



第6章 线性模型

线性模型： 感知机： $f(x)=w^T x+b$

线性回归： $f(x_i)=w^T x_i+b$

对数线性回归： $\ln y=w^T x+b$

对数几率回归

$$\frac{p}{1-p} \quad p(y=1 | \mathbf{x}) = \frac{e^{w^T \mathbf{x}+b}}{1+e^{w^T \mathbf{x}+b}}$$
$$p(y=0 | \mathbf{x}) = \frac{1}{1+e^{w^T \mathbf{x}+b}}$$




6.1 线性回归

什么是回归？

回归是监督学习的一个重要问题，回归用于预测输入变量和输出变量之间的关系。

回归模型是表示输入变量到输出变量之间映射的函数。

回归问题的学习等价于函数拟合：使用一条函数曲线使其很好的拟合已知函数且很好的预测未知数据。



回归问题分为模型的学习和预测两个过程。基于给定的训练数据集构建一个模型，根据新的输入数据预测相应的输出。

回归问题按照输入变量的个数可以分为一元回归和多元回归；按照输入变量和输出变量之间关系的类型，可以分为：线性回归和非线性回归。

一元线性回归

回归分析只涉及到两个变量的，称一元回归分析

一元回归的主要任务是从两个相关变量中的一个变量去估计另一个变量，被估计的变量，称因变量，可设为Y；估计出的变量，称自变量，设为X。回归分析就是要找出一个数学模型 $Y=f(X)$

