

# 人工智能原理-作业4

Author: 夏弘宇 2023011004

## T2

在一个线性回归问题中，有 $n$ 个点 $(x_i, y_i)$ ,  $i = 1, 2, \dots, n$ , 通过最小二乘法求得的线性回归方程为 $\hat{y} = \hat{w}x + b$ 。需要证明以下等式成立：

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

其中 $\bar{y} = \frac{1}{n} \sum_{i=1}^n y_i$ 表示 $y$ 的样本均值。

$$\begin{aligned} \text{MSE} &\Rightarrow \hat{y} = \hat{w}x + b. \text{ 证 } \sum_{i=1}^n (y_i - \bar{y})^2 = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 + \sum_{i=1}^n (y_i - \hat{y}_i)^2 \\ \text{左} &= \sum_{i=1}^n (y_i - \bar{y})^2 = \sum_{i=1}^n (y_i - \hat{y} + \hat{y} - \bar{y})^2 = \sum_{i=1}^n (y_i - \hat{y})^2 + \sum_{i=1}^n (\hat{y} - \bar{y})^2 + 2 \sum_{i=1}^n (y_i - \hat{y})(\hat{y} - \bar{y}) \\ &\text{即只需证 } \sum_{i=1}^n (y_i - \hat{y})(\hat{y} - \bar{y}) = 0. \\ \text{由 MSE 最小} &\Leftrightarrow f(w, b) = \sum_{i=1}^n (wx_i + b - y_i)^2 \text{ 最小} \\ \frac{\partial f}{\partial w} &= 2 \sum_{i=1}^n (wx_i + b - y_i) \cdot x_i = 0 \Rightarrow \sum_{i=1}^n (\hat{y}_i - y_i) x_i = 0 \Rightarrow \sum_{i=1}^n (\hat{y}_i - y_i) \cdot wx_i = 0 \\ \frac{\partial f}{\partial b} &= 2 \sum_{i=1}^n (wx_i + b - y_i) = 0, \Rightarrow \sum_{i=1}^n (\hat{y}_i - y_i) = 0 \Rightarrow \sum_{i=1}^n (\hat{y}_i - y_i)(b - \bar{y}) = 0 \\ &\Rightarrow \sum_{i=1}^n (\hat{y}_i - y_i)(\hat{y} - \bar{y}) = 0. \text{ 即 } \sum_{i=1}^n (y_i - \hat{y})(\hat{y} - \bar{y}) = 0. \# \end{aligned}$$

## T4

设在一个 $K$ 分类问题中，一个样例预测为第 $k$ 类的概率建模为如下的对数线性模型：

$$\log P(Y = k) = \beta_k x - \log Z$$

其中：

- $P(Y = k)$  表示样例预测为第 $k$ 类的概率
- $x$  是输入的样例数据（特征向量）
- $\beta_k$  为第 $k$ 类的权重向量
- $-\log Z$  是归一化项，保证所有类别的概率之和为1

证明通过该对数线性模型，预测概率的表达式为Softmax形式：

$$P(Y = k) = \frac{e^{\beta_k x}}{\sum_{j=1}^K e^{\beta_j x}}$$

$$P(Y=k) = e^{\beta_k x - \log Z} = \frac{e^{\beta_k x}}{Z}$$

设一共有 $K$ 类

$$\text{则 } \sum_{i=1}^K P(Y=i) = \sum_{i=1}^K \frac{e^{\beta_i x}}{Z} = \frac{1}{Z} \cdot \sum_{i=1}^K e^{\beta_i x} \stackrel{\Delta}{=} 1.$$

