机器学习

Machine Learning

主讲人：张敏 清华大学长聘副教授

机器学习

MACHINE LEARNING-MIN ZHANG

Unit.04

回归分析

*图片均来自网络或已发表刊物

目录

什么是回归分析

- Regression
- 回归分析是描述变量间关系的一种统计分析方法
- 例：在线教育场景
  - 因变量 Y：在线学习课程满意度
  - 自变量 X：平台交互性、教学资源、课程设计
- 前面提到过的西洋跳棋系统目标函数的设计也是一个回归问题
- 预测性的建模技术，通常用于预测分析
- 预测的结果多为连续值（但也可以是离散值，甚至是二值）

目录

线性回归 (Linear regression)

- 因变量和自变量之间是线性关系，就可以使用线性回归来建模

$$\text{在线课程满意度} = \boxed{1.6} + \boxed{0.11} \times \text{平台交互性} + \boxed{0.15} \times \text{教学资源} + \boxed{0.27} \times \text{课程设计}$$

常数项 (截距)　　**系数 (斜率)**

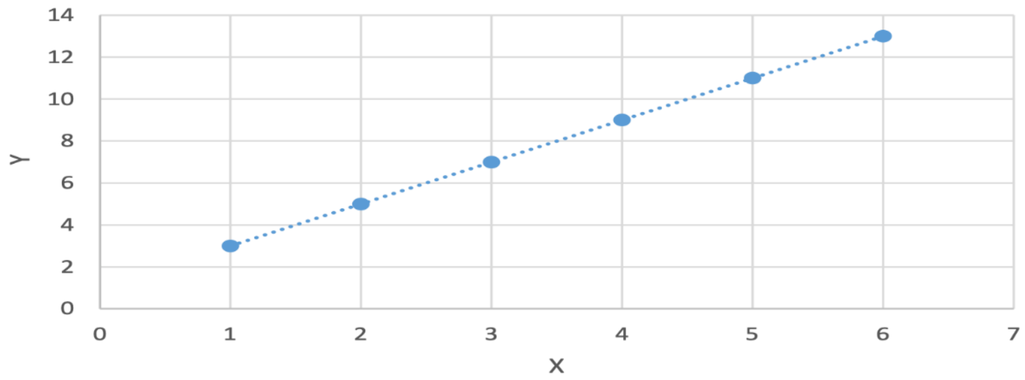
- 线性回归的目的即找到最能匹配(解释)数据的截距和斜率

线性假设

- 线性：有些变量间的线性关系是确定性的

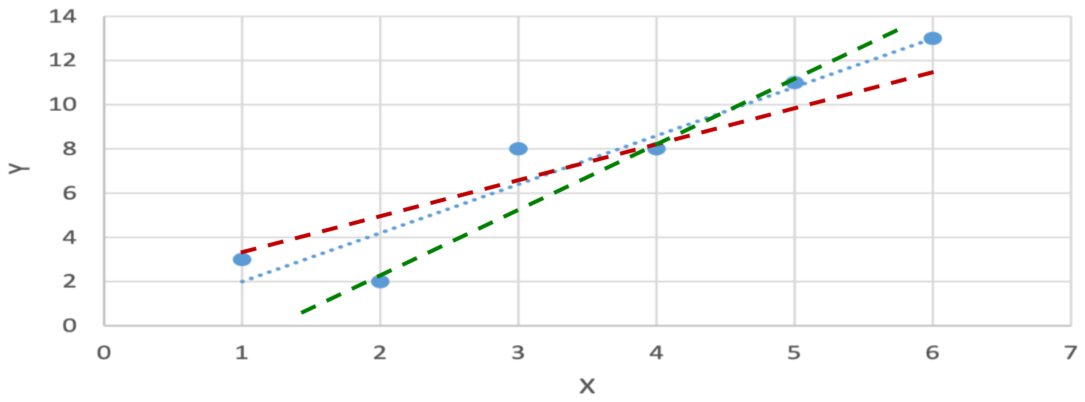
X	1	2	3	4	5	6
Y	3	5	7	9	11	13

$$Y = 1 + 2X$$



线性假设

- 线性：然而通常情况下，变量间是近似的线性关系



如何得到一条直线能够**最好地解释数据**？

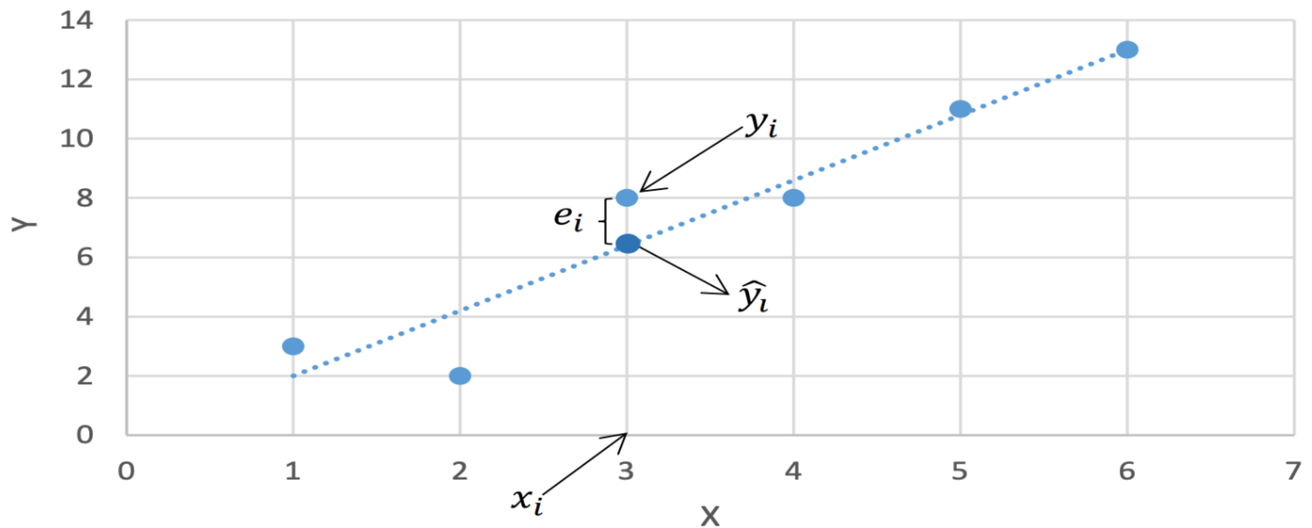
如何拟合数据

- 假设只有一个因变量和自变量，每个训练样例表示 (x_i, y_i)
- 用 $\hat{y}_i$ 表示根据拟合直线和 x_i 对 y_i 的预测值

$$\hat{y}_i = b_1 + b_2 x_i$$

- 定义 $\mathrm{e}_i = y_i - \hat{y}_i$ 为误差项

如何拟合数据



- 目标：得到一条直线使得对于所有训练样例的误差项尽可能小

线性回归的基本假设

- 1 自变量与因变量间存在线性关系;
- 2 数据点之间独立;
- 3 自变量之间无共线性，相互独立;
