Python 3 玩儿转机器学习

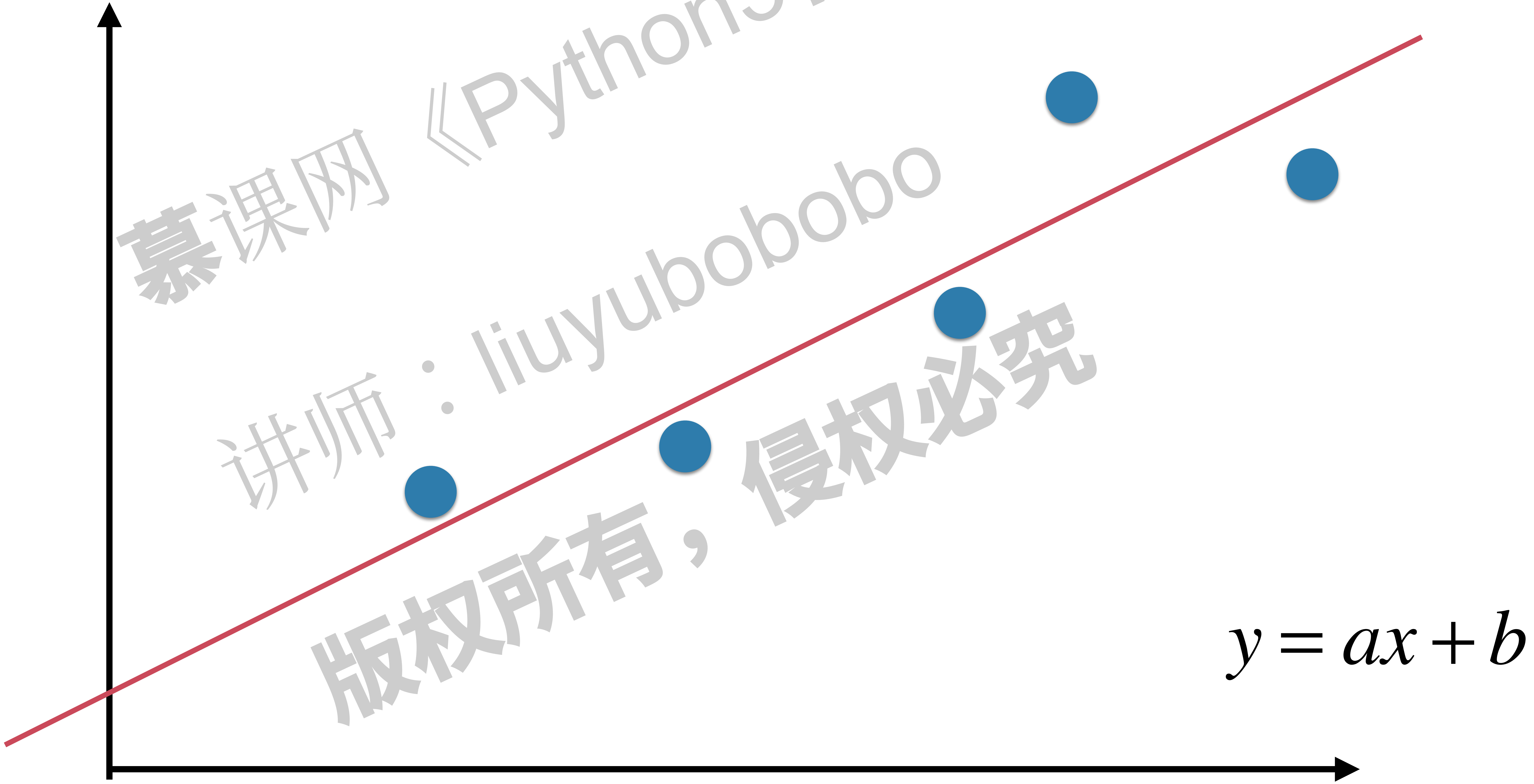
liuyubobobo

多项式回归与模型泛化

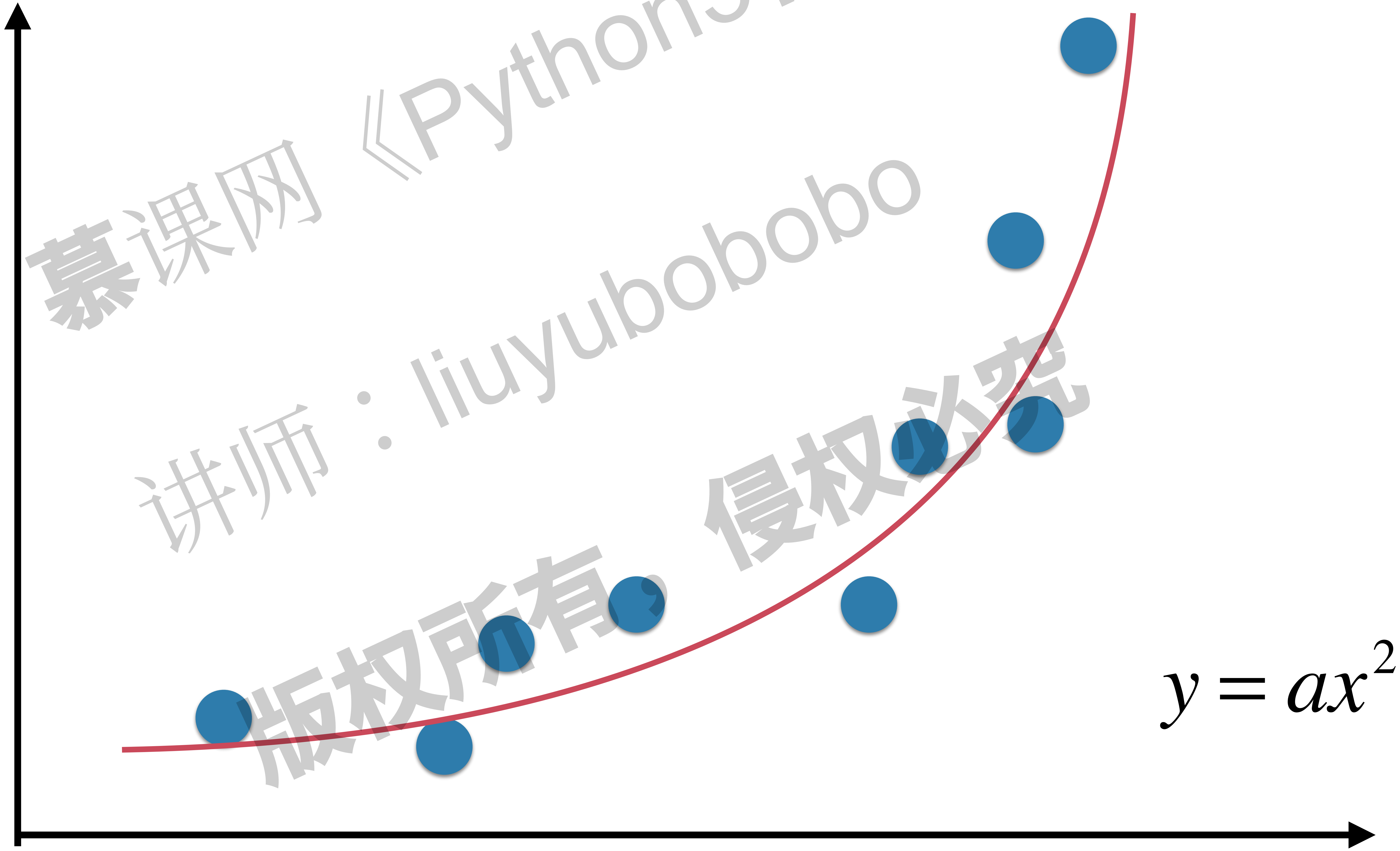
Polynomial Regression and Model Generalization

多项式回归

线性回归



多项式回归



$$y = ax^2 + bx + c$$

实践：多项式回归

scikit-learn中的多项式回归

实践：scikit-learn中的多项式回归

PolynomialFeatures(degree=3)