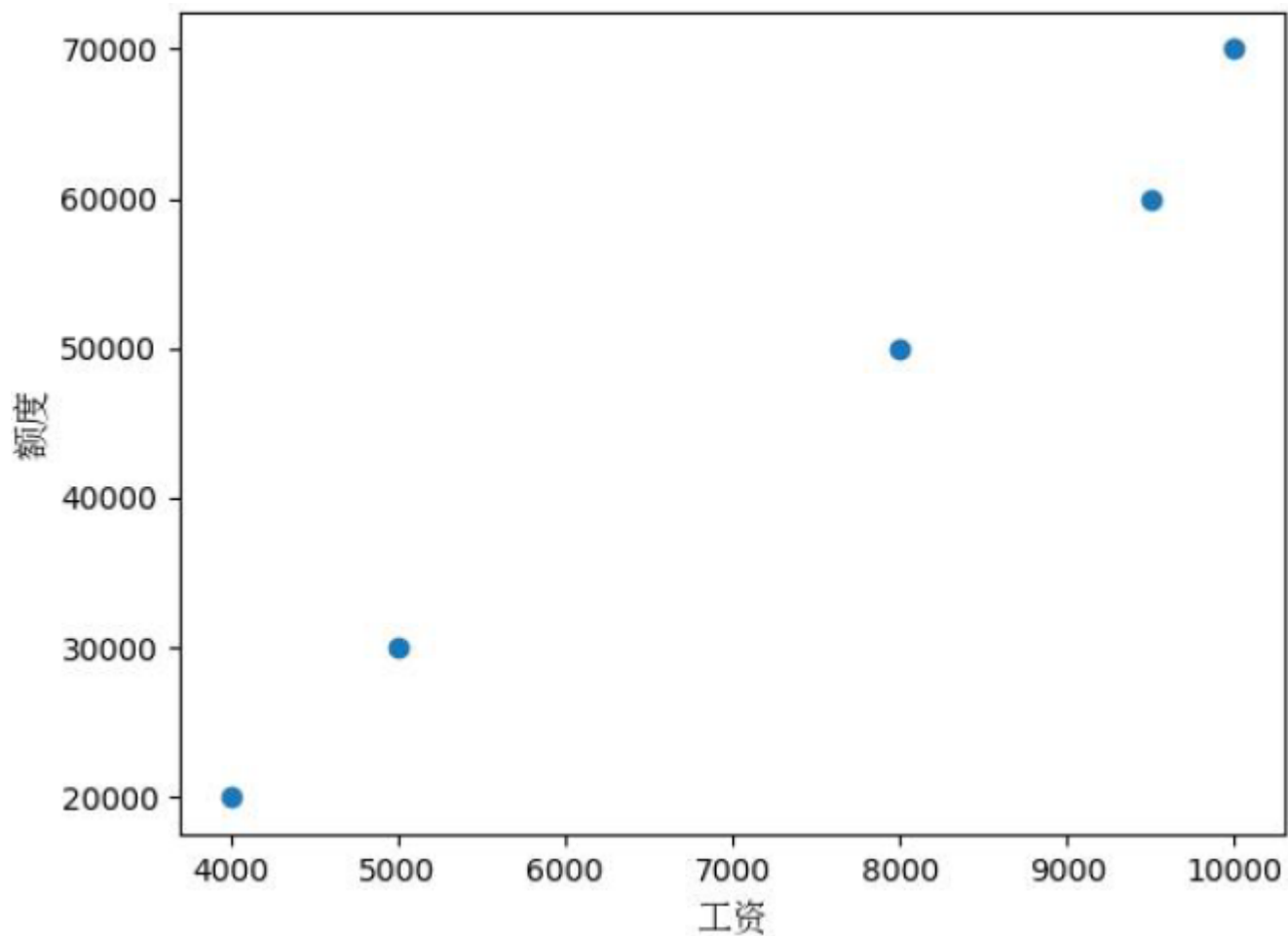
$$y = ax + b$$

案例

编号	工资	额度
1	4000	20000
2	8000	50000
3	5000	30000
4	10000	70000
5	12000	60000
6	15000	?

