

第三章(4) 正则化线性回归 方法

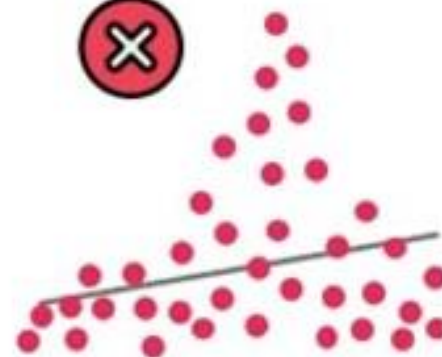
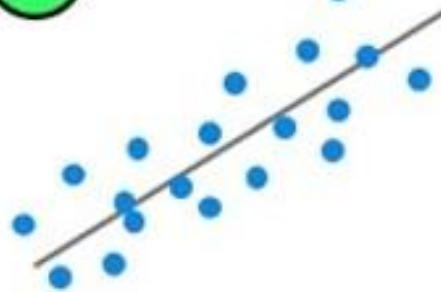
Lianghai Xiao

<https://github.com/styluck/mlb>

作业邮箱: alswhfx@126.com

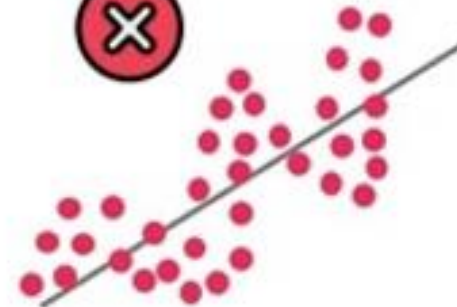
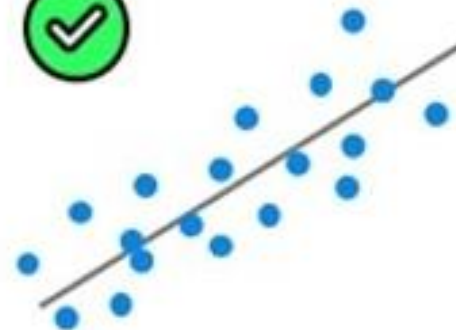
线性回归的假设

- **线性关系假设：**因变量 y 与自变量 x 之间存在线性关系。
- **独立性假设：**误差项 ϵ_i 之间相互独立。自变量 x 和误差项 ϵ 相互独立。
- **同方差性假设：**误差项 ϵ 的方差在不同的 x 值上是相同的，称为同方差性。
- **正态性假设：**误差项 ϵ 服从正态分布。
- **多重共线性：**解释变量之间没有相关关系。



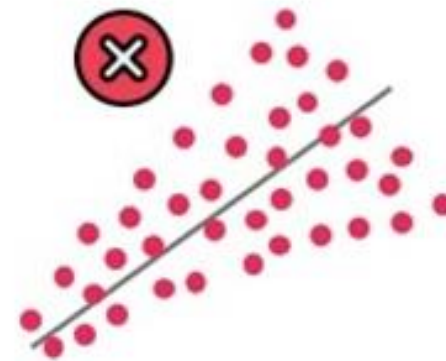
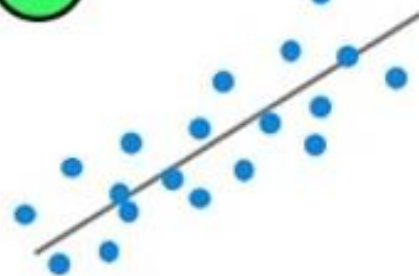
线性回归的假设

- **线性关系假设：**因变量 y 与自变量 x 之间存在线性关系。
- **独立性假设：**误差项 ϵ_i 之间相互独立。自变量 x 和误差项 ϵ 相互独立。
- **同方差性假设：**误差项 ϵ 的方差在不同的 x 值上是相同的，称为同方差性。
- **正态性假设：**误差项 ϵ 服从正态分布。
- **多重共线性：**解释变量之间没有相关关系。