$$\text{则 } Z = \sum_{i=1}^K e^{\beta_i x}.$$

$$\text{故 } P(Y=k) = \frac{e^{\beta_k x}}{\sum_{i=1}^K e^{\beta_i x}}$$

T1

## 题目描述

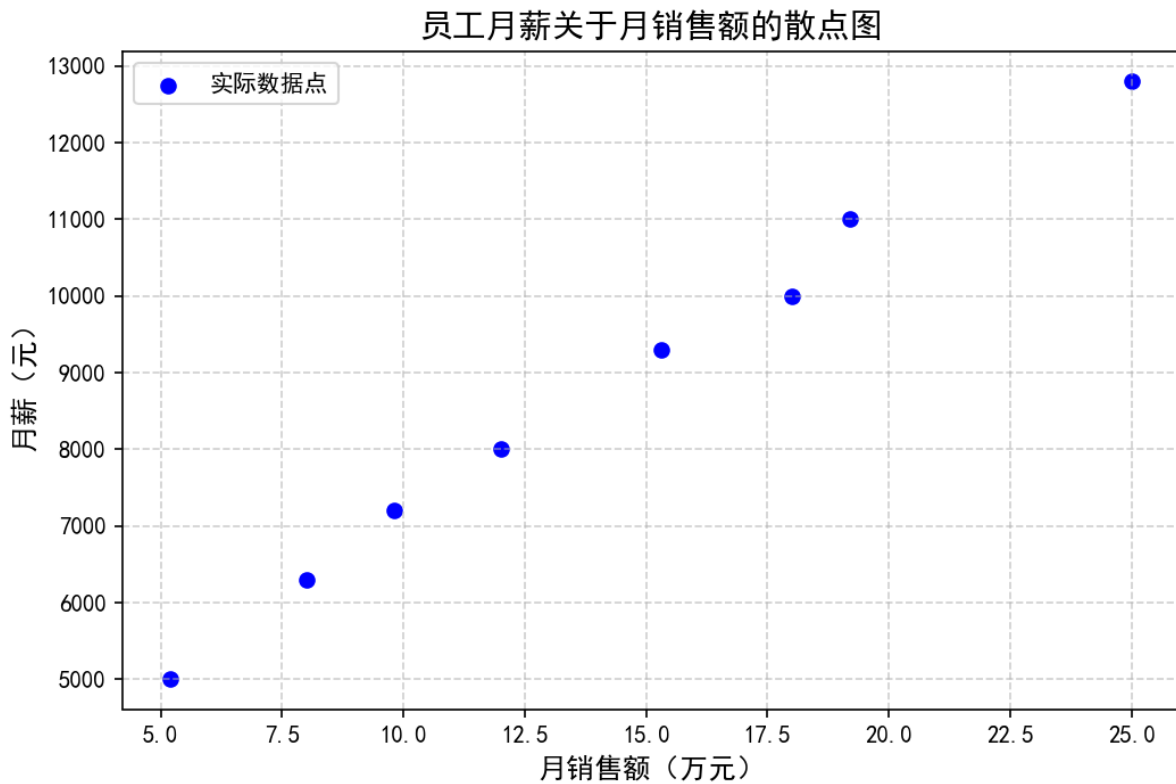
某销售公司收集了8名员工的月销售额（万元）和对应月薪（元）数据如下：

月销售额（万元）	5.2	9.8	15.3	19.2	25	8	12	18
月薪（元）	5000	7200	9300	11000	12800	6300	8000	10000

需要完成以下分析任务：

1. 绘制散点
2. 计算线性回归方程和回归系数 $r^2$
3. 计算MAE和MSE评估模型

## (1) 散点图绘制



## (2) 线性回归分析

### 1. 基础统计量计算：

- $\sum x = 112.5$
- $\sum y = 69600$
- $n = 8$
- $\bar{x} = 14.0625$
- $\bar{y} = 8700$

### 2. 协方差计算： $\sum xy = 1096450$

$$\sum x^2 = 1882.81$$

$$\text{Cov}(x,y) = (1096450 - 8 \times 14.0625 \times 8700) / 8 = 14712.5$$

### 3. 方差计算： $\text{Var}(x) = (1882.81 - 8 \times 14.0625^2) / 8 = 37.5973$

### 4. 回归系数： $b = \text{Cov}(x,y) / \text{Var}(x) = 14712.5 / 37.5973 \approx 391.32$

$$a = \bar{y} - b \cdot \bar{x} \approx 3197$$

### 5. 回归方程： $\hat{y} = 391.3x + 3197$

6. 回归系数 $r^2$ 计算:  $SST = \sum(y-\bar{y})^2 = 46340000$   
 $SSE = \sum(y-\hat{y})^2 \approx 281703.73$   
 $r^2 = 1 - SSE/SST \approx 0.9939$

7. 最终结果

- 回归方程:  $y = 391.3x + 3197$
- 决定系数:  $r^2 = 0.9939$

(3) 误差分析

x	y	$\hat{y}$		$y-\hat{y}$
5.2	5000	5232	232	53824
9.8	7200	7032	168	28224
15.3	9300	9185	115	13225
19.2	11000	10710	290	84100
25	12800	12980	180	32400
8	6300	6327	27	729
12	8000	7893	107	11449
18	10000	10240	240	57600

$MAE = (232+168+115+290+180+27+107+240)/8 \approx 170.0$   
 $MSE = (53824+...+57600)/8 \approx 35190$

- MAE: 170.0元
- MSE: 35190

结论

1. 数据呈现强线性相关性 ( $r^2=0.9939$ )
2. 回归方程 $y=391.3x+3197$ 能很好解释月薪变化
3. 误差指标显示模型预测精度较高 (MAE=170, MSE=35190)