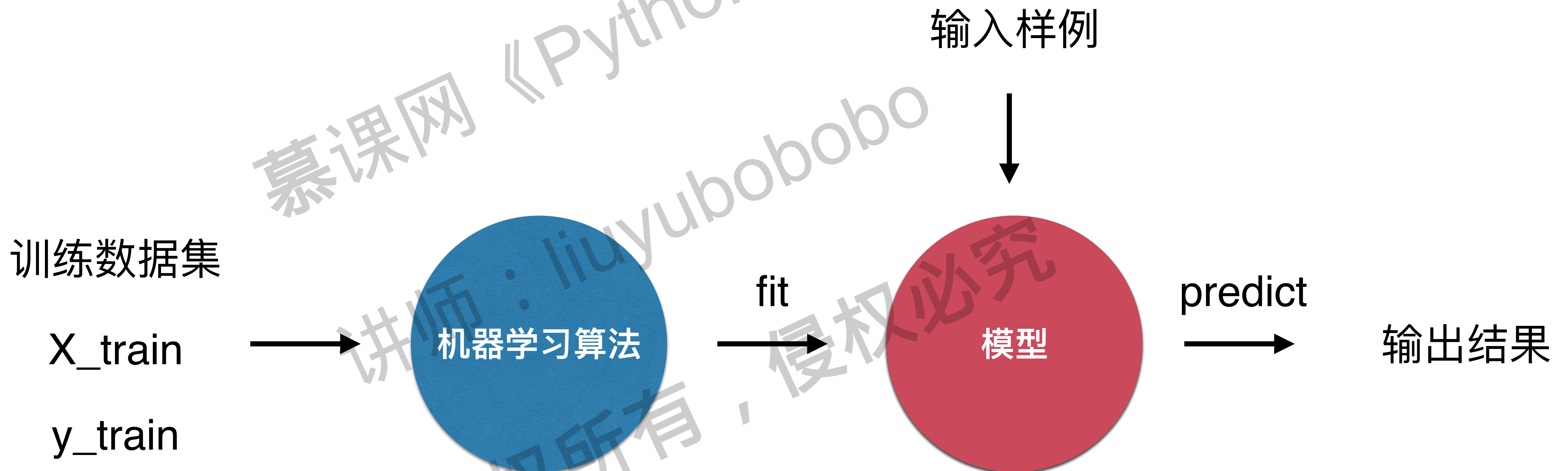
可以说kNN是一个不需要训练过程的算法

k近邻算法

k近邻算法是非常特殊的，可以被认为是没有模型的算法

为了和其他算法统一，可以认为训练数据集就是模型本身

什么是机器学习



对于kNN来说，训练集就是模型

演示：scikit-learn中kNN的使用

讲师：liuyunbobo
版权所有，侵权必究

演示：我们自己的kNN算法封装

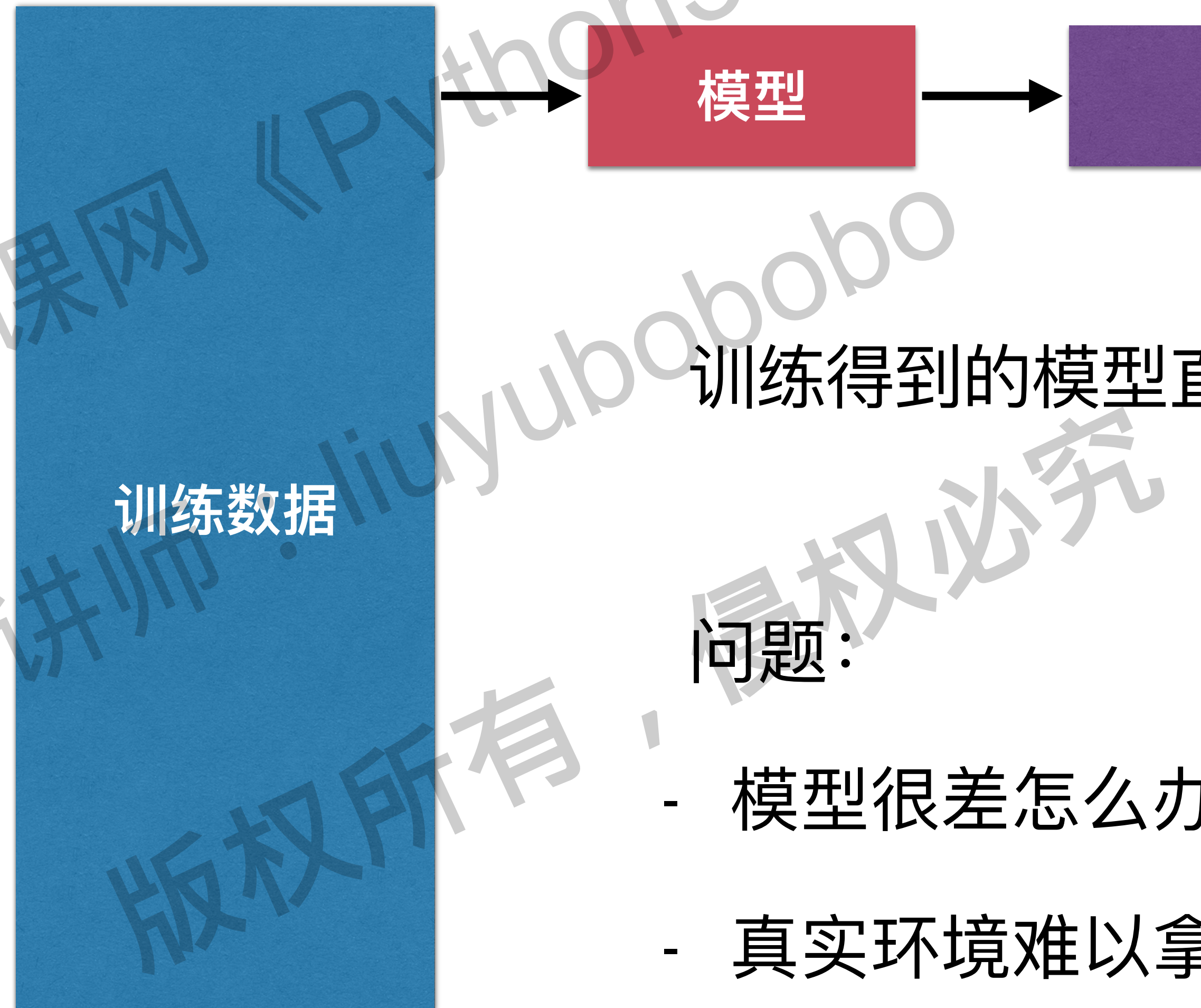