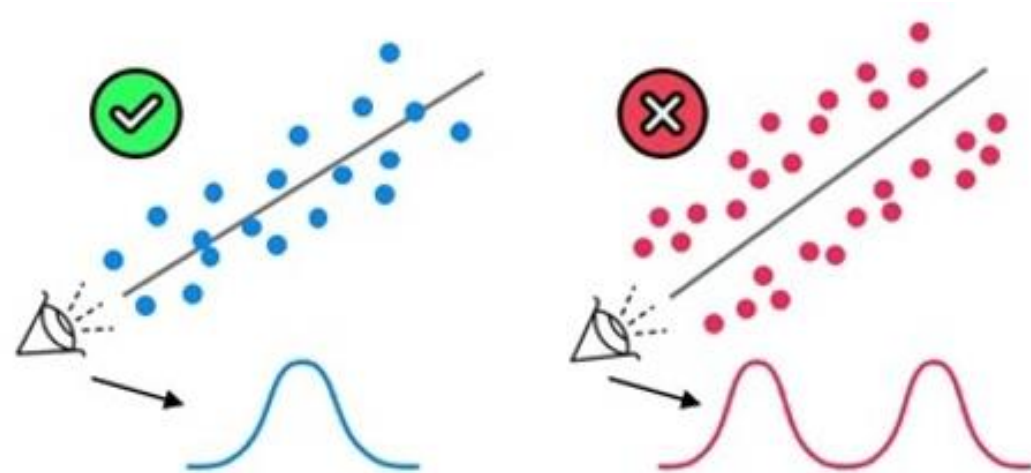
线性回归的假设

- **线性关系假设：**因变量 y 与自变量 x 之间存在线性关系。
- **独立性假设：**误差项 ϵ_i 之间相互独立。自变量 x 和误差项 ϵ 相互独立。
- **同方差性假设：**误差项 ϵ 的方差在不同的 x 值上是相同的，称为同方差性。
- **正态性假设：**误差项 ϵ 服从正态分布。
- **多重共线性：**解释变量之间没有相关关系。



线性回归的假设

- **线性关系假设：**因变量 y 与自变量 x 之间存在线性关系。
- **独立性假设：**误差项 ϵ_i 之间相互独立。自变量 x 和误差项 ϵ 相互独立。
- **同方差性假设：**误差项 ϵ 的方差在不同的 x 值上是相同的，称为同方差性。
- **正态性假设：**误差项 ϵ 服从正态分布。
- **多重共线性：**解释变量之间没有相关关系。



线性回归的假设

- **线性关系假设：**因变量 y 与自变量 x 之间存在线性关系。
- **独立性假设：**误差项 ϵ_i 之间相互独立。自变量 x 和误差项 ϵ 相互独立。
- **同方差性假设：**误差项 ϵ 的方差在不同的 x 值上是相同的，称为同方差性。
- **正态性假设：**误差项 ϵ 服从正态分布。
- **多重共线性：**解释变量之间没有相关关系。