$$x_1, x_2 \Rightarrow 1, x_1, x_2$$

$$x_1^2, x_2^2, x_1 x_2$$

$$x_1^3, x_2^3, x_1^2 x_2, x_1 x_2^2$$

实践：Pipeline

欠拟合和过拟合

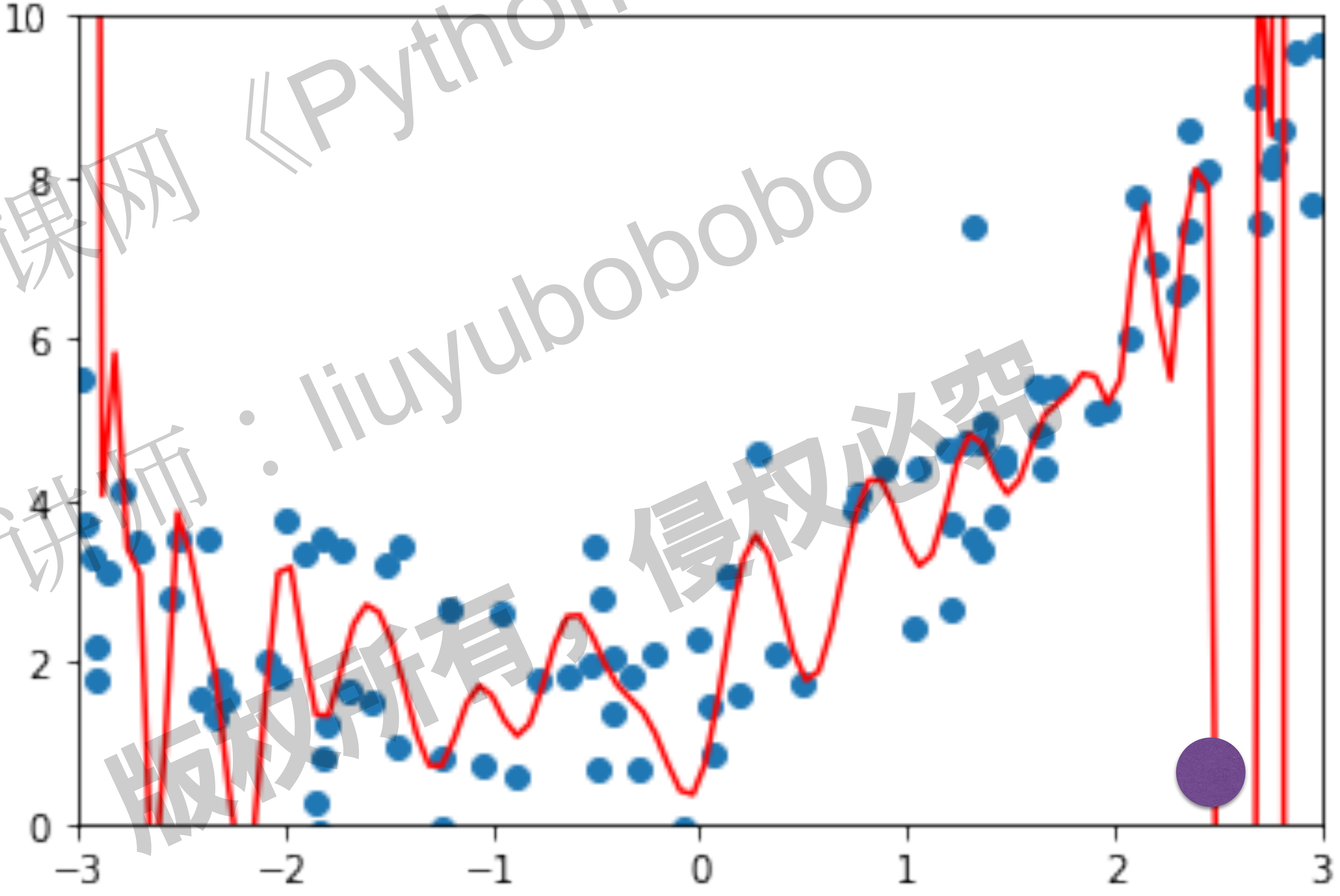
欠拟合和过拟合

- 欠拟合 underfitting
- 过拟合 overfitting

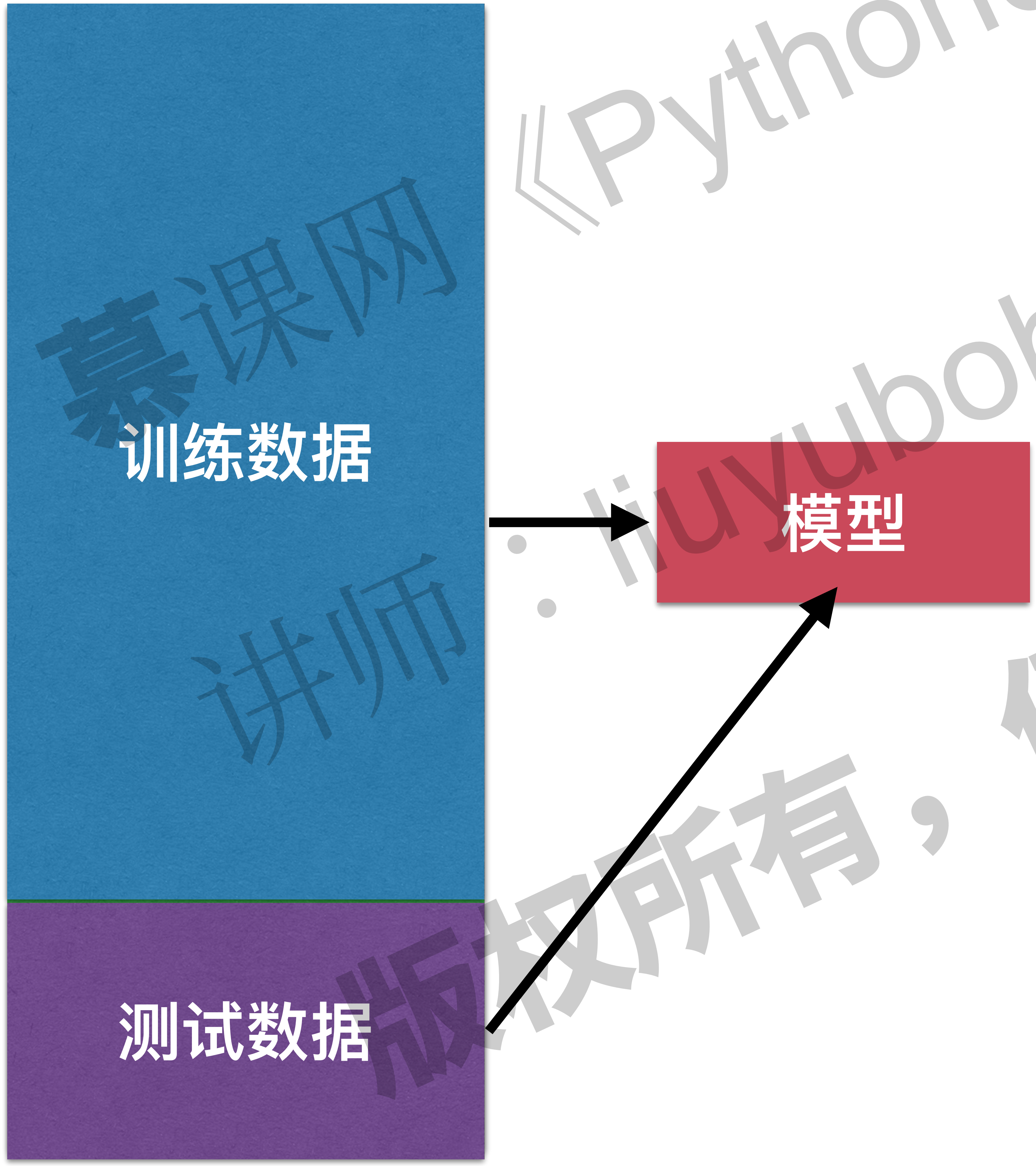
实践：欠拟合和过拟合

模型的泛化能力

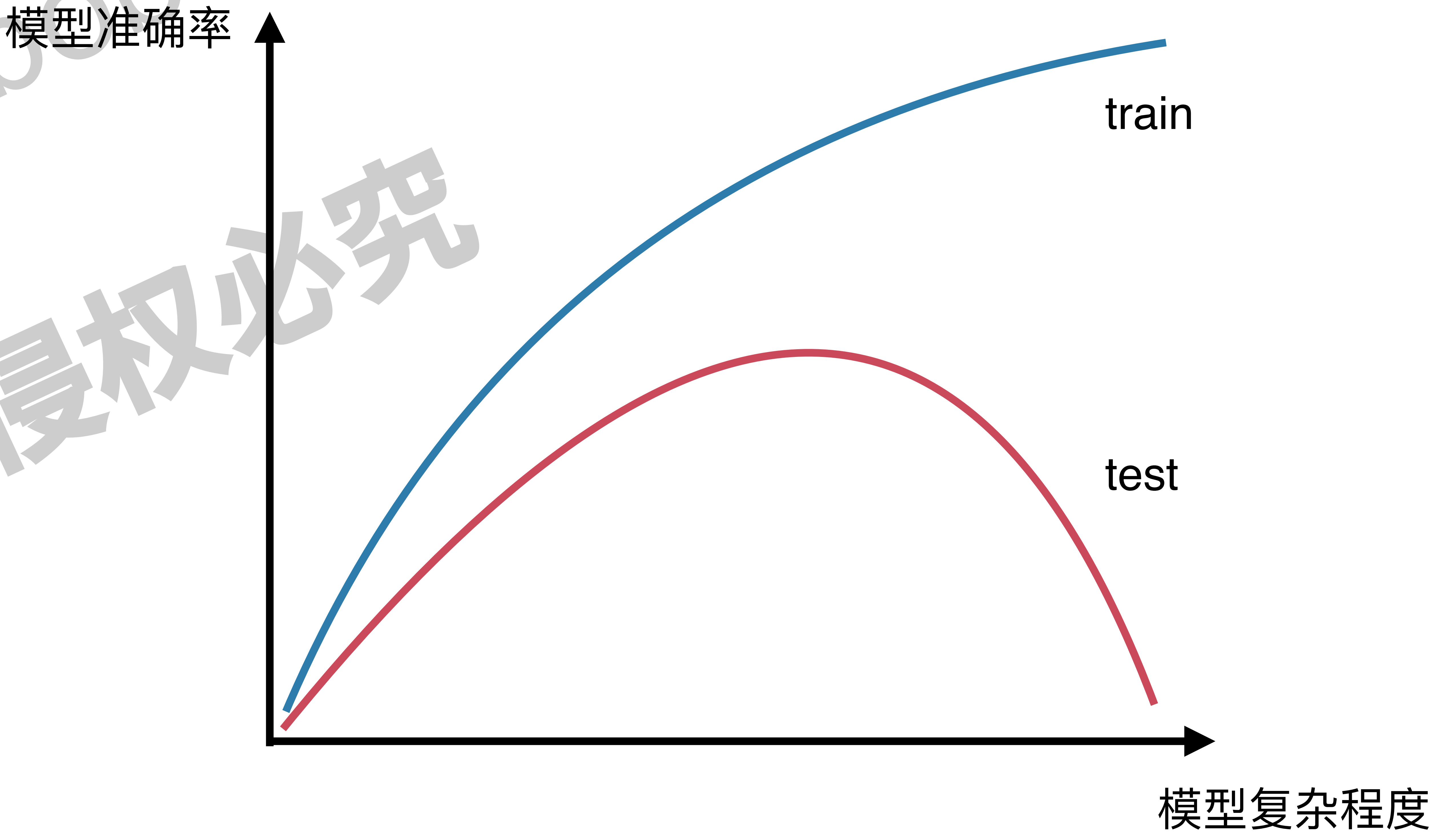
模型的泛化能力



测试数据集的意义



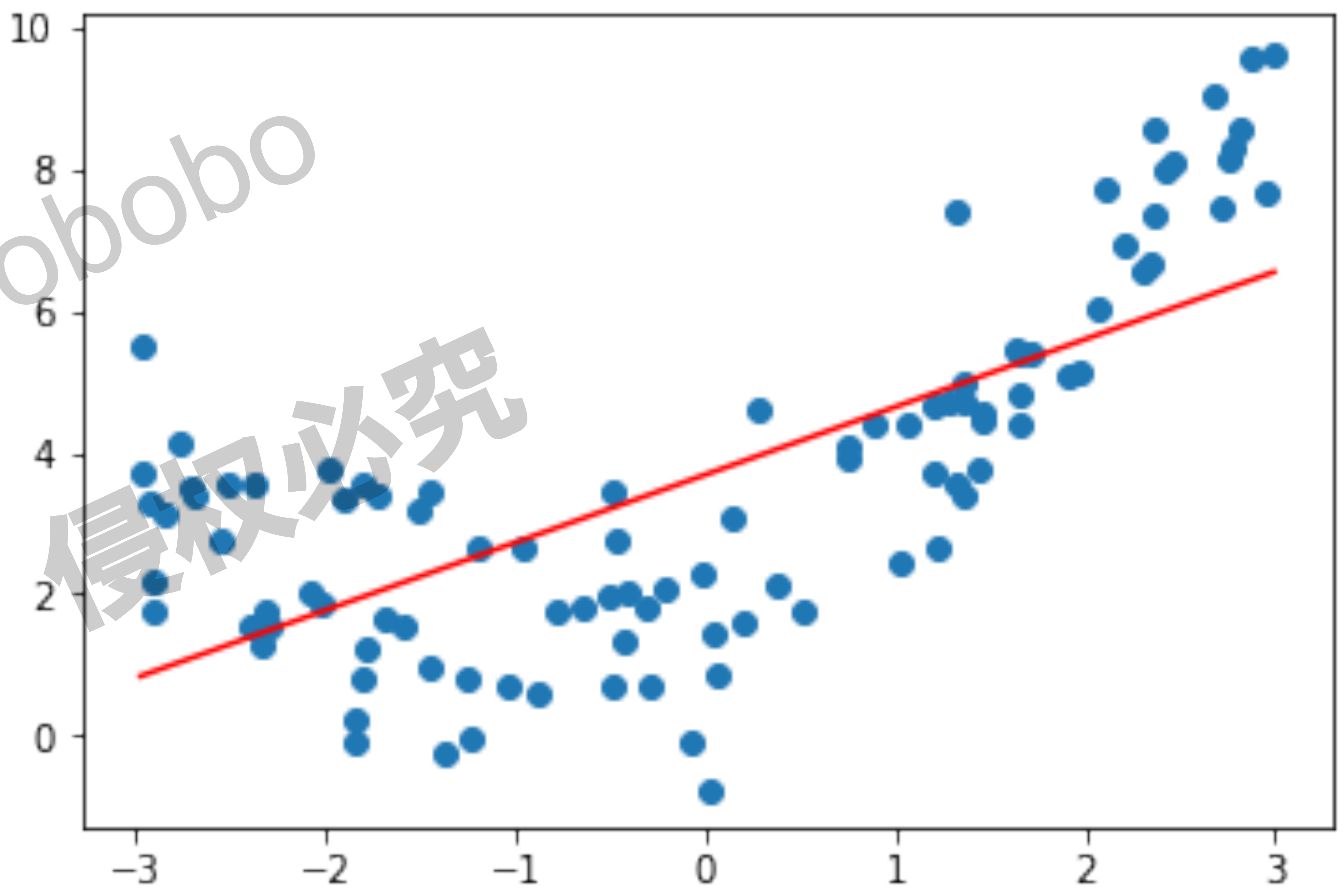
模型复杂度



欠拟合和过拟合

欠拟合 underfitting

