



回归模型

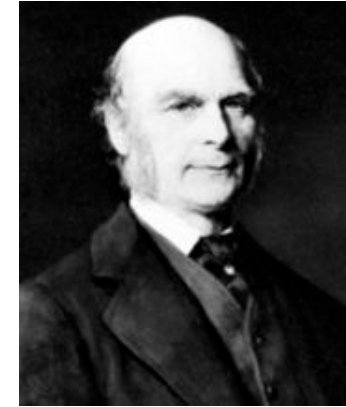
目录

第一部分	回归简介
第二部分	线性回归和正则化
第三部分	逻辑回归



回归的概念

- 回归(*Regression*)这一概念最早由英国生物统计学家高尔顿和他的学生皮尔逊在研究父母亲和子女的身高遗传特性时提出。
- “子女身高趋向于高于父母的身高的平均值，但一般不会超过父母的身高。”-- 《遗传的身高向平均数方向的回归》
- 如今，我们所讨论的“回归”和这种趋中效应已经没有任何瓜葛了，它只是指源于高尔顿工作的那样——**用一个或多个自变量来预测因变量的数学方法。**

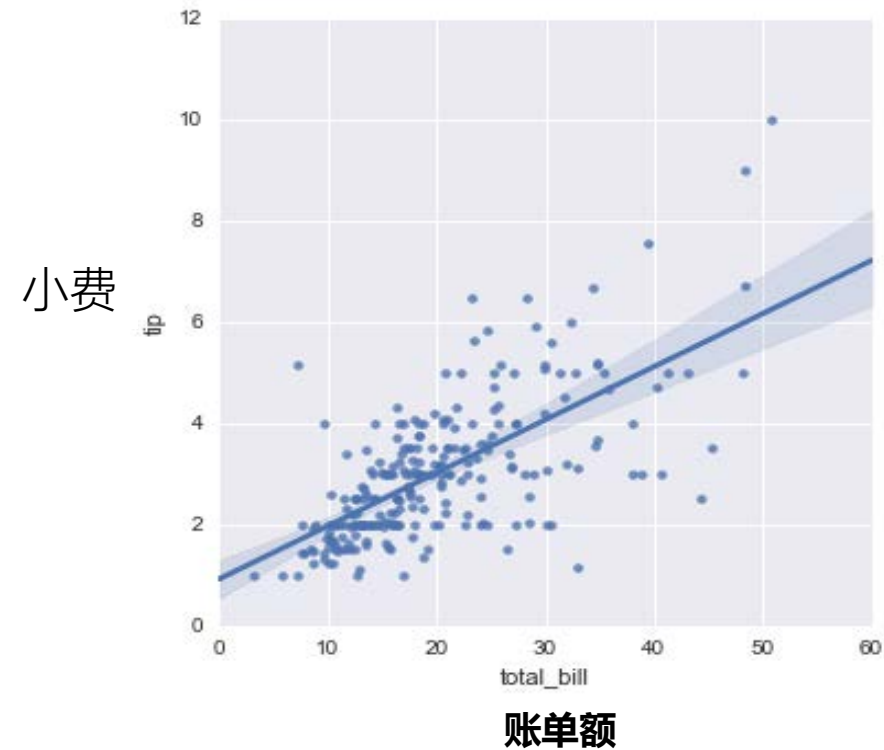
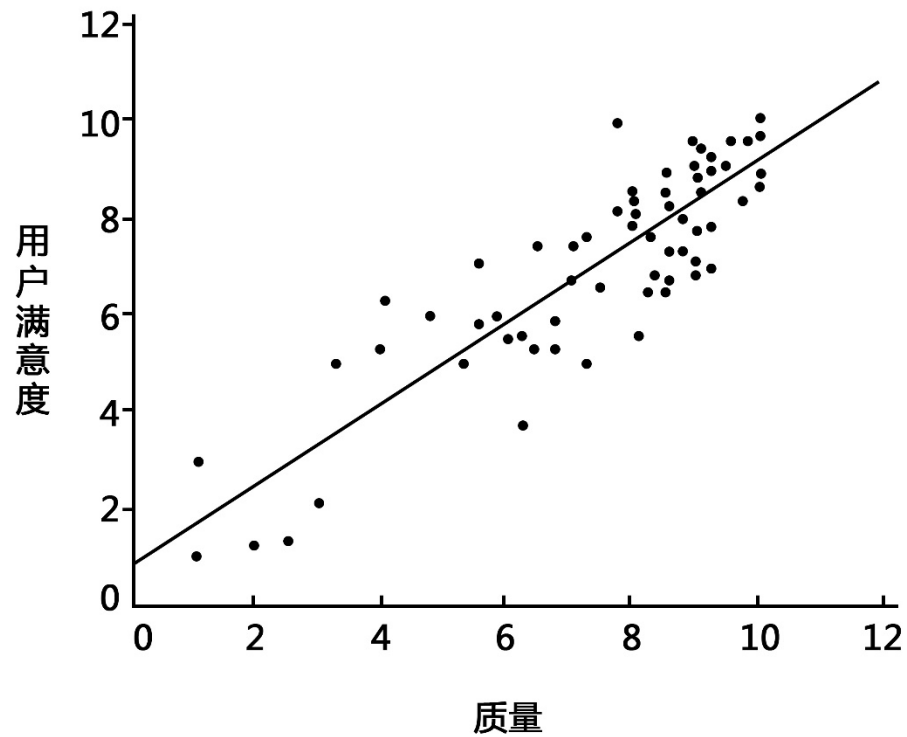


高尔顿



皮尔逊

- 在回归模型中，需要关注或预测的变量叫做因变量（响应变量或结果变量）
- 用来解释因变量变化的变量叫做自变量（解释变量或预测变量）



目录

第一部分  回归简介

第二部分 线性回归和正则化

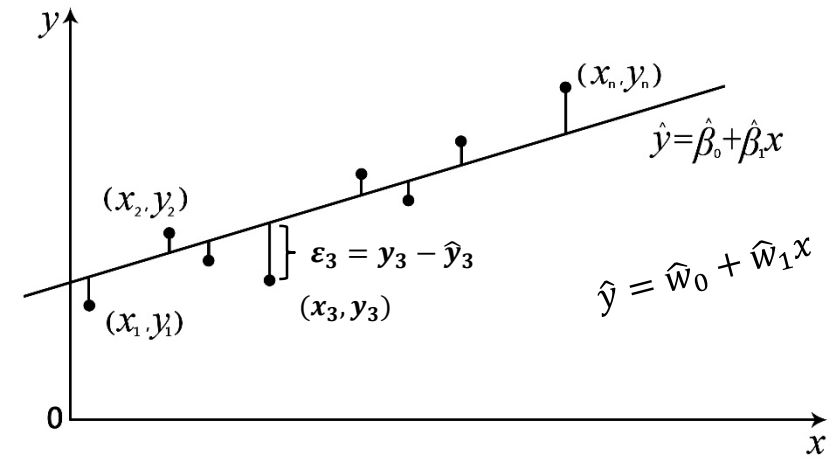
第三部分  逻辑回归

线性模型

一元线性回归模型

- $y = w_0 + w_1x + \varepsilon$, 其中 w_0, w_1 为回归系数, ε 为随机误差项 (Noise), 假设 $\varepsilon \sim N(0, \sigma^2)$, 则随机变量 $\mathbf{y} \sim N(\mathbf{w}_0 + \mathbf{w}_1\mathbf{x}, \sigma^2)$ 。
- 给定样本集合, 即 X, Y 的观测值是 $(x_1, y_1), \dots, (x_n, y_n)$, 我们的目标是找到一条直线 $y = w_0 + w_1x$ 使得所有样本点尽可能落在它的附近
- 换句话说就是在某种意义上极小化残差 ε , 即

$$(\hat{w}_0, \hat{w}_1) = \arg \min_{(w_0, w_1)} \sum_{i=1}^n (y_i - w_0 - w_1x_i)^2$$