慕课网《Python3机器学习》

判断机器学习算法的性能

讲师：liuyubobobo

版权所有，侵权必究

判断机器学习算法的性能

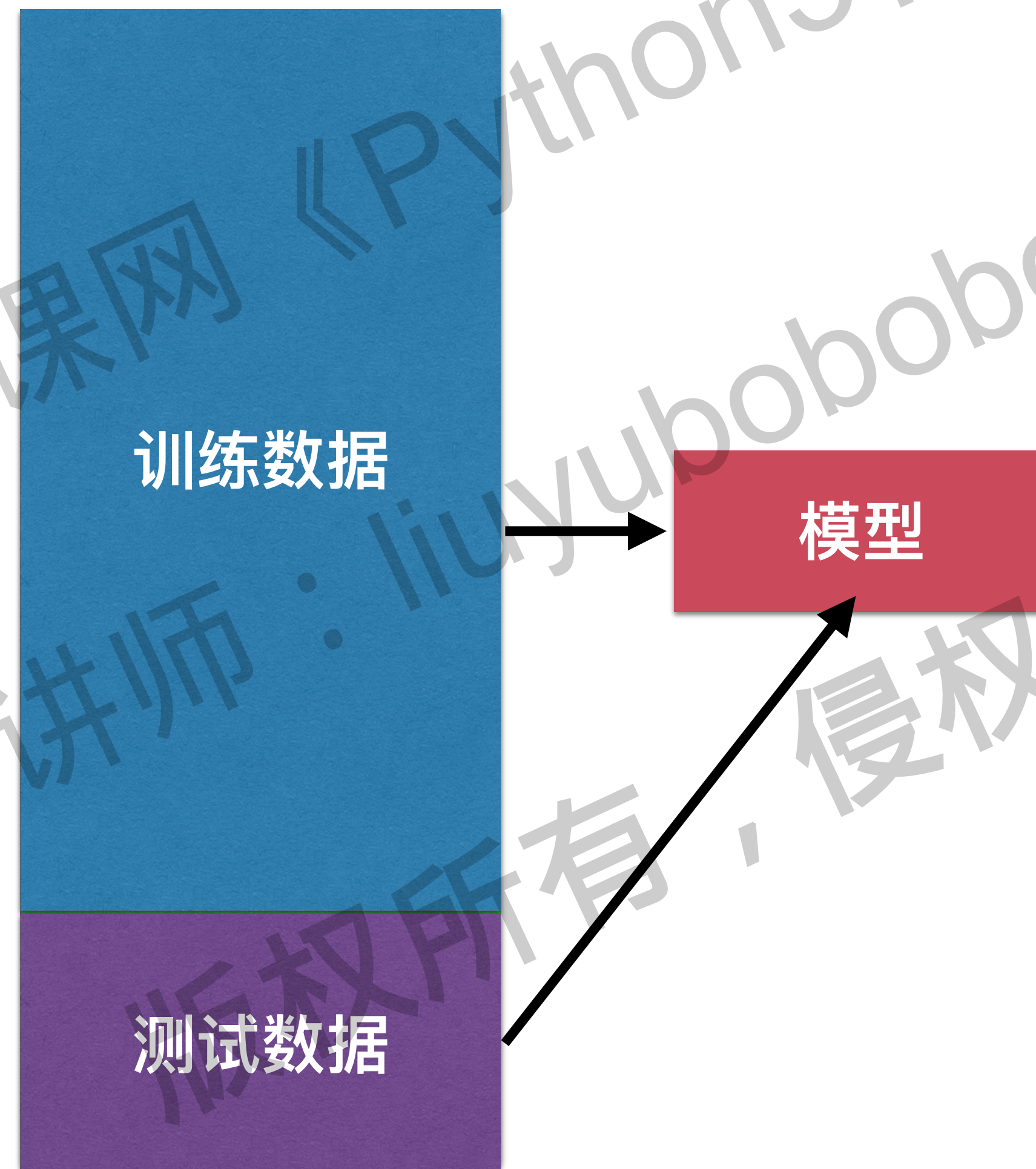


训练得到的模型直接在真实环境中使用。

问题：

- 模型很差怎么办？ 真实损失。
- 真实环境难以拿到真实label？


判断机器学习算法的性能



通过测试数据直接判断模型好坏

在模型进入真实环境前改进模型

判断机器学习算法的性能



训练数据