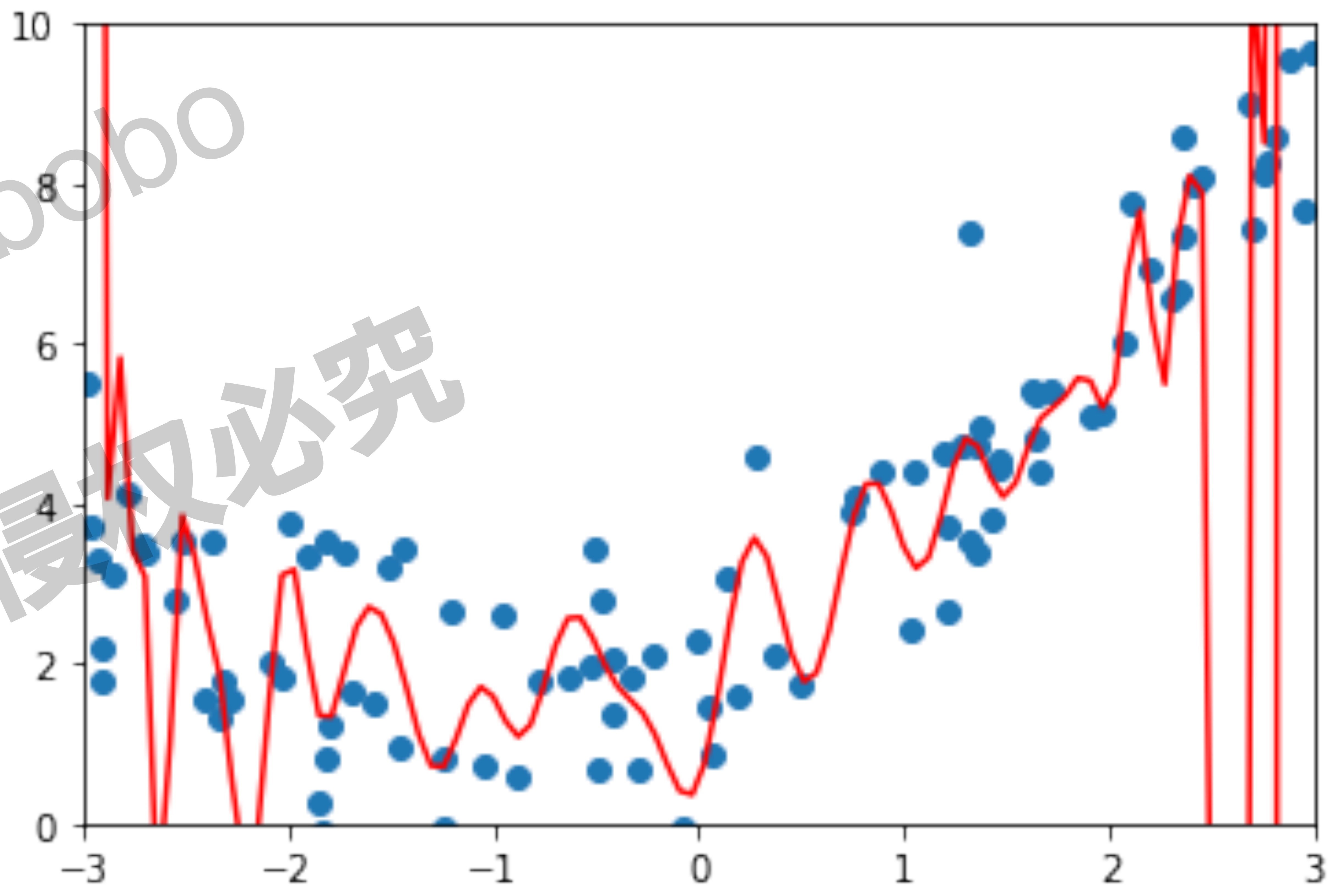
算法所训练的模型不能完整表述数据关系



欠拟合和过拟合

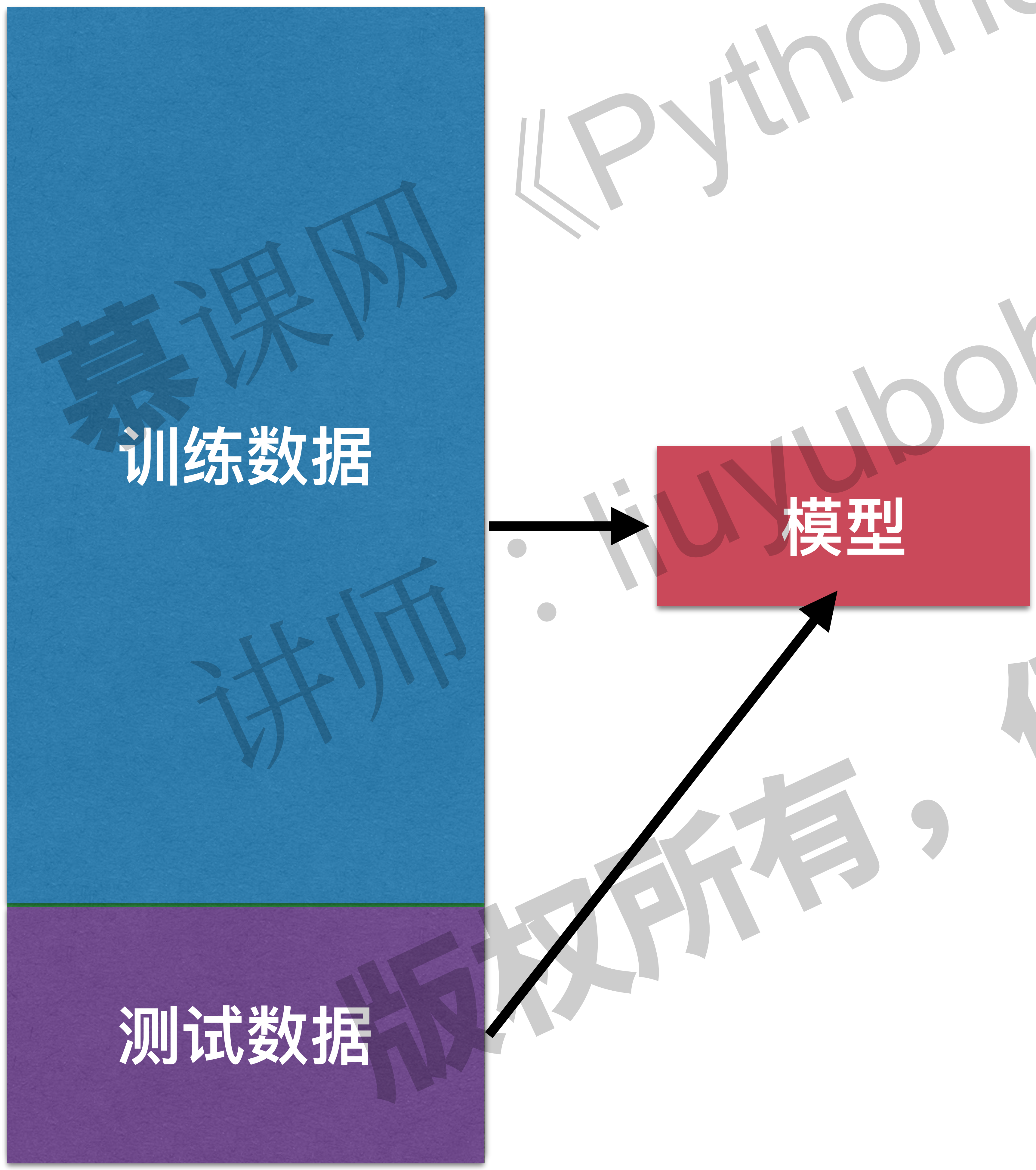
过拟合 overfitting

算法所训练的模型过多地表达了数据间的噪音关系

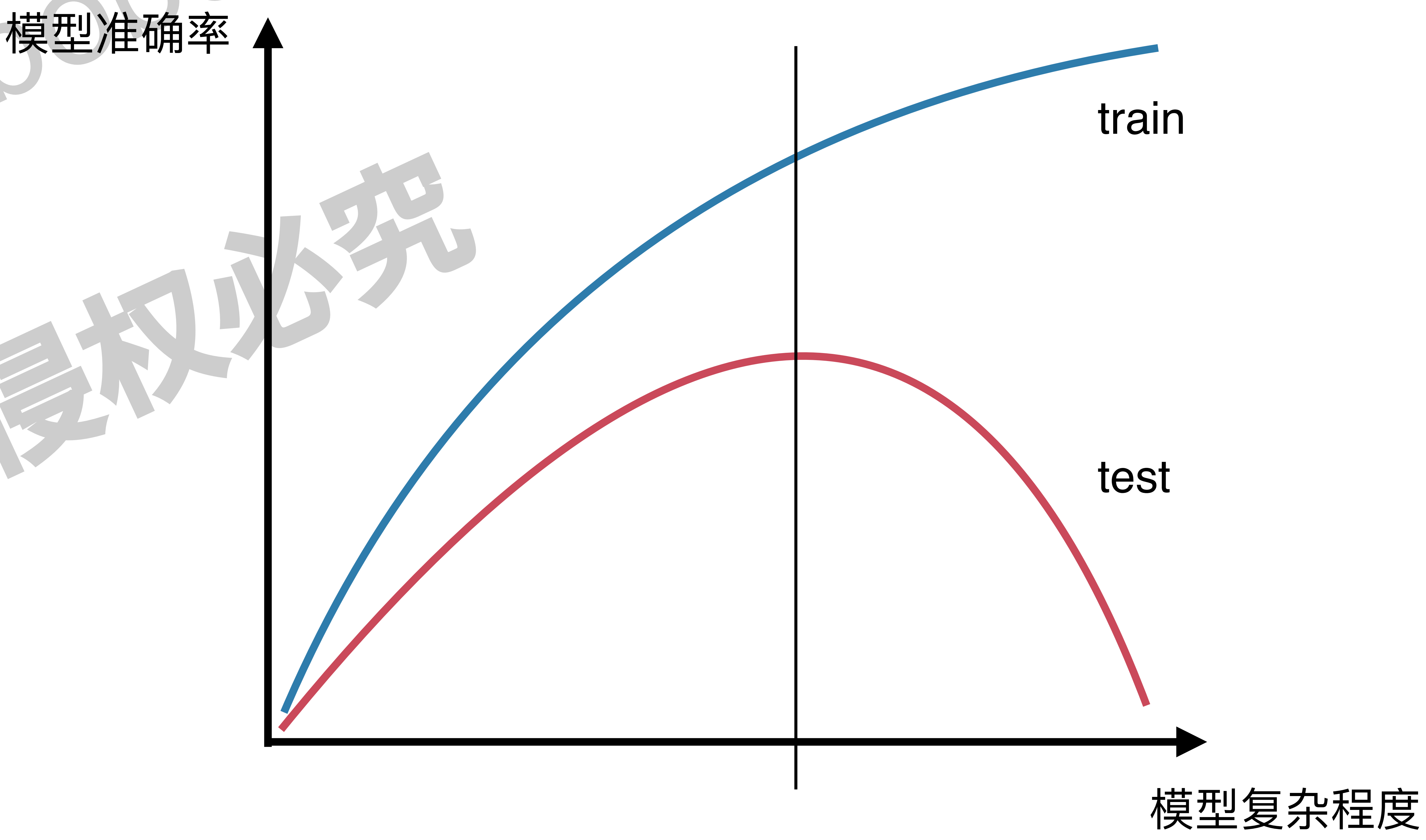


一个生活中的例子

测试数据集的意义

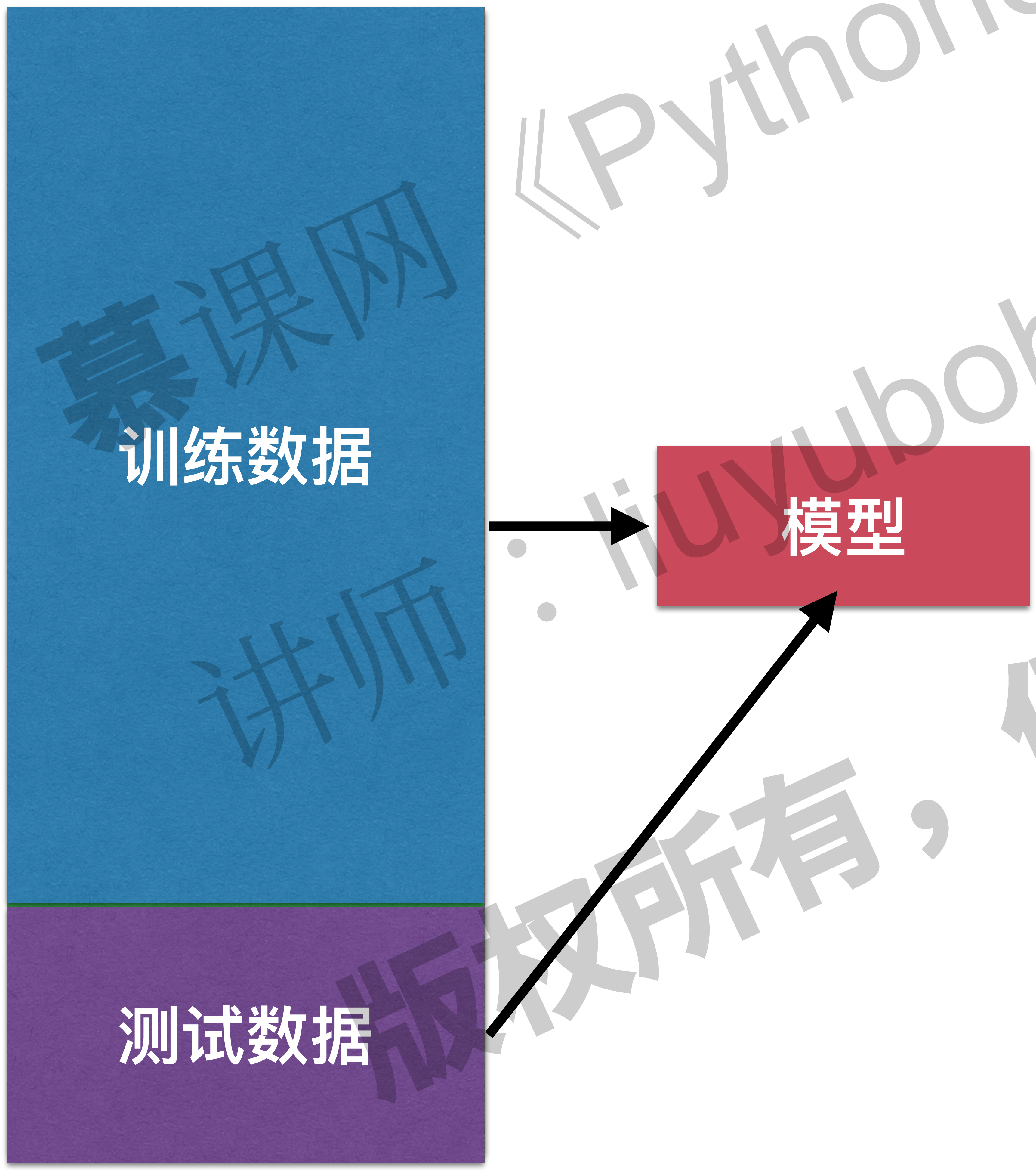


寻找泛化能力最好的地方

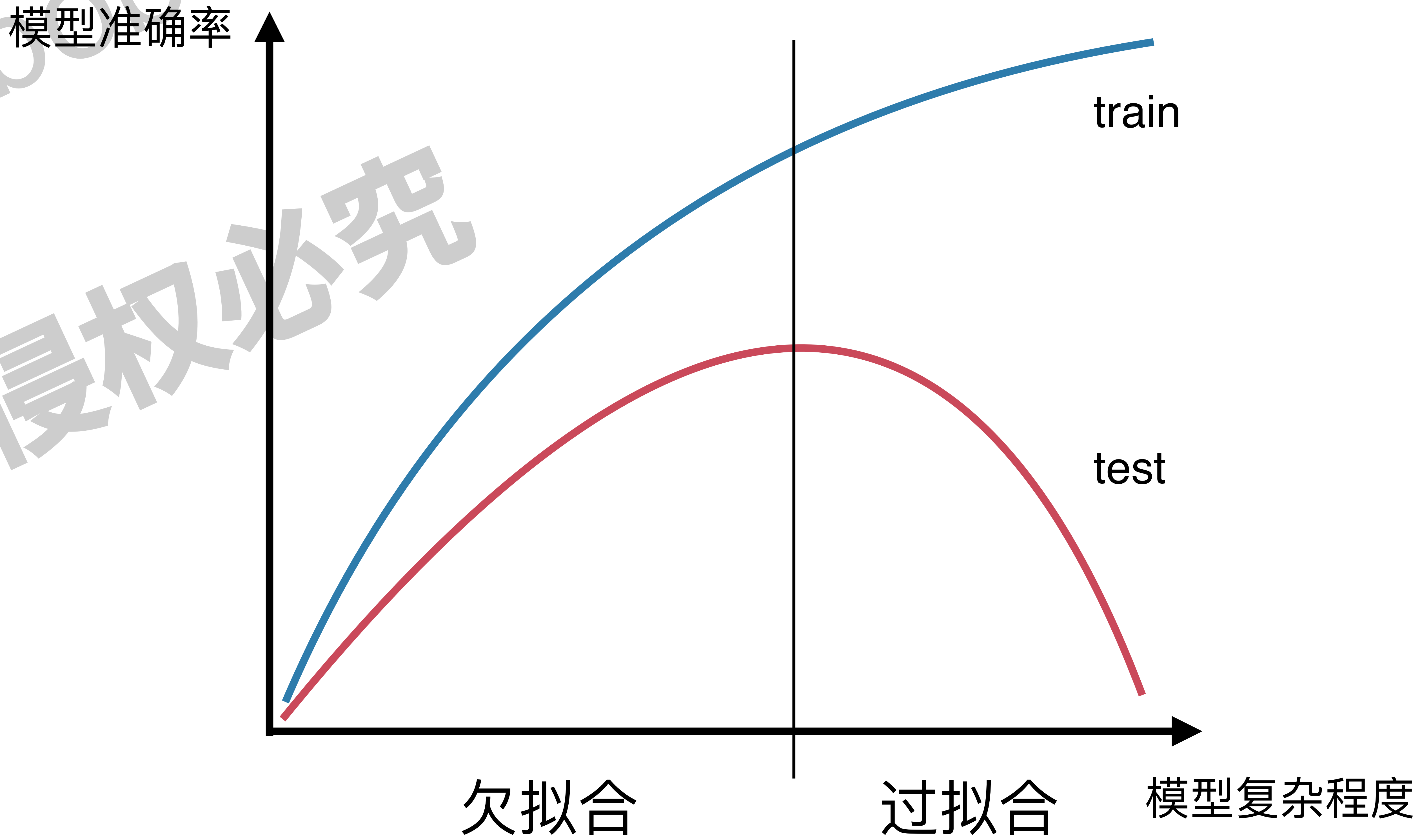


学习曲线

模型复杂度曲线



寻找泛化能力最好的地方

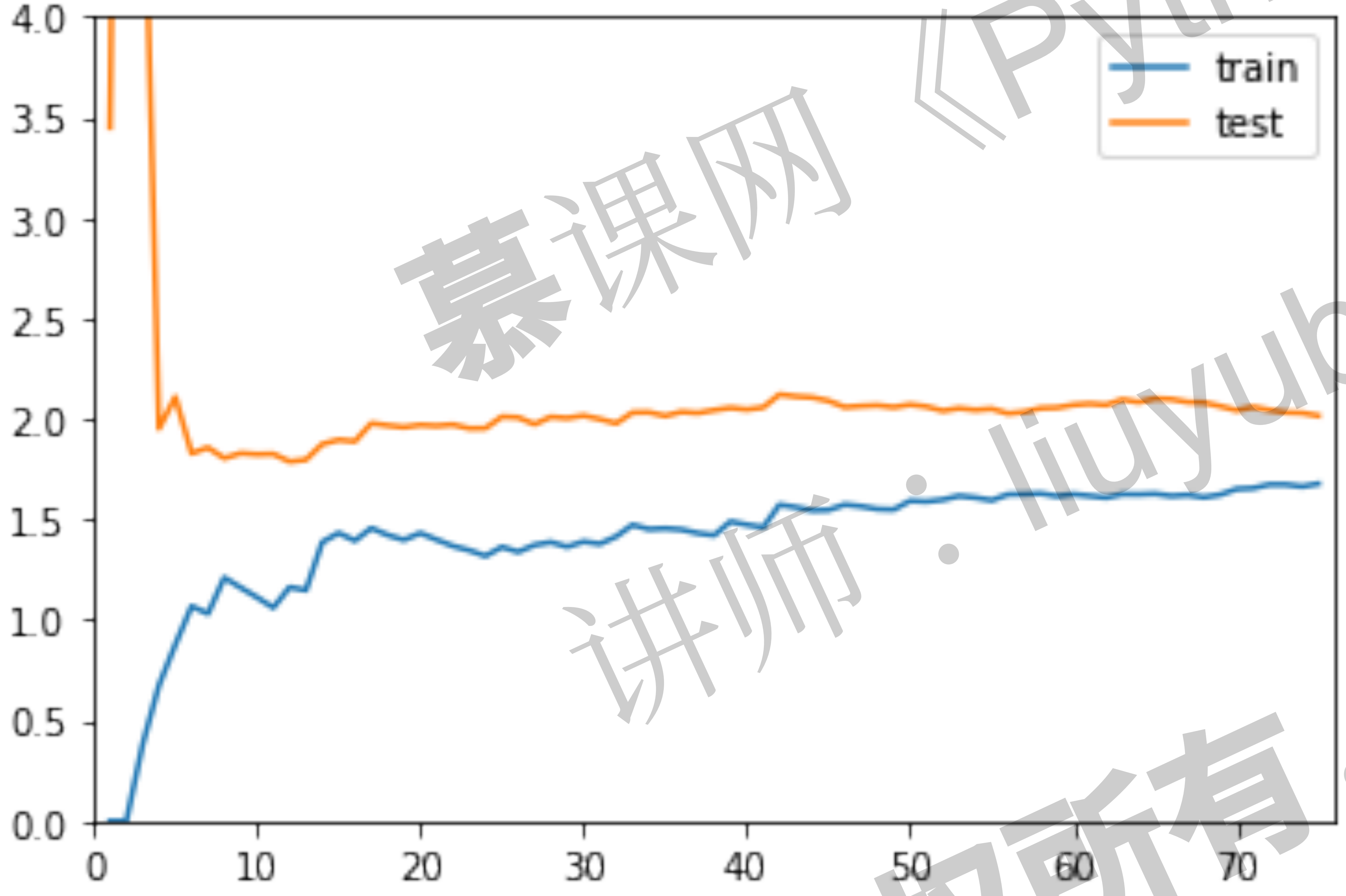


学习曲线