- 4 残差独立,等方差,且符合正态分布。

目录

- 什么是回归分析
- 简单线性回归
- 损失函数
  - 最小二乘法
  - 梯度下降法
- 多元线性回归
- 相关系数与决定系数

损失函数(loss function)的定义

- 多种损失函数都是可行的，凭直觉就可以想到：
  - 所有误差项的加和 $\sum_{i=1}^{n} e_i = \sum_{i=1}^{n}(y_i - \hat{y}_i)$
  - 所有误差项绝对值的加和 $\sum_{i=1}^{n} |e_i| = \sum_{i=1}^{n} |(y_i - \hat{y}_i)|$
- 考虑到优化等问题，最常用的是基于误差平方和的损失函数

$$
\begin{aligned}
\min_{b_1,b_2}: \sum_{i=1}^{n} e_i^2 &= \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \\
&= \sum_{i=1}^{n}(y_i - b_1 - b_2 x_i)^2
\end{aligned}
$$

最小二乘法（Least Square, LS)

- 为了求解最优的截距和斜率，可以转化为一个针对损失函数的凸优化问题，称为最小二乘法

$$\min_{b_1,b_2}: \quad \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n}(y_i - b_1 - b_2x_i)^2$$

$$\frac{\partial \sum_{i=1}^{n} e_i^2}{\partial b_1} = -2\sum_{i=1}^{n}(y_i - b_1 - b_2x_i) = 0 \qquad (1)$$

$$\frac{\partial \sum_{i=1}^{n} e_i^2}{\partial b_2} = -2\sum_{i=1}^{n}x_i(y_i - b_1 - b_2x_i) = 0 \qquad (2)$$

- 求解得到:

$$b_2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

$$b_1 = \bar{y} - b_2\bar{x}$$

- $\bar{x}$ 和 $\bar{y}$ 分别表示自变量和因变量的均值

目录

梯度下降法(Gradient Descent, GD)

