华院
www.UniDT.com

线性回归

华院 www.UniDT.com

冷饮批发商一周销售数据

	周一	周二	周三	周四	周五	周六	周日
气温x（摄氏度）	32	38	40	40	39	37	35
冷饮销售y（箱）	97	114	123	118	117	112	107

均值

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{32+38+40+40+39+37+35}{7} = 37.28$$

冷饮批发商一周销售数据

	周一	周二	周三	周四	周五	周六	周日
气温x（摄氏度）	32	38	40	40	39	37	35
冷饮销售y（箱）	97	114	123	118	117	112	107

方差

$$S_x^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{(32-37.28)^2 + \cdots + (35-37.28)^2}{7} = 7.35$$

冷饮批发商一周销售数据

	周一	周二	周三	周四	周五	周六	周日
气温x（摄氏度）	32	38	40	40	39	37	35
冷饮销售y（箱）	97	114	123	118	117	112	107

标准差

$$S_x = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2} = \sqrt{\frac{(32-37.28)^2 + \cdots + (35-37.28)^2}{7}} = 2.71$$

华院 www.UniDT.com

冷饮批发商一周销售数据

	周一	周二	周三	周四	周五	周六	周日
气温x（摄氏度）	32	38	40	40	39	37	35
冷饮销售y（箱）	97	114	123	118	117	112	107

协方差

$$\mathrm{Cov} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

$$= \frac{(32-37.28)(97-112.57)+\cdots+(35-37.28)(107-112.57)}{7} = 20.98$$

华院 www.UniDT.com

冷饮批发商一周销售数据

	周一	周二	周三	周四	周五	周六	周日
气温x（摄氏度）	32	38	40	40	39	37	35
冷饮销售y（箱）	97	114	123	118	117	112	107

相关系数

$$\rho = \frac{\mathrm{Cov}}{S_x S_y} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2}\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i-\bar{y})^2}} = 0.98$$

