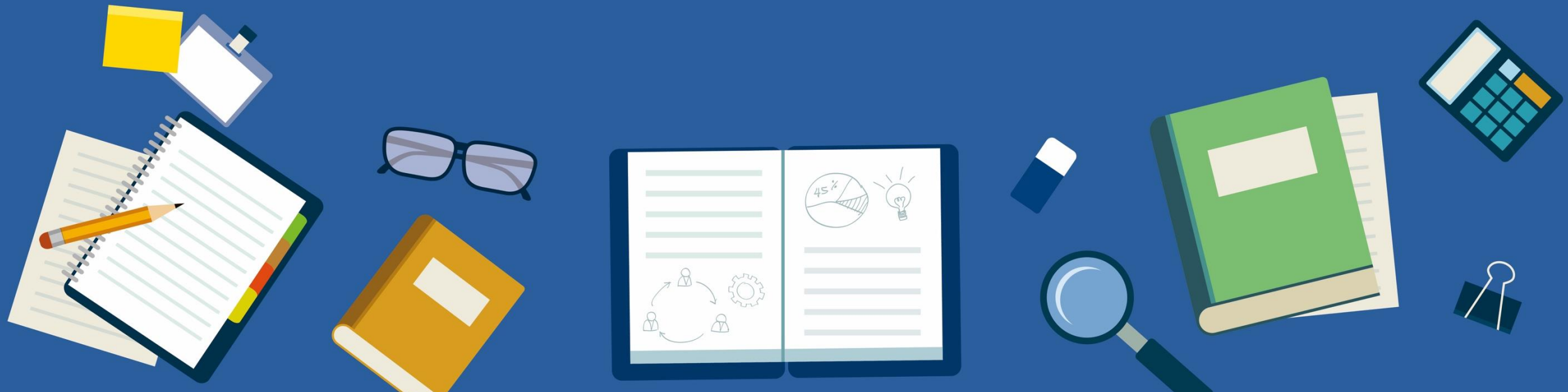