train test split

问题？ 后续分解

测试数据

实践：自己编写train_test_split

实践：使用scikit-learn的train_test_split

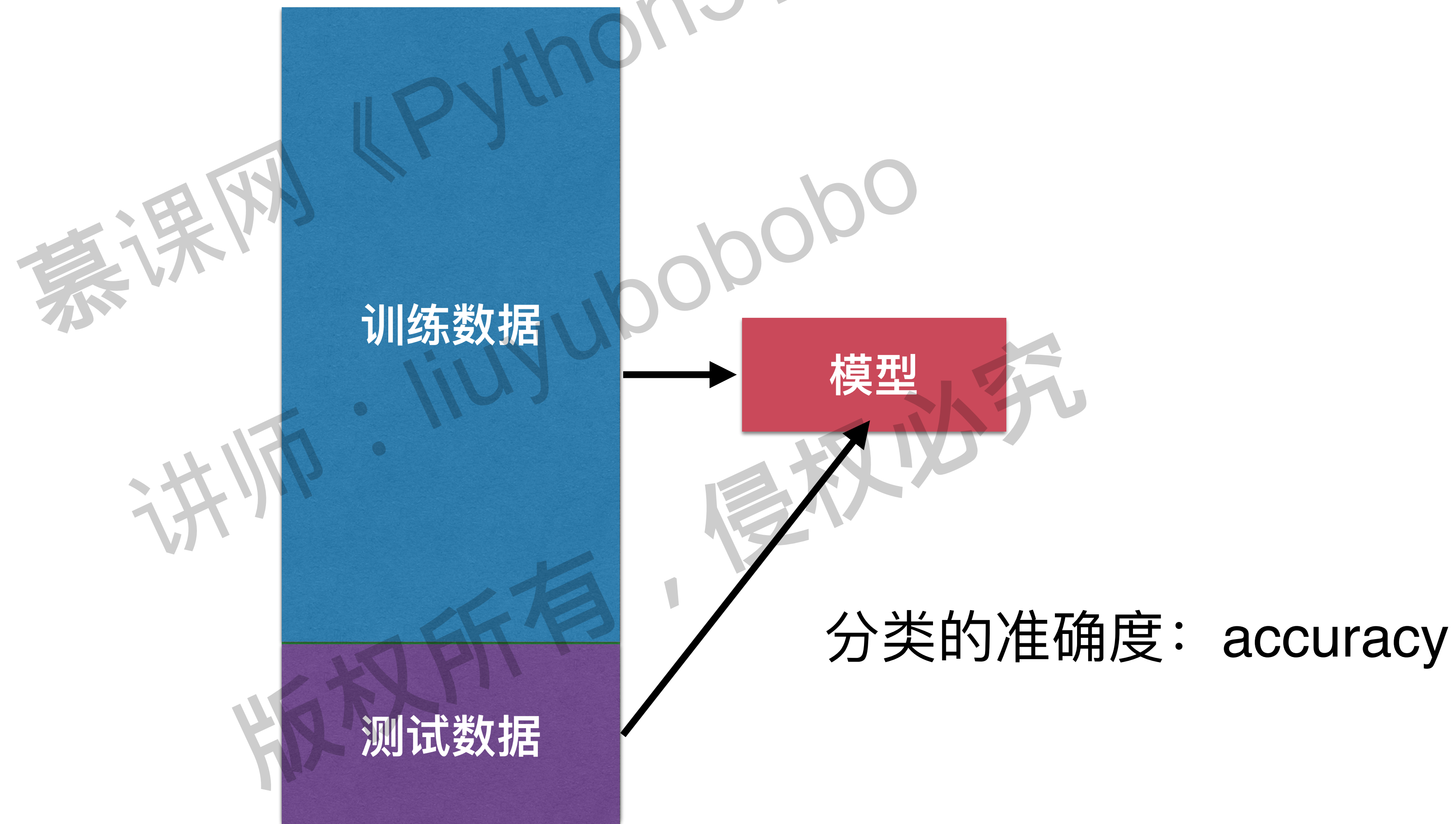
慕课网《Python3机器学习》

分类准确度

讲师：liuyubobobo

版权所有，侵权必究

判断机器学习算法的性能



慕课网《Python3机器学习》

超参数

讲师：liuyubobobo

版权所有，侵权必究

超参数和模型参数