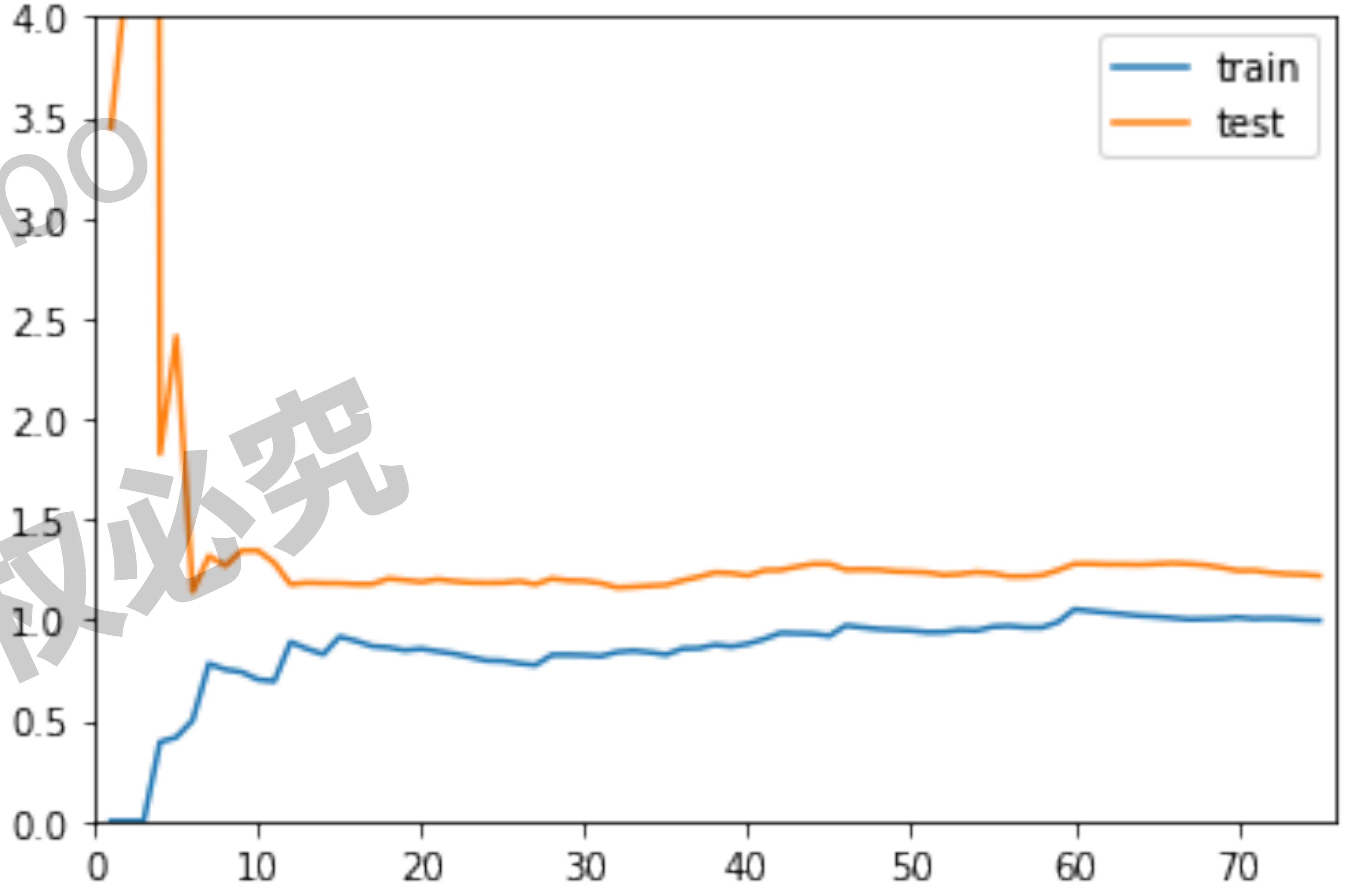
随着训练样本的逐渐增多，算法训练出的模型的表现能力。

实践：学习曲线

学习曲线

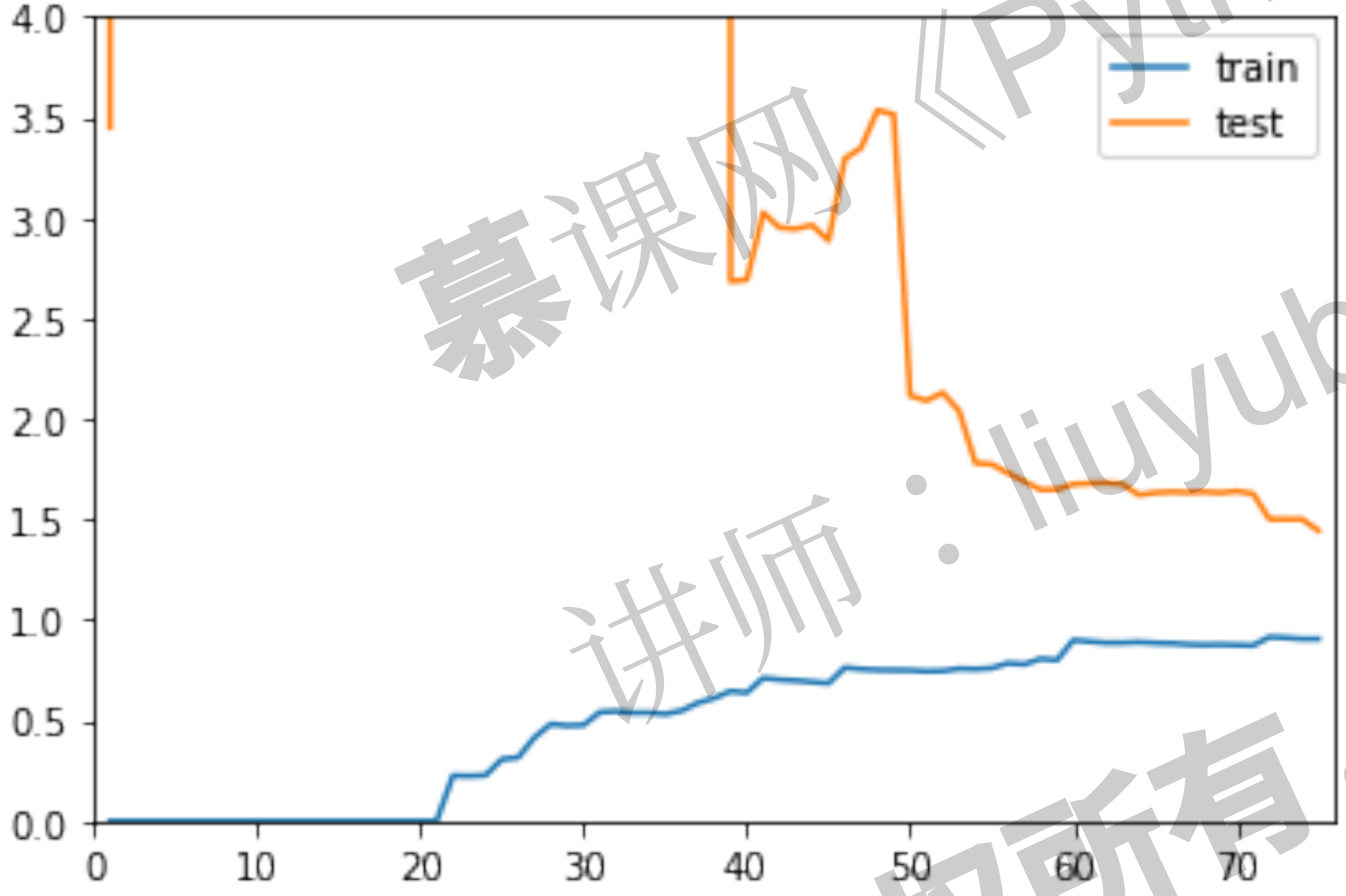


欠拟合



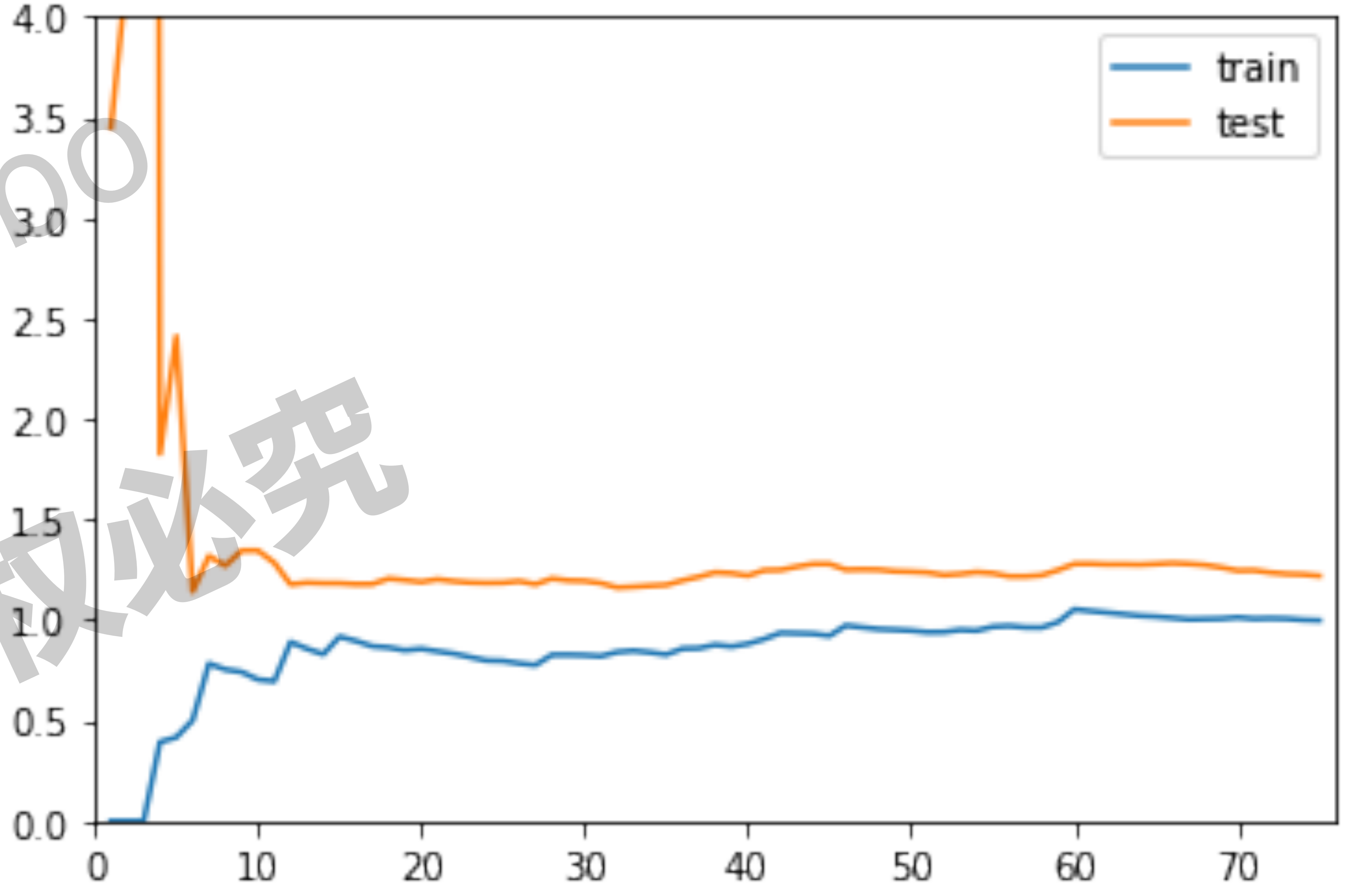
最佳

学习曲线



4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
0
10
20
30
40
50
60
70
train
test

过拟合

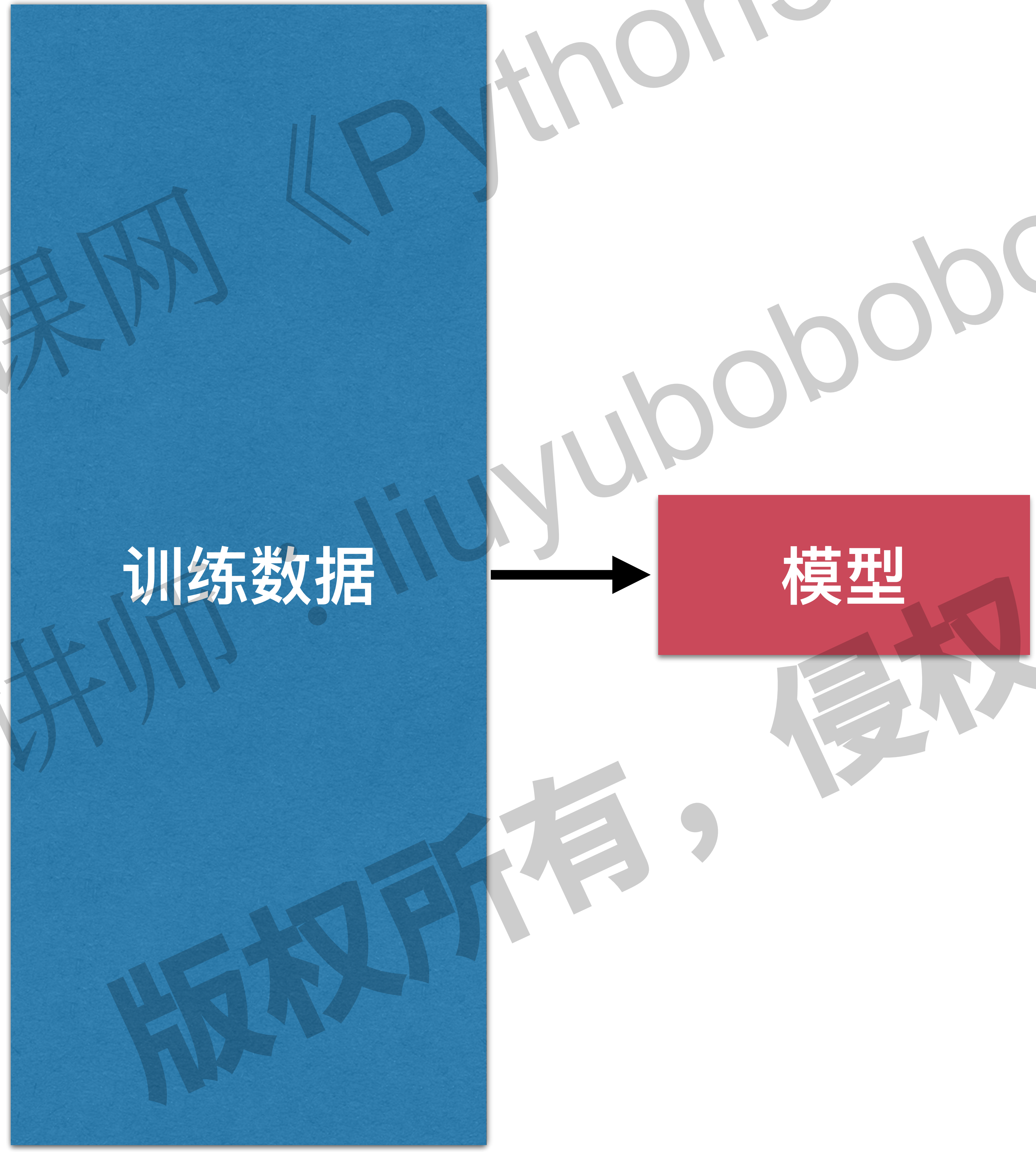


4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
0
10
20
30
40
50
60
70
train
test