多元线性回归模型

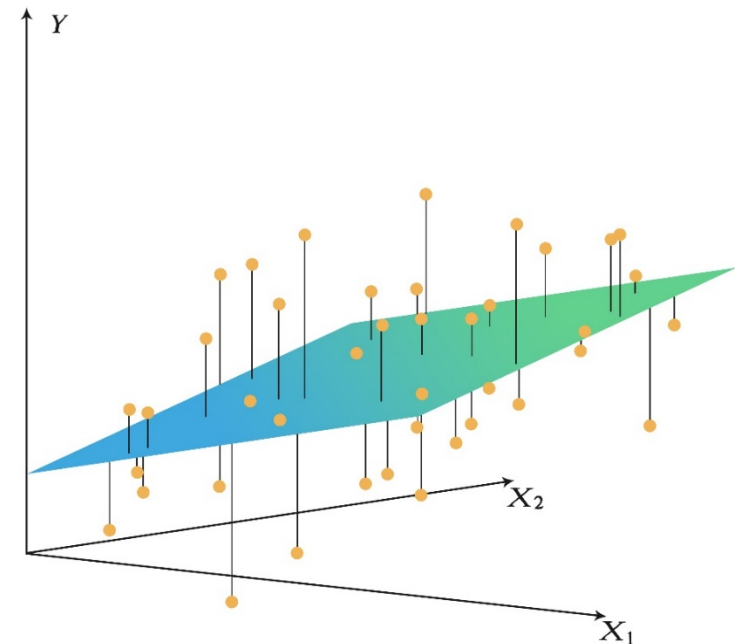
$$y = \mathbf{w}^T \mathbf{x} + w_0 + \varepsilon$$

其中, $\mathbf{x} = (x_1, x_2, \dots, x_d)^T$ 为自变量, $\mathbf{w} = (w_1; w_2; \dots; w_d)$ 为自变量的回归系数。

- 考虑到 n 个观察值的更一般的情况, $\mathbf{y} = (y_1, y_2, \dots, y_n)^T$ 为因变量, 将 w_0 并入 \mathbf{w} 中, $\mathbf{w} = (w_1; w_2; \dots; w_d; w_0)$, \mathbf{X} 为 $n \times (d+1)$ 大小的矩阵, ε 为残差, 假设 ε_i 是独立同分布的随机变量, 且服从均值为 0, 方差为 σ^2 的正态分布, 有

$$\mathbf{y} = \mathbf{X}\mathbf{w} + \varepsilon$$

- 其中, $\mathbf{X} = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1d} & 1 \\ x_{21} & x_{22} & \cdots & x_{2d} & 1 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nd} & 1 \end{pmatrix}$



参数估计——最小二乘估计

- 数学原理——误差平方和最小
- 按照这个法则，最好地拟合于各数据点的最佳曲线应使各数据点与曲线偏差的平方和为最小。
- 多元线性回归模型为： $\mathbf{y} = \mathbf{X}\mathbf{w} + \varepsilon$

$\mathbf{y} = (y_1 \dots, y_n)^T$ 为 n 维列向量； $\mathbf{X} = (x_{ij})_{n \times (d+1)}$ 为 $n \times (d+1)$ 矩阵， $\mathbf{w} = (w_1; w_2; \dots; w_d; w_0)$ 为 $(d+1)$ 维列向量

- 最小化目标函数为： $E(\mathbf{w}) = (\mathbf{y} - \mathbf{X}\mathbf{w})^T(\mathbf{y} - \mathbf{X}\mathbf{w})$ ，(\mathbf{y} 为真实值)
- 当矩阵 $\mathbf{X}^T\mathbf{X}$ 满秩时，根据二元函数求极值法，对 \mathbf{w} 求偏导数：

$$\text{令 } \frac{\partial E(\mathbf{w})}{\partial \mathbf{w}} = -2\mathbf{X}^T(\mathbf{y} - \mathbf{X}\mathbf{w}) = 0, \text{ 可得: } \hat{\mathbf{w}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$$

参数估计——极大似然估计

- **数学思想**：在已经得到试验结果的情况下，寻找使这个结果出现可能性最大的那个 $\hat{\mathbf{w}}$ 作为真实 \mathbf{w} 的估计
- 多元线性回归模型为： $\mathbf{y} = \mathbf{X}\mathbf{w} + \varepsilon$
- 最大化目标函数为（对数似然）： $l(\mathbf{w}; \mathbf{y}) = \log L(\mathbf{w}; \mathbf{y}) = \sum_{i=1}^n \log P(y_i | \mathbf{x}, \mathbf{w})$
其中，似然函数 $L(\mathbf{w}; \mathbf{y}) = P(\mathbf{y} | \mathbf{x}, \mathbf{w}) = \prod_{i=1}^n P(y_i | \mathbf{x}, \mathbf{w})$
- 当矩阵 $\mathbf{X}^T \mathbf{X}$ 非奇异时，求解该目标函数可得： $\hat{\mathbf{w}} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}$ （推导见下一页）

- 样本集合: $D = \{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \dots, (\mathbf{x}_n, y_n)\}$, $\mathbf{x}_i = (x_{i1}, x_{i2}, \dots, x_{id})^T$, $\mathbf{y} = (y_1, y_2, \dots, y_n)^T$ 为 n 维列向量,

$\mathbf{X} = (x_{ij})_{n \times (d+1)}$ 为 $n \times (d+1)$ 矩阵

- 参数: $\mathbf{w} = (w_1; w_2; \dots; w_d; w_0)$ 为 $(d+1)$ 维列向量

假设 $P(y_i | \mathbf{x}, \mathbf{w})$ 服从高斯分布

$$P(y_i | \mathbf{x}, \mathbf{w}) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(y_i - \sum_{j=1}^{d+1} x_{ij}w_j)^2}{2\sigma^2}}$$

则对数似然函数为:

$$\begin{aligned} l(\mathbf{w}; \mathbf{y}) &= \log L(\mathbf{w}; \mathbf{y}) = \log \prod_{i=1}^n P(y_i | \mathbf{x}, \mathbf{w}) = \sum_{i=1}^n \log P(y_i | \mathbf{x}, \mathbf{w}) = \sum_{i=1}^n \log \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(y_i - \sum_{j=1}^{d+1} x_{ij}w_j)^2}{2\sigma^2}} \\ &= n \log \frac{1}{\sqrt{2\pi}\sigma} - \frac{1}{2\sigma^2} \sum_{i=1}^n \left(y_i - \sum_{j=1}^{d+1} x_{ij}w_j \right)^2 \end{aligned}$$

$$\Leftrightarrow \nabla_{\mathbf{w}} l(\mathbf{w}; \mathbf{y}) = -\frac{1}{\sigma^2} \mathbf{X}^T (\mathbf{y} - \mathbf{X}\mathbf{w})$$

令 $\nabla_{\mathbf{w}} l(\mathbf{w}; \mathbf{y}) = 0$, 则 $\hat{\mathbf{w}} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}$