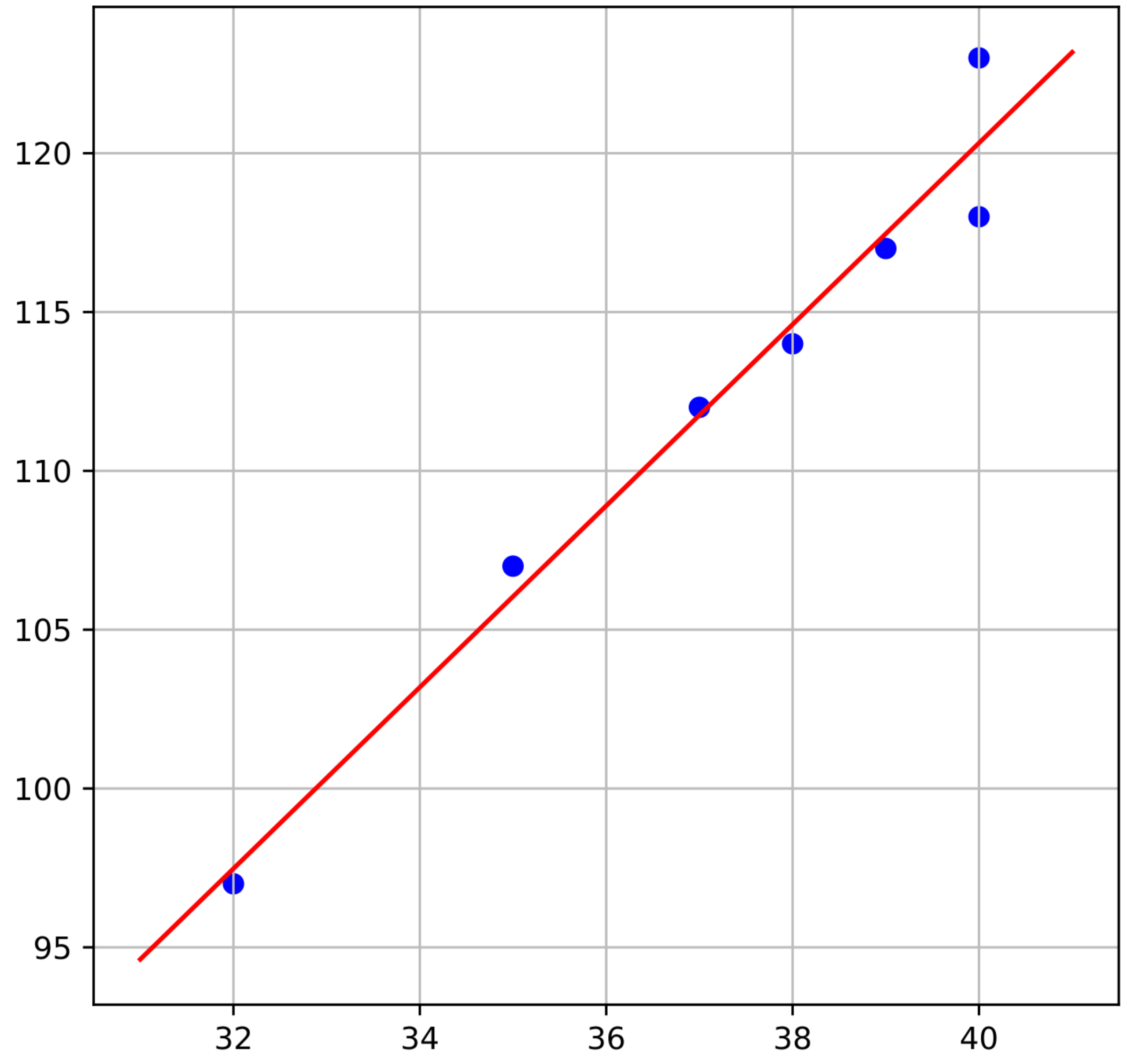
一元线性回归

找一条直线，拟合数据点

$$y = \beta_0 + \beta_1 x$$

最小二乘法

$$\min_{\beta_0,\beta_1} \sum_{i=1}^{n} \left((\beta_0 + \beta_1 x_i) - y_i\right)^2$$



最小二乘法求解：$\boldsymbol{\beta_0}$

$$f(\beta_0,\beta_1)=\sum_{i=1}^{n}((\beta_0+\beta_1 x_i)-y_i)^2$$

$$\frac{\partial f}{\partial \beta_0}=2\sum_{i=1}^{n}(\beta_0+\beta_1 x_i-y_i)=0$$

$$\Rightarrow n\beta_0+\beta_1\sum_{i=1}^{n}x_i-\sum_{i=1}^{n}y_i=0$$

$$\Rightarrow \beta_0+\beta_1\bar{x}-\bar{y}=0$$

其中 $$\bar{x}=\frac{1}{n}\sum_{i=1}^{n}x_i,\bar{y}=\frac{1}{n}\sum_{i=1}^{n}y_i$$

最小二乘法求解：$\boldsymbol{\beta_1}$

$$\frac{\partial f}{\partial \beta_1} = 2\sum_{i=1}^{n} x_i(\beta_0 + \beta_1 x_i - y_i) = 0$$

$$\sum_{i=1}^{n} x_i(\beta_0 + \beta_1 x_i - y_i) = 0$$

注意到

$$\sum_{i=1}^{n} \bar{x}(\beta_0 + \beta_1 x_i - y_i) = \bar{x}\sum_{i=1}^{n}(\beta_0 + \beta_1 x_i - y_i) = 0$$

所以

$$\sum_{i=1}^{n}(x_i - \bar{x})(\beta_0 + \beta_1 x_i - y_i) = 0$$

最小二乘法求解：$\boldsymbol{\beta_1}$

而

$$\sum_{i=1}^{n}(x_i - \bar{x})(\beta_0 + \beta_1\bar{x} - \bar{y}) = (\beta_0 + \beta_1\bar{x} - \bar{y})\sum_{i=1}^{n}(x_i - \bar{x}) = 0$$

所以

$$\sum_{i=1}^{n}(x_i - \bar{x})(\beta_1(x_i - \bar{x}) - (y_i - \bar{y})) = 0$$

$$\beta_1 = \frac{\sum\limits_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum\limits_{i=1}^{n}(x_i - \bar{x})^2} = \frac{\frac{1}{n}\sum\limits_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\frac{1}{n}\sum\limits_{i=1}^{n}(x_i - \bar{x})^2}$$

协方差

方差

最小二乘法求解：$\boldsymbol{\beta_0}, \boldsymbol{\beta_1}$

$$\beta_1 = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2}\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i-\bar{y})^2}} \frac{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i-\bar{y})^2}}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2}} = \rho\frac{S_y}{S_x}$$