最佳

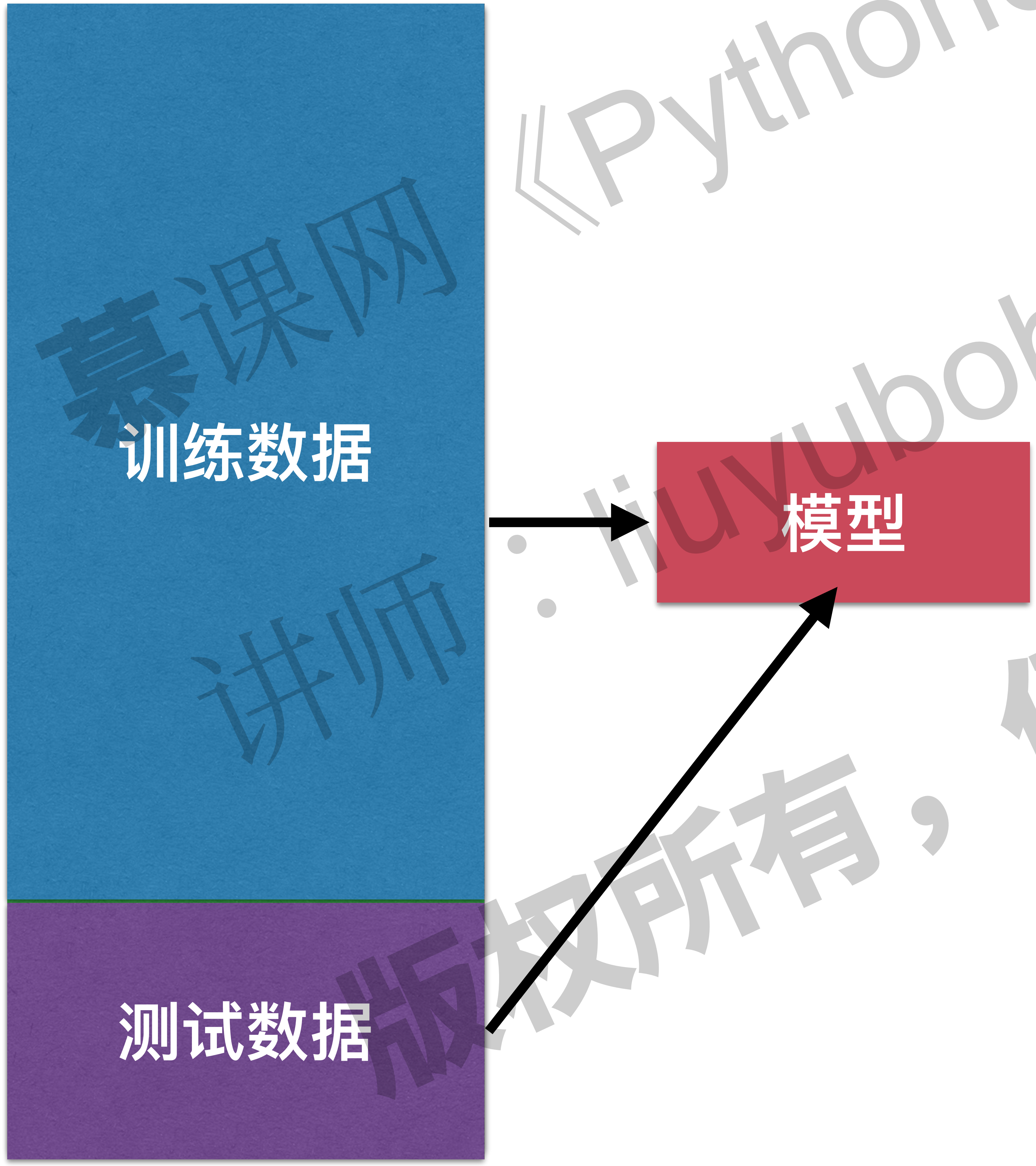
验证集和交叉验证

测试数据集的意义

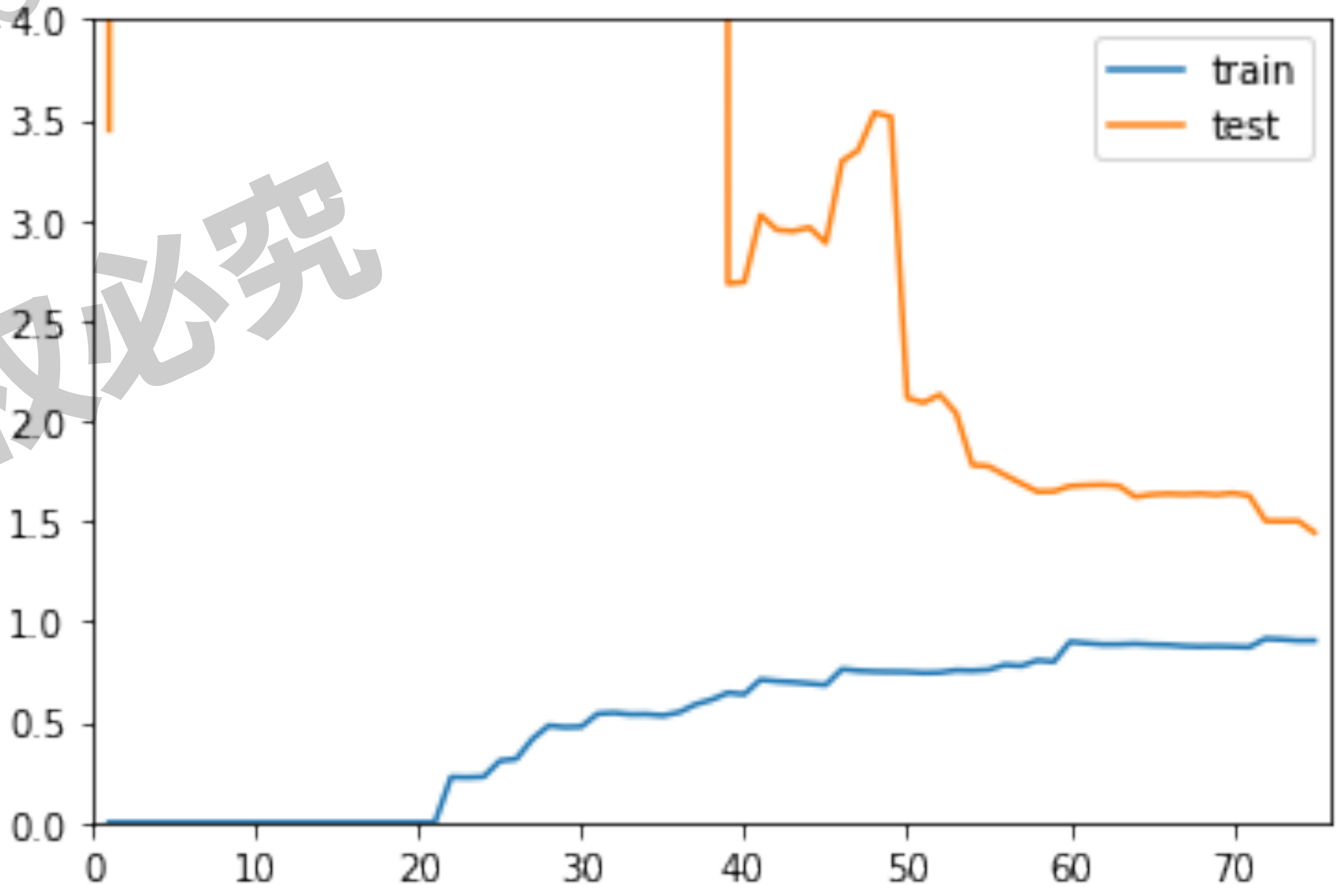


发生了过拟合却不自知

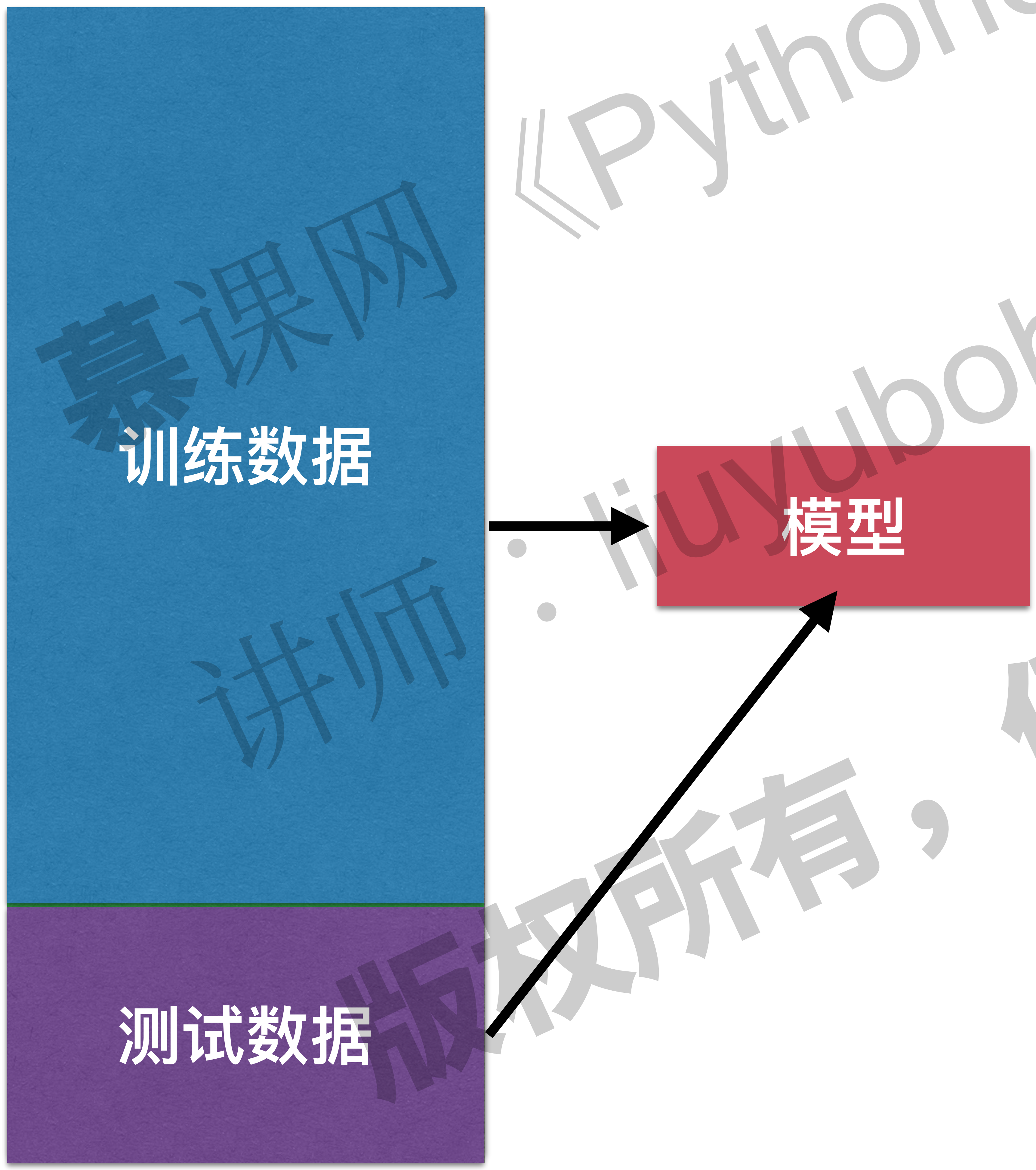
测试数据集的意义



通过测试数据集判断模型好坏

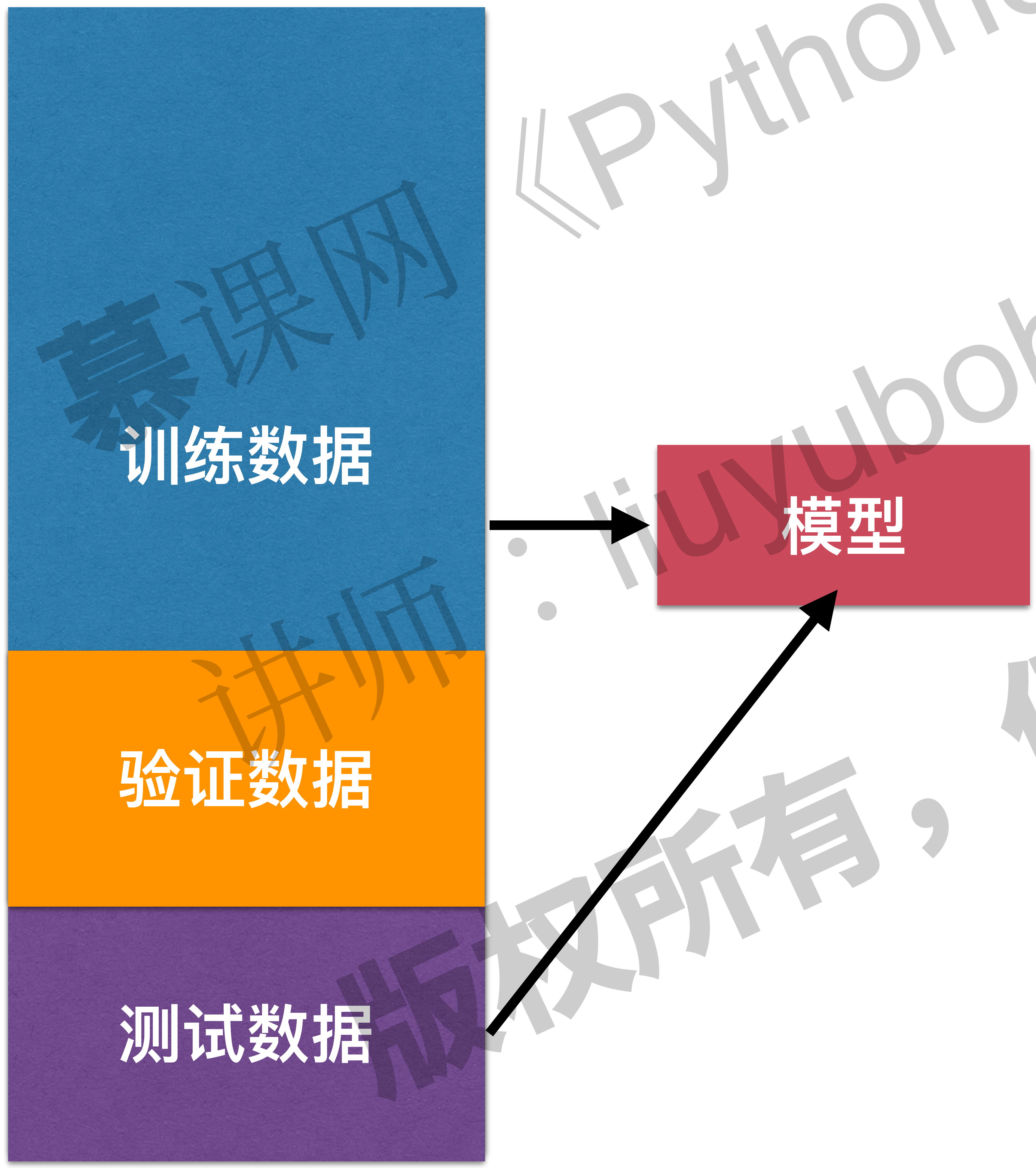


测试数据集的意义



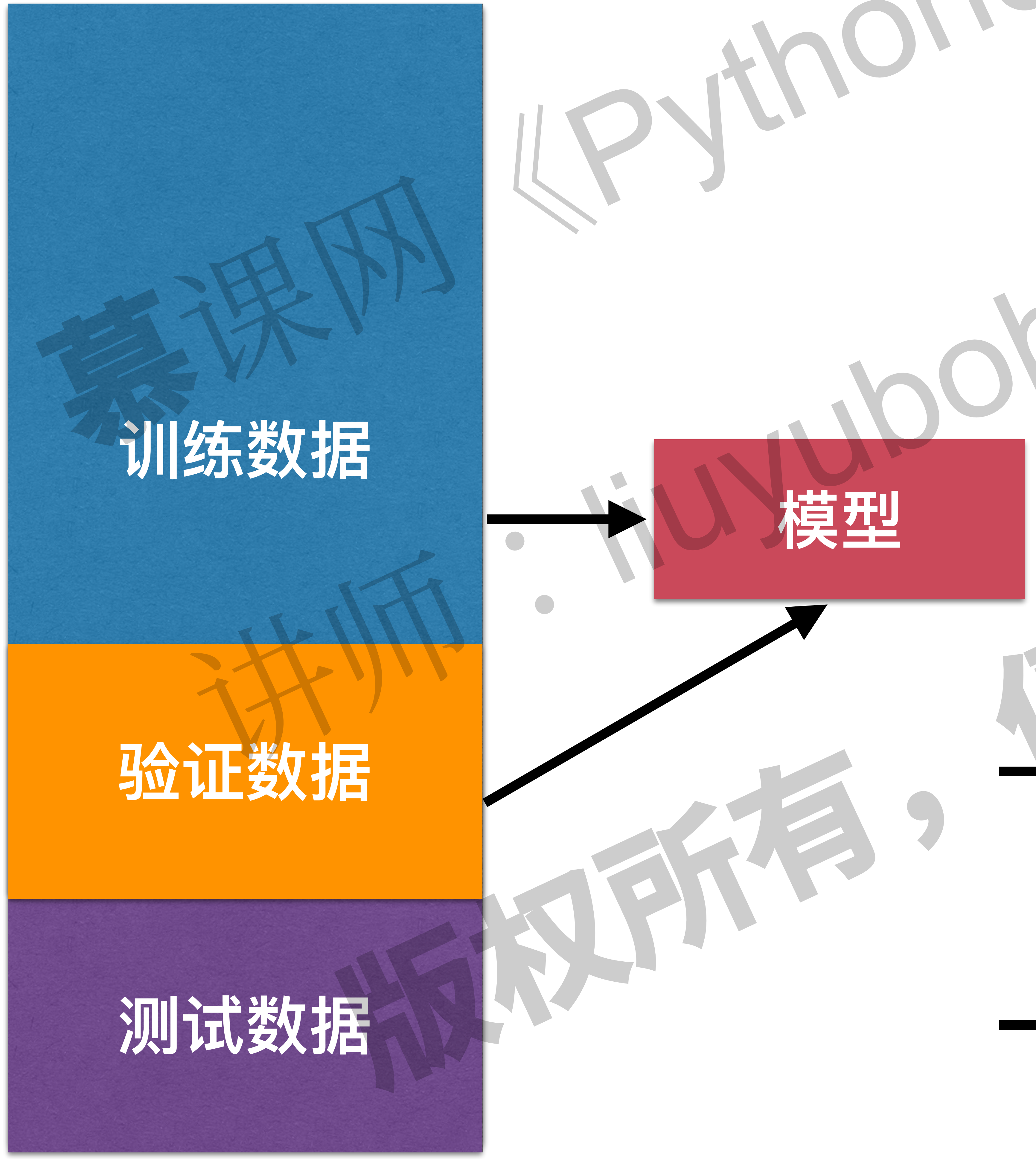
问题：针对特定测试数据集过拟合?