- 即对于多元线性模型而言, 参数的最小二乘估计量和极大似然估计量是一致的

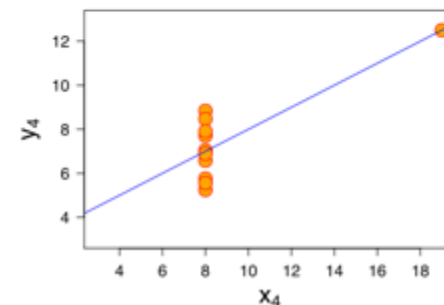
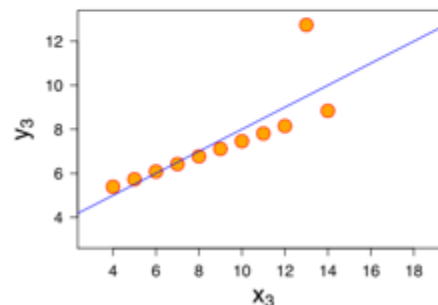
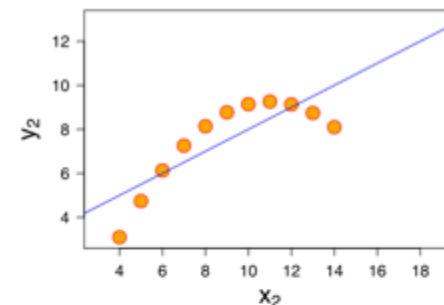
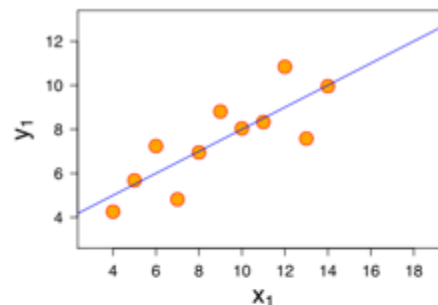
线性回归的问题

1. 实际数据可能不是线性的

解决方案：进行模型评估，使用 R^2 等指标

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST} = \frac{\sum_i (\hat{y}_i - \bar{y})^2}{\sum_i (y_i - \bar{y})^2} = 1 - \frac{\sum_i (\hat{y}_i - y_i)^2}{\sum_i (y_i - \bar{y})^2}$$

其中 y_i 为真实值， \bar{y} 为真实值的平均值， \hat{y}_i 为模型估计值



2. 多重共线性

最小二乘的参数估计为 $\hat{\mathbf{w}} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}$ ，如果变量之间存在较强的共线性，则

$|\mathbf{X}^T \mathbf{X}| \approx 0$ ，从而引起 $(\mathbf{X}^T \mathbf{X})^{-1}$ 对角线上的值很大，导致 参数估计值 $\hat{\mathbf{w}}$ 的方差 增大，对参数的估计变得不准确

解决方法： 岭回归、主成分回归、偏最小二乘回归

例：假设已知 x_1, x_2 与 y 的关系服从线性回归模型

$$y = 10 + 2x_1 + 3x_2 + \varepsilon$$



但参数估计量 方差 的表达式为

$$\text{Var}(\hat{\beta}_j) = \sigma^2 c_{jj}$$

c_{jj} 是矩阵 $(\mathbf{X}'\mathbf{X})^{-1}$ 中第 j 行第 j 列位置上的元素

- 给定 x_1, x_2 的10个值及 ε 的值，则有：

序号	1	2	3	4	5	6	7	8	9	10
x_1	1.1	1.4	1.7	1.7	1.8	1.8	1.9	2.0	2.3	2.4
x_2	1.1	1.5	1.8	1.7	1.9	1.8	1.8	2.1	2.4	2.5
ε_i	0.8	-0.5	0.4	-0.5	0.2	1.9	1.9	0.6	-1.5	-1.5
y_i	16.3	16.8	19.2	18.0	19.5	20.9	21.1	20.9	20.3	22.0