- 除了最小二乘法，还可以用基于梯度的方法迭代更新截距和斜率
- 梯度下降法
  - 初始化 b_1, b_2
  - 重复：
    - $b_1 = b_1 - \alpha \dfrac{\partial \sum_{i=1}^{n} e_i^2}{\partial b_1}$
    - $b_2 = b_2 - \alpha \dfrac{\partial \sum_{i=1}^{n} e_i^2}{\partial b_2}$

对比LS:

回忆西洋跳棋系统设计：

$w_i \leftarrow w_i + c \cdot f_i \cdot error(b)$

目录

多元线性回归(Multiple Linear Regression)

- 当因变量有多个时，我们可以用矩阵方式表达

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{12} & x_{13} & \cdots & x_{1k} \\ 1 & x_{22} & x_{23} & \cdots & x_{2k} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n2} & x_{n3} & \cdots & x_{nk} \end{bmatrix} \boxed{\begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix}} + \boxed{\begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}}$$

回归系数

误差项 / 残差

- 基于以上矩阵表示，可以写为

$$Y = X\beta + \epsilon$$

$$Y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \ X = \begin{bmatrix} 1 & x_{12} & \cdots & x_{1k} \\ 1 & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n2} & \cdots & x_{nk} \end{bmatrix}, \ \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix} \text{ and } \epsilon = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

多元线性回归（续）

$$Y = X\beta + \epsilon$$

- 此时误差项

$$e = \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix} = y - X\beta$$

- 损失函数

$$\sum_{i=1}^{n} e_i^2 = e'e$$

e' 表示转置

- 求解

$$\frac{\partial e'e}{\partial \beta} = -2X'Y + 2X'X\beta = 0$$

- 得到

$$\beta = (X'X)^{-1}X'Y$$

多元线性回归参数估计的推导(法二)

$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_{i1} - \ldots - \beta_p x_{ik})^2$$

对每一个需要估计的参数β_i求偏导:

$$\sum(y_i - \beta_0 - \beta_1 x_{i1} - \ldots - \beta_k x_{ik}) = 0$$
$$\sum(y_i - \beta_0 - \beta_1 x_{i1} - \ldots - \beta_k x_{ik})x_{i1} = 0$$
$$\ldots$$
$$\sum(y_i - \beta_0 - \beta_1 x_{i1} - \ldots - \beta_k x_{ik})x_{ik} = 0$$

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

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

实例：家庭花销预测

- 记录了 25 个家庭每年在快销品和日常服务上
  - 总开销（Y）
  - 每年固定收入（X_2）、持有的流动资产（X_3）
- 可以构建如下线性回归模型：

$$y_i = \beta_1 + \beta_2 x_{i2} + \beta_3 x_{i3} + \epsilon_i; \quad i = 1, \cdots, 25.$$

实例：家庭花销预测（续）

$$\beta = (X'X)^{-1}X'Y$$

$$X = [\,1\ X_{i2}\ X_{i3}\,]$$

- 计算系数的表达式

$$(X'X)^{-1}X'y = \begin{bmatrix} n & \sum_{i=1}^{n} x_{i2} & \sum_{i=1}^{n} x_{i3} \\ \sum_{i=1}^{n} x_{i2} & \sum_{i=1}^{n} x_{i2}^2 & \sum_{i=1}^{n} x_{i2}x_{i3} \\ \sum_{i=1}^{n} x_{i3} & \sum_{i=1}^{n} x_{i2}x_{i3} & \sum_{i=1}^{n} x_{i3}^2 \end{bmatrix}^{-1} \begin{bmatrix} \sum_{i=1}^{n} y_i \\ \sum_{i=1}^{n} x_{i2}y_i \\ \sum_{i=1}^{n} x_{i3}y_i \end{bmatrix}$$

- 得到

$$\begin{bmatrix} b_1 \\ b_2 \\ b_3 \end{bmatrix} = \begin{bmatrix} 25 & 4080.1 & 14379.1 \\ 4080.1 & 832146.8 & 2981925 \\ 14379.1 & 2981925 & 11737267 \end{bmatrix}^{-1} \begin{bmatrix} 4082.34 \\ 801322.7 \\ 2994883 \end{bmatrix} = \begin{bmatrix} 36.789 \\ 0.332 \\ 0.125 \end{bmatrix}$$

