Python 3 玩儿转机器学习

liuyubobobo

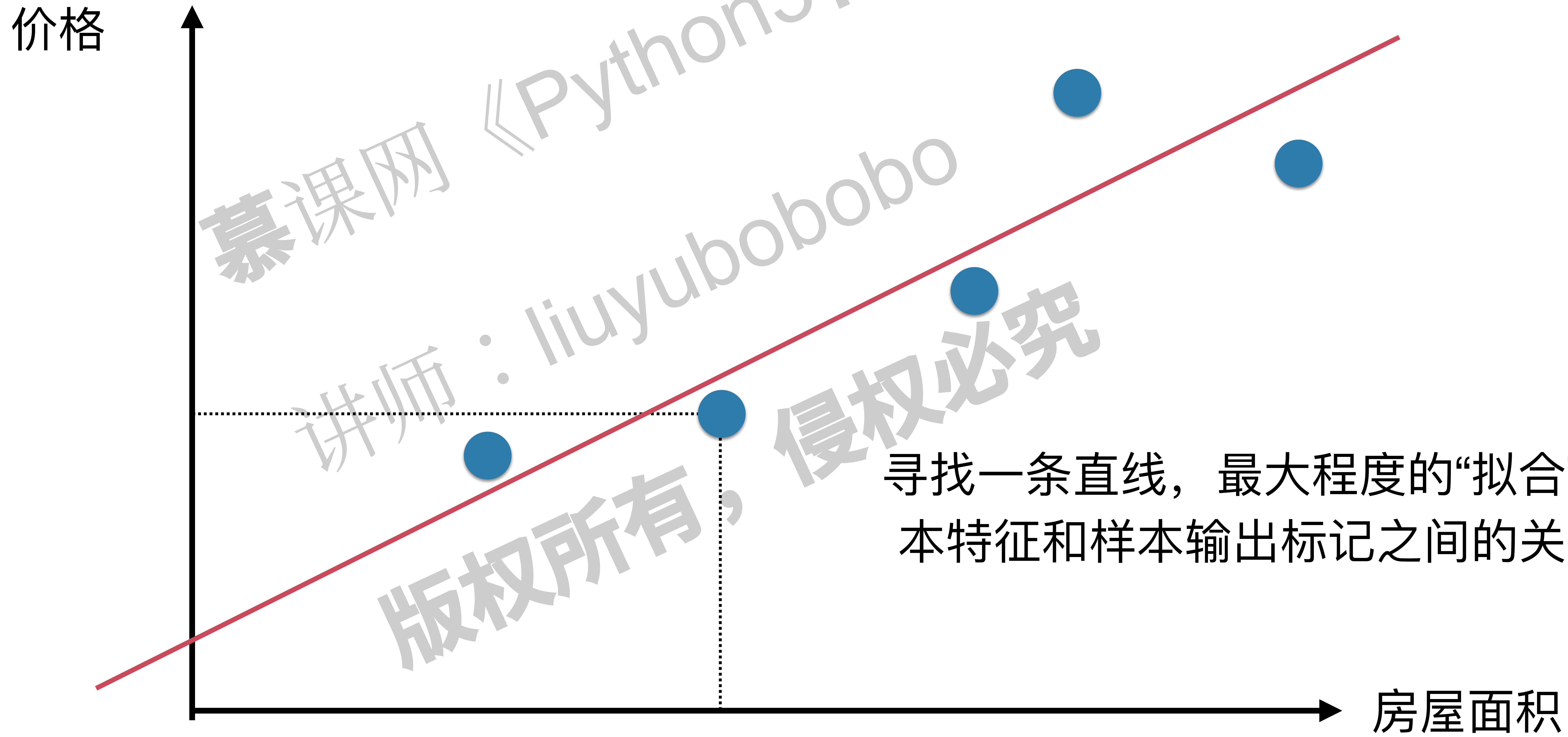
线性回归法

Linear Regression

线性回归算法

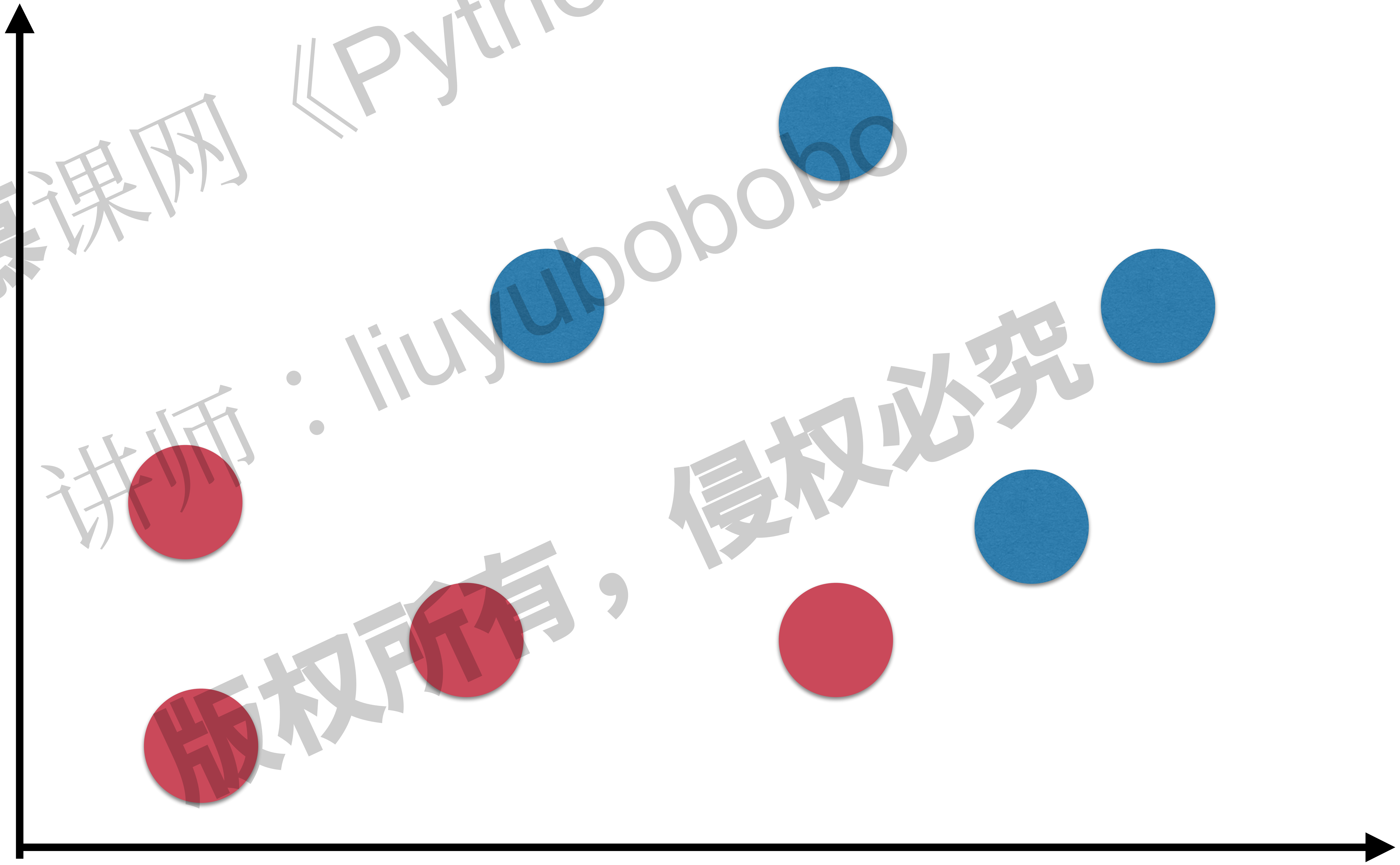
- 解决回归问题
- 思想简单，实现容易
- 许多强大的非线性模型的基础
- 结果具有很好的可解释性
- 蕴含机器学习中的很多重要思想