- 超参数：在算法运行前需要决定的参数
- 模型参数：算法过程中学习的参数

kNN算法没有模型参数

kNN算法中的k是典型的超参数

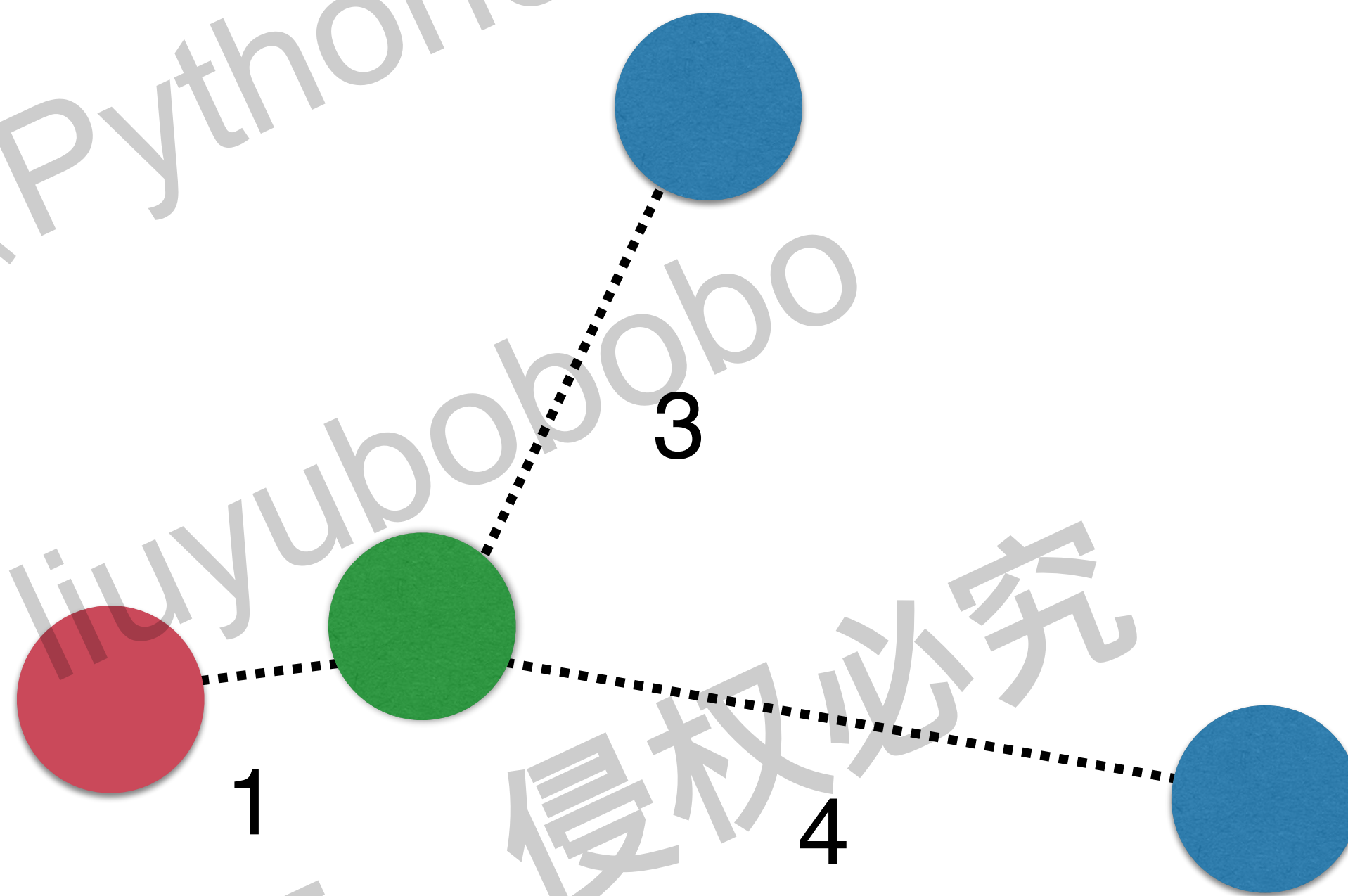
寻找好的超参数

- 领域知识
- 经验数值
- 实验搜索

kNN中的另一个超参数 - 距离

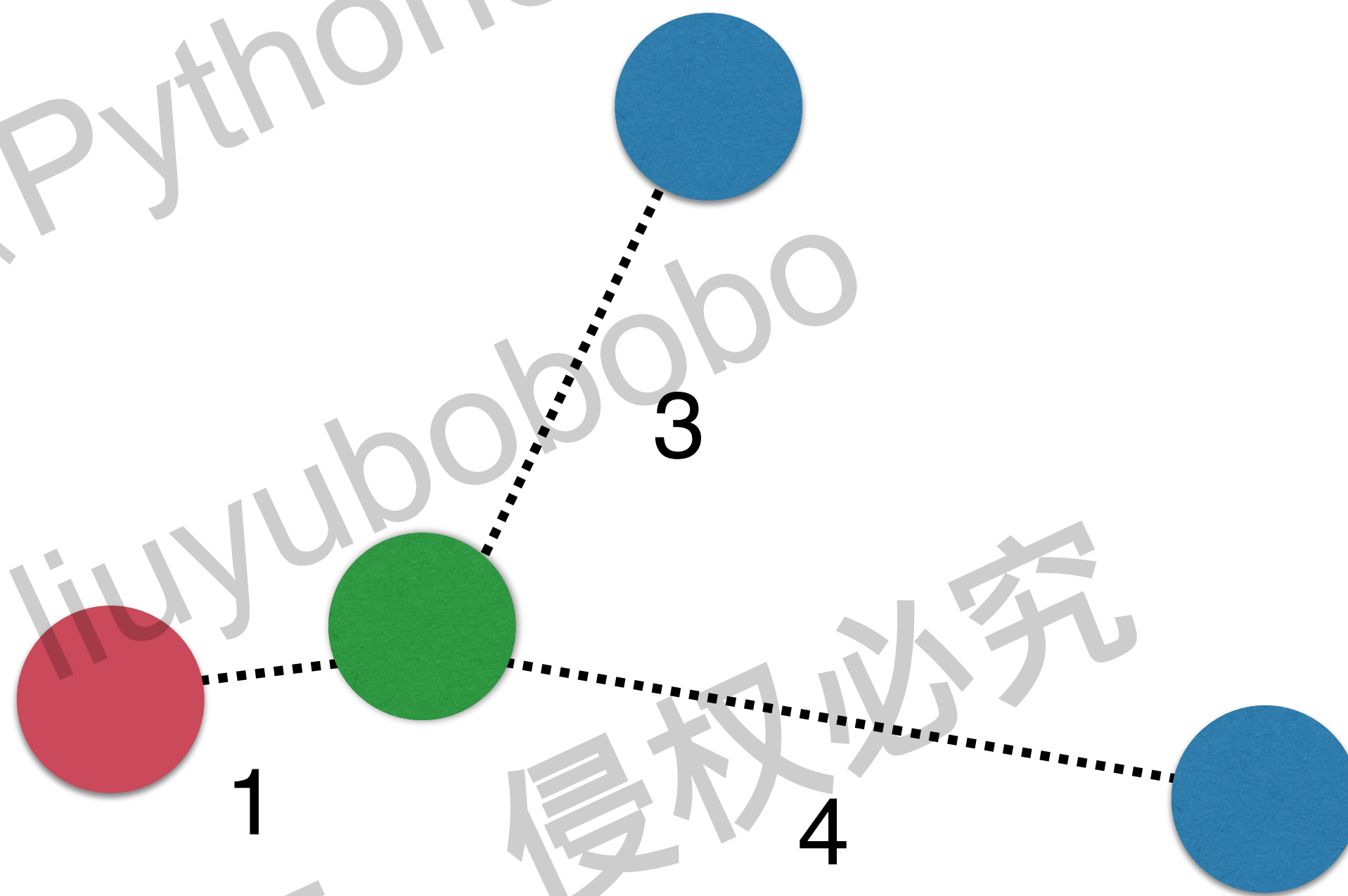
讲师：liuyubobobo
版权所有，侵权必究

k近邻算法