实例：家庭花销预测（续）

- 最终的预测模型为

$$\hat{y}_i = 36.79 + 0.3318 x_{i2} + 0.1258 x_{i3}$$

- 如果一个家庭每年固定收入为 50K$、持有流动资产 100K$，则预计一年将会花费

$$\begin{aligned}\hat{y}_i &= 36.79 + 0.3318(50) + 0.1258(100) \\ &= 65.96 \text{ K\$}\end{aligned}$$

以“误差平方和”为损失函数的优缺点

- 用误差平方和作为损失函数有很多优点
  - 损失函数是严格的凸函数，有唯一解
  - 求解过程简单且容易计算
- 同时也伴随着一些缺点
  - 结果对数据中的“离群点”(outlier)非常敏感
    - 解决方法：提前检测离群点并去除
  - 损失函数对于超过和低于真实值的预测是等价的
    - 但有些真实情况下二者带来的影响是不同的

目录

线性回归的相关系数

- 定义因变量和自变量之间的相关系数 r

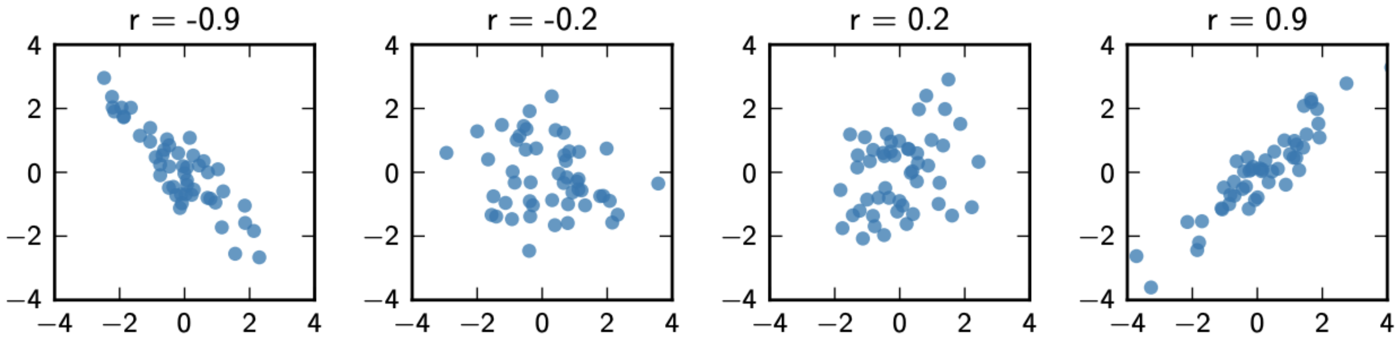
$$r = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{s_x}\right)\left(\frac{y_i - \bar{y}}{s_y}\right)$$

$\bar{x}$: X 的均值

s_x: X 的标准差

$$\sqrt{\frac{1}{n-1}\sum(xi - \bar{x})^2}$$

协方差，描述两个变量 X 和 Y 的线性相关程度



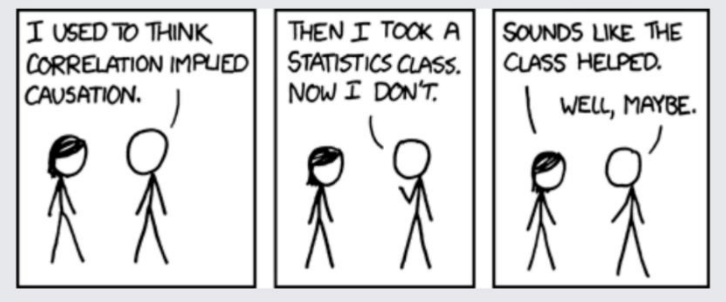
线性回归的决定系数(coefficient of determination)

- 决定系数R^2,也称作判定系数、拟合优度

$$R^2 = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2} \qquad R^2 = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2 / n}{\sum_i (y_i - \bar{y})^2 / n} = 1 - \frac{MSE}{VAR}$$

y_i：真实值，$\hat{y}_i$：预测值，$\bar{y}$均值

- 注意：有可能<0, R^2不是r^2
- 衡量了模型对数据的解释程度
  - y的波动有多少百分比能被x的波动所描述
  - R^2越接近1，表示回归分析中自变量对因变量的解释越好



- 特别注意：变量相关 ≠ 存在因果关系

总结

- 回归分析：描述变量间关系的统计分析方法
- 线性回归：最常用，基本假设
- 基于误差平方和的损失函数
  - 最小二乘法
  - 梯度下降法
- 扩展到多元线性回归
- 相关系数与决定系数：相关 ≠ 因果