线性回归算法

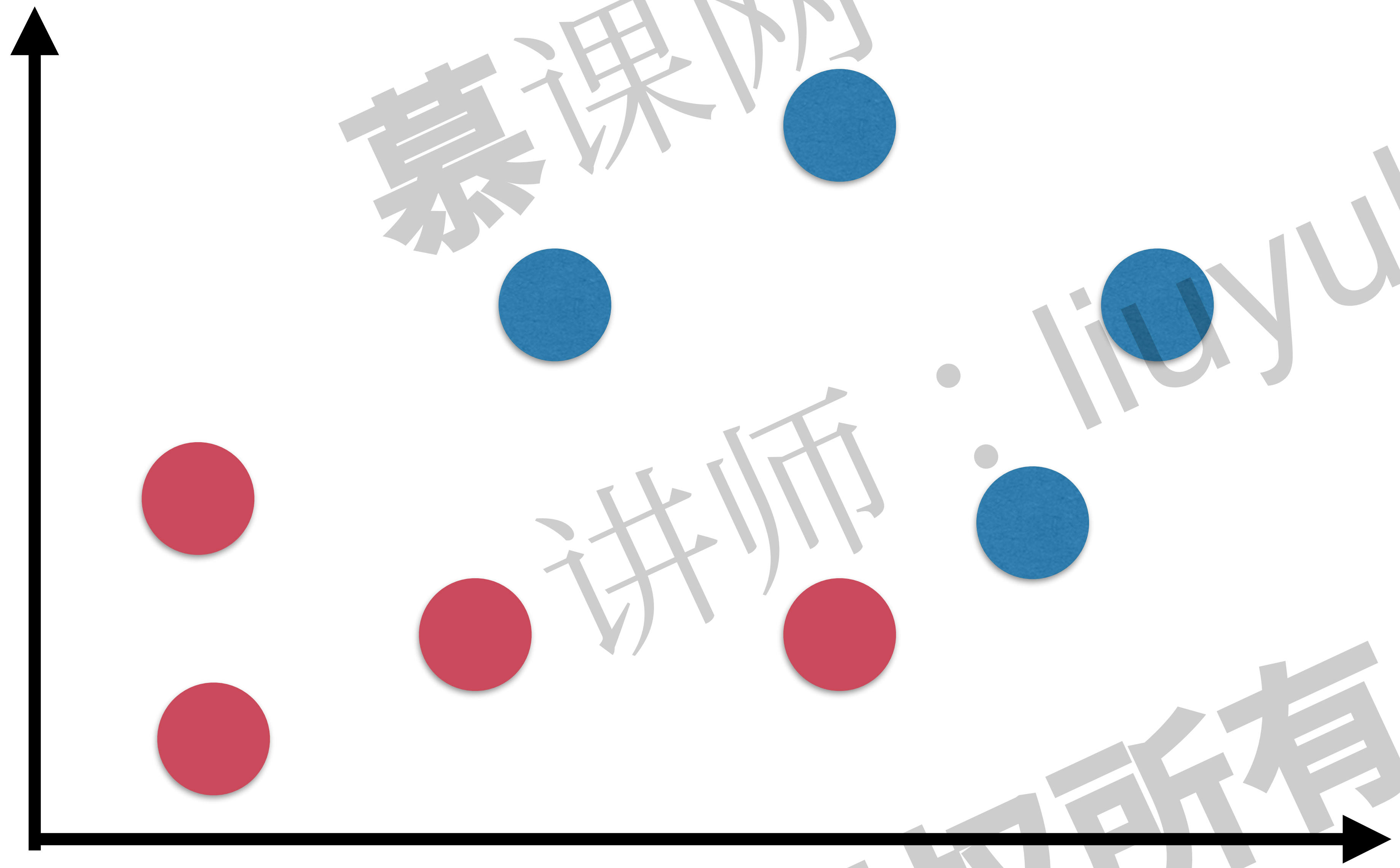


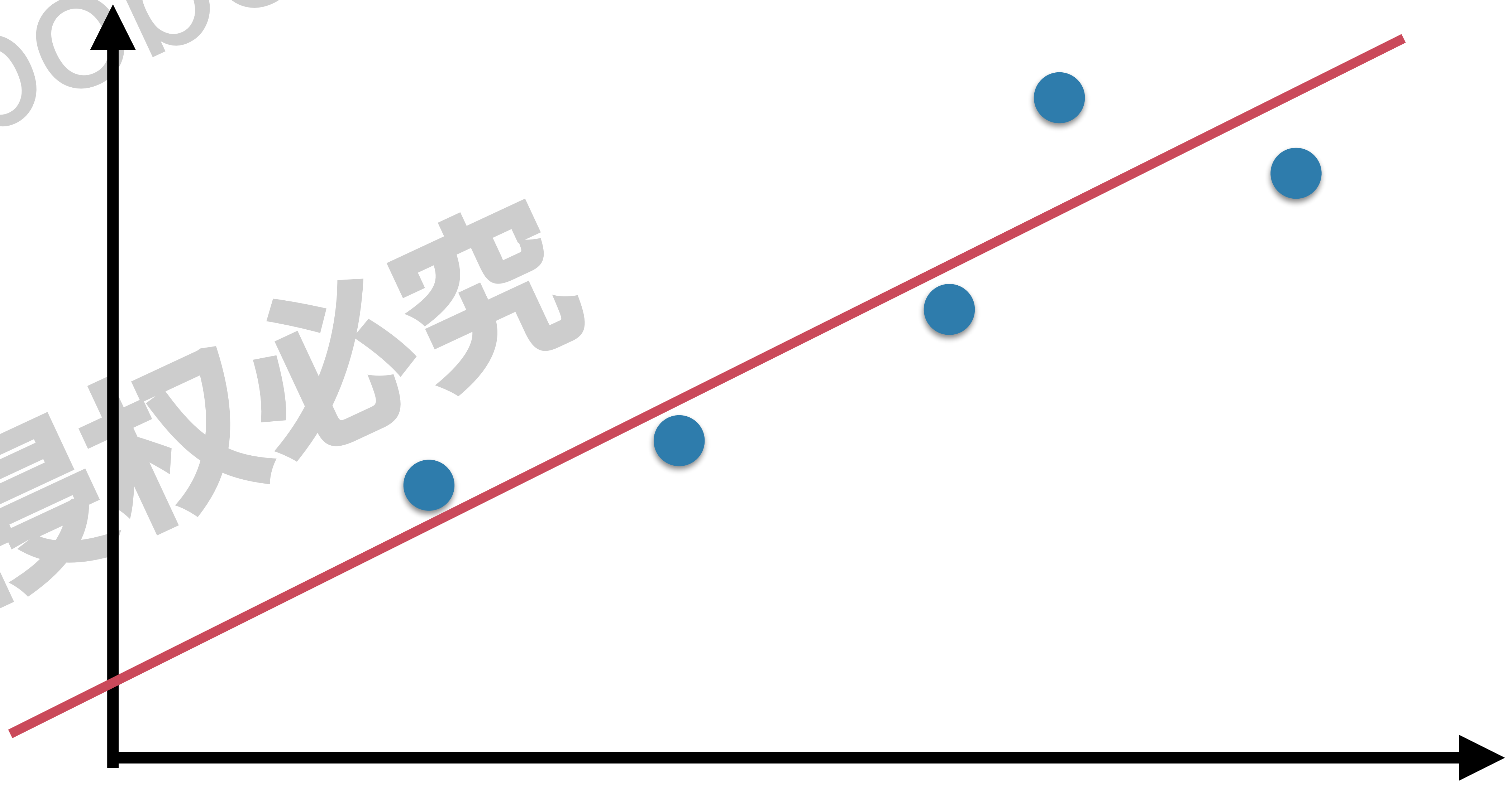
寻找一条直线，最大程度的“拟合”样本特征和样本输出标记之间的关系

分类问题

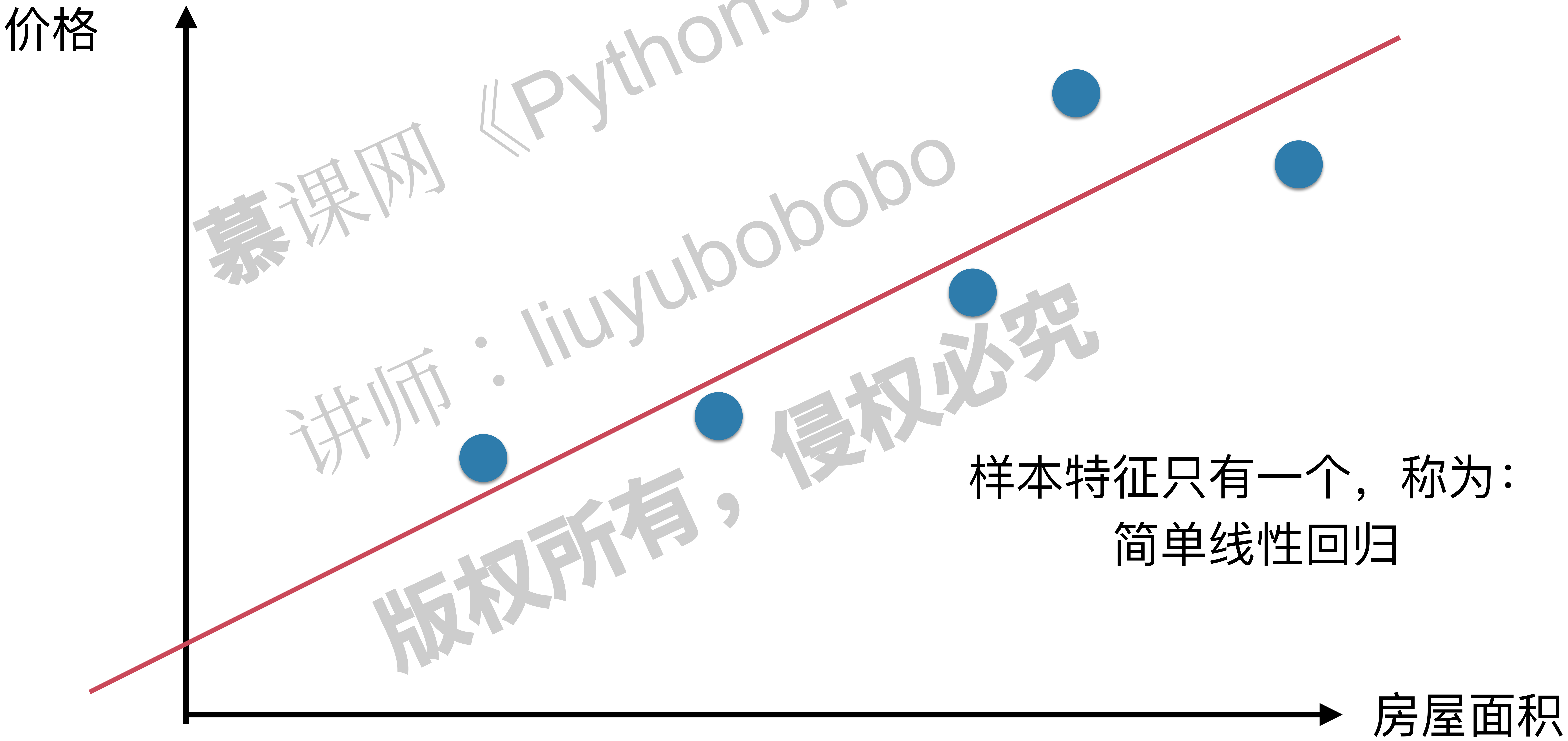


分类问题和回归问题



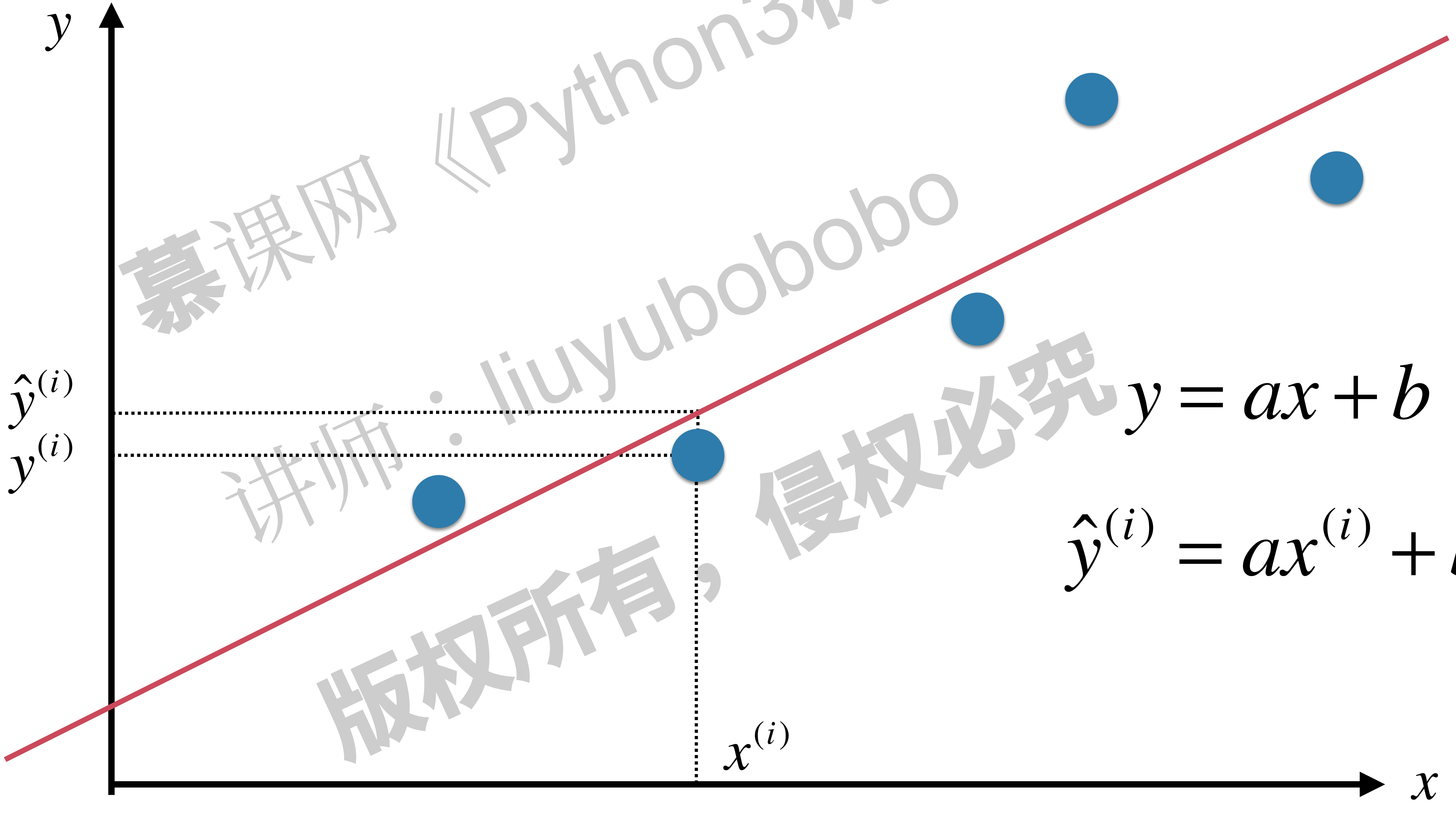


简单线性回归



样本特征只有一个，称为：
简单线性回归

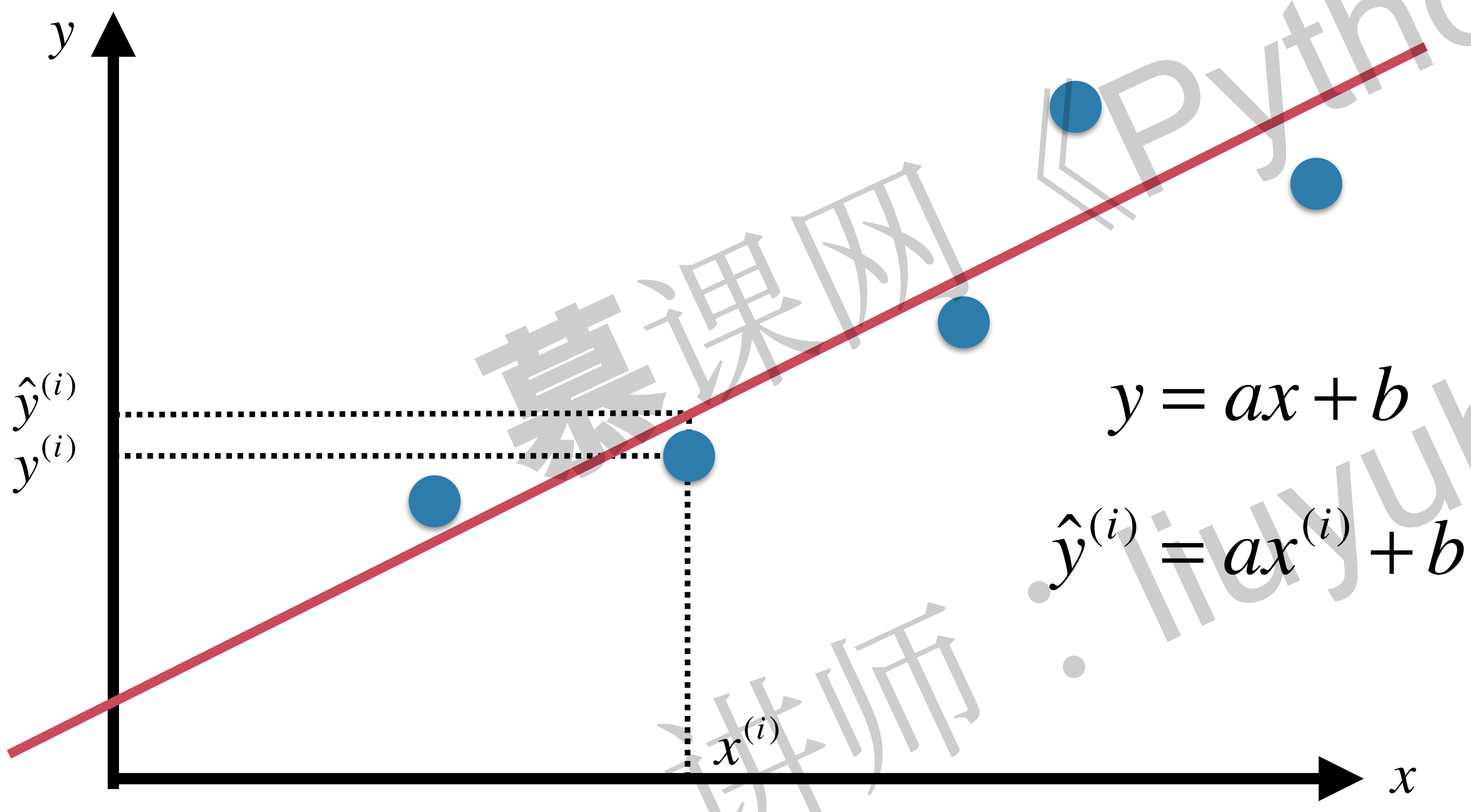
简单线性回归



$$y = ax + b$$

$$\hat{y}^{(i)} = ax^{(i)} + b$$

简单线性回归



假设我们找到了最佳拟合的直线方程：

$$y = ax + b$$

则对于每一个样本点 $x^{(i)}$

根据我们的直线方程，预测值为：

$$\hat{y}^{(i)} = ax^{(i)} + b$$

真值为： $y^{(i)}$

简单线性回归

假设我们找到了最佳拟合的直线方程：

$$y = ax + b$$

则对于每一个样本点 $x^{(i)}$

根据我们的直线方程，预测值为：