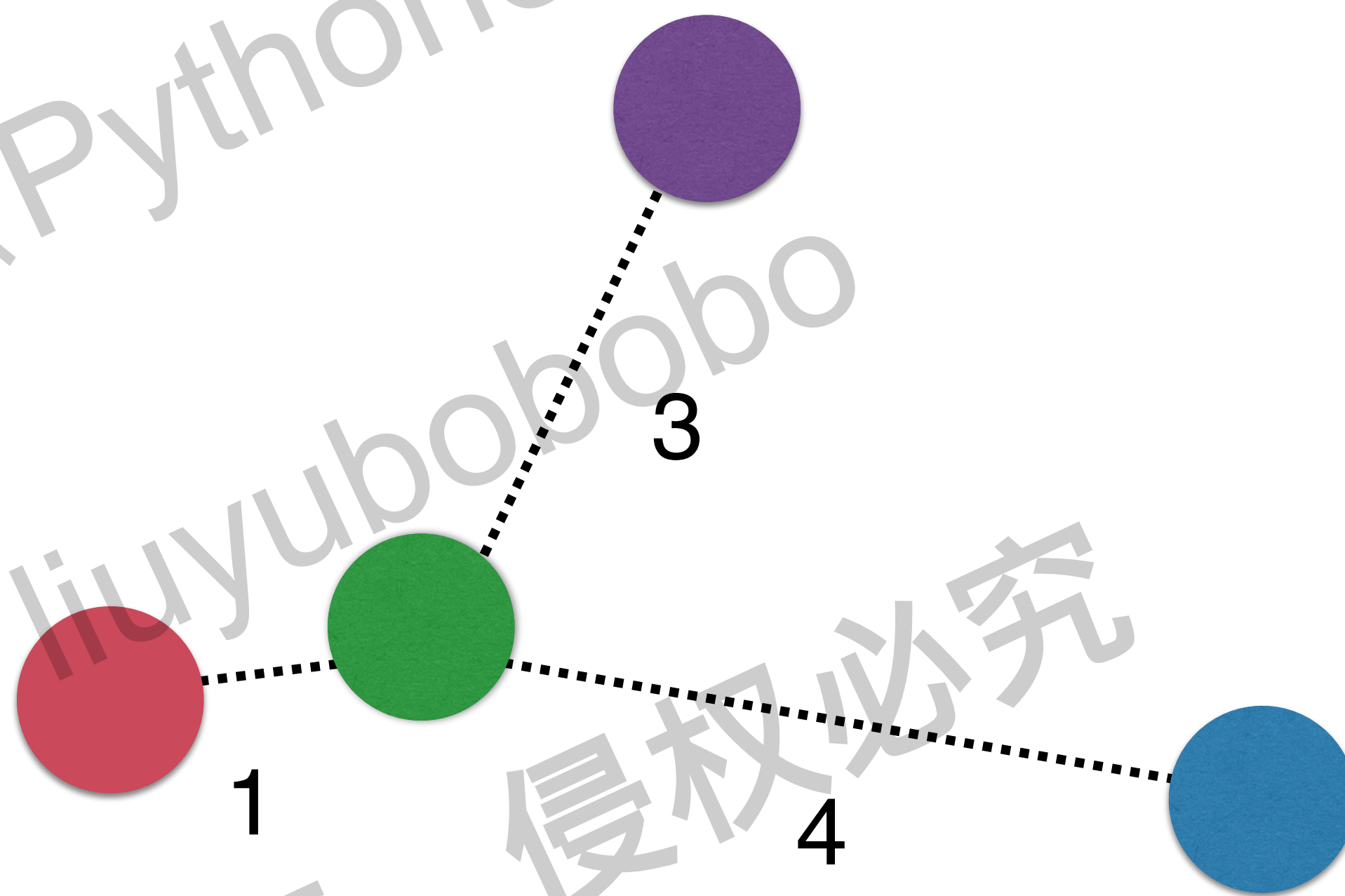
普通的k近邻算法：蓝色获胜

k近邻算法



考虑距离：红色：1 蓝色： $\frac{1}{3} + \frac{1}{4} = \frac{7}{12}$ 红色胜

k近邻算法



另一个好处：解决平票的情况

sklearn kNN文档

参数 weights

<http://scikit-learn.org/stable/modules/generated/sklearn.neighbors.KNeighborsClassifier.html>