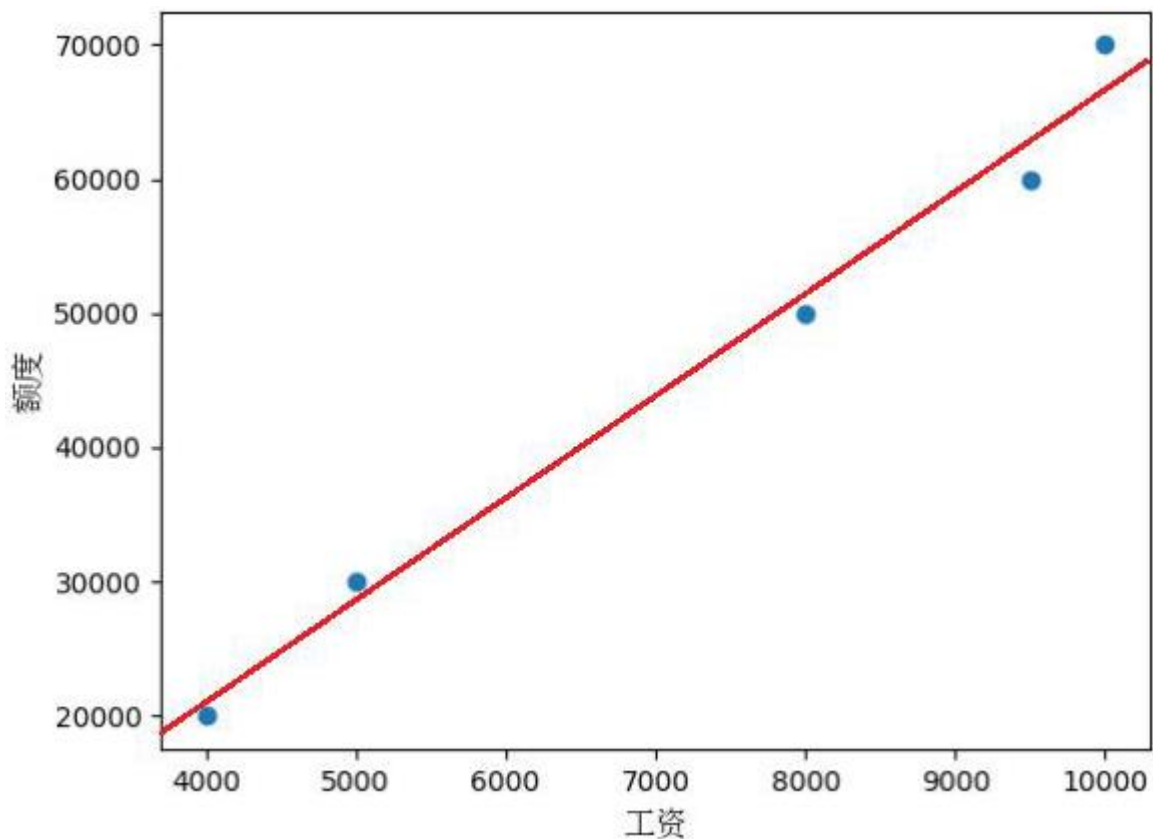
试图学到一个线性模型尽可能准确地预测新样本的输出值？

数据点可视化



找出一条最合适的线来拟合所有的数据点

误差 真实值和预测值之间肯定存在差异，用 ε 来表示



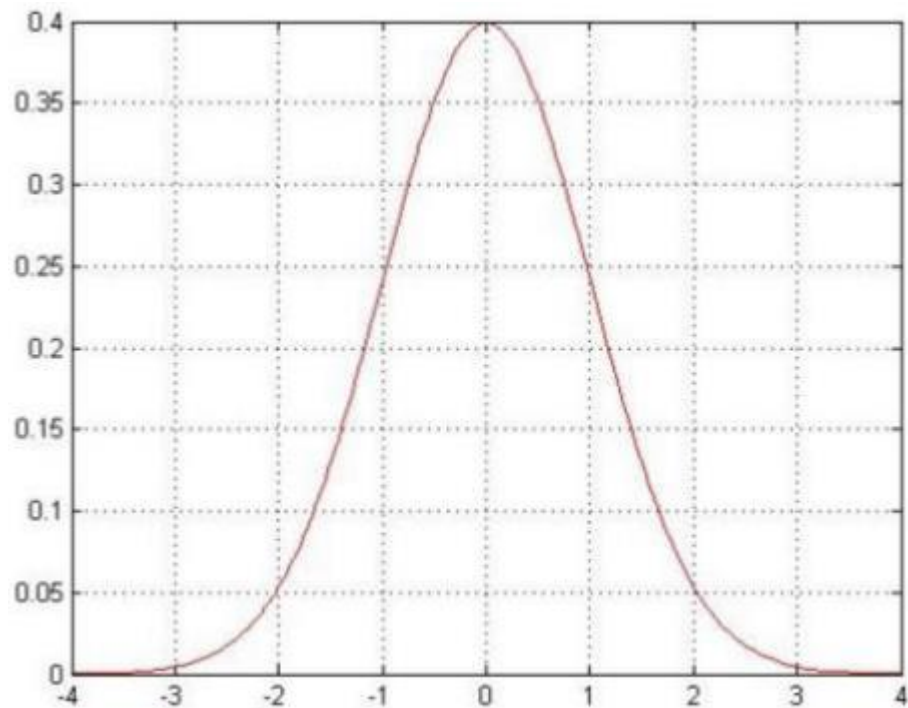
对于每个样本: $\hat{y}_i = wx_i + b + \varepsilon_i$

误差 是独立并且具有相同分布，并且服从均值为0方差为 σ^2 的高斯分布

独立：张三和李四一起来办卡，他们两没有关系。

同分布：他们来的是同一家银行

高斯分布：银行可能会多给，也可能会少给，但绝大多数情况下浮动不会太大，极小的情况下浮动较大，符合正常情况。



预测值与误差：

$$\hat{y}_i = wX_i + b + \varepsilon_i \quad (1)$$

由于误差服从高斯分布：

$$p(\varepsilon_i) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{\varepsilon_i^2}{2\sigma^2}\right) \quad (2)$$

将(1)式代入(2)式

$$p(y_i \mid x_i; w) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(y_i - (wX_i + b))^2}{2\sigma^2}\right)$$

似然函数

$$L(w) = \prod_{i=1}^m \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(y_i - (wX_i + b))^2}{2\sigma^2}\right)$$