$$\hat{y}^{(i)} = ax^{(i)} + b$$

真值为： $y^{(i)}$

我们希望 $y^{(i)}$ 和 $\hat{y}^{(i)}$ 的差距尽量小

表达 $y^{(i)}$ 和 $\hat{y}^{(i)}$ 的差距：

$$\cancel{y^{(i)} - \hat{y}^{(i)}}$$

$$\cancel{|y^{(i)} - \hat{y}^{(i)}|}$$

$$(y^{(i)} - \hat{y}^{(i)})^2$$

考虑所有样本： $\sum_{i=1}^{m}(y^{(i)} - \hat{y}^{(i)})^2$

简单线性回归

目标：使 $\sum_{i=1}^{m}(y^{(i)}-\hat{y}^{(i)})^2$ 尽可能小

$$\hat{y}^{(i)}=ax^{(i)}+b$$

目标：找到a和b，使得 $\sum_{i=1}^{m}(y^{(i)}-ax^{(i)}-b)^2$ 尽可能小

一类机器学习算法的基本思路

目标：找到a和b，使得 $\sum_{i=1}^{m}(y^{(i)} - ax^{(i)} - b)^2$ 尽可能小

损失函数(loss function)　效用函数(utility function)

通过分析问题，确定问题的损失函数或者效用函数；
通过最优化损失函数或者效用函数，获得机器学习的模型。

一类机器学习算法的基本思路

通过分析问题，确定问题的损失函数或者效用函数；
通过最优化损失函数或者效用函数，获得机器学习的模型。

近乎所有参数学习算法都是这样的套路

线性回归

多项式回归

逻辑回归

SVM

神经网络

……

最优化原理

凸优化

简单线性回归

目标：找到a和b，使得 $\sum_{i=1}^{m}(y^{(i)} - ax^{(i)} - b)^2$ 尽可能小

典型的最小二乘法问题：最小化误差的平方

$$a = \frac{\sum_{i=1}^{m}(x^{(i)} - \overline{x})(y^{(i)} - \overline{y})}{\sum_{i=1}^{m}(x^{(i)} - \overline{x})^2} \qquad b = \overline{y} - a\overline{x}$$

最小二乘法

最小二乘法

目标：找到a和b，使得 $\sum_{i=1}^{m}(y^{(i)} - ax^{(i)} - b)^2$ 尽可能小

$$\downarrow$$

$$J(a,b)$$

$$\frac{\partial J(a,b)}{\partial a} = 0 \qquad \frac{\partial J(a,b)}{\partial b} = 0$$