慕课网《Python3机器学习》

更多关于距离的定义

讲师：liuyubobobo

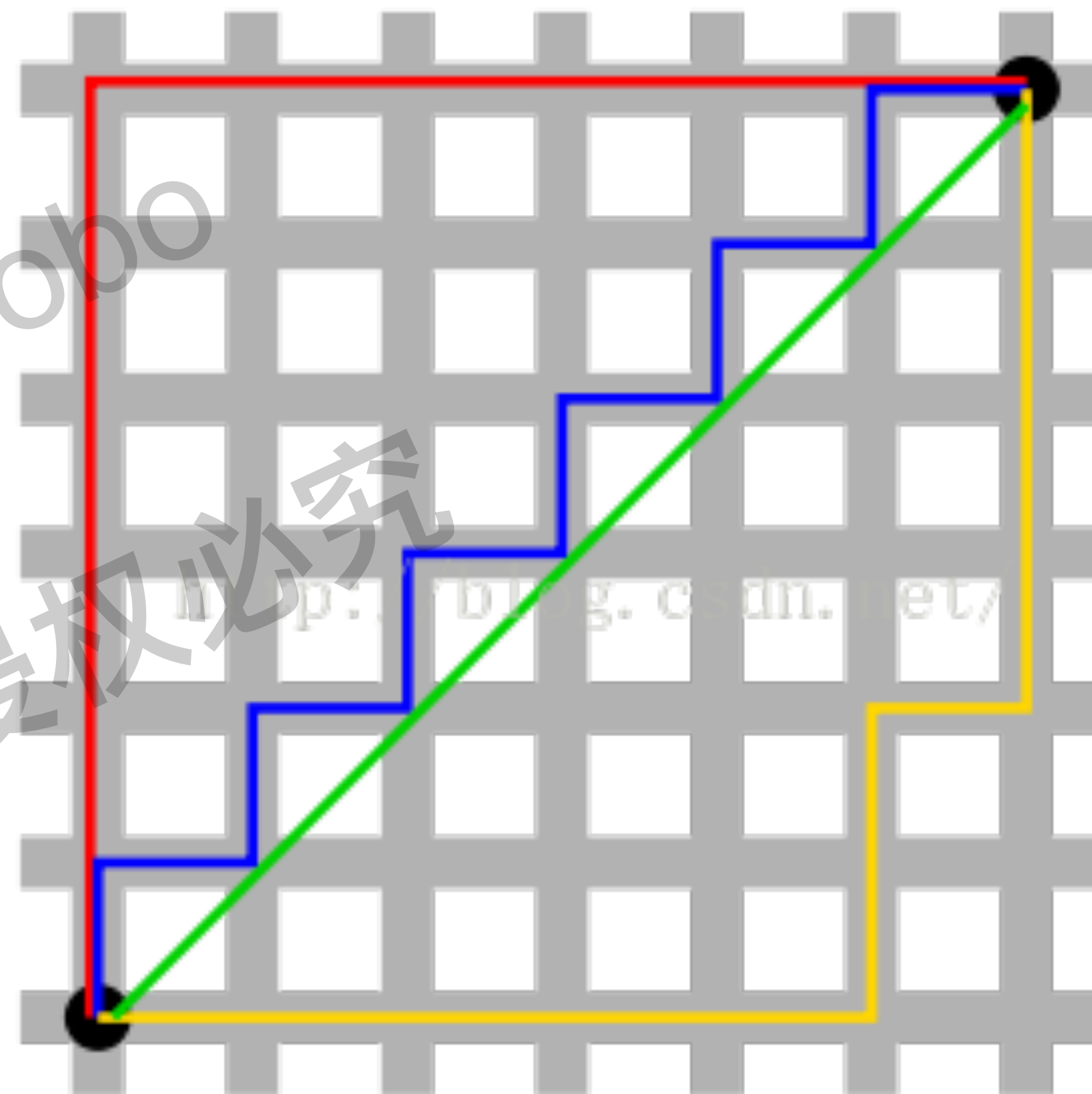
版权所有，侵权必究

欧拉距离

$$\sqrt{\sum_{i=1}^n (X_i^{(a)} - X_i^{(b)})^2}$$

曼哈顿距离

$$\sum_{i=1}^n |X_i^{(a)} - X_i^{(b)}|$$



距离

$$\sum_{i=1}^n |X_i^{(a)} - X_i^{(b)}|$$

$$\sqrt{\sum_{i=1}^n (X_i^{(a)} - X_i^{(b)})^2}$$

距离

$$\sum_{i=1}^n |X_i^{(a)} - X_i^{(b)}|$$

$$\sqrt{\sum_{i=1}^n |X_i^{(a)} - X_i^{(b)}|^2}$$

距离

$$\sum_{i=1}^n |X_i^{(a)} - X_i^{(b)}|$$

$$\left(\sum_{i=1}^n |X_i^{(a)} - X_i^{(b)}|^2 \right)^{\frac{1}{2}}$$

距离

$$\left(\sum_{i=1}^n |X_i^{(a)} - X_i^{(b)}| \right)^{\frac{1}{1}}$$

$$\left(\sum_{i=1}^n |X_i^{(a)} - X_i^{(b)}|^2 \right)^{\frac{1}{2}}$$

距离

$$\left(\sum_{i=1}^n |X_i^{(a)} - X_i^{(b)}| \right)^{\frac{1}{1}}$$

$$\left(\sum_{i=1}^n |X_i^{(a)} - X_i^{(b)}|^2 \right)^{\frac{1}{2}}$$

$$\left(\sum_{i=1}^n |X_i^{(a)} - X_i^{(b)}|^p \right)^{\frac{1}{p}}$$

明可夫斯基距离

Minkowski Distance

$$\left(\sum_{i=1}^n |X_i^{(a)} - X_i^{(b)}|^p \right)^{\frac{1}{p}}$$

获得了一个超参数 p

kNN中更多超参数与网格搜索

讲师：liuyubob0
版权所有，侵权必究

慕课网《Python3机器学习》

演示：网格搜索

讲师：liuyubobobo

版权所有，侵权必究

更多的距离定义

- 向量空间余弦相似度 Cosine Similarity
- 调整余弦相似度 Adjusted Cosine Similarity
- 皮尔森相关系数 Pearson Correlation Coefficient
- Jaccard相似系数 Jaccard Coefficient

sklearn kNN文档

参数 metric

<http://scikit-learn.org/stable/modules/generated/sklearn.neighbors.KNeighborsClassifier.html>

sklearn DistanceMetric文档

<http://scikit-learn.org/stable/modules/generated/sklearn.neighbors.DistanceMetric.html>

数据归一化 Feature Scaling

讲师：liuylubobobo
版权所有，侵权必究

数据归一化

	肿瘤大小 (厘米)	发现时间 (天)
样本1	1	200
样本2	5	100

样本间的距离被发现时间所主导

数据归一化

	肿瘤大小 (厘米)	发现时间 (年)
样本1	1	200天 = 0.55年
样本2	5	100 = 0.27年

数据归一化

解决方案：将所有数据映射到同一尺度

最值归一化：把所有数据映射到0-1之间

$$x_{scale} = \frac{x - x_{\min}}{x_{\max} - x_{\min}}$$

最值归一化 normalization

最值归一化：把所有数据映射到0-1之间

$$x_{scale} = \frac{x - x_{min}}{x_{max} - x_{min}}$$

适用于分布有明显边界的情况；受outlier影响较大

均值方差归一化 standardization

数据分布没有明显的边界；有可能存在极端数据值

均值方差归一化：把所有数据归一到均值为0方差为1的分布中

$$x_{scale} = \frac{x - x_{mean}}{s}$$

慕课网《Python3机器学习》

演示：实现两种归一化

讲师：liuyubobobo

版权所有，侵权必究

慕课网《Python3机器学习》

Scikit-Learn中的Scaler

讲师：liuyubobobo

版权所有，侵权必究

对测试数据集如何归一化？

训练数据

mean_train

std_train

测试数据

~~mean_test~~

~~std_test~~

对测试数据集如何归一化？



A vertical rectangle is divided into two horizontal sections. The top section is blue and labeled '训练数据' (Training Data). The bottom section is purple and labeled '测试数据' (Test Data).