- 假设回归系数与误差项未知，则最小二乘法估计得到的回归系数为：

$$\hat{w}_0 = 11.292, \hat{w}_1 = 11.307, \hat{w}_2 = -6.591$$

- 而原模型的系数为：

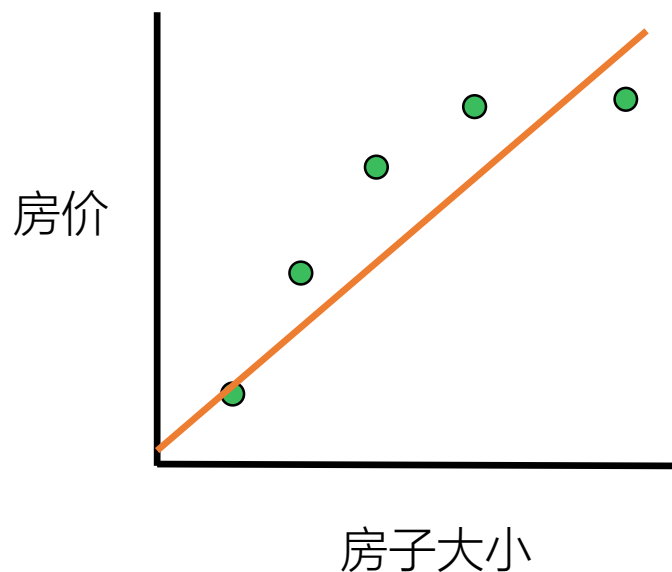
$$w_0 = 10, w_1 = 2, w_2 = 3$$

- 相差如此大，这是为什么呢？ x_1, x_2 样本相关系数为 $r_{12} = 0.986$ ！

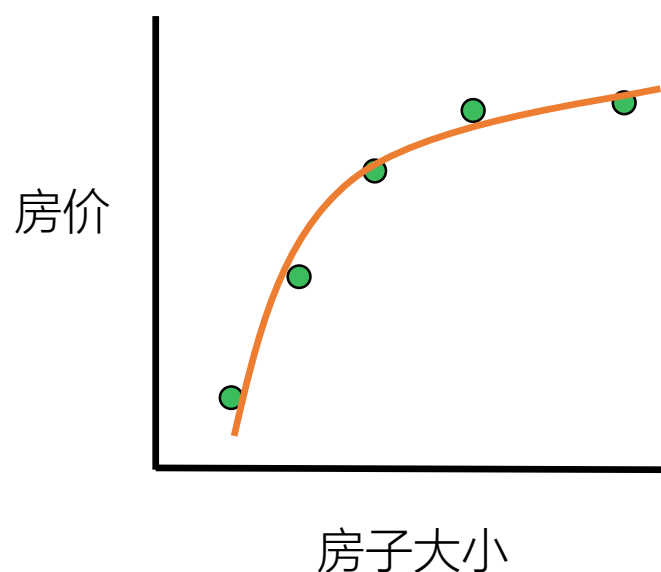
3. 过拟合问题：当模型的变量过多时候，线性回归可能会出现过拟合问题。

房子大小与房价的回归

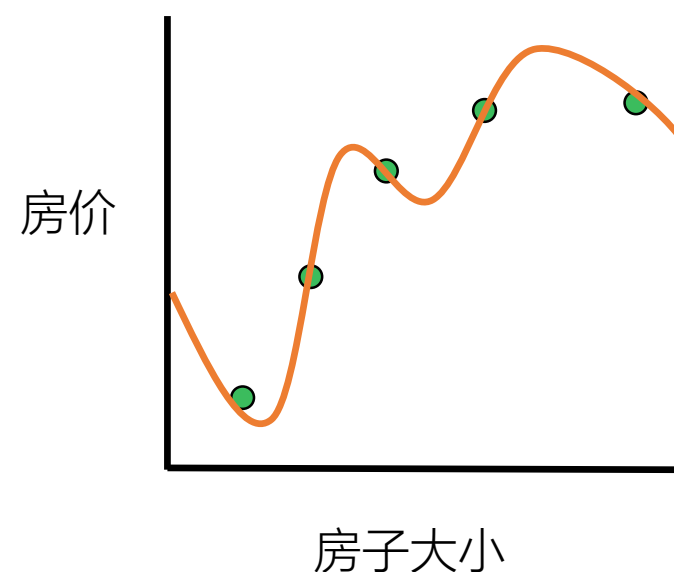
拟合函数 $w_0 + w_1x$:
欠拟合 高偏差, 训练误差大



拟合函数 $w_0 + w_1x + w_2x^2$:
拟合很好



拟合函数 $w_0 + w_1x + w_2x^2 + w_3x^3 + w_4x^4$:
过拟合 高方差, 泛化能力差

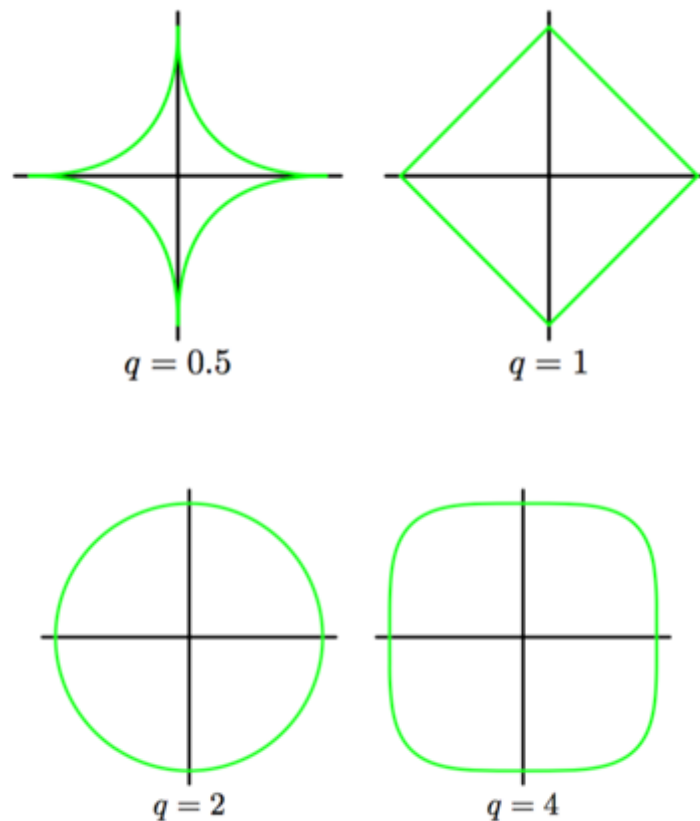


正则化

正则化可以解决线性回归的过拟合和多重共线性等问题,同时能够完成变量选择,使得模型解释性更强。

$$\sum_{i=1}^n \left(y_i - \sum_{j=1}^{d+1} x_{ij} w_j \right)^2 + \lambda ||\mathbf{w}||^q$$

- $q = 2$: 岭回归(Ridge)
- $q = 1$: LASSO



岭回归

- 岭回归 (Ridge Regression) 思路:

共线性会导致参数估计值变得非常大, 那么在最小二乘法的目标函数基础上加上一个对 \mathbf{w} 的惩罚函数, 最小化新的目标函数的时候也需要考虑到 \mathbf{w} 值的大小, \mathbf{w} 不能过大

惩罚项即正则化项, 岭回归采用含有正则化参数的 L_2 范数正则化: $\lambda ||\mathbf{w}||_2^2$

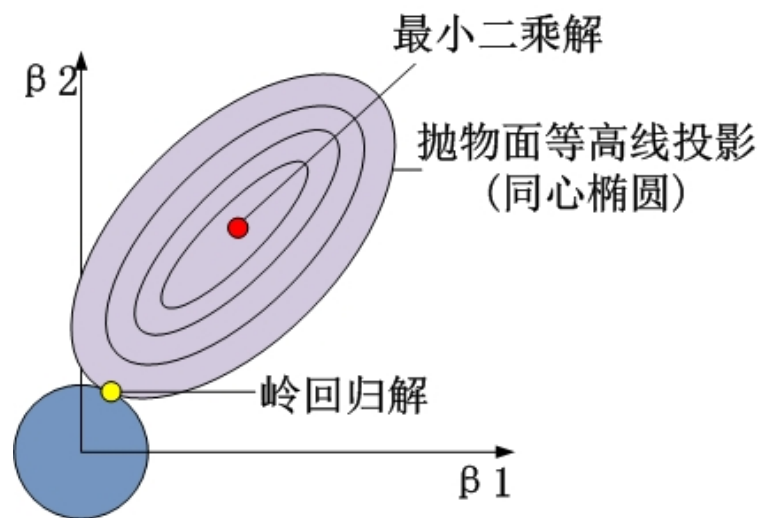
- 线性回归的优化目标函数:

$$\hat{\mathbf{w}} = \arg \min_{\mathbf{w}} \sum_{i=1}^n \left(y_i - \sum_{j=1}^{d+1} x_{ij} w_j \right)^2$$

- 岭回归的目标函数变为:

$$\hat{\mathbf{w}}^{ridge} = \arg \min_{\mathbf{w}} \left\{ \sum_{i=1}^n \left(y_i - \sum_{j=1}^{d+1} x_{ij} w_j \right)^2 + \lambda ||\mathbf{w}||_2^2 \right\}$$

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



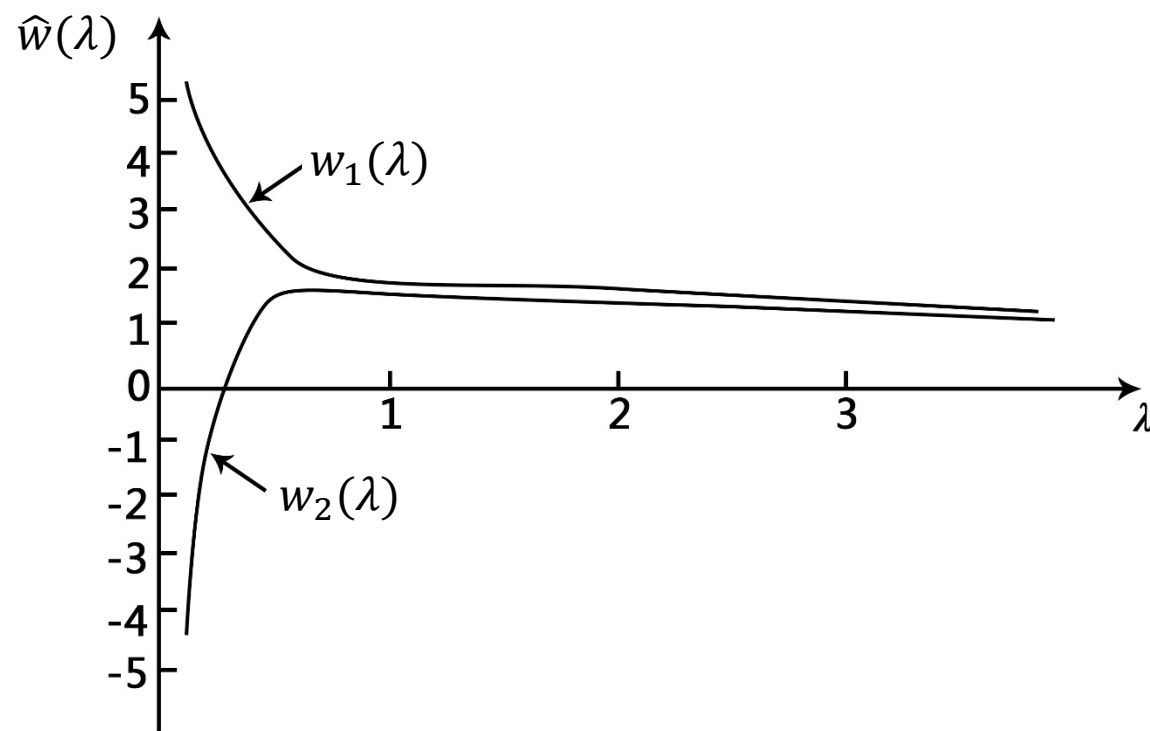
- 岭迹分析:当不断增大正则化参数 λ , 估计参数 $\hat{\mathbf{w}}^{ridge}(\lambda)$ (也称岭回归系数) 在坐标系上的变化曲线称为**岭迹**。
- 岭迹波动很大, 说明该变量参数有共线性