测试数据集的意义



问题：针对特定测试数据集过拟合？

测试数据集的意义

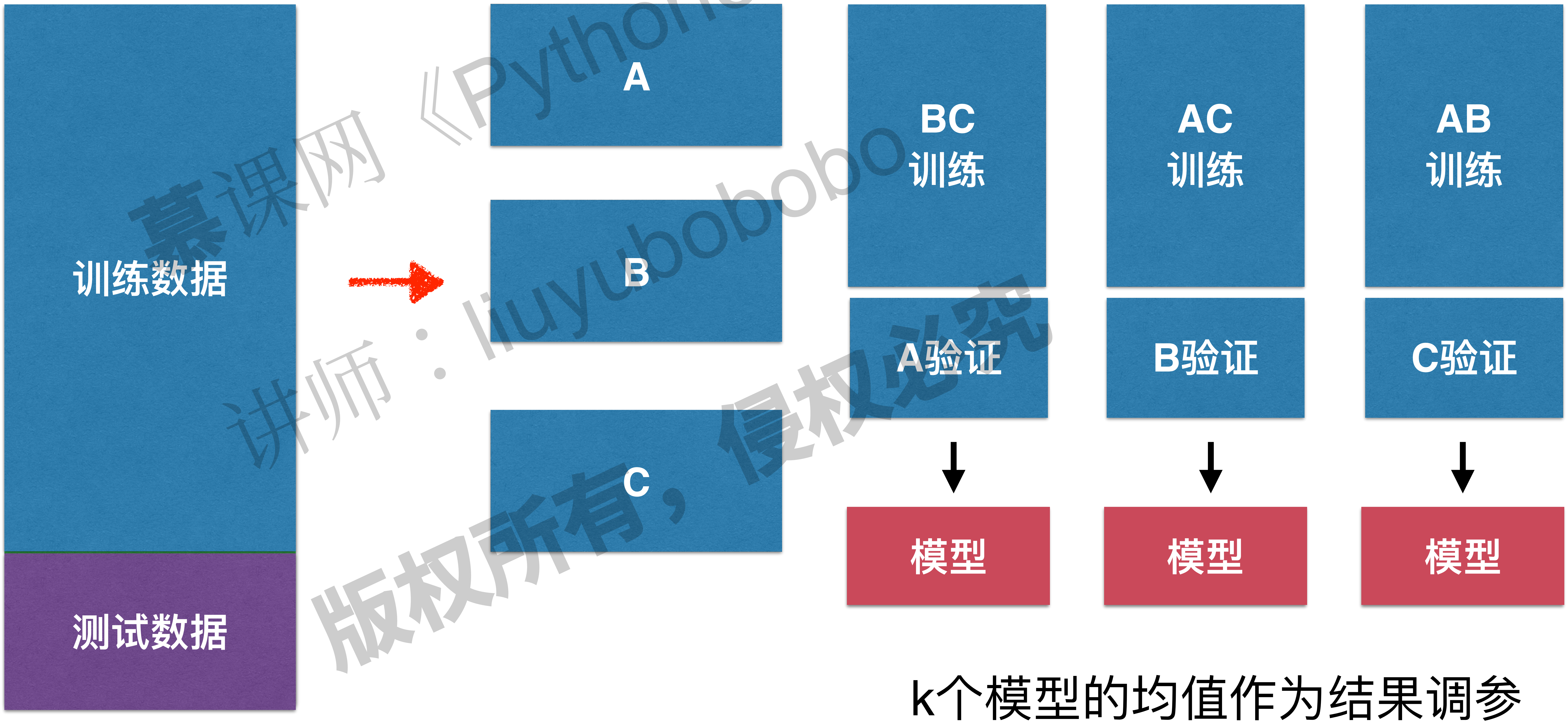


问题：随机？

→ 调整超参数使用的数据集

→ 作为衡量最终模型性能的数据集

交叉验证 Cross Validation



k个模型的均值作为结果调参

实践：交叉验证

k-folds 交叉验证

把训练数据集分成k份，

称为k-folds cross validation

缺点，每次训练k个模型，相当于整体性能慢了k倍

留一法 LOO-CV

把训练数据集分成m份，称为留一法

Leave-One-Out Cross Validation

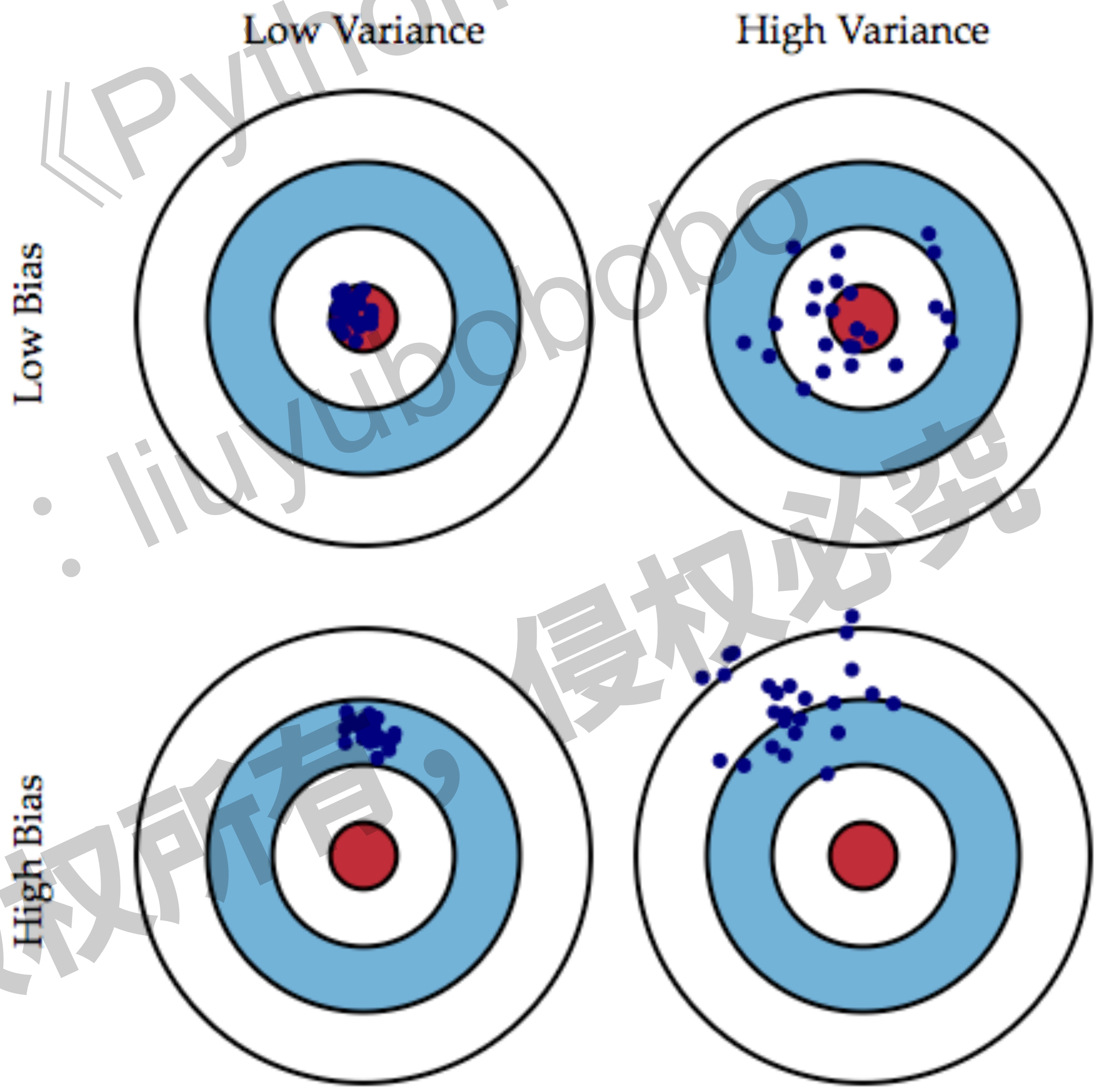
完全不受随机的影响，最接近模型真正的性能指标

缺点：计算量巨大

偏差方差权衡

Bias Variance Trade off

偏差和方差



模型误差

模型误差 = 偏差(Bias) + 方差(Variance) + 不可避免的误差

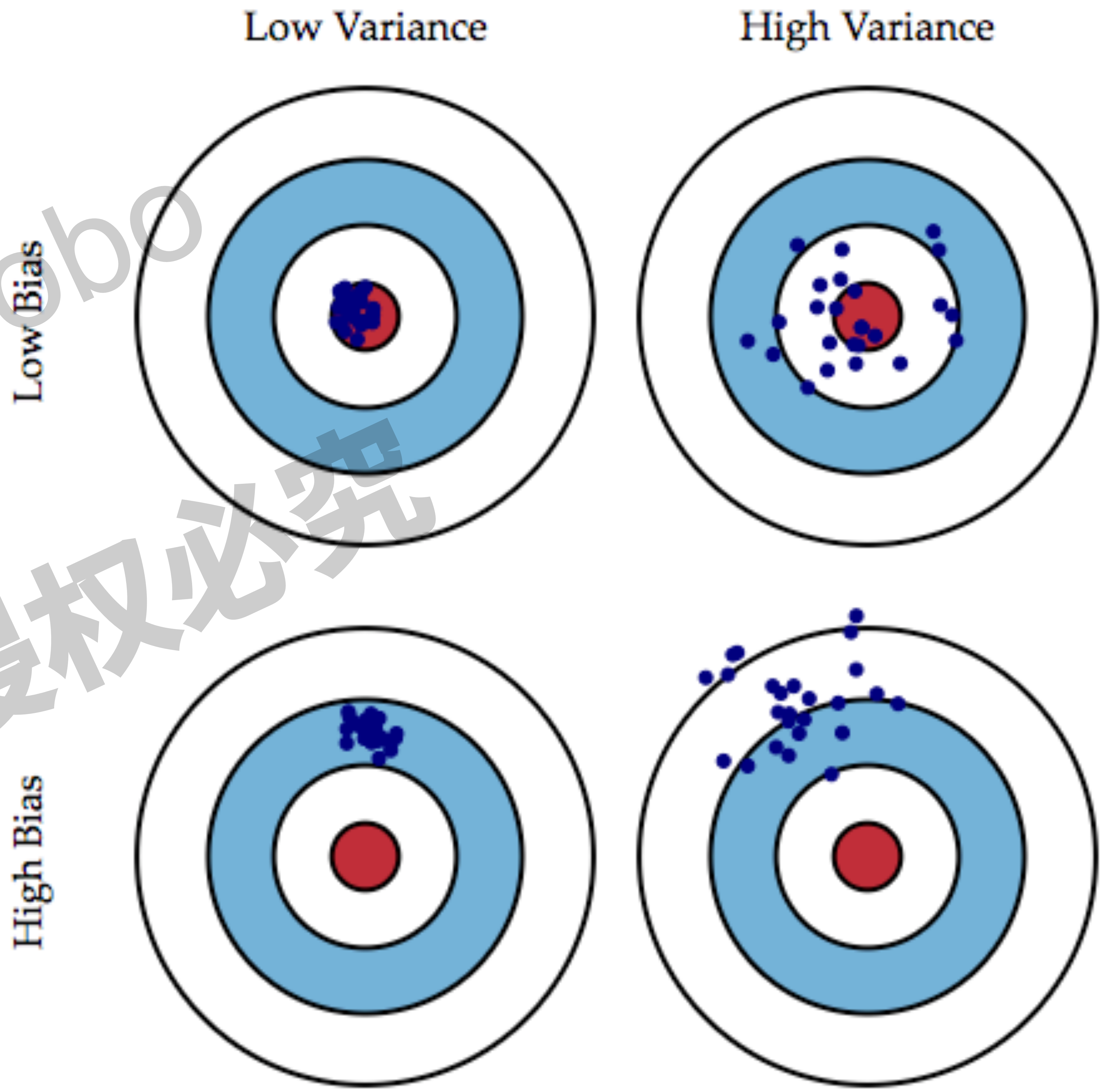
偏差（Bias）

导致偏差的主要原因：

对问题本身的假设不正确！

如：非线性数据使用线性回归

欠拟合 underfitting



方差（Variance）

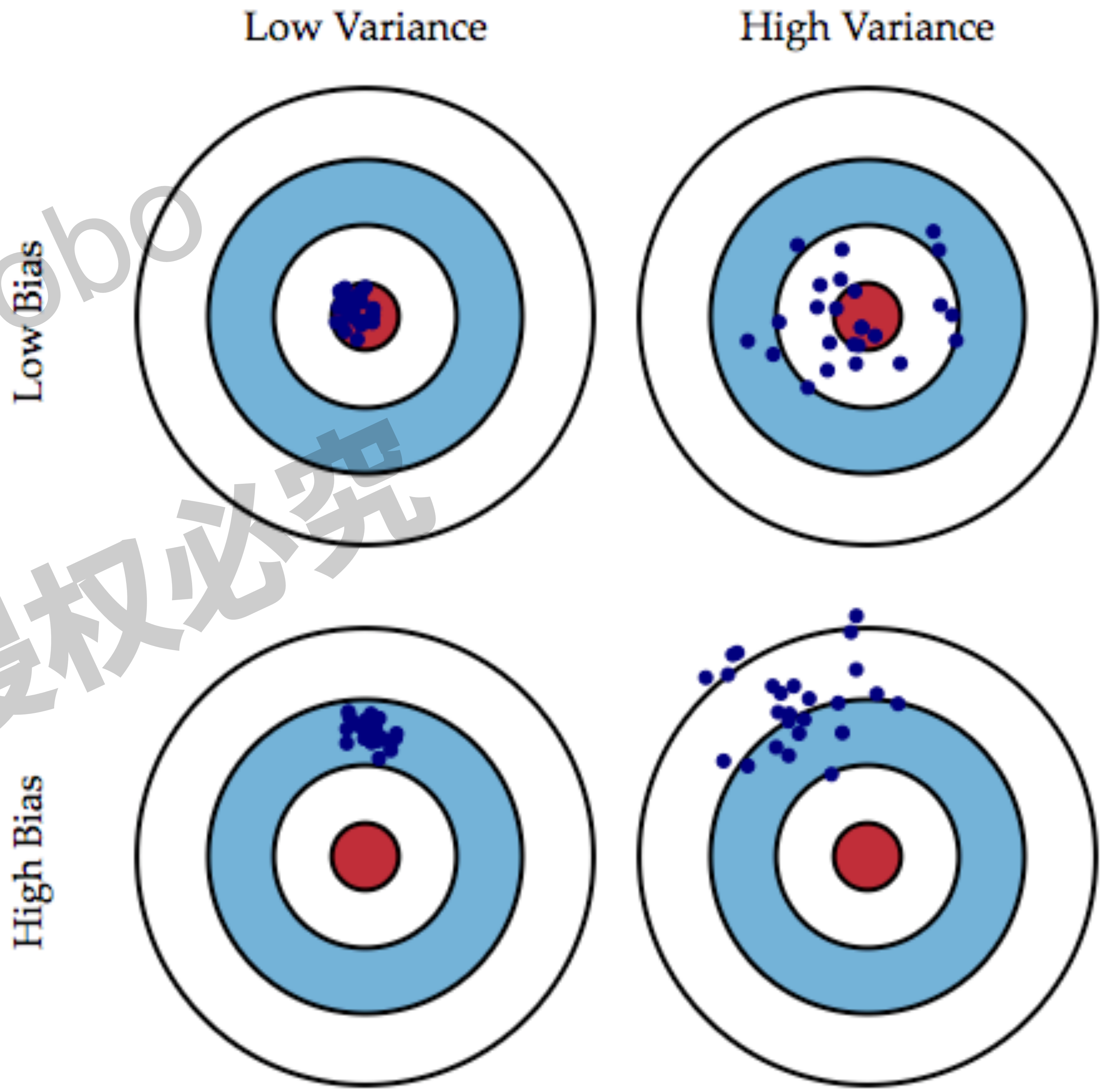
数据的一点点扰动都会

较大地影响模型。

通常原因，使用的模型太复杂。

如高阶多项式回归。

过拟合 overfitting



偏差和方差

有一些算法天生是高方差的算法。如kNN。

非参数学习通常都是高方差算法。因为不对数据进行任何假设

有一些算法天生是高偏差算法。如线性回归。

参数学习通常都是高偏差算法。因为对数据具有极强的假设

偏差和方差

大多数算法具有相应的参数，可以调整偏差和方差

如kNN中的k。

如线性回归中使用多项式回归。

偏差和方差

偏差和方差通常是矛盾的。

降低偏差，会提高方差。

降低方差，会提高偏差。

方差

机器学习的主要挑战，来自于方差！

解决高方差的通常手段：

1 降低模型复杂度

2 减少数据维度；降噪

3 增加样本数

4 使用验证集

方差