而

$$\beta_0 = \bar{y} - \beta_1\bar{x}$$

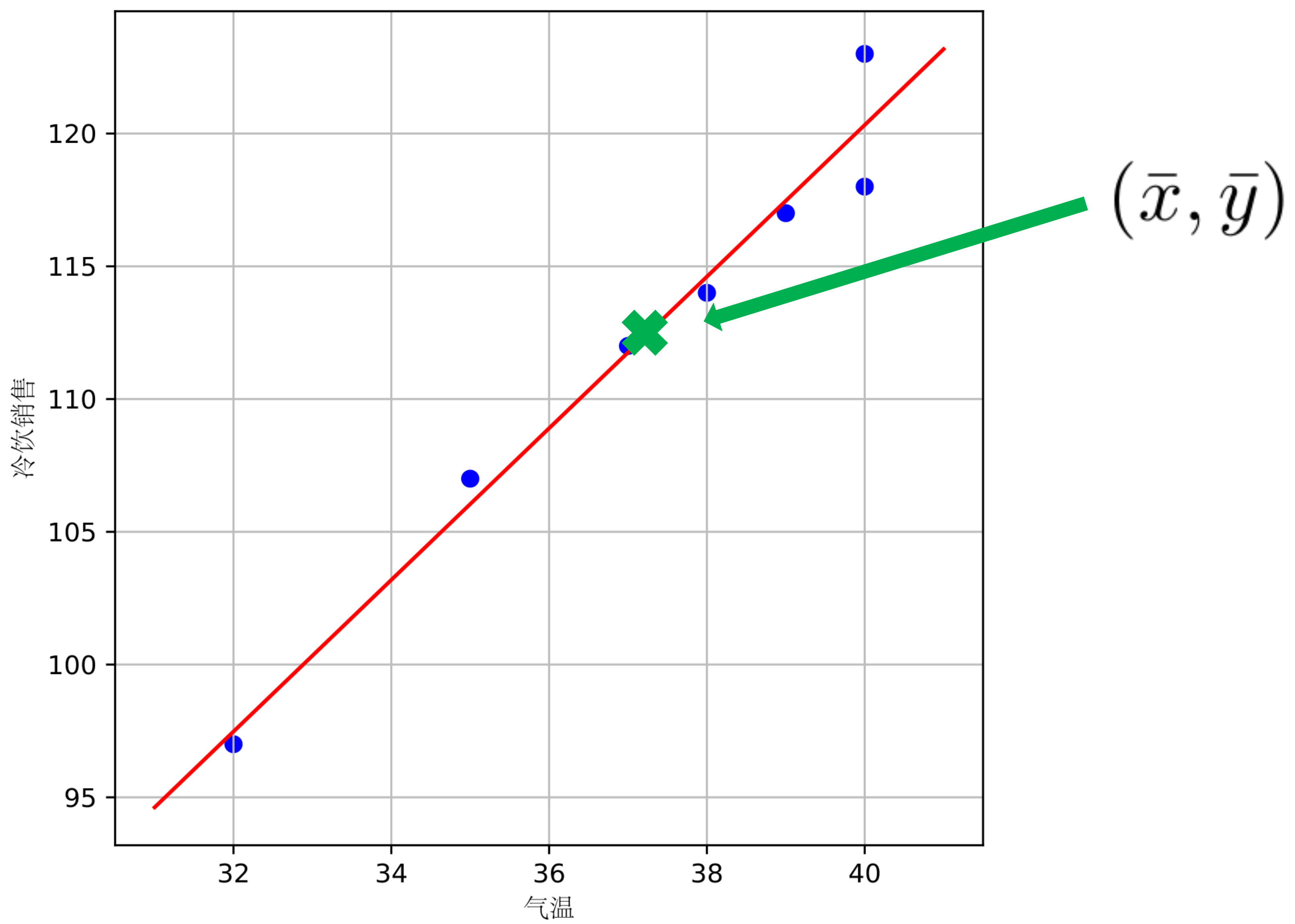
所以

$$y = \bar{y} - \beta_1\bar{x} + \beta_1 x \Rightarrow y - \bar{y} = \beta_1(x-\bar{x}) = \rho\frac{S_y}{S_x}(x-\bar{x})$$

一元线性回归直线

$$y-\bar{y}=\rho\frac{S_y}{S_x}(x-\bar{x}) \Leftrightarrow \boxed{\frac{y-\bar{y}}{S_y}=\rho\frac{x-\bar{x}}{S_x}}$$