训练数据

mean_train

std_train

测试数据

$(x_{\text{test}} - \text{mean_train}) / \text{std_train}$

对测试数据集如何归一化？



测试数据是模拟真实环境

- 真实环境很有可能无法得到所有测试数据的均值和方差
- 对数据的归一化也是算法的一部分

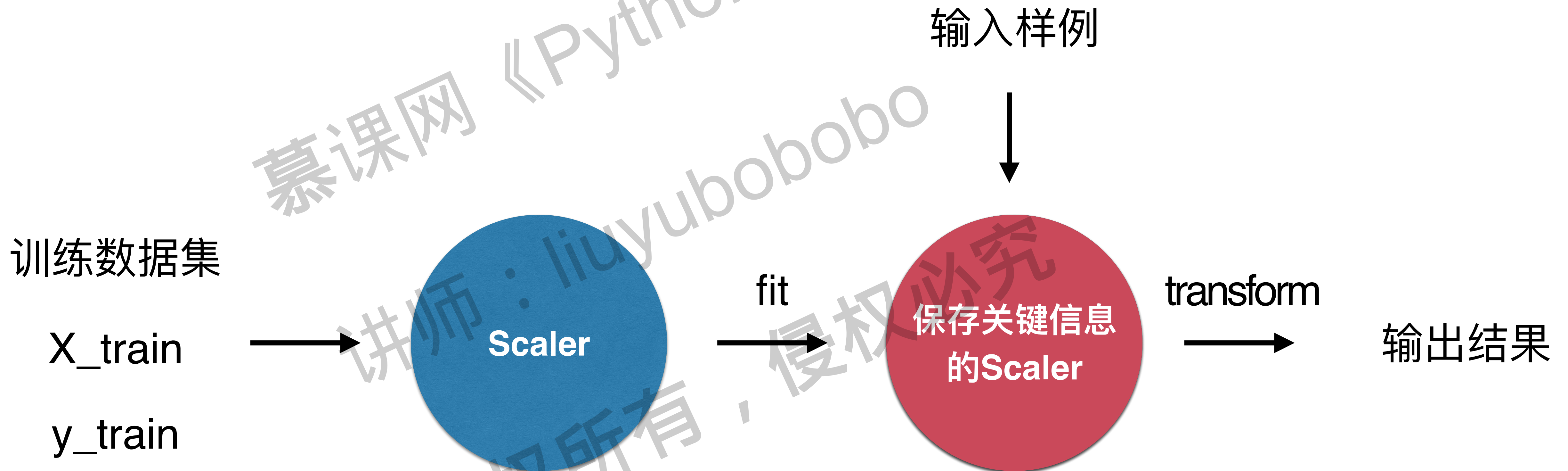
$$(x_test - mean_train) / std_train$$

对测试数据集如何归一化？

要保存训练数据集得到的均值和方差

scikit-learn中使用Scaler

对测试数据集如何归一化？



演示：scikit-learn中的Scaler

讲师：liuyubobobo

版权所有，侵权必究

演示：创建我们自己的Scaler

讲师：liuyunbo

版权所有，侵权必究

慕课网《Python3机器学习》

更多有关k近邻算法

讲师：liuyubobobo

版权所有，侵权必究

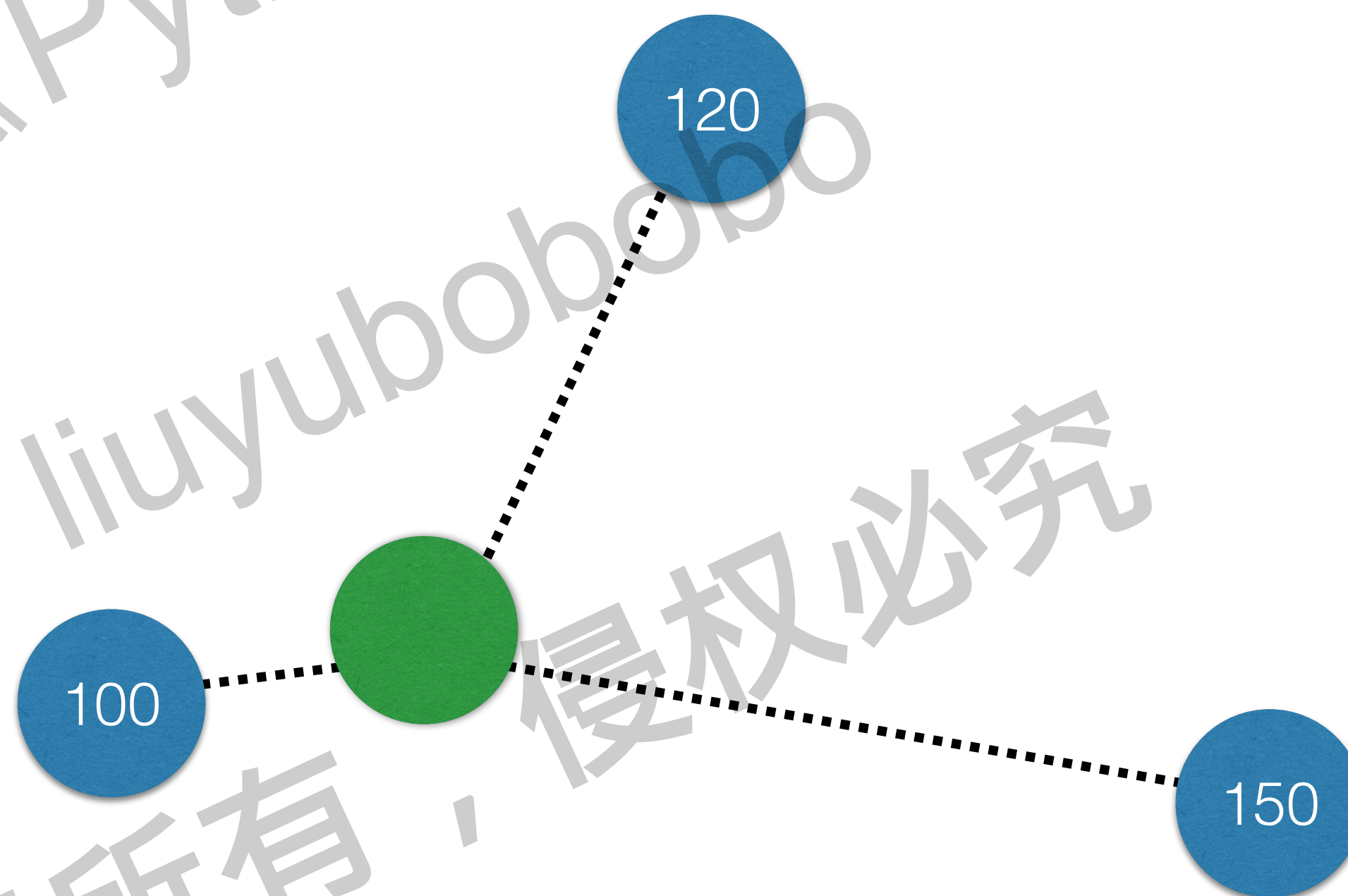
更多有关k近邻算法

解决分类问题

天然可以解决多分类问题

思想简单，效果强大

使用k近邻算法解决回归问题