解决高方差的通常手段：

1 降低模型复杂度

2 减少数据维度；降噪

3 增加样本数

4 使用验证集

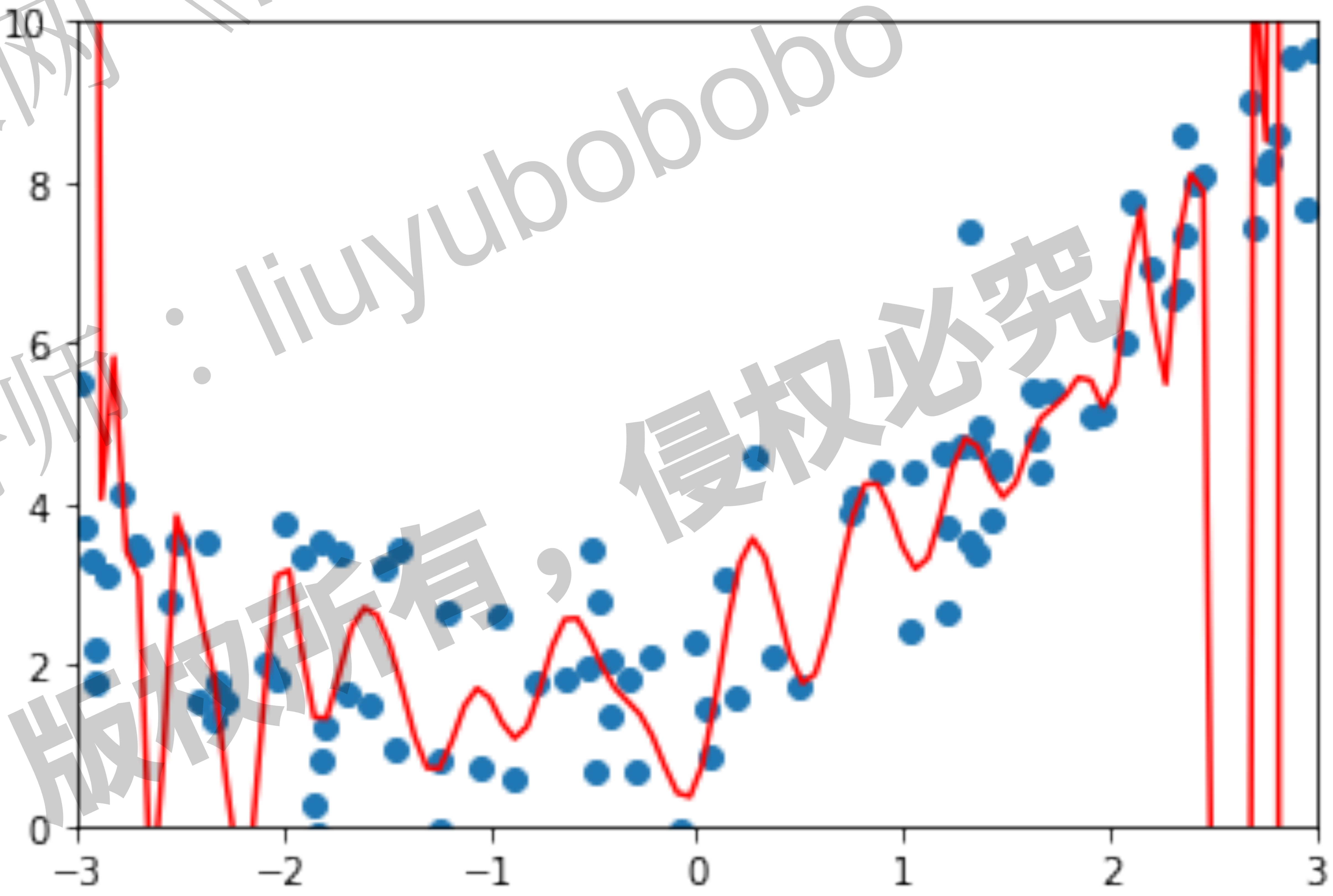
5 模型正则化

模型正则化

Regularization

模型正则化 Regularization

模型正则化：限制参数的大小



模型正则化 Regularization

目标：使 $\sum_{i=1}^{m}(y^{(i)} - \theta_0 - \theta_1 X_1^{(i)} - \theta_2 X_2^{(i)} - \ldots - \theta_n X_n^{(i)})^2$ 尽可能小

目标：使 $J(\theta) = MSE(y, \hat{y}; \theta)$ 尽可能小

加入模型正则化，目标：使 $J(\theta) = MSE(y, \hat{y}; \theta) + \alpha \frac{1}{2} \sum_{i=1}^{n} \theta_i^2$ 尽可能小

岭回归 Ridge Regression

目标：使 $J(\theta) = MSE(y, \hat{y}; \theta) + \alpha \frac{1}{2} \sum_{i=1}^{n} \theta_i^2$ 尽可能小

实践：岭回归

LASSO Regularization

岭回归 Ridge Regression

目标：使 $J(\theta)=MSE(y,\hat{y};\theta)+\alpha\frac{1}{2}\sum_{i=1}^{n}\theta_i^2$ 尽可能小

LASSO Regression

目标：使 $J(\theta)=MSE(y,\hat{y};\theta)+\alpha\sum_{i=1}^{n}|\theta_i|$ 尽可能小

LASSO Regression

目标：使 $J(\theta) = MSE(y, \hat{y}; \theta) + \alpha \sum_{i=1}^{n} |\theta_i|$ 尽可能小

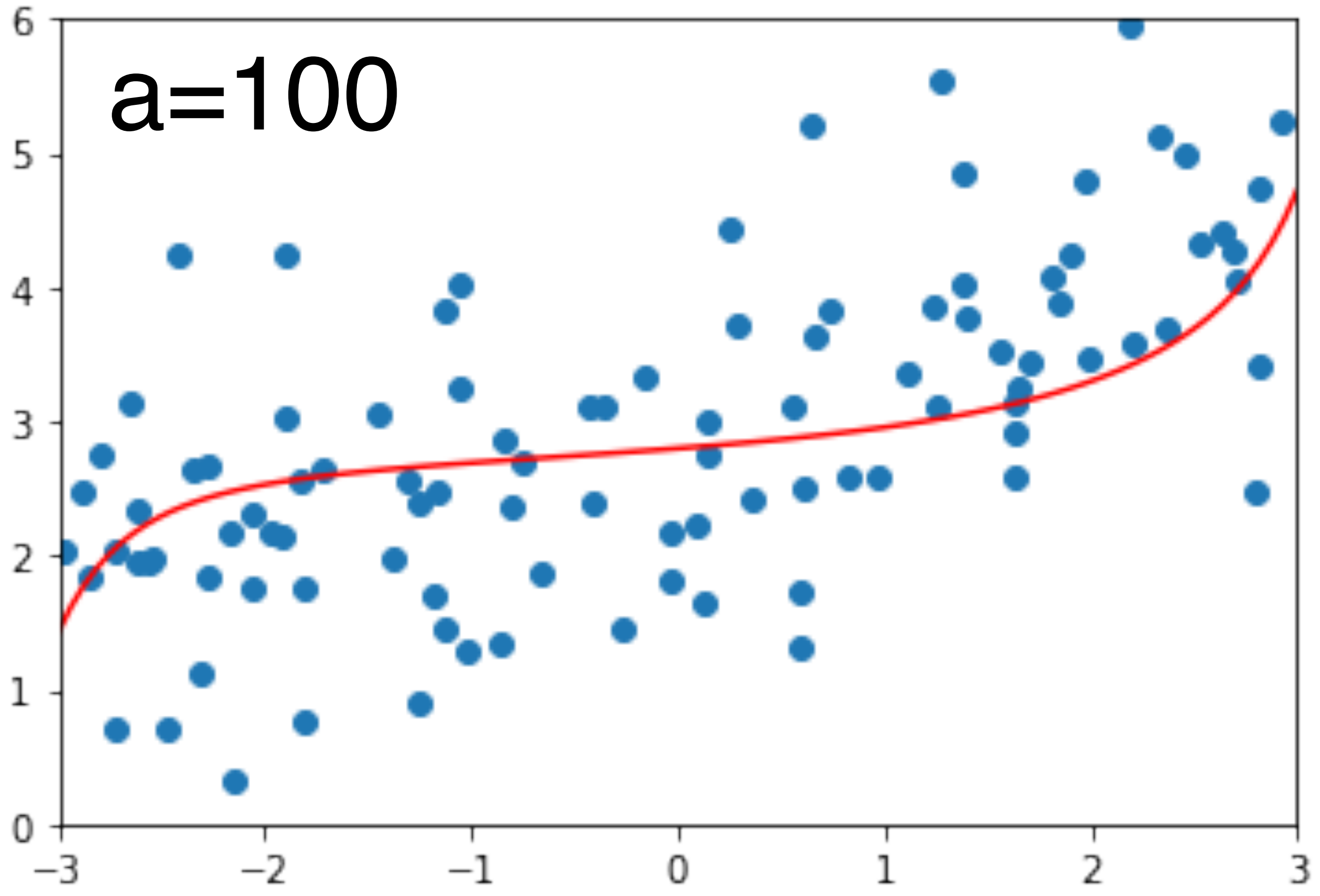
Least Absolute Shrinkage and Selection Operator Regression