$$X_1 \neq X_2$$



$$X_1 \sim X_2$$

多重共线性

- 多重共线性是指在多元线性回归模型中，自变量之间存在较强的线性相关关系。元线性回归模型为

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \epsilon$$

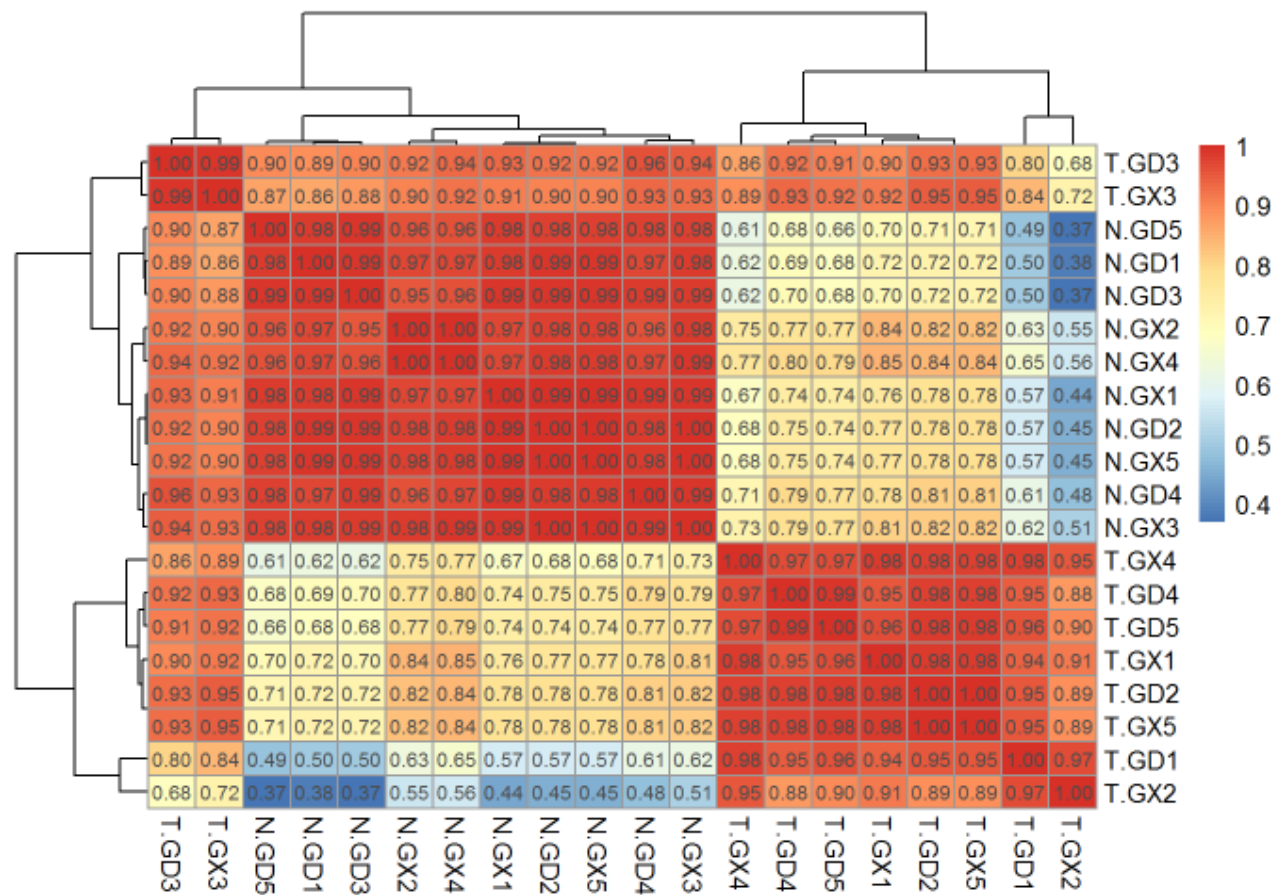
- 如果存在一组不全为零的数 c_1, c_2, \cdots, c_p ，使得

$$c_1 x_1 + c_2 x_2 + \cdots + c_p x_p \approx 0$$

- 那么就说自变量之间存在多重共线性。

多重共线性检验

- **相关系数矩阵法：** 计算自变量之间的简单相关系数矩阵。
- 如果某些自变量之间的相关系数绝对值较高（通常大于0.7），则可能存在多重共线性。



多重共线性检验

- **方差膨胀因子（VIF）法：**对于每个自变量计算其方差膨胀因子

$$VIF_i = \frac{1}{1 - R_i^2}$$

- 这里的 R_i^2 是自变量 x_i 对其余自变量进行回归得到的决定系数。
- VIF 值反映了由于多重共线性的存在，使得估计回归系数的方差比不存在共线性时增大了多少倍。例如，VIF=4意味着由于多重共线性，该自变量回归系数的方差是无共线性情况下方差的4倍。
- 一般来说，如果值大于10，就表示存在严重的多重共线性；在一些较为严格的情况下，当大于5时，也被认为存在值得关注的共线性问题。

逐步回归

- 逐步回归 (Stepwise Regression) 是一种自动从众多自变量中选择重要变量来构建回归模型的方法。
- 因变量 y 通常受到许多因素的影响，回归方程却不一定是最好的。回归方程引入无意义的变量，会使得误差方差的估计值变大，降低预测的精确性以及回归方程的稳定性。
- 逐步回归通过逐步添加或删除自变量的方式，尝试找出对因变量有显著预测能力的自变量集合。在逐步引入变量的同时，也会检查已引入变量的显著性，并且根据一定的标准来决定是否剔除变量。