对数似然

$$\begin{aligned} \log L(w) &= \log \prod_{i=1}^m \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(y_i - (wX_i + b))^2}{2\sigma^2}\right) \\ &= m \log \prod_{i=1}^m \frac{1}{\sqrt{2\pi}\sigma} - \frac{1}{\sigma^2} \frac{1}{2} \sum_{i=1}^m (y_i - (wX_i + b))^2 \end{aligned}$$

让似然函数越大越好

$$J(w) = \min_{w,b} \sum_{i=1}^m (y_i - (wX_i + b))^2 \quad \text{求式子最小时候的w和 b值}$$

求解

$$\frac{\partial J(w)}{\partial w} = -2 \sum_{i=1}^m (y_i - (wX_i + b)) X_i$$

$$\frac{\partial J(w)}{\partial b} = -2 \sum_{i=1}^m (y_i - (wX_i + b))$$

$$b = \frac{1}{m} \sum_{i=1}^m (y_i - wX_i) = \bar{y} - w\bar{X}$$

$$w = \frac{\sum_{i=1}^m (X_i - \bar{X})(y_i - \bar{y})}{\sum_{i=1}^m (X_i - \bar{X})^2}$$

多元线性回归

在回归分析中，如果有两个或两个以上的自变量，就称为多元回归

$$f(x_i) = w^T x_i + b$$

把w和b吸收入向量， $b=w_0$ ，则X可写成

$$X = \begin{pmatrix} X_1^{(1)} & X_1^{(2)} & \dots & X_1^{(d)} & 1 \\ X_2^{(1)} & X_2^{(2)} & \dots & X_2^{(d)} & 1 \\ \dots & \dots & \dots & \dots & \dots \\ X_n^{(1)} & X_n^{(2)} & \dots & X_n^{(d)} & 1 \end{pmatrix} = \begin{pmatrix} X_1^T & 1 \\ X_2^T & 1 \\ \dots & \dots \\ X_n^T & 1 \end{pmatrix}$$

目标函数

$$J(w) = \min_w \sum_{i=1}^n (w^T x_i - y_i)^2 = (Xw - y)^T (Xw - y)$$

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

$$\because (A-B)^T = A^T - B^T, \quad (AB)^T = B^T A^T$$

$$\frac{\partial J(w)}{\partial w} = \frac{\partial (y^T y - w^T X^T y - y^T Xw + w^T X^T Xw)}{\partial w}$$

$$\because \frac{\partial A^T B^T C}{\partial A} = B^T C, \quad \frac{\partial C^T BA}{\partial A} = B^T C, \quad \frac{\partial A^T BA}{\partial A} = (B + B^T)A$$

$$\frac{\partial J(w)}{\partial w} = -X^T y - X^T y + 2X^T Xw = 2(X^T Xw - X^T y)$$

若 $X^T X$ 是满秩： $2(X^T X w - X^T y) \xrightarrow{\text{求驻点}} 0$

$$w = (X^T X)^{-1} X^T y$$

令 $\hat{x}_i = (x_i; 1)$ ，则线性回归模型：

$$f(\hat{x}_i) = \hat{x}_i^T (X^T X)^{-1} X^T y$$

若 $X^T X$ 是不满秩，则有多多个 \hat{w}

选择哪一个解作为输出呢？需要引入正则化项，
参见P252 11.4嵌入式选择与L1正则化



6.2 逻辑回归

逻辑回归

【定义】逻辑回归分析是对定性变量的回归分析

线性回归模型中，我们处理的因变量都是数值型区间变量，建立的模型描述的是因变量的期望与自变量之间的线性关系。

然而，在许多实际问题中，我们需要研究的因变量不是区间变量而是顺序变量或名义变量这样的属性变量。

比如在致癌因素的研究中，我们收集了若干人的健康记录，包括年龄、性别、抽烟史、日常饮食以及家庭病史等变量的数据。因变量在这里是一个两点（**0-1**）分布变量，**Y=1**（一个人得了癌症），**Y=0**（没得癌症）。

线性回归模型：
$$\hat{y}_i = w_0 + w_1 X_i^{(1)} + w_2 X_i^{(2)} + \dots + w_d X_i^{(d)} \quad (1)$$