最小二乘法

$$J(a,b)=\sum_{i=1}^{m}(y^{(i)}-ax^{(i)}-b)^2$$

$$\frac{\partial J(a,b)}{\partial a}=0 \qquad \frac{\partial J(a,b)}{\partial b}=0$$

$$\frac{\partial J(a,b)}{\partial b}=\sum_{i=1}^{m}2(y^{(i)}-ax^{(i)}-b)(-1)=0$$

$$\sum_{i=1}^{m}(y^{(i)}-ax^{(i)}-b)=0$$

最小二乘法

$$\sum_{i=1}^{m}(y^{(i)} - ax^{(i)} - b) = 0$$

$$\sum_{i=1}^{m} y^{(i)} - a\sum_{i=1}^{m} x^{(i)} - \sum_{i=1}^{m} b = 0 \longrightarrow \sum_{i=1}^{m} y^{(i)} - a\sum_{i=1}^{m} x^{(i)} - mb = 0$$

$$mb = \sum_{i=1}^{m} y^{(i)} - a\sum_{i=1}^{m} x^{(i)}$$

$$b = \overline{y} - a\overline{x}$$

最小二乘法

$$J(a,b)=\sum_{i=1}^{m}(y^{(i)}-ax^{(i)}-b)^2$$

$$\frac{\partial J(a,b)}{\partial a}=0 \qquad \frac{\partial J(a,b)}{\partial b}=0$$

最小二乘法

$$J(a,b)=\sum_{i=1}^{m}(y^{(i)}-ax^{(i)}-b)^2 \qquad \frac{\partial J(a,b)}{\partial a}=0 \qquad b=\bar{y}-a\bar{x}$$

$$\frac{\partial J(a,b)}{\partial a}=\sum_{i=1}^{m}2(y^{(i)}-ax^{(i)}-b)(-x^{(i)})=0$$

$$\sum_{i=1}^{m}(y^{(i)}-ax^{(i)}-b)x^{(i)}=0$$

$$\sum_{i=1}^{m}(y^{(i)}-ax^{(i)}-\bar{y}+a\bar{x})x^{(i)}=0$$

最小二乘法

$$\sum_{i=1}^{m}(y^{(i)} - ax^{(i)} - \bar{y} + a\bar{x})x^{(i)} = 0$$

$$\sum_{i=1}^{m}(x^{(i)}y^{(i)} - a(x^{(i)})^2 - x^{(i)}\bar{y} + a\bar{x}x^{(i)})$$

$$\sum_{i=1}^{m}(x^{(i)}y^{(i)} - x^{(i)}\bar{y} - a(x^{(i)})^2 + a\bar{x}x^{(i)})$$

$$\sum_{i=1}^{m}(x^{(i)}y^{(i)} - x^{(i)}\bar{y}) - \sum_{i=1}^{m}(a(x^{(i)})^2 - a\bar{x}x^{(i)})$$

$$\sum_{i=1}^{m}(x^{(i)}y^{(i)} - x^{(i)}\bar{y}) - a\sum_{i=1}^{m}((x^{(i)})^2 - \bar{x}x^{(i)}) = 0$$

$$a = \frac{\sum_{i=1}^{m}(x^{(i)}y^{(i)} - x^{(i)}\bar{y})}{\sum_{i=1}^{m}((x^{(i)})^2 - \bar{x}x^{(i)})}$$

最小二乘法

$$a=\frac{\sum_{i=1}^{m}(x^{(i)}y^{(i)}-x^{(i)}\bar{y})}{\sum_{i=1}^{m}((x^{(i)})^2-\bar{x}x^{(i)})}$$

$$\sum_{i=1}^{m}x^{(i)}\bar{y} = \bar{y}\sum_{i=1}^{m}x^{(i)} = m\bar{y}\cdot\bar{x} = \bar{x}\sum_{i=1}^{m}y^{(i)} = \sum_{i=1}^{m}\bar{x}y^{(i)}$$

$$=\sum_{i=1}^{m}\bar{x}\cdot\bar{y}$$

$$=\frac{\sum_{i=1}^{m}(x^{(i)}y^{(i)}-x^{(i)}\bar{y}-\bar{x}y^{(i)}+\bar{x}\cdot\bar{y})}{\sum_{i=1}^{m}((x^{(i)})^2-\bar{x}x^{(i)}-\bar{x}x^{(i)}+\bar{x}^2)} = \frac{\sum_{i=1}^{m}(x^{(i)}-\bar{x})(y^{(i)}-\bar{y})}{\sum_{i=1}^{m}(x^{(i)}-\bar{x})^2}$$

简单线性回归

目标：找到a和b，使得 $\sum_{i=1}^{m}(y^{(i)} - ax^{(i)} - b)^2$ 尽可能小

$$a = \frac{\sum_{i=1}^{m}(x^{(i)} - \bar{x})(y^{(i)} - \bar{y})}{\sum_{i=1}^{m}(x^{(i)} - \bar{x})^2} \qquad b = \bar{y} - a\bar{x}$$

实现简单线性回归法

实践：实现简单线性回归法

向量化运算

简单线性回归

目标：找到a和b，使得 $\sum_{i=1}^{m}(y^{(i)} - ax^{(i)} - b)^2$ 尽可能小

$$a = \frac{\sum_{i=1}^{m}(x^{(i)} - \bar{x})(y^{(i)} - \bar{y})}{\sum_{i=1}^{m}(x^{(i)} - \bar{x})^2} \qquad b = \bar{y} - a\bar{x}$$

简单线性回归

$$a = \frac{\sum_{i=1}^{m}(x^{(i)} - \bar{x})(y^{(i)} - \bar{y})}{\sum_{i=1}^{m}(x^{(i)} - \bar{x})^2}$$