对上一个例子算得不同 λ 时的 $\hat{\mathbf{w}}^{ridge}(\lambda)$ 有:

λ	0	0.1	0.15	0.2	0.3	0.4	0.5	1.0	1.5	2.0	3.0
$\hat{w}_1^{ridge}(\lambda)$	11.31	3.48	2.99	2.71	2.39	2.20	2.06	1.66	1.43	1.27	1.03
$\hat{w}_2^{ridge}(\lambda)$	-6.59	0.63	1.02	1.21	1.39	1.46	1.49	1.41	1.28	1.17	0.98

岭迹分析

- 岭迹图
- 在 $\lambda \sim (0, 0.5)$ 的范围内波动较大, 故需要加入正则化项重新进行参数估计, 可选 $\lambda = 1$

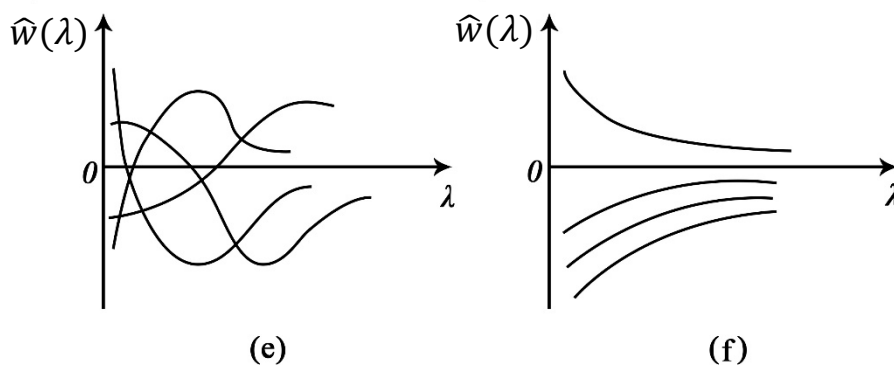
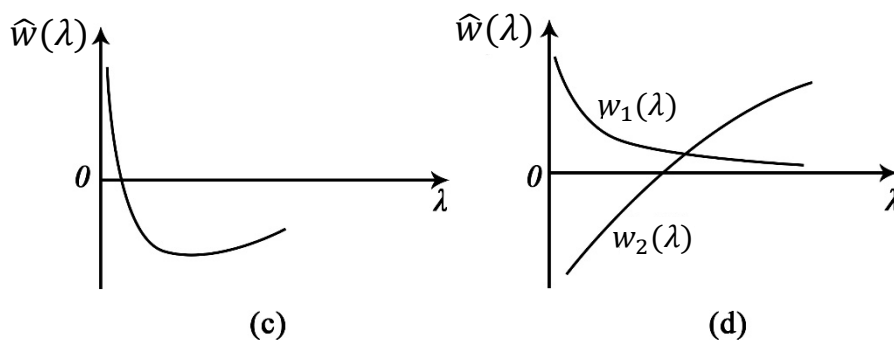
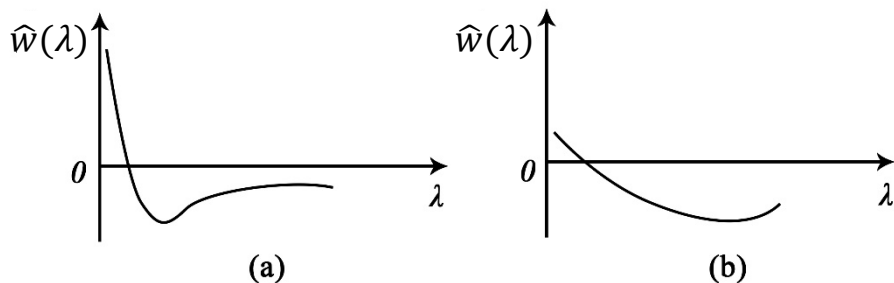


● 如何判断是否要剔除某参数

1. λ 很小时, w 很大, 稍微增大, w 迅速变小, 肯定有多重共线性。
2. 多重共线性的岭迹图一般呈喇叭口状。选喇叭附近的 k 值。
3. 岭迹图可以筛选变量, 有个变量的系数长期很接近于0, 可以剔除这些变量。

● 选择 k 或者 λ 使得

1. 各岭回归系数的岭基本稳定
2. 正负直接穿梭时, 不合乎实际意义
3. 残差平方和增大不太多。



岭回归缺点

- 建模时同时引入 p 个自变量，罚约束项可以收缩这些自变量的待估系数接近0，但并非恰好是0。这个缺点对模型精度影响不大，但给模型的解释造成了困难。
- 这个缺点可以由lasso来克服
 - 当 λ 充分大时可以把某些待估系数精确地收缩到0
- 岭回归虽然减少了模型的复杂度，并没有真正解决变量选择的问题



LASSO回归

- 斯坦福教授 Robert Tibshirani 于 1996 年发表了名为《Regression shrinkage and selection via the lasso》的论文，其中介绍了著名的 LASSO (Least Absolute Shrinkage and Selection Operator) 方法。该论文至今已经被引用了 20484 次，该论文在学术界和业界中都产生了深远的影响。

Regression shrinkage and selection via the lasso

[R Tibshirani](#) - [Journal of the Royal Statistical Society. Series B](#) (..., 1996 - JSTOR

We propose a new method for estimation in linear models. The lasso minimizes the residual sum of squares subject to the sum of the absolute value of the coefficients being less than a constant. Because of the nature of this constraint it tends to produce some coefficients that are exactly 0 and hence gives interpretable models. Our simulation studies suggest that the lasso enjoys some of the favourable properties of both subset selection and ridge ...

被引用次数: 20484 相关文章 所有 75 个版本 引用 保存

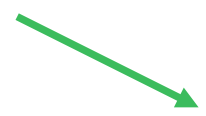
LASSO引言

- 构建线性回归模型时为减少因缺失重要特征而出现的模型偏差，通常会选择尽可能多的特征
- 但包含过多特征不仅增加模型的复杂度，而且还会导致模型产生过拟合现象
- 当数据集中包含非常多的特征时，我们应该如何提高模型的解释性和预测精度呢？
- 应用 LASSO 方法可以精简模型中的特征项，降低复杂度，提高模型的预测精度。



LASSO，中文名为“套索”