KNeighborsRegressor

<http://scikit-learn.org/stable/modules/generated/sklearn.neighbors.KNeighborsRegressor.html>

更多有关k近邻算法

最大的缺点：效率低下

如果训练集有 m 个样本， n 个特征，则预测

每一个新的数据，需要 $O(m*n)$

优化，使用树结构：KD-Tree, Ball-Tree

更多有关k近邻算法

缺点2：高度数据相关

缺点3：预测结果不具有可解释性

更多有关k近邻算法

缺点4：维数灾难

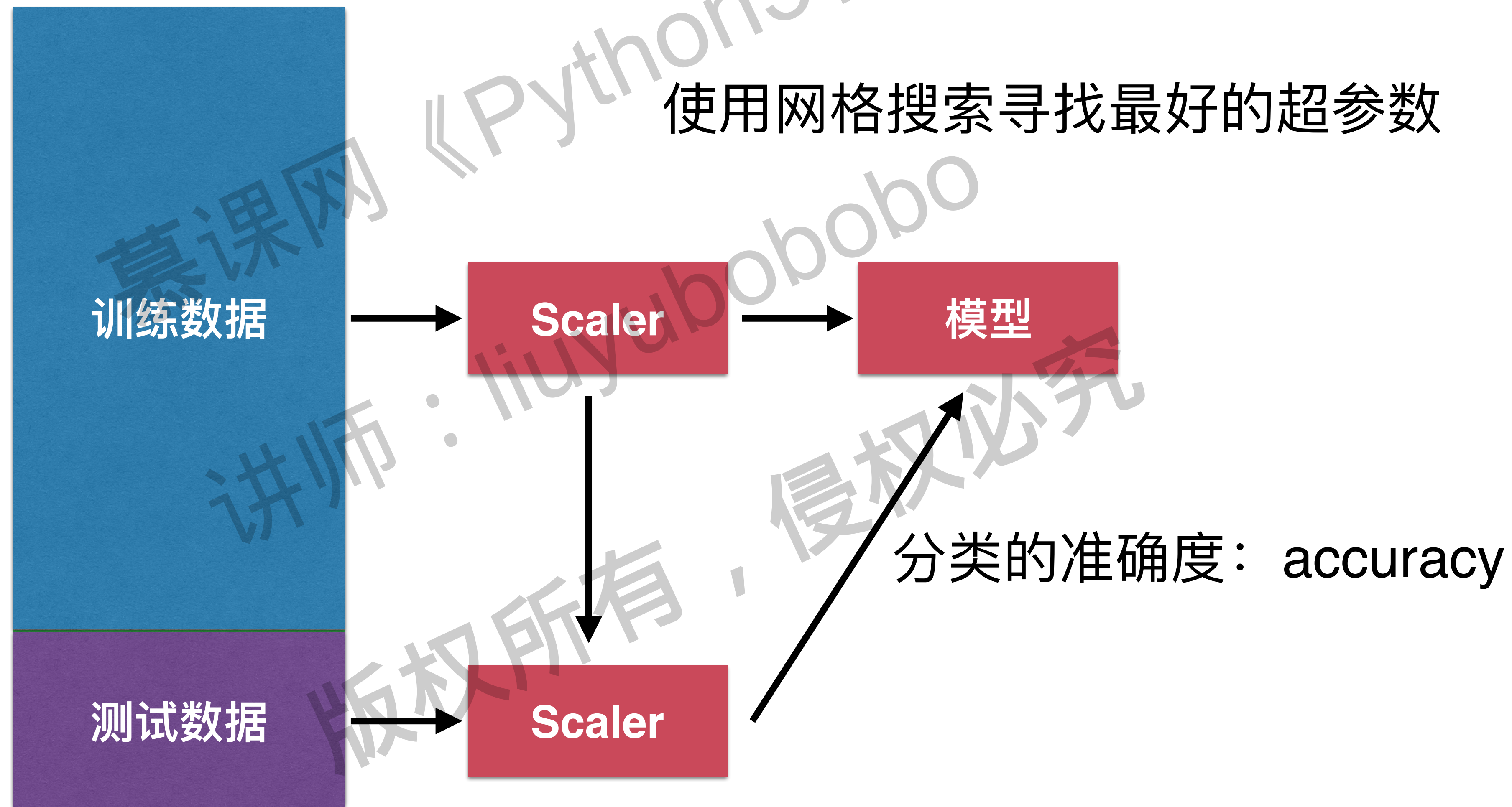
维数灾难

随着维度的增加，“看似相近”的两个点之间的距离越来越大

1维	0到1的距离	1
2维	(0,0)到(1,1)的距离	1.414
3维	(0,0,0)到(1,1,1)的距离	1.73
64维	(0,0,...0)到(1,1,...,1)	8
10000维	(0,0,...0)到(1,1,...,1)	100

解决方法：降维

机器学习流程回顾



其他

欢迎大家关注我的个人公众号：是不是很酷



Python 3 玩儿转机器学习

讲师：liuyubobobo

版权所有 侵权必究
liuyubobobo