$$\sum_{i=1}^{m} w^{(i)} \cdot v^{(i)}$$

向量化运算

$$\sum_{i=1}^{m} w^{(i)} \cdot v^{(i)}$$

$$w = (w^{(1)}, w^{(2)}, \ldots, w^{m})$$

$$v = (v^{(1)}, v^{(2)}, \ldots, v^{m})$$

$$w \cdot v$$

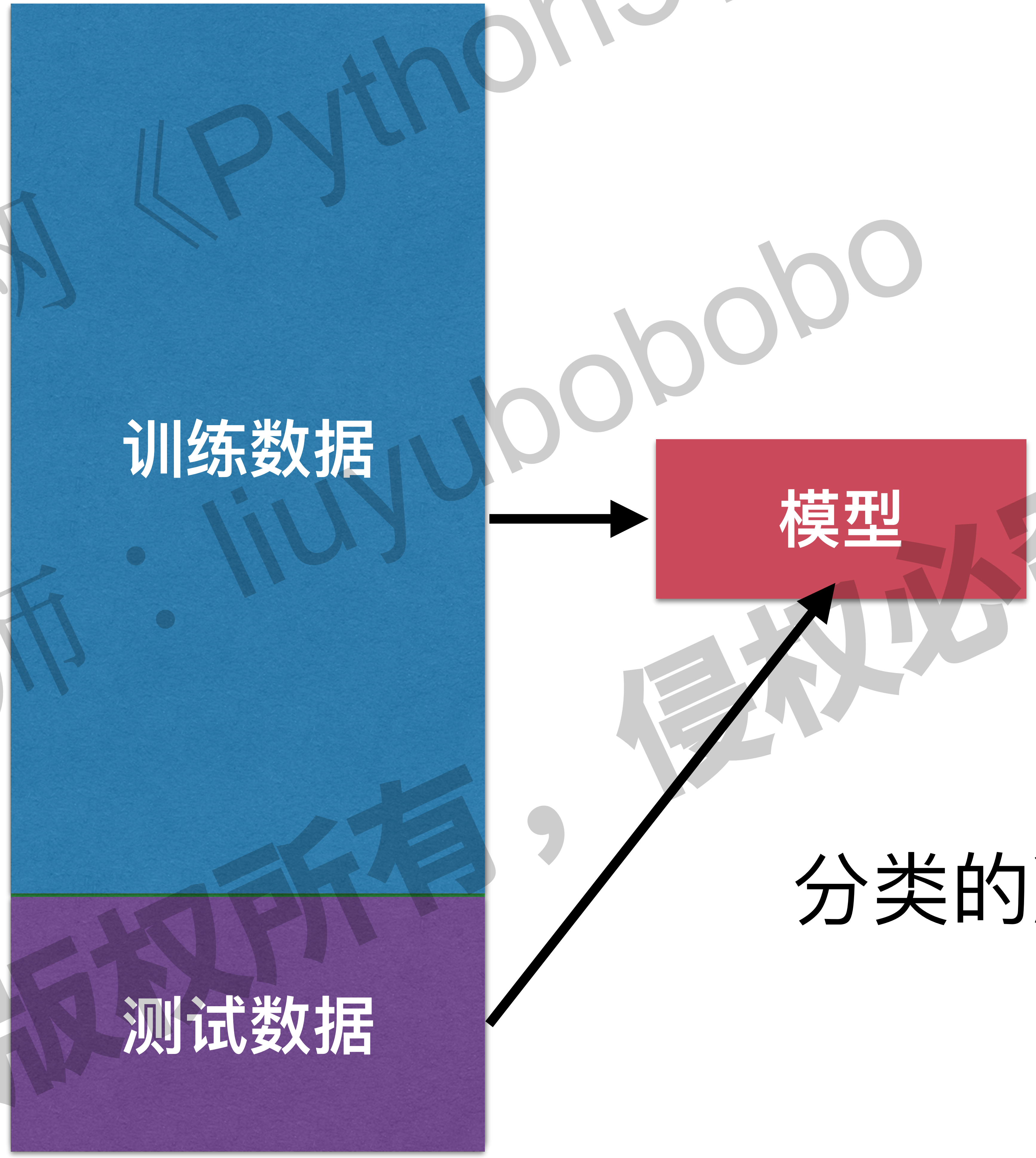
向量化运算

$$a = \frac{\sum_{i=1}^{m}(x^{(i)} - \overline{x})(y^{(i)} - \overline{y})}{\sum_{i=1}^{m}(x^{(i)} - \overline{x})^2}$$

实践：向量化实现简单线性回归法

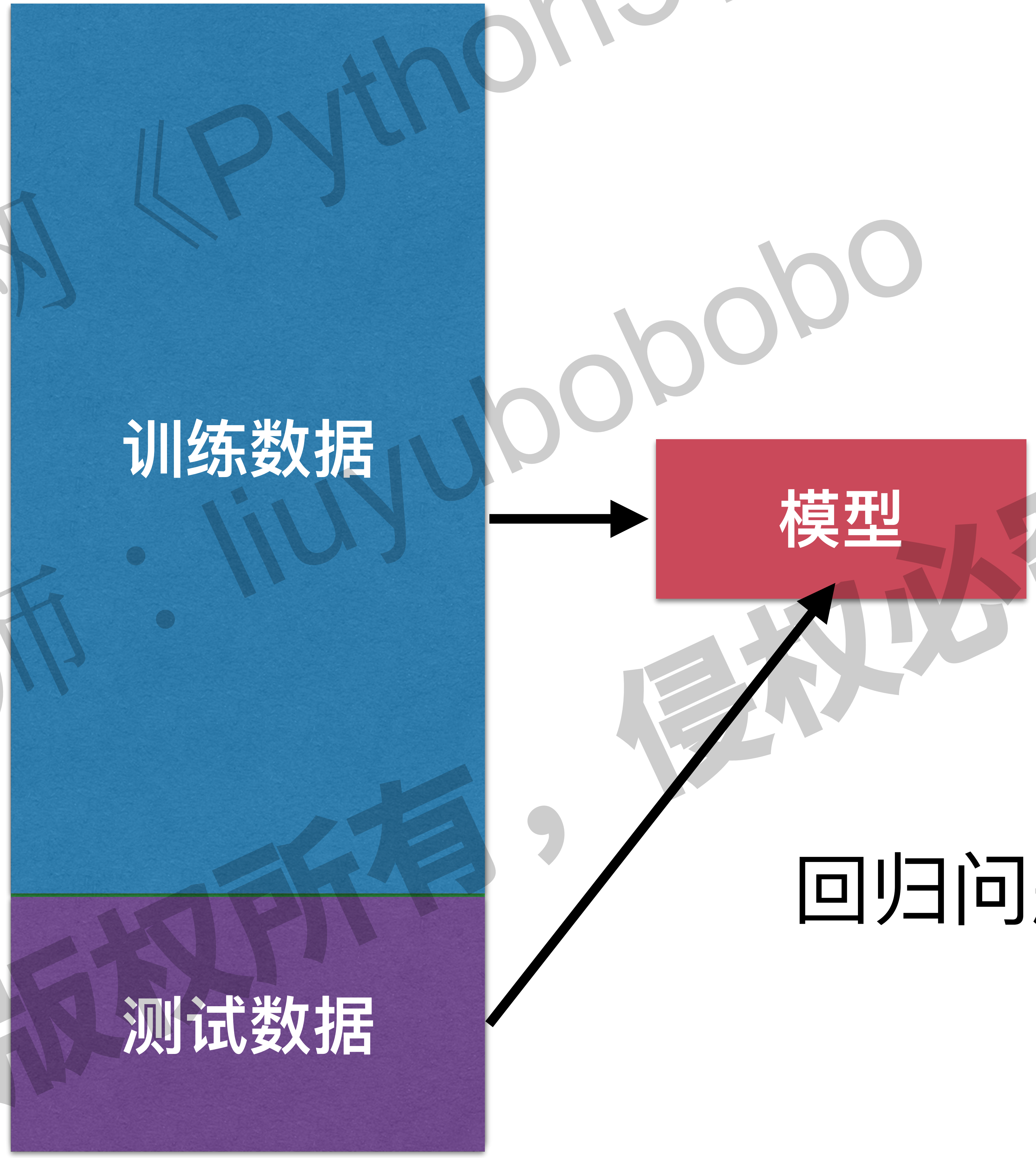
回归算法的衡量 MSE vs. MAE

回归算法的评价



分类的准确度： accuracy

回归算法的评价



回归问题如何评价?

简单线性回归

目标：找到a和b，使得 $\sum_{i=1}^{m}(y_{train}^{(i)} - ax_{train}^{(i)} - b)^2$ 尽可能小

$$\hat{y}_{test}^{(i)} = ax_{test}^{(i)} + b$$

$$\sum_{i=1}^{m}(y_{train}^{(i)} - \hat{y}_{train}^{(i)})^2$$

衡量标准：$\sum_{i=1}^{m}(y_{test}^{(i)} - \hat{y}_{test}^{(i)})^2$

线性回归算法的评测

衡量标准：$\sum_{i=1}^{m}(y_{test}^{(i)} - \hat{y}_{test}^{(i)})^2$

问题：和m相关？

线性回归算法的评测

$$\frac{1}{m}\sum_{i=1}^{m}(y_{test}^{(i)} - \hat{y}_{test}^{(i)})^2$$

均方误差 MSE
（Mean Squared Error）

问题：量纲?