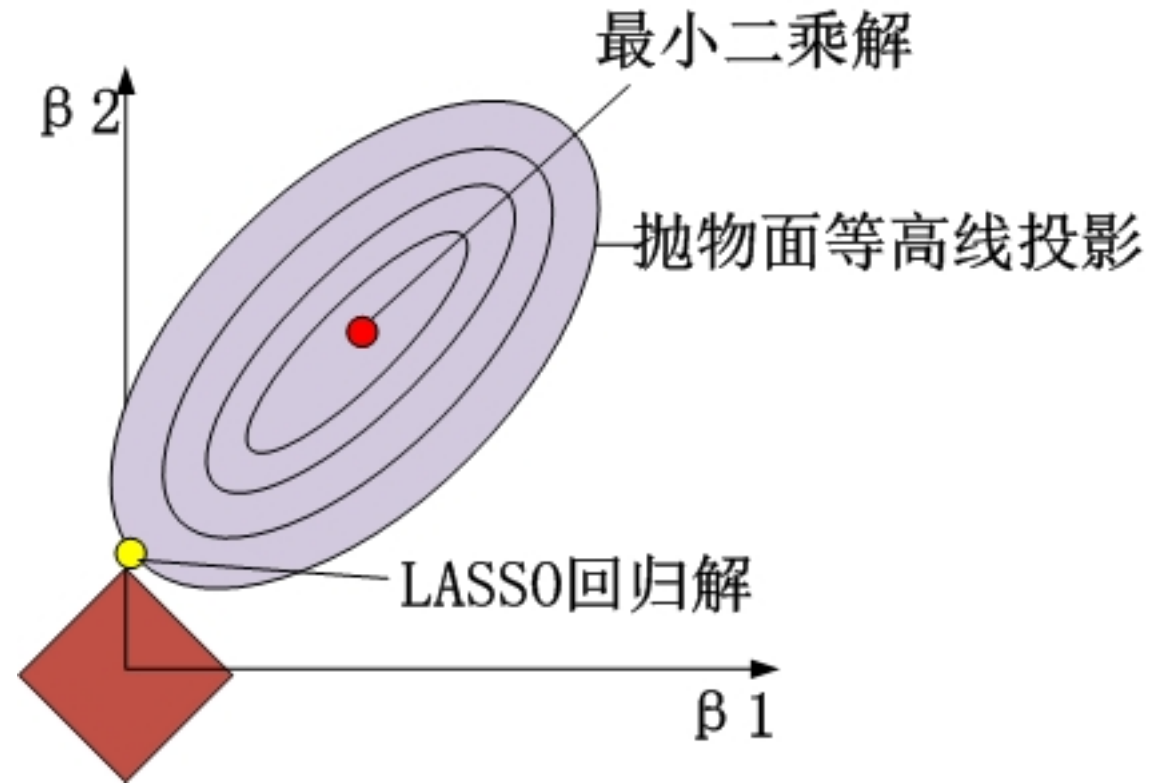
- LASSO 是一种系数压缩估计方法，它的基本思想是：在回归系数 w 的绝对值之和小于某个常数的条件下，最小化模型的残差平方和，从而能够得到一些严格等于0的回归系数，得到一个较为精简的模型。
- LASSO 线性回归模型的目标函数（残差平方和函数）：

$$\sum_{i=1}^n \left(y_i - \sum_{j=1}^{d+1} x_{ij} w_j \right)^2 + \lambda \sum_{j=1}^{d+1} |w_j|$$


L_1 惩罚函数

L_1 惩罚函数形式

- 如右图所示：图中蓝色区域表示约束区域，红色线为普通线性回归模型的残差平方和函数的等高线
- 通过添加 L_1 惩罚函数，LASSO 方法可以得到角点解，即稀疏的最优解 $\hat{\mathbf{w}}$ ，此时 $\hat{w}_1 = 0$ ，我们可以将对应的特征项从模型中剔除
- 要找到第一个落到限制区域上的等高线的那个位置的坐标。因为菱形带尖角，所以更有可能使得某个变量的系数为0。当回归变量增多时，lasso的尖角也会变得更多，从而增大更多系数变0的可能性。而光滑的高维球面的显然不可能有这样的概率。



回归系数求解

- 则LASSO 的目标函数为:

$$E(\mathbf{w}) = \frac{1}{2} \sum_{i=1}^n \left(y_i - \sum_{j=1}^{d+1} x_{ij} w_j \right)^2 + \sum_{j=1}^{d+1} \left\{ \frac{1}{2} (\hat{w}_j - w_j)^2 + \lambda |w_j| \right\}$$

其中 \hat{w} 是系数 w 的最小二乘估计

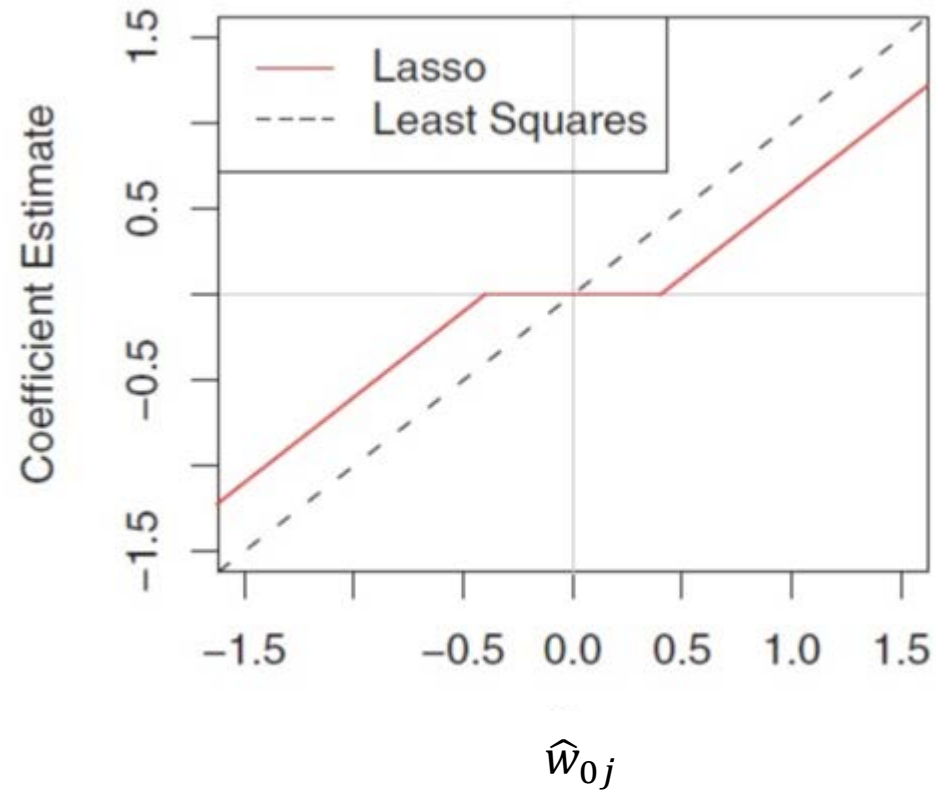
- 求得参数 w 的 LASSO 估计值为:

$$\hat{w}_j^{lasso} = \text{sign}(\hat{w}_j) (|\hat{w}_j| - \lambda)_+$$

其中, sign 函数表示符号函数, $(|\hat{w}_j| - \lambda)_+ = \begin{cases} |\hat{w}_j| - \lambda & \text{if } |\hat{w}_j| - \lambda > 0 \\ 0 & \text{if } |\hat{w}_j| - \lambda \leq 0 \end{cases}$