线性模型的取值是连续的，而Y只能取0或1

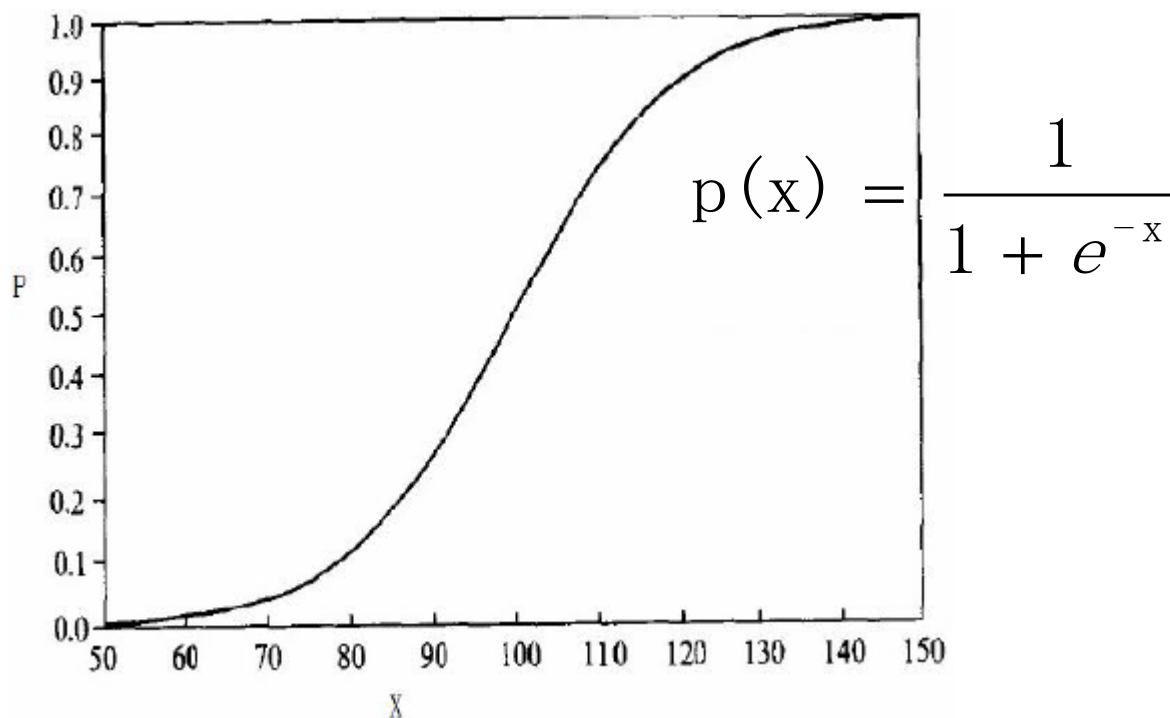
我们注意到，对于0-1型变量，可以用 \hat{y}_i 预测Y=1的概率。

$$p(Y = 1 \mid X) = w_0 + w_1 X^{(1)} + w_2 X^{(2)} + \dots + w_d X^{(d)} \quad (2)$$


问题1. Y=1的概率与自变量之间的关系到底是不是线性的？

例如：我们分析一个人是否买车与其年收入的关系。对于年薪5000元、5万元、50万元三个人，让他们的年薪分别增加5000元对于其买车的可能性影响是不一样的。

概率与自变量之间的关系曲线




问题2. 概率的取值应该在0~1之间，但是（2）式的概率线性模型并不能满足这一点，于是对 P 进行一种变换（logit变换），令 $\text{logit}(p) = \ln(p/(1-p))$ ，使得 $\text{logit}(p)$ 与自变量之间存在线性相关的关系。


$$\ln \frac{p}{1-p} = w_0 + w_1 x^{(1)} + w_2 x^{(2)} + \dots + w_d x^{(d)} = w^T x$$

$$p(y = 1 \mid x) = \frac{\exp(w^T x)}{1 + \exp(w^T x)}$$

$$p(y = 0 \mid x) = \frac{1}{1 + \exp(w^T x)}$$

经过变换得到的模型也解决了（2）中，概率的预测值可能是[0,1]之外的数的缺陷。上式建立的模型，我们称为**logistic模型**（**逻辑回归模型**）



最终，我们关心的是根据自变量的值来对 Y 的取值**0**或**1**进行预测。而我们的逻辑回归模型得到的只是关于 **$P\{Y=1|x\}$** 的预测概率。