线性回归算法的评测

$$\sqrt{\frac{1}{m}\sum_{i=1}^{m}(y_{test}^{(i)} - \hat{y}_{test}^{(i)})^2} = \sqrt{MSE_{test}}$$

均方根误差 RMSE

（Root Mean Squared Error）

线性回归算法的评测

$$\frac{1}{m}\sum_{i=1}^{m}|y_{test}^{(i)}-\hat{y}_{test}^{(i)}|$$

平均绝对误差 MAE
(Mean Absolute Error)

RMSE vs MAE

$$RMSE = \sqrt{\frac{1}{m}\sum_{i=1}^{m}(y_{test}^{(i)} - \hat{y}_{test}^{(i)})^2}$$

$$MAE = \frac{1}{m}\sum_{i=1}^{m}|\, y_{test}^{(i)} - \hat{y}_{test}^{(i)}\, |$$

实践：实现MSE,RMSE和MAE

评价回归算法 R Square

RMSE vs MAE

$$RMSE = \sqrt{\frac{1}{m}\sum_{i=1}^{m}(y_{test}^{(i)} - \hat{y}_{test}^{(i)})^2}$$

$$MAE = \frac{1}{m}\sum_{i=1}^{m}| y_{test}^{(i)} - \hat{y}_{test}^{(i)} |$$

问题：分类的准确度：1最好，0最差

RMSE? MAE?

R Squared

$$R^2 = 1 - \frac{SS_{residual}}{SS_{total}}$$

(Residual Sum of Squares)

(Total Sum of Squares)

$$R^2 = 1 - \frac{\sum_i (\hat{y}^{(i)} - y^{(i)})^2}{\sum_i (\overline{y} - y^{(i)})^2}$$

R Squared

$$R^2 = 1 - \frac{\sum_i (\hat{y}^{(i)} - y^{(i)})^2}{\sum_i (\overline{y} - y^{(i)})^2}$$

使用我们的模型预测产生的错误

使用 $y = \overline{y}$ 预测产生的错误

Baseline Model

R Squared

$$R^2 = 1 - \frac{\sum_i (\hat{y}^{(i)} - y^{(i)})^2}{\sum_i (\bar{y} - y^{(i)})^2}$$

- $R^2 <= 1$
- R^2 越大越好。当我们的预测模型不犯任何错误是，R^2得到最大值1
- 当我们的模型等于基准模型时，R^2为0
- 如果$R^2 < 0$，说明我们学习到的模型还不如基准模型。此时，很有可能我们的数据不存在任何线性关系。

R Squared

$$R^2 = 1 - \frac{\sum_i (\hat{y}^{(i)} - y^{(i)})^2}{\sum_i (\overline{y} - y^{(i)})^2} = 1 - \frac{(\sum_{i=1}^{m} (\hat{y}^{(i)} - y^{(i)})^2)/m}{(\sum_{i=1}^{m} (y^{(i)} - \overline{y})^2)/m}$$

$$= 1 - \frac{MSE(\hat{y}, y)}{Var(y)}$$

实践：实现 R Square

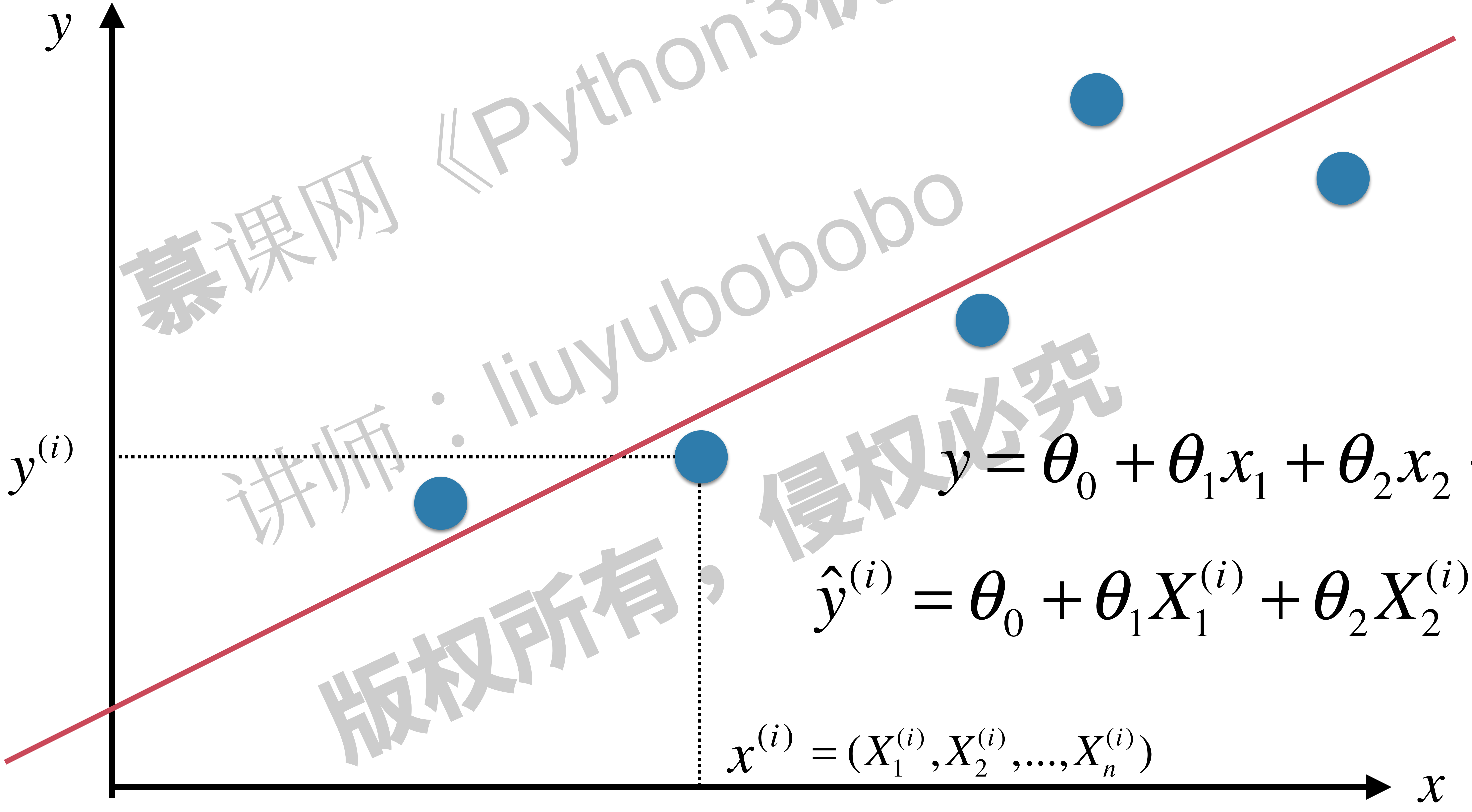
R Squared

Scikit-Learn中的线性回归法，
score默认为R Squared

http://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LinearRegression.html#sklearn.linear_model.LinearRegression.score

多元线性回归

多元线性回归



$$y = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \dots + \theta_n x_n$$

$$\hat{y}^{(i)} = \theta_0 + \theta_1 X_1^{(i)} + \theta_2 X_2^{(i)} + \dots + \theta_n X_n^{(i)}$$

多元线性回归

目标：使 $\sum_{i=1}^{m}(y^{(i)} - \hat{y}^{(i)})^2$ 尽可能小

$$\hat{y}^{(i)} = \theta_0 + \theta_1 X_1^{(i)} + \theta_2 X_2^{(i)} + \ldots + \theta_n X_n^{(i)}$$