逐步回归

- **向前选择（只进不出，Forward Selection）：**
 - 从没有任何自变量的模型开始。
 - 计算每个尚未进入模型的自变量与因变量之间的相关性或者进行 F 检验（或 t 检验）。
 - 选择与因变量相关性最强（或者通过 F 检验或 t 检验最显著）的自变量进入模型。
- **向后消除（只出不进，Backward Elimination）：**
 - 从包含所有候选自变量的模型开始。
 - 检查模型中已有的每个自变量的显著性（例如， p 值 >0.05 ），如果某个自变量的显著性水平超过了预先设定的阈值，则考虑将这个自变量从模型中剔除。
 - 每一步，移除最不显著的一个自变量。
- **逐步选择（Stepwise Selection）：** 每一步结合向前选择和向后消除。

逐步回归

- **逐步回归的优点：**

- **自动化：** 逐步回归提供了一种自动化的变量选择方法。
- **模型简化：** 可以帮助识别和构建一个更简洁且有预测能力的模型。

- **逐步回归的缺点：**

- **数据挖掘：** 可能会因为过度拟合而捕捉到数据中的随机噪声。
- **稳定性：** 逐步回归的结果可能对数据集非常敏感，不同的样本可能导致不同的模型。
- **多重比较问题：** 由于进行多次统计检验，可能会增加第一类错误（假阳性）的风险。

逐步回归的局限性

- 逐步回归依赖于预先设定的统计显著性标准（如p值）来选择变量，这种机械的选择过程可能会错过一些虽然单个变量不显著，但多个变量组合起来对因变量有重要影响的情况。
- 例如，两个自变量 x_1 和 x_2 单独对因变量的影响不显著，但它们的交互作用 x_1x_2 可能对 y 有很强的影响，逐步回归可能会将这两个变量都剔除。

多项式回归

- 多项式回归：建立自变量 x 和因变量 y 之间的非线性关系模型。
- 简单的二次多项式回归模型：

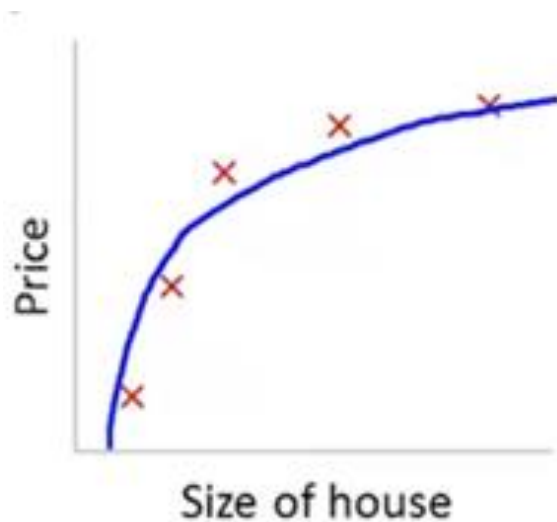
$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \epsilon$$

- 多元多项式回归模型的一般形式是

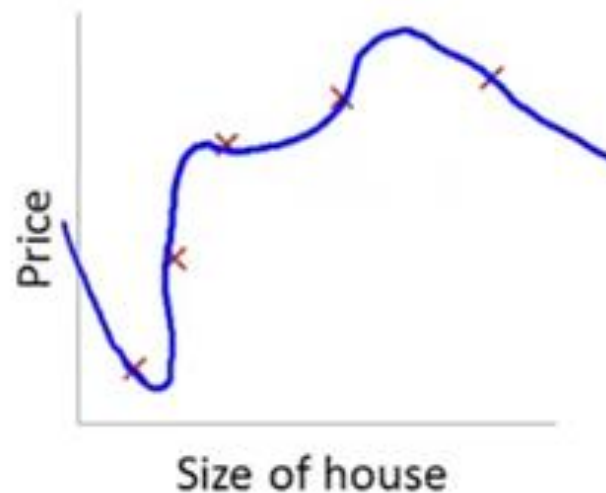
$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_n X_n + \beta_{n+1} X_1^2 + \beta_{n+2} X_1 X_2 + \dots + \beta_{2n} X_n^2 + \epsilon$$

- 通过包括高次项和交互项，多元多项式回归能够捕捉自变量与因变量之间的非线性关系。

多项式回归遇到的问题

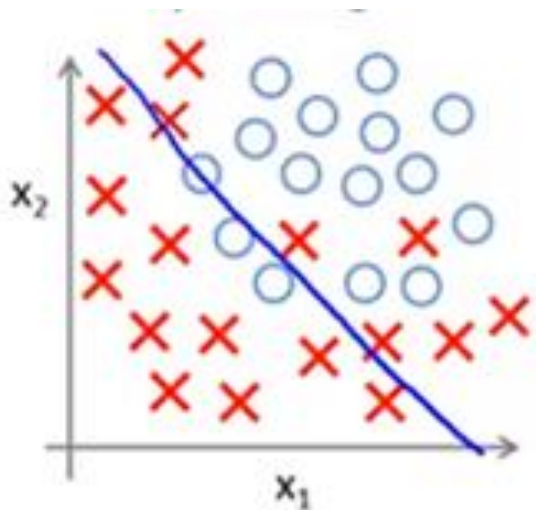


$$\theta_0 + \theta_1 x + \theta_2 x^2$$



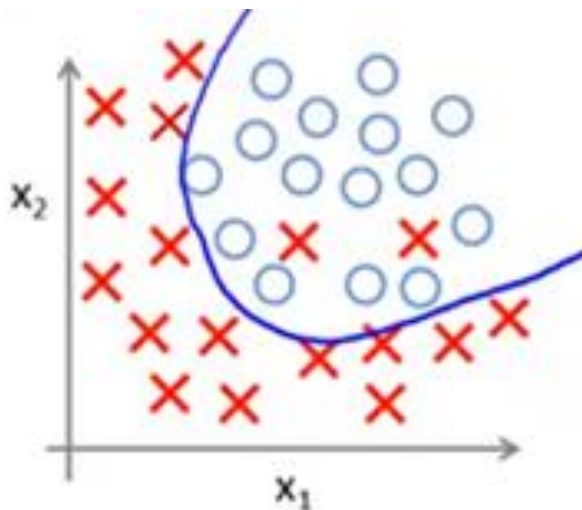
$$\theta_0 + \theta_1 x + \theta_2 x^2 + \theta_3 x^3 + \theta_4 x^4$$

一般线性回归遇到的问题

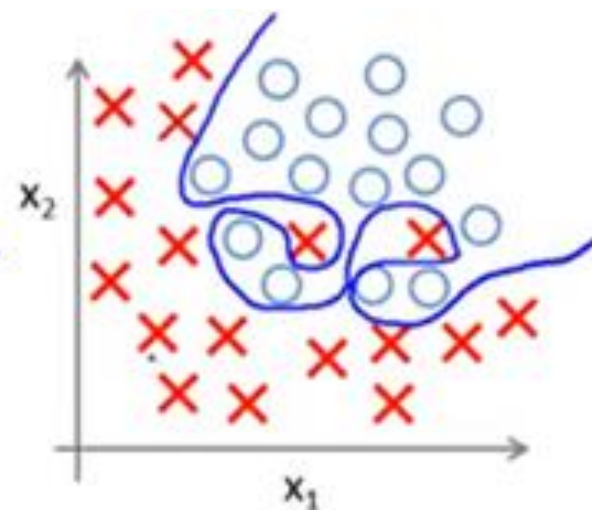


$$h_{\theta}(x) = g(\theta_0 + \theta_1 x_1 + \theta_2 x_2)$$

(g = sigmoid function)



$$g(\theta_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_1^2 + \theta_4 x_2^2 + \theta_5 x_1 x_2)$$



$$g(\theta_0 + \theta_1 x_1 + \theta_2 x_1^2 + \theta_3 x_1^2 x_2 + \theta_4 x_1^2 x_2^2 + \theta_5 x_1^2 x_2^3 + \theta_6 x_1^3 x_2 + \dots)$$

一般线性回归遇到的问题

- 预测精度：样本量 n 和自变量的数量 p
- $n \gg p$ 时，最小二乘回归会有较小的方差
- $n \approx p$ 时，最小二乘容易产生过拟合
- $n < p$ 时，最小二乘回归得不到有意义的结果

正则化线性回归方法

- Ridge 回归（岭回归）
- Lasso 回归（套索回归）
- 弹性网络（Elastic Net）回归

向量的范数

- 向量范数可以看作是向量长度概念在更一般情况下的推广。
- L1 范数（曼哈顿距离）： 向量各个元素绝对值之和。

$$\|\mathbf{x}\|_1 = \sum_{i=1}^n |x_i|$$

- L2 范数（欧几里得距离）：

$$\|\mathbf{x}\|_2 = \sqrt{\sum_{i=1}^n x_i^2}$$

Ridge 回归（岭回归）

- Ridge 回归在损失函数中添加了 L2 正则化项。假设线性回归的目标是最小化残差平方和

$$\min_{\beta} ((y - X\beta)^T (y - X\beta) + \lambda \beta^T \beta)$$

- 这里 $\lambda > 0$ 是正则化参数。
- 岭回归模型有一个等价的优化问题，即考虑一个带有约束的线性回归问题，其中约束条件是将系数向量 β 限制在一个半径为 t 的超球内部，即下面这个模型是与原问题等价的：

$$\beta^* = \operatorname{argmin}_{\beta} \frac{1}{n} \|\mathbf{y} - \mathbf{X}\beta\|_2^2$$

$$\text{s.t. } \|\beta\|_2^2 \leq t$$

Ridge 回归（岭回归）

- 展开 $J(\beta)$ 可得：
- 对 $J(\beta)$ 关于 β 求导并令导数为0，可得：
- 求解 β ，得到岭回归系数的估计值为：

Ridge 回归（岭回归）

- 展开 $J(\beta)$ 可得：

$$\begin{aligned} J(\beta) &= y^T y - 2\beta^T X^T y + \beta^T X^T X \beta + \lambda \beta^T \beta \\ &= y^T y - 2\beta^T X^T y + (\beta^T X^T X + \lambda \beta^T) \beta \\ &= y^T y - 2\beta^T X^T y + (\lambda I + X^T X) \beta^T \beta \end{aligned}$$

- 对 $J(\beta)$ 关于 β 求导并令导数为，可得：

$$-2X^T y + 2(\lambda I + X^T X)\beta = 0$$

- 求解 β ，得到岭回归系数的估计值为：

$$\hat{\beta}_{ridge} = (\lambda I + X^T X)^{-1} X^T y$$

Ridge 回归（岭回归）

- Ridge 回归: $\hat{\beta}_{ridge} = (\lambda I + X^T X)^{-1} X^T y$
- 线性回归: $\hat{\beta} = (X^T X)^{-1} X^T y$
- 它与普通线性回归系数估计值的区别在于多了 λI 这一项。随着 λ 的增大，正则化的作用越强，系数会被更多地向0收缩，从而避免模型过拟合。
- **L2 范数正则化项的作用：**
 - 矩阵 $X^T X$ 非正定时，模型回归系数存在唯一解。
 - 使得回归系数的绝对值不会过大，它会将系数向量向原点收缩，从而防止模型过拟合。在处理多重共线性问题时也很有效，它可以使系数估计更加稳定。

Lasso回归

- Lasso 回归（Least Absolute Shrinkage and Selection Operator）：与普通线性回归试图最小化残差平方和（RSS）不同，Lasso 回归在 RSS 的基础上添加了一个 L1 正则化项

$$\min_{\beta} \left(\frac{1}{2n} (y - X\beta)^T (y - X\beta) + \lambda \sum_{j=1}^p |\beta_j| \right)$$

- 这里 $\lambda > 0$ 是正则化参数。同理地，Lasso 也有一个带约束形式的等价优化模型：

$$\beta^* = \operatorname{argmin}_{\beta} \frac{1}{n} \|\mathbf{y} - \mathbf{X}\beta\|_2^2$$

$$\text{s.t. } \|\beta\|_1 \leq t$$

Lasso回归

- 与Ridge回归不同，Lasso回归问题的难点在于L1正则项不可微，需要采用非光滑优化算法进行求解。
- Lasso回归问题的最优解是一个稀疏解，因为L1范数在原点处不可导，优化过程中会倾向于产生稀疏解，系数向量中有较多的0元素。
- 因此。Lasso回归会促使一些系数 β_i 变为0，这有助于筛选出对模型真正有贡献的特征，从而实现自动的特征选择。

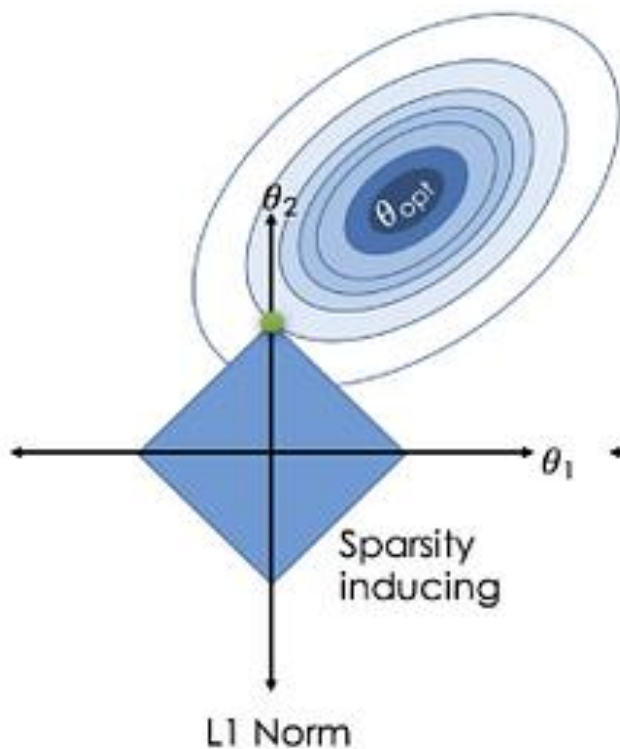
弹性网络 (Elastic Net) 回归

- Ridge 回归几乎没有特征选择能力。因为 L2 正则化只是将系数向原点收缩，不会使系数变为0。所有的特征在模型中都会保留一定的权重，即使某些特征可能对因变量的贡献相对较小。
- Lasso 回归在处理多重共线性方面相对较弱。当特征之间存在较强的线性关系时，Lasso 回归可能会在多个相关特征中随机选择一个或几个，而不是综合考虑这些相关特征，因此可能会出现过度筛选的情况。
- **弹性网络回归**：结合了 Lasso 回归 (L1 正则化) 和 Ridge 回归 (L2 正则化) 的特点：

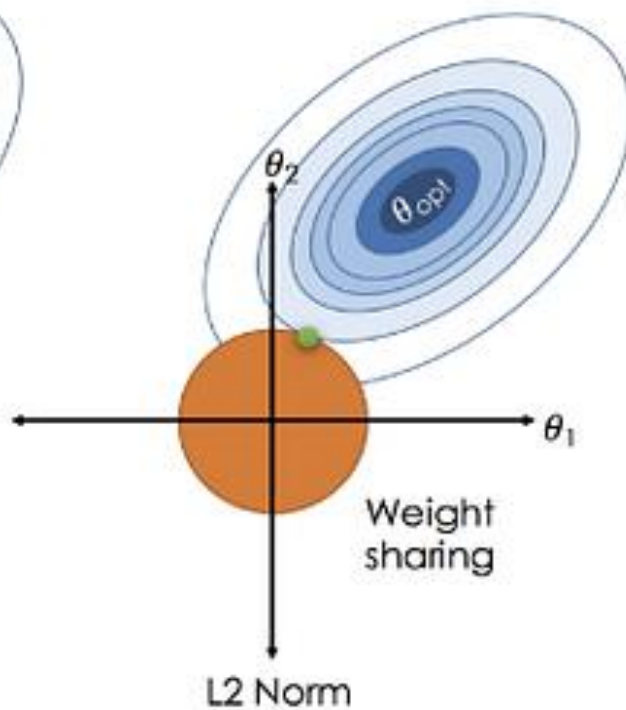
$$J(\beta) = (y - X\beta)^T (y - X\beta) + \lambda_1 \|\beta\|_1 + \lambda_2 \beta^T \beta$$

Ridge回归、Lasso回归和Elastic Net回归

Lasso回归



Ridge回归



Elastic Net回归