于是我们根据模型给出的 $Y=1$ 的概率（可能性）的大小来判断预测 Y 的取值。

一般，以**0.5**为界限，预测 **p** 大于**0.5**时，我们判断此时 Y 更可能为**1**，否则认为 $Y=0$ 。

对于逻辑模型

$$\ln \frac{p}{1-p} = w^T X$$

模型系数的估计不能适用最小二乘估计（**OLS**）

这里，我可以运用最大似然估计（**MLE**）的方法。

OLS通过使得样本观测数据的残差平方和最小来选择参数，

而**MLE**通过最大化对数似然值来估计参数。

设 y 是0-1型变懒， $x^{(1)}, x^{(2)}, \dots, x^{(k)}$ 是与 y 相关的自变量， n 组观测数据为 $(x_i^{(1)}, x_i^{(2)}, \dots, x_i^{(k)}; y_i), i=1,2,\dots, n$ 。

设 $\phi(z)$ 表示预测为1的概率，则有

$$p(y = 1 \mid X; w) = \phi(w^T X) = \phi(z)$$

$$p(y = 0 \mid X; w) = 1 - \phi(z)$$

$P(y=1|x;w)$ 表示给定 w 情况下， x 点 $y=1$ 的概率大小。上式可写为一般形式

$$p(y \mid X; w) = \phi(z)^y (1 - \phi(z))^{(1-y)}$$

于是 y_1, y_2, \dots, y_n 的似然函数为

$$L = \prod_{i=1}^N \phi(w^T X_i)^{y_i} [1 - \phi(w^T X_i)]^{1-y_i}$$

对数似然函数

$$\begin{aligned} L(w) &= \sum_{i=1}^N [y_i \log \phi(w^T x_i) + (1 - y_i) \log (1 - \phi(w^T x_i))] \\ &= \sum_{i=1}^N [y_i \log \frac{\phi(w^T x_i)}{1 - \phi(w^T x_i)} + \log (1 - \phi(w^T x_i))] \\ &= \sum_{i=1}^N [y_i (w^T x_i) - \log (1 + \exp(w^T x_i))] \end{aligned}$$

对 $L(w)$ 求极大值，得到 w 的估计值。

$$\phi(w^T x) = \frac{\exp(w^T x)}{1 + \exp(w^T x)}$$

对数似然函数 $L(w) = \sum_{i=1}^N [y_i(w^T x_i) - \log(1 + \exp(w^T x_i))]$

对L(w)取随机梯度为

$$\begin{aligned} \frac{\partial L}{\partial w_j} &= \sum_{i=1}^N [y_i x_i^{(j)} - \frac{1}{1 + \exp(w^T x_i)} \bullet \exp(w^T x_i) \bullet x_i^{(j)}] \\ &= \sum_{i=1}^N [y_i - \frac{\exp(w^T x_i)}{1 + \exp(w^T x_i)}] \bullet x_i^{(j)} \end{aligned}$$

Logisti回归参数w的求解过程为（类似梯度下降方法，往正梯度方向迭代）

$$w_j \leftarrow w_j + \alpha \sum_{i=1}^N [y_i - \frac{\exp(w^T x_i)}{1 + \exp(w^T x_i)}] \bullet x_i^{(j)}$$

多项逻辑斯蒂回归

假设离散型随机变量 Y 的取值集合 $\{1, 2, \dots, K\}$ ，那么多项逻辑斯蒂回归模型

$$P(Y = k \mid x) = \frac{\exp(\hat{w}_k \bullet x)}{1 + \sum_{k=1}^{K-1} \exp(\hat{w}_k \bullet x)} \quad k = 1, 2, \dots, K-1$$

$$P(Y = K \mid x) = \frac{1}{1 + \sum_{k=1}^{K-1} \exp(\hat{w}_k \bullet x)}$$