实践：LASSO

比较 Ridge 和 LASSO

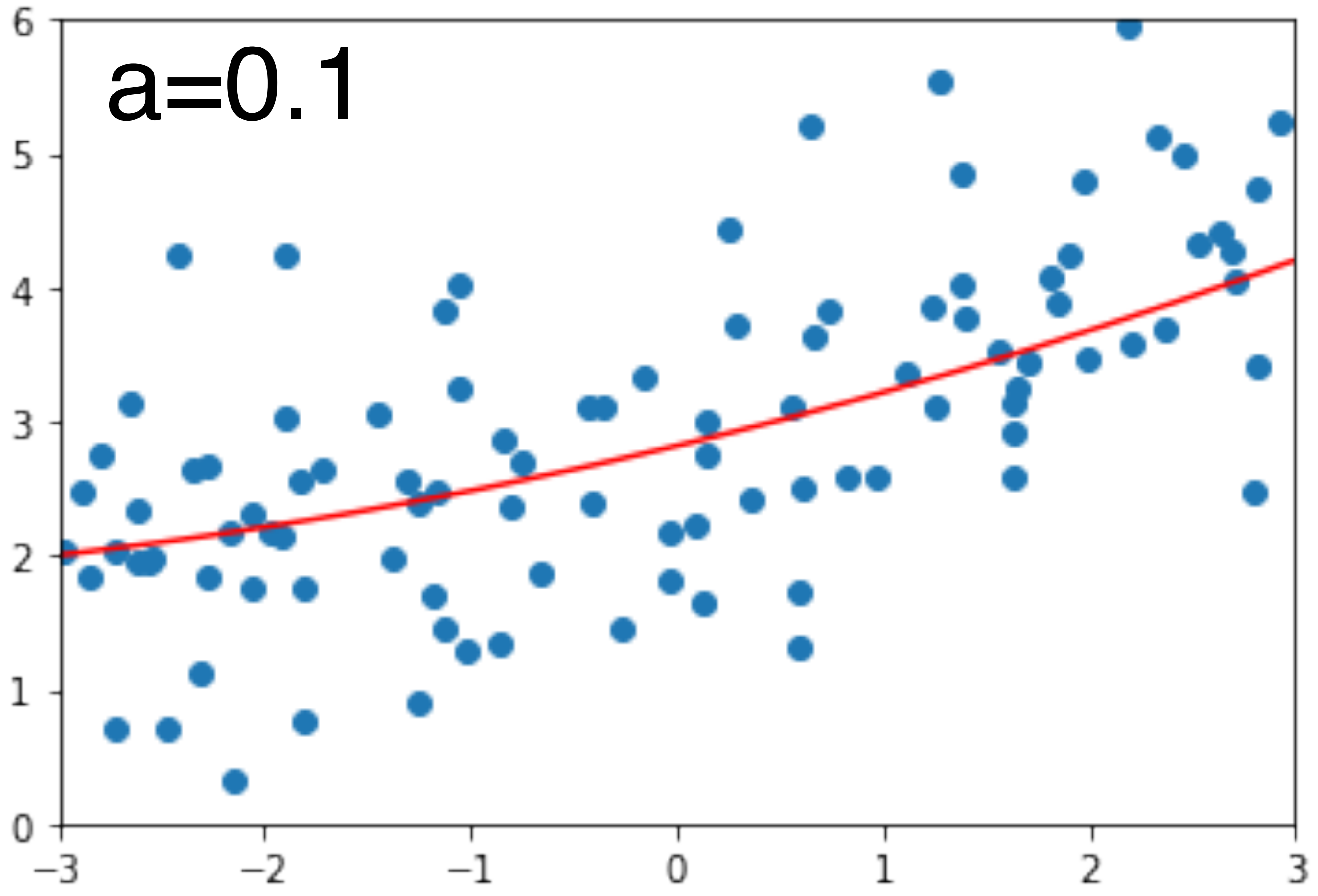
Ridge

目标：使 $J(\theta)=MSE(y,\hat{y};\theta)+\alpha\frac{1}{2}\sum_{i=1}^{n}\theta_i^2$



LASSO

目标：使 $J(\theta)=MSE(y,\hat{y};\theta)+\alpha\sum_{i=1}^{n}|\theta_i|$



LASSO Regression

目标：使 $J(\theta) = MSE(y, \hat{y}; \theta) + \alpha \sum_{i=1}^{n} |\theta_i|$ 尽可能小

Least Absolute Shrinkage and Selection Operator Regression

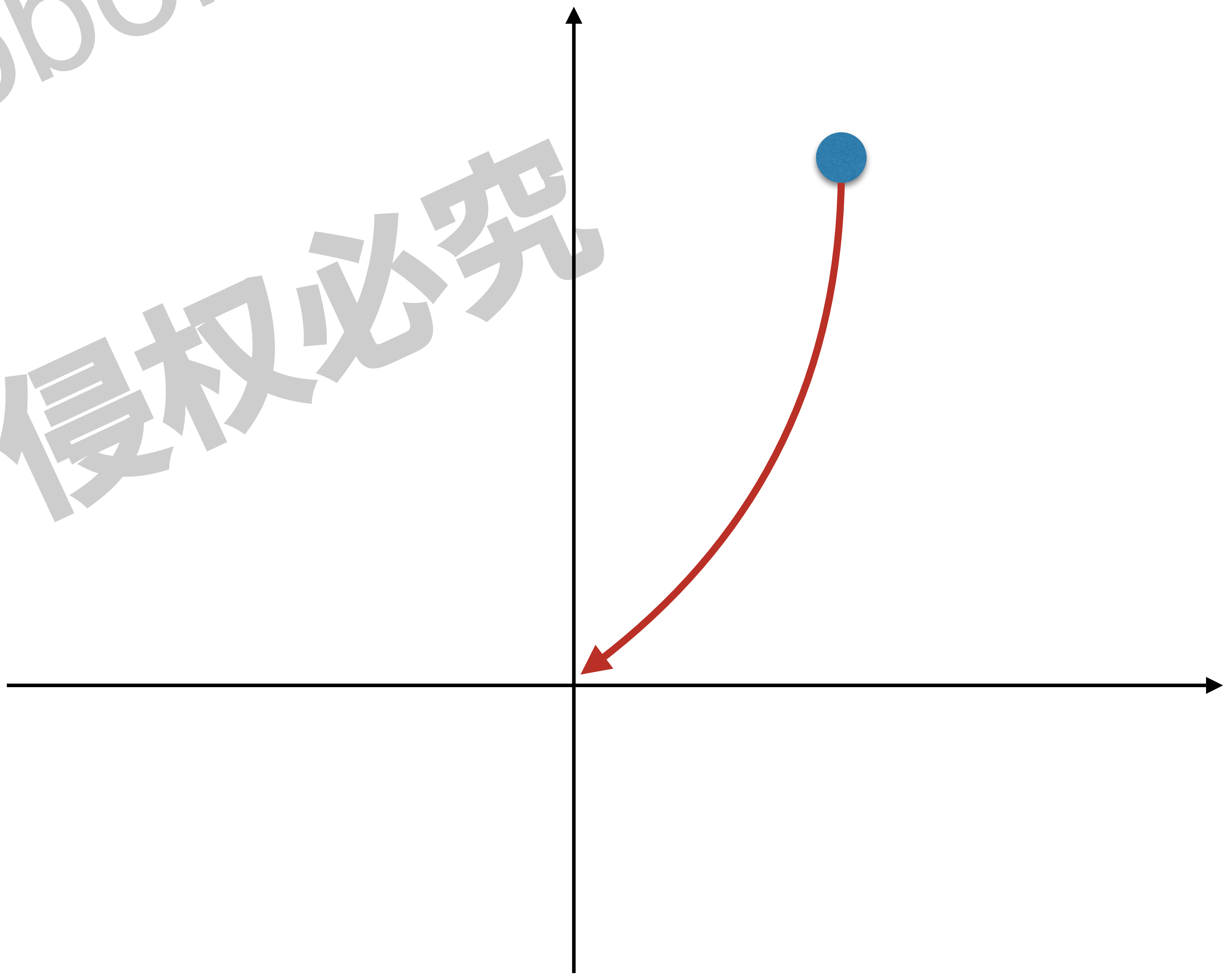
LASSO趋向于使得一部分theta值变为0。所以可作为特征选择用。

Ridge Regression

Ridge Regression: $J(\theta) = MSE(y, \hat{y}; \theta) + \alpha \frac{1}{2} \sum_{i=1}^{n} \theta_i^2$

alpha趋近于无穷时

$$\nabla = \alpha \begin{pmatrix} \theta_1 \\ \theta_2 \\ \dots \\ \theta_n \end{pmatrix}$$

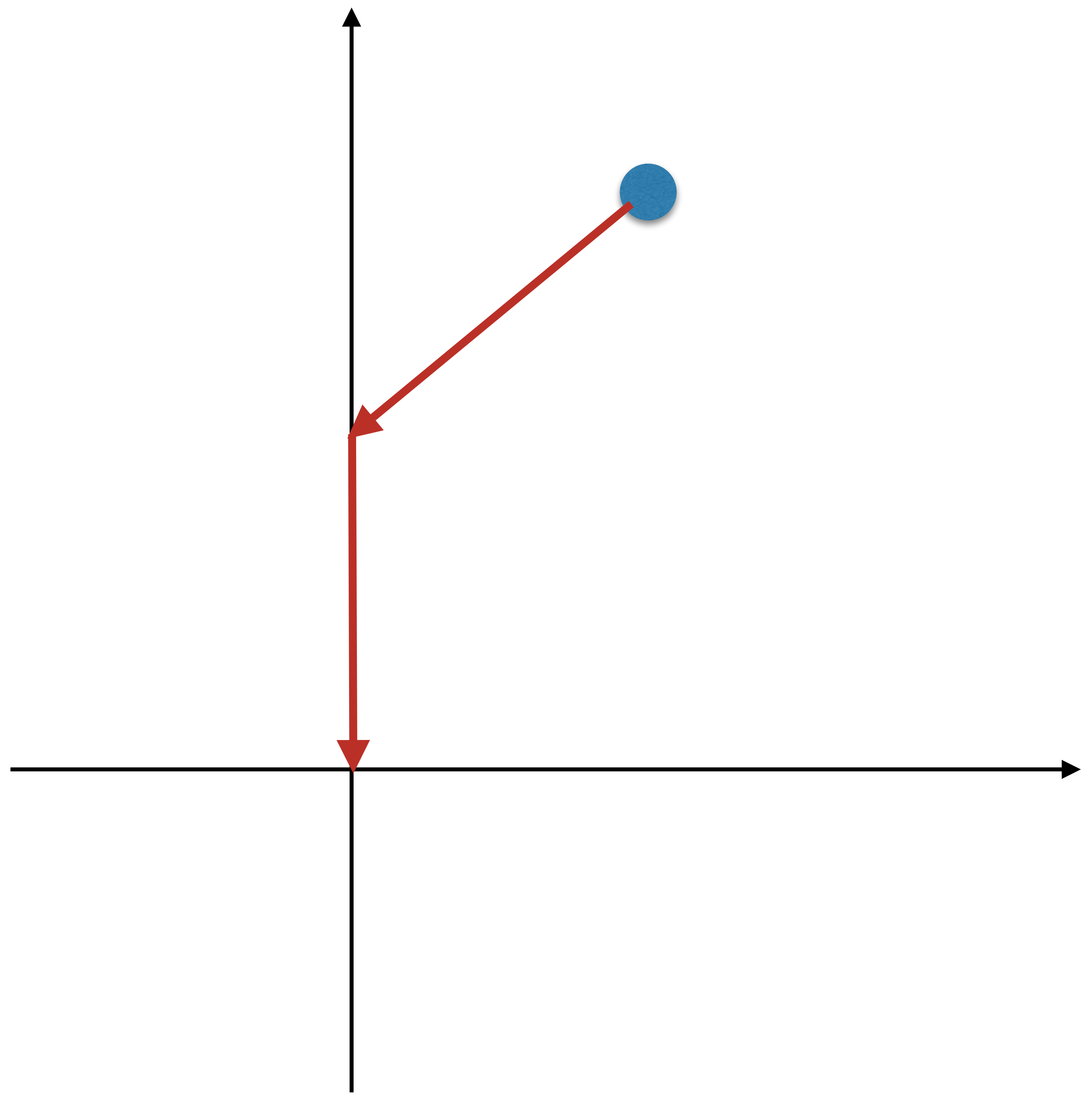


LASSO Regression

LASSO Regression: $J(\theta) = MSE(y, \hat{y}; \theta) + \alpha \sum_{i=1}^{n} |\theta_i|$

alpha趋近于无穷时

$$\nabla = \alpha \begin{pmatrix} sign(\theta_1) \\ sign(\theta_2) \\ \dots \\ sign(\theta_n) \end{pmatrix} \quad sign(x) = \begin{cases} 1, x > 0 \\ 0, x = 0 \\ -1, x < 0 \end{cases}$$



LASSO Regression

目标：使 $J(\theta) = MSE(y, \hat{y}; \theta) + \alpha \sum_{i=1}^{n} |\theta_i|$ 尽可能小

Least Absolute Shrinkage and Selection Operator Regression

LASSO趋向于使得一部分theta值变为0。所以可作为特征选择用。

L1 正则，L2正则 和 弹性网 Elastic Net

岭回归 Ridge Regression

目标：使 $J(\theta) = MSE(y, \hat{y}; \theta) + \alpha \frac{1}{2} \sum_{i=1}^{n} \theta_i^2$ 尽可能小

LASSO Regression

目标：使 $J(\theta) = MSE(y, \hat{y}; \theta) + \alpha \sum_{i=1}^{n} |\theta_i|$ 尽可能小

比较 Ridge 和 LASSO

Ridge

$$\frac{1}{2}\sum_{i=1}^{n}\theta_i^2$$

MSE

$$\frac{1}{n}\sum_{i=1}^{n}(y_i-\hat{y}_i)^2$$

欧拉距离

$$\sqrt{\sum_{i=1}^{n}(x_i^{(1)}-x_i^{(2)})^2}$$

LASSO

$$\sum_{i=1}^{n}|\theta_i|$$

MAE

$$\frac{1}{n}\sum_{i=1}^{n}|y_i-\hat{y}_i|$$

曼哈顿距离

$$\sum_{i=1}^{n}|x_i^{(1)}-x_i^{(2)}|$$

L1正则，L2正则

明可夫斯基距离

Minkowski Distance

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

L1正则，L2正则

$$\|x\|_p = (\sum_{i=1}^{n} |x_i|^p)^{\frac{1}{p}}$$

Lp范数

Ridge	$\sum_{i=1}^{n} \theta_i^2$	L2正则项
LASSO	$\sum_{i=1}^{n} \lvert\theta_i\rvert$	L1正则项
		Ln正则项

L1正则，L2正则

Ridge Regression: $J(\theta) = MSE(y, \hat{y}; \theta) + \alpha \frac{1}{2} \sum_{i=1}^{n} \theta_i^2$ L2正则

LASSO Regression: $J(\theta) = MSE(y, \hat{y}; \theta) + \alpha \sum_{i=1}^{n} |\theta_i|$ L1正则

L0正则

$$J(\theta) = MSE(y, \hat{y}; \theta) + \min\{number - of - non - zero - \theta\}$$

实际用L1取代，因为L0正则的优化是一个NP难的问题

弹性网 Elastic Net

$$J(\theta) = MSE(y, \hat{y}; \theta) + r\alpha \sum_{i=1}^{n} |\theta_i| + \frac{1-r}{2} \alpha \sum_{i=1}^{n} \theta_i^2$$

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

liuyubobobo