- 回归直线经过 $(\bar{x},\bar{y})$
- ρ为$\{x_i\}$和$\{y_i\}$的相关系数
- S_x ，S_y分别是$\{x_i\}$和$\{y_i\}$的标准差



例子求解

$$\rho=0.98, S_x=2.71, S_y=7.87, \bar{x}=37.28, \bar{y}=112.57, \rho\frac{S_y}{S_x}=2.85$$

$$y=2.85x+6.322$$

方差分解

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

$$\begin{aligned}\frac{2}{n}\sum_{i=1}^{n}(y_i-\hat{y}_i)(\hat{y}_i-\bar{y}) &= \rho\frac{S_y}{S_x}\frac{2}{n}\sum_{i=1}^{n}\left(y_i-\bar{y}-\rho\frac{S_y}{S_x}(x_i-\bar{x})\right)(x_i-\bar{x})\\ &= \rho\frac{S_y}{S_x}\frac{2}{n}\sum_{i=1}^{n}\left((y_i-\bar{y})(x_i-\bar{x})-\rho\frac{S_y}{S_x}(x_i-\bar{x})^2\right)\\ &= 2\rho\frac{S_y}{S_x}(\mathrm{cov}-\rho\frac{S_y}{S_x}S_x^2)=2\rho\frac{S_y}{S_x}(\mathrm{cov}-\rho S_x S_y)=0\end{aligned}$$

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

数据集方差 MSE 被解释方差

线性回归的衡量指标

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}|\hat{y}_i - y_i|^2$$

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}|\hat{y}_i - y_i|^2}$$

$$R^2 = \frac{\sum_{i=1}^{n}|\hat{y}_i - \bar{y}|^2}{\sum_{i=1}^{n}|y_i - \bar{y}|^2}$$

$$0 \le R^2 \le 1$$

案例（多变量）

华院
www.UniDT.com

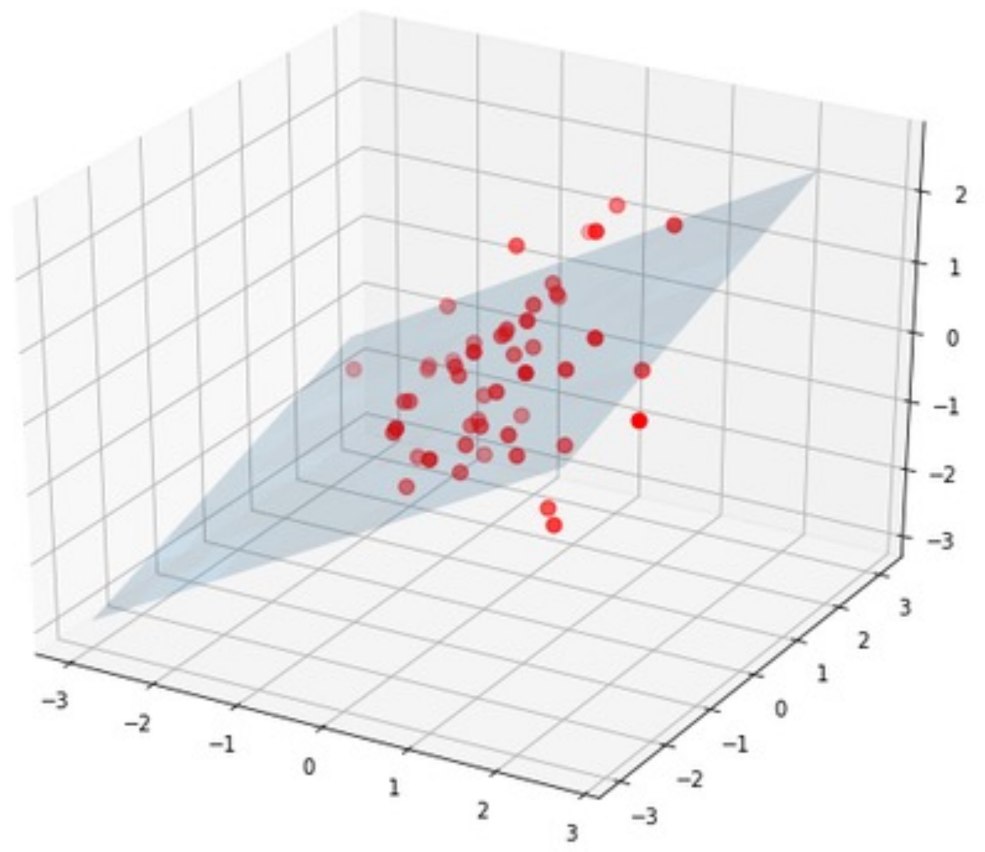
冷饮销量预测

	品种一	品种二	品种三	品种四	品种五	品种六	品种七
冷饮单价	3	4	10	3	8	7	5
广告投入	20	18	17	30	52	41	25
冷饮销量	37	30	21	51	95	80	45