回归系数示意图

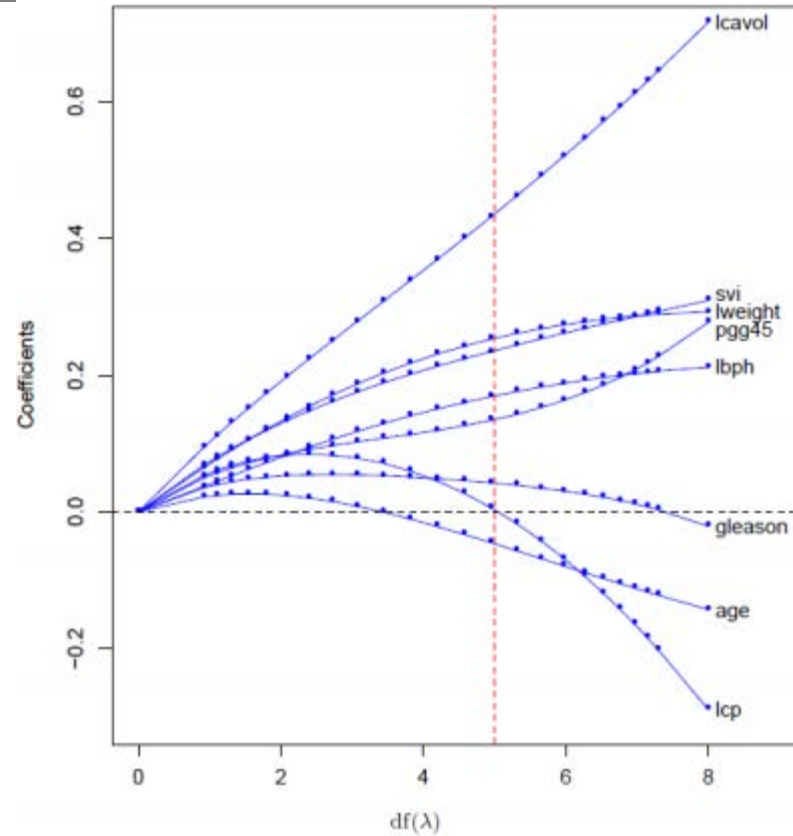
$$\hat{w}_{lasso,j} = \text{sign}(\hat{w}_j)(|\hat{w}_j| - \lambda)_+$$



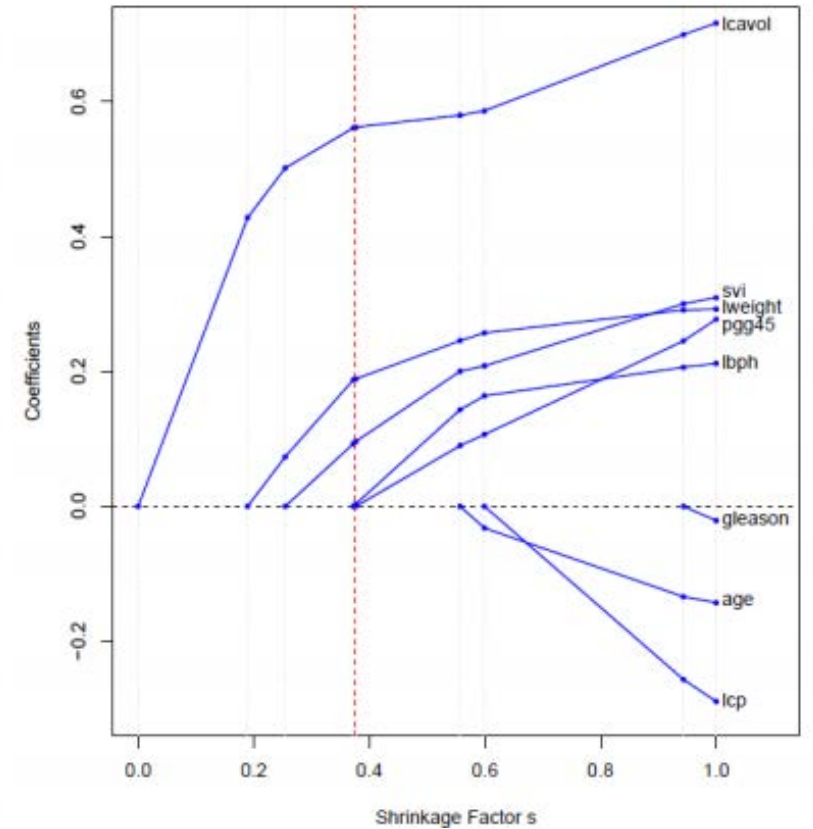
参数路径图

- 如右图所示，选定不同的参数值 λ 可以获得不同的参数估计值，当 λ 很大时， $t = \sum_j |w_j|$ 的取值趋于0，此时模型中所有特征的系数都被压缩为0。

岭回归岭迹图



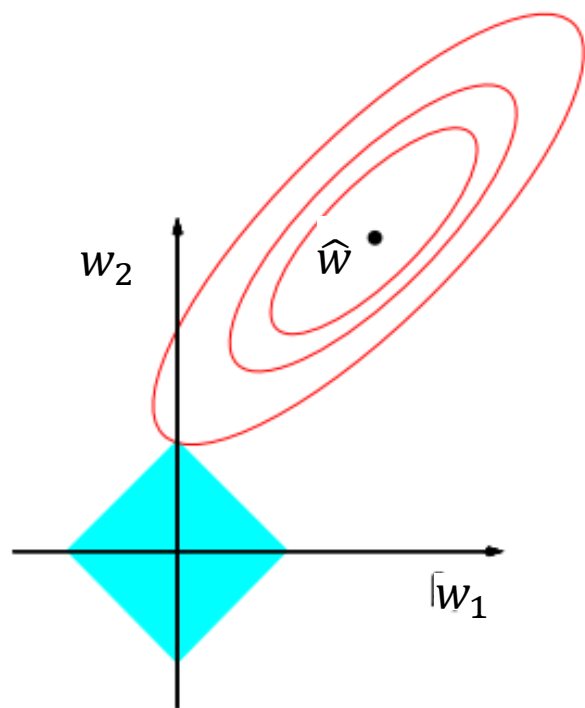
Lasso的岭迹图



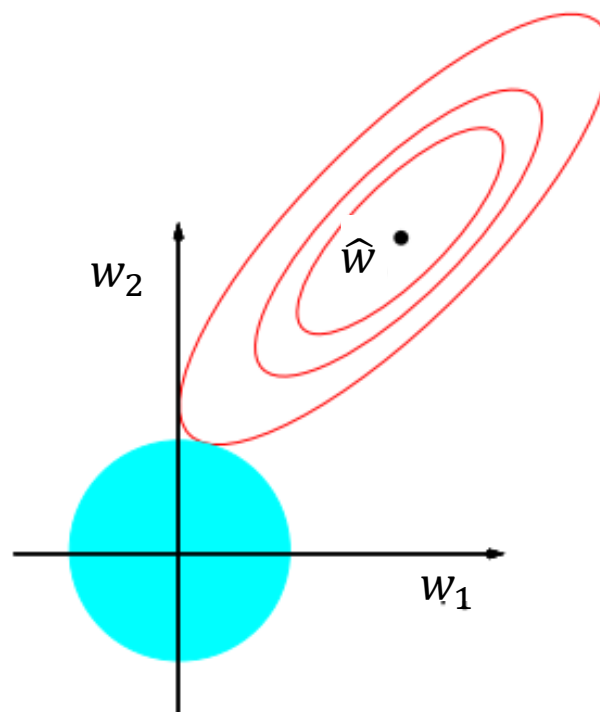
横轴越往左， λ 越大，系数（即纵轴）会越趋于0。但是岭回归没有系数真正为0，但lasso的不断有系数变为0.

