Regression

Mathematical Modeling

Prof. Dr. Jingzhi Li

Department of Mathematics,
Southern University of Science and Technology

2025 Spring

Overview

回归（Regression）是一种用于预测连续数值变量的监督学习方法。回归分析的目标是建立输入特征（自变量）与输出变量（因变量）之间的数学关系，从而能够根据新的输入预测对应的数值输出。

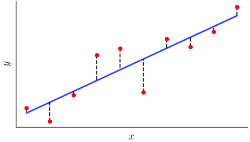
线性回归

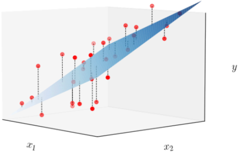
在线性回归问题中，我们假设输入和输出成线性关系。设输入 $\mathbf{x} \in \mathbb{R}^d$，那么线性映射关系可以写为：

$$\hat{y}(\omega, x) = \omega_0 + \omega_1 x_1 + \cdots + \omega_d x_d$$

其中，$\omega \in \mathbb{R}^d$ 是模型的参数，包括偏置项 ω_0 和特征权重 $(\omega_1, \ldots, \omega_d)$.

图中展示了 1 维和 2 维情况下数据点和线性回归模型的结果。在 d 维输入和 1 维输出的情况下，线性回归模型有 $d+1$ 个参数，从而生成了 $d+1$ 维空间中一个 d 维的超平面。





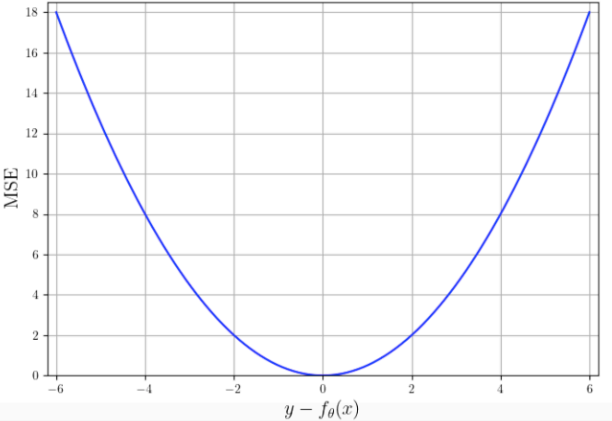
损失函数

- 损失函数 $\mathcal{L}(y_i, \hat{y}_i)$ 测量预测值和真实值之间的误差，越小越好
- 具体损失函数的定义依赖于具体的数据和任务
- 最常用的损失函数之一：均方误差（mean squared error, MSE）

$$J(\theta) = \frac{1}{2N} \sum_{i=1}^{N} (y_i - f_\theta(x_i))^2$$

- 线性回归问题的优化目标为：

$$\min_\theta J(\theta)$$



- 对预测误差大的有更大的惩罚
- 容忍很小的预测误差

评价回归模型的指标

- 均方根误差（RMSE）
  - $\text{RMSE} = \sqrt{\text{MSE}}$
  - 与输出具有相同的量纲，从直观上易于比较
  - 越小越好
- 决定系数（R^2）
  - $R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$
  - 越接近 1 说明模型越好
- p 值：用于检验自变量是否对因变量有显著影响
  - 如果 p 值很小（通常 < 0.05），说明该自变量对回归模型有统计学意义。

1. 回归
2. 使用 MATLAB 进行回归分析
   - 使用 \ 运算符进行简单的线性回归
   - 使用 polyfit 函数实现一元线性回归
   - 使用 fitlm 函数实现线性回归
3. 梯度下降法
4. 多项式拟合
5. 岭回归与 Lasso 回归

1. 回归
2. 使用 MATLAB 进行回归分析
   - 使用 \ 运算符进行简单的线性回归
   - 使用 polyfit 函数实现一元线性回归
   - 使用 fitlm 函数实现线性回归
3. 梯度下降法
4. 多项式拟合
5. 岭回归与 Lasso 回归

\ 运算符可以用来求解最小二乘问题，是一种基于矩阵运算的方法，本质上等价于求解：

$$\min_{\theta} J(\theta) = \frac{1}{2}(y - X\theta)^T(y - X\theta)$$

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

- 注意要添加一列常数列

```
% 添加常数列
X = [ones(length(x),1) x]; % 第一列全为1，表示截距项
```

1 回归

2 使用 MATLAB 进行回归分析

- 使用 \ 运算符进行简单的线性回归
- 使用 polyfit 函数实现一元线性回归
- 使用 fitlm 函数实现线性回归

3 梯度下降法

4 多项式拟合

5 岭回归与 Lasso 回归

polyfit 适用于多项式拟合，一次多项式对应线性回归。

- p = polyfit(x,y,1)
  - x 为输入数据特征
  - y 为输出数据特征
  - 1 指定一维
  - p(1) 是斜率，P(2) 是截距
- 可以使用 polyval 函数生成拟合值

- 该函数会返回一个 LinearModel 对象，该对象包含拟合模型的详细信息，包括系数、统计检验结果、残差分析等。
- model.Coefficients：输出模型系数
- model.Rsquared：输出模型的 R 方信息
- plot(model)：模型可视化