目标：找到 $\theta_0, \theta_1, \theta_2, \ldots, \theta_n$，使得 $\sum_{i=1}^{m}(y^{(i)} - \hat{y}^{(i)})^2$ 尽可能小

多元线性回归

$$\hat{y}^{(i)} = \theta_0 + \theta_1 X_1^{(i)} + \theta_2 X_2^{(i)} + \ldots + \theta_n X_n^{(i)}$$

$$\theta = (\theta_0, \theta_1, \theta_2, \ldots, \theta_n)^T$$

$$\hat{y}^{(i)} = \theta_0 X_0^{(i)} + \theta_1 X_1^{(i)} + \theta_2 X_2^{(i)} + \ldots + \theta_n X_n^{(i)} \quad , X_0^{(i)} \equiv 1$$

$$X^{(i)} = (X_0^{(i)}, X_1^{(i)}, X_2^{(i)}, \ldots, X_n^{(i)})$$

$$\hat{y}^{(i)} = X^{(i)} \cdot \theta$$

多元线性回归

$$X_b = \begin{pmatrix} 1 & X_1^{(1)} & X_2^{(1)} & \dots & X_n^{(1)} \\ 1 & X_1^{(2)} & X_2^{(2)} & \dots & X_n^{(2)} \\ \dots & & & & \dots \\ 1 & X_1^{(m)} & X_2^{(m)} & \dots & X_n^{(m)} \end{pmatrix} \qquad \theta = \begin{pmatrix} \theta_0 \\ \theta_1 \\ \theta_2 \\ \dots \\ \theta_n \end{pmatrix}$$

$$\hat{y} = X_b \cdot \theta$$

多元线性回归

$$\hat{y} = X_b \cdot \theta$$

目标：使 $\sum_{i=1}^{m}(y^{(i)} - \hat{y}^{(i)})^2$ 尽可能小

$$\downarrow$$

目标：使 $(y - X_b \cdot \theta)^T (y - X_b \cdot \theta)$ 尽可能小

多元线性回归

目标：使 $(y - X_b \cdot \theta)^T (y - X_b \cdot \theta)$ 尽可能小

$$\theta = (X_b^T X_b)^{-1} X_b^T y$$

多元线性回归

多元线性回归的正规方程解 (Normal Equation)

$$\theta = (X_b^T X_b)^{-1} X_b^T y$$

问题：时间复杂度高：O(n^3) (优化O(n^2.4))

优点：不需要对数据做归一化处理

多元线性回归正规解实现

多元线性回归

多元线性回归的正规方程解 (Normal Equation)

$$\theta = (X_b^T X_b)^{-1} X_b^T y$$

多元线性回归

$$\theta = (X_b^T X_b)^{-1} X_b^T y$$

$$\theta = \begin{pmatrix} \theta_0 \\ \theta_1 \\ \theta_2 \\ \dots \\ \theta_n \end{pmatrix}$$

截距 intercept

系数 coefficients

实践：多元线性回归的正规解实现

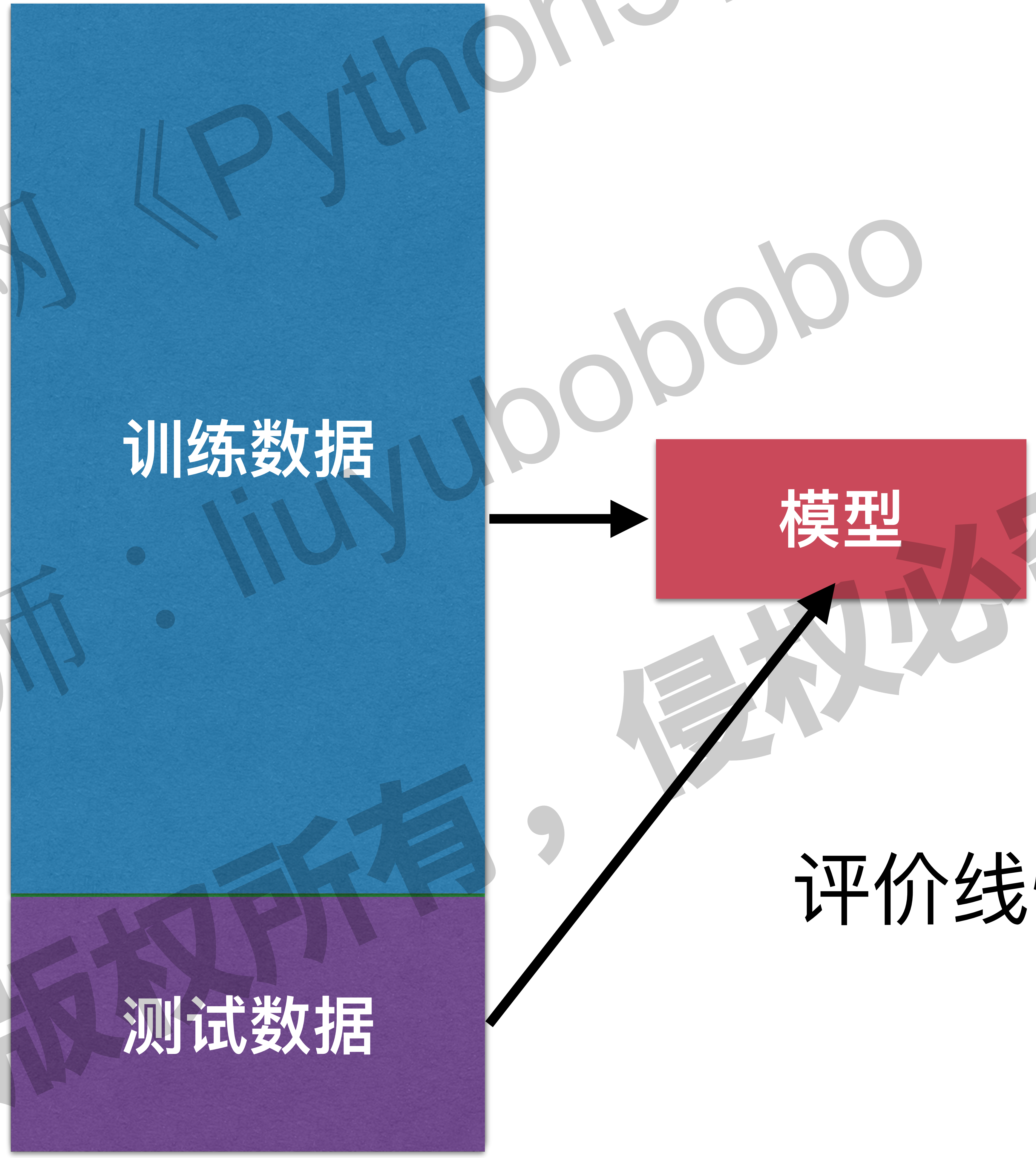
实践：scikit-learn中的多元线性回归

实践：使用kNN Regressor解决回归问题

线性回归算法总结

实践：使用Linear Regressor对数据解释

线性回归算法总结



$$y = \theta^T x$$

评价线性回归算法：R Squared

线性回归算法总结

- 典型的参数学习

  对比kNN：非参数学习

- 只能解决回归问题

  虽然很多分类方法中，线性回归是基础（如逻辑回归）

  对比kNN：既可以解决分类问题，又可以解决回归问题

线性回归算法总结

- 对数据有假设：线性

  对比kNN 对数据没有假设

- 优点：对数据具有强解释性

多元线性回归

多元线性回归的正规方程解 (Normal Equation)

$$\theta = (X_b^T X_b)^{-1} X_b^T y$$

问题：时间复杂度高：O(n^3) (优化O(n^2.4))

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