多元线性回归

自变量m个

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_m x_m$$



最小二乘法

$$\min_{\beta_0,\beta_1,\ldots,\beta_m} \sum_{i=1}^{n} \left((\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_m x_{im}) - y_i\right)^2$$

最小二乘形式改写

令

$$\beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_m \end{pmatrix} \qquad X = \begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1m} \\ 1 & x_{21} & x_{22} & \cdots & x_{2m} \\ 1 & x_{31} & x_{32} & \cdots & x_{3m} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nm} \end{pmatrix} \qquad y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}$$

最小二乘形式

$$\min_{\beta} \|X\beta - y\|^2$$

梯度计算

$$g(\beta) = \langle w, \beta \rangle = w^T \beta = \sum_{i=0}^{m} w_i \beta_i$$

$$\nabla g = \begin{pmatrix} \frac{\partial g}{\partial \beta_0} \\ \frac{\partial g}{\partial \beta_1} \\ \dots \\ \frac{\partial g}{\partial \beta_m} \end{pmatrix} = \begin{pmatrix} w_0 \\ w_1 \\ \dots \\ w_m \end{pmatrix} = w$$

梯度计算

假设 $A = A^T$

$$h(\beta) = \langle A\beta, \beta \rangle = \beta^T A\beta = \sum_{i,j} a_{ij}\beta_i\beta_j$$

定义 $p(u,v) = \langle Au, v \rangle = \langle Av, u \rangle$

令 $u(\beta) = \beta, v(\beta) = \beta \Rightarrow h(\beta) = p(u(\beta), v(\beta))$

$$\nabla h = \frac{\partial p}{\partial u}\frac{\partial u}{\partial \beta} + \frac{\partial p}{\partial v}\frac{\partial v}{\partial \beta} = Av(\beta) + Au(\beta) = 2A\beta$$

最小二乘法求解

$$
\begin{aligned}
f(\beta) &= (X\beta - y)^T (X\beta - y) \\
&= (\beta^T X^T - y^T)(X\beta - y) \\
&= \beta^T X^T X\beta - \beta^T X^T y - y^T X\beta + y^T y
\end{aligned}
$$

$$\nabla_\beta f = 2X^T X\beta - X^T y - X^T y = 2(X^T X\beta - X^T y) = 0$$

Normal Equation $$X^T X\beta = X^T y$$

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

例子求解

	品种一	品种二	品种三	品种四	品种五	品种六	品种七
冷饮单价x_1	3	4	10	3	8	7	5
广告投入x_2	20	18	17	30	52	41	25
冷饮销量y	37	30	21	51	95	80	45

$$X = \begin{pmatrix} 1 & 3 & 20 \\ 1 & 4 & 18 \\ 1 & 10 & 17 \\ 1 & 3 & 30 \\ 1 & 8 & 52 \\ 1 & 7 & 41 \\ 1 & 5 & 25 \end{pmatrix}, y = \begin{pmatrix} 37 \\ 30 \\ 21 \\ 51 \\ 95 \\ 80 \\ 45 \end{pmatrix}$$

$$\beta = (X^T X)^{-1} X^T y = (-4.08, -0.85, 2.07)^T$$

$$y = -4.08 - 0.85x_1 + 2.07x_2$$

延伸

$$\beta = (X^TX)^{-1}X^Ty$$

$(X^TX)^{-1}$ 是否一定存在？

- 在实际应用中，通常X的行数远大于列数， X列满秩，所以X^TX可逆
- 为了增加模型鲁棒性，通常会最小化如下目标函数

$$\|X\beta - y\|^2 + \lambda\|\beta\|^2 \quad (\lambda > 0)$$

此时

$$\beta = (X^TX + \lambda I)^{-1}X^Ty$$

谢 谢！

让世界更智慧！

华院计算技术（上海）股份有限公司

上 海 · 北 京 · 成 都 · 西 安 · 杭 州

地址 · 上海市静安区万荣路1268号云立方大厦A座9楼

电话 · 021-63617288 传真 · 021-63617299

网址 · www.UniDT.com