最小二乘法的局限性

- 对异常值敏感
  - 误差平方放大了异常值的影响
  - 使用稳健回归或去除异常值
- 多重共线性问题
  - 当自变量之间高度相关时，会使 X^TX 接近奇异矩阵或不可逆，导致回归系数不稳定
  - 使用岭回归或 Lasso 回归
- 计算复杂度高
  - 计算 $(X^TX)^{-1}X^Ty$ 的时间复杂度大约是 $O(Nd^2 + d^3)$
  - 使用梯度下降算法

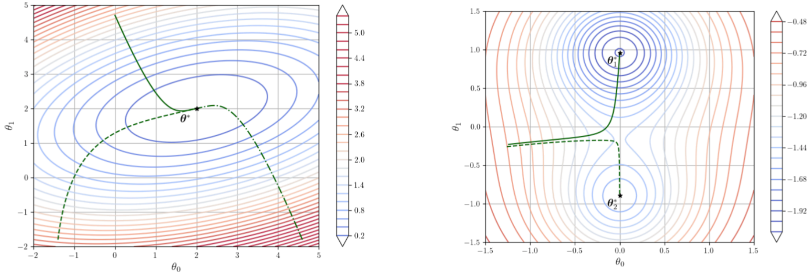
梯度下降法

梯度下降法（Gradient Descent Method）通过向最陡峭的下降方向迭代移动来最小化损失函数，并沿途更新参数。公式为

$$\theta = \theta - \eta \nabla_\theta J(\theta)$$

其中 η 是参数更新的步长，称为学习率（learning rate），将之前的 MSE 代入可得：

$$\begin{aligned}
\theta &= \theta - \eta \nabla_\theta \left(\frac{1}{2N} \sum_{i=1}^{N} (y_i - f_\theta(x_i))^2 \right) \\
&= \theta - \frac{\eta}{N} \sum_{i=1}^{N} (f_\theta(x_i) - y_i) \nabla_\theta f_\theta(x_i) \\
&= \theta - \frac{\eta}{N} \sum_{i=1}^{N} (f_\theta(x_i) - y_i) x_i
\end{aligned}$$

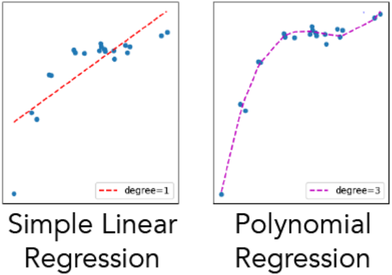


θ^*
θ_0
θ_1
θ_1^*
θ_2^*

多项式拟合

当自变量和因变量之间的关系是非线性时，我们使用多项式回归。

$$\hat{y} = \theta_0 + \theta_1 x_1 + \theta_2 x_1^2 + \cdots + \theta_n x_1^n$$



注意：多项式特征的数值范围可能相差很大，需要对数据进行标准化或归一化。

选择合适的阶数

- 选择过低的阶数：可能欠拟合（Underfitting），模型无法捕捉数据的非线性趋势。
- 选择过高的阶数：可能过拟合（Overfitting），模型在训练集上表现很好，但泛化能力差。

解决方案：

- 正则化
- 交叉验证

岭回归

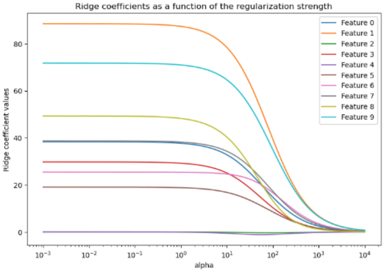
岭回归是在最小二乘法回归的损失函数中增加了 L2 正则化项，其优化目标如下：

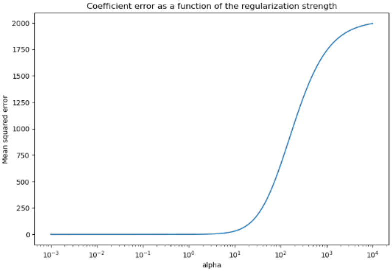
$$\min_{w} \|Xw - y\|_2^2 + \alpha\|w\|_2^2$$

其中，α 是正则化超参数，控制 L2 惩罚项权重

- 避免模型过拟合
- 适用于多重共线性问题
- 在 MATLAB 通过 ridge 函数实现岭回归
- 并不能选择特征

正则化强度对回归系数的影响





Lasso 回归

Lasso（least absolute shrinkage and selection operator）回归的损失函数在最小二乘回归的损失函数的基础上增加了 L1 正则化项：

$$\min_{w} \|Xw - y\|_2^2 + \alpha\|w\|_1$$

- 参数收缩与特征选择：通过 L1 正则化项，Lasso 回归可以将某些回归系数精确地压缩到 0，从而实现特征选择的目的。这使得模型更为简洁，减少了模型的复杂度。
- 防止过拟合：在拟合过于复杂的模型时，Lasso 回归通过正则化项对系数进行惩罚，有助于防止过拟合现象。
- 适用于高维数据：对于特征数多于样本数的高维数据，Lasso 回归能够有效地进行参数估计和变量筛选。
- 使用 lasso 函数实现