- 岭回归与LASSO回归最优解思路

惩罚函数（蓝色）与目标函数（红色）



LASSO回归



岭回归

模型评估

- 评价回归模型效果的优劣是回归分析的重要内容之一。
- 常用的评价指标有：
 - 决定系数
 - 校正决定系数

决定系数

- 决定系数 (coefficient of determination) 表示回归平方和占总平方和的比例, 反映各自变量对因变量回归贡献的大小, 用 R^2 表示。

- 设 y_i 为真实值, 均值为 \bar{y} , 拟合值为 \hat{y}_i , 则有

总平方和 (SST, sum of squares for total) : $\sum_i (y_i - \bar{y})^2$

回归平方和 (SSR, sum of squares for regression) : $\sum_i (\hat{y}_i - \bar{y})^2$

残差平方和 (SSE, sum of squares for error) : $\sum_i (\hat{y}_i - y_i)^2$

且 $SST = SSR + SSE$

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST} = \frac{\sum_i (\hat{y}_i - \bar{y})^2}{\sum_i (y_i - \bar{y})^2} = 1 - \frac{\sum_i (\hat{y}_i - y_i)^2}{\sum_i (y_i - \bar{y})^2}$$

- R^2 无单位, 取值在 0~1 之间。值越大, 说明回归平方和在总平方和中所占的比重越大, 残差平方和所占比重越小, 回归效果越好。

目录

第一部分 回归简介

第二部分 线性回归和正则化

第三部分 逻辑回归

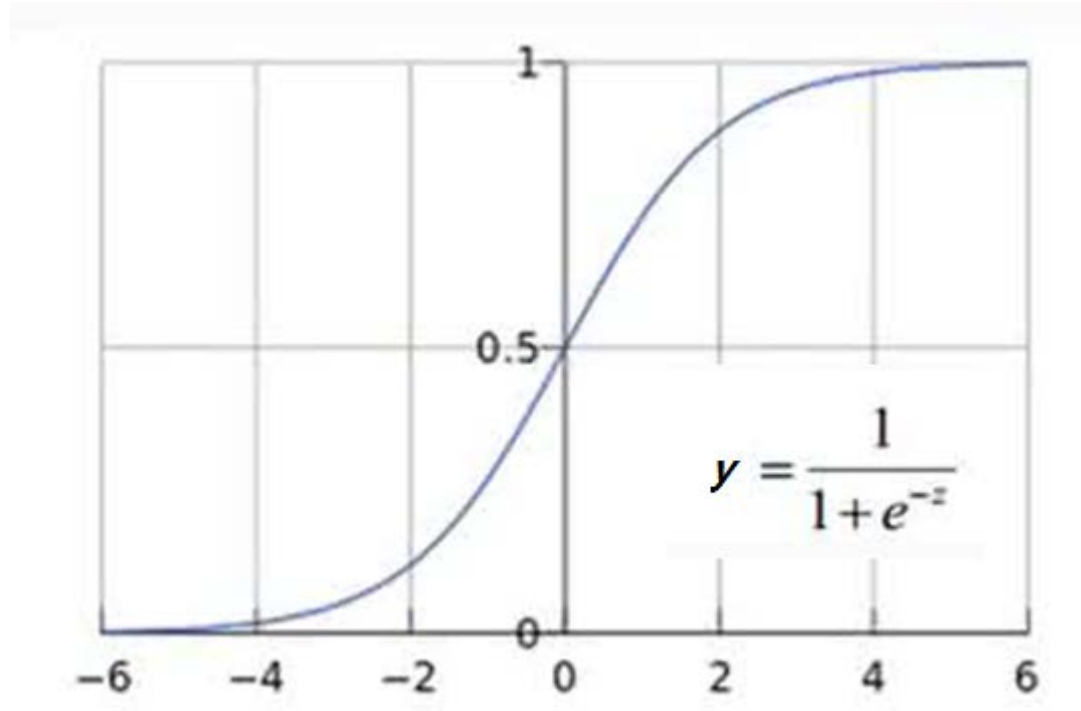


逻辑回归模型

- 逻辑回归，也称 Logistic Regression，主要区别于一般的线性回归模型。
- 一般的线性回归模型都是处理因变量是连续变量的问题，如果因变量是定性变量，一般线性回归模型就不再适用。
- 逻辑回归模型是用于处理因变量为分类变量的回归问题，最常见的就是因变量为二分类的情况，比如常见的信用评分模型，该模型用于评估个人的信用违约概率。
- 注：逻辑回归用于解决分类问题

SIGMOID函数

- 考虑二分类问题，输出标记为 $y \in \{0, 1\}$ ，而线性回归模型产生的预测值 $z = w^T x + b$ 是实数值，因此我们需要将 z 转换为0/1值。
- Sigmoid函数



- 该函数将 z 值转化为一个接近0或者1的 y 值，并且其输出值在 $z=0$ 附近变化的很陡。

逻辑回归模型

- 将z函数带入Sigmoid函数可得

$$y = \frac{1}{1 + e^{-(\mathbf{w}^T \mathbf{x} + b)}}$$

- 将y视作样本取正例的可能性，1-y作为取反例的可能性，两者的比值称为“**几率**”，反应了 \mathbf{x} 作为正例的相对可能性，取对数得到对数几率函数：

$$\ln \frac{y}{1-y} = \ln \frac{p(y=1|\mathbf{x})}{p(y=0|\mathbf{x})} = \mathbf{w}^T \mathbf{x} + b$$

得到对数几率模型，他无需事先假设数据分布。

- Sigmoid 函数很好地体现了概率 p 与解释变量之间的非线性关系。

参数估计

- 极大似然估计函数

$$\mathbf{w} = \arg \max_{\mathbf{w}} \sum_l \ln P(y^l | \mathbf{x}^l, \mathbf{w})$$

$$l(\mathbf{w}) = \sum_l y^l \ln P(y^l = 1 | \mathbf{x}^l, \mathbf{w}) + (1 - y^l) \ln P(y^l = 0 | \mathbf{x}^l, \mathbf{w})$$

- 梯度下降法

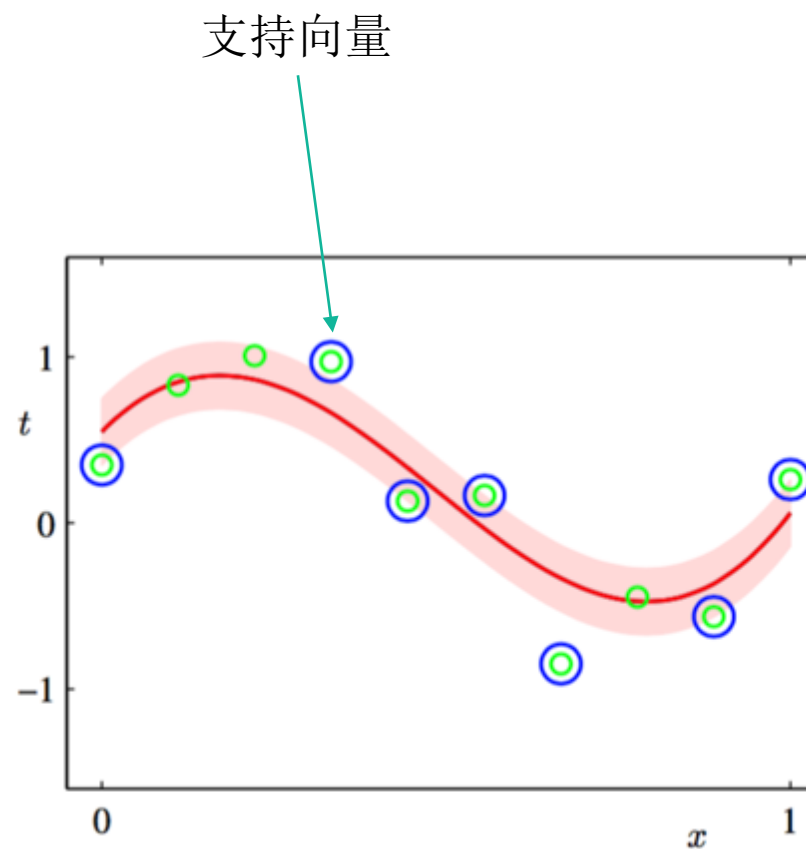
$$\frac{\partial}{\partial w_i} l(\mathbf{w}) = \sum_l x_i^l (y^l - \hat{P}(y^l = 1 | \mathbf{x}^l, \mathbf{w}))$$

$$w_i \leftarrow w_i + \eta \sum_l x_i^l (y^l - \hat{P}(y^l = 1 | \mathbf{x}^l, \mathbf{w}))$$



其他回归方法

- SVR: SVM for regression
- Regression Tree
- ...





「 THANK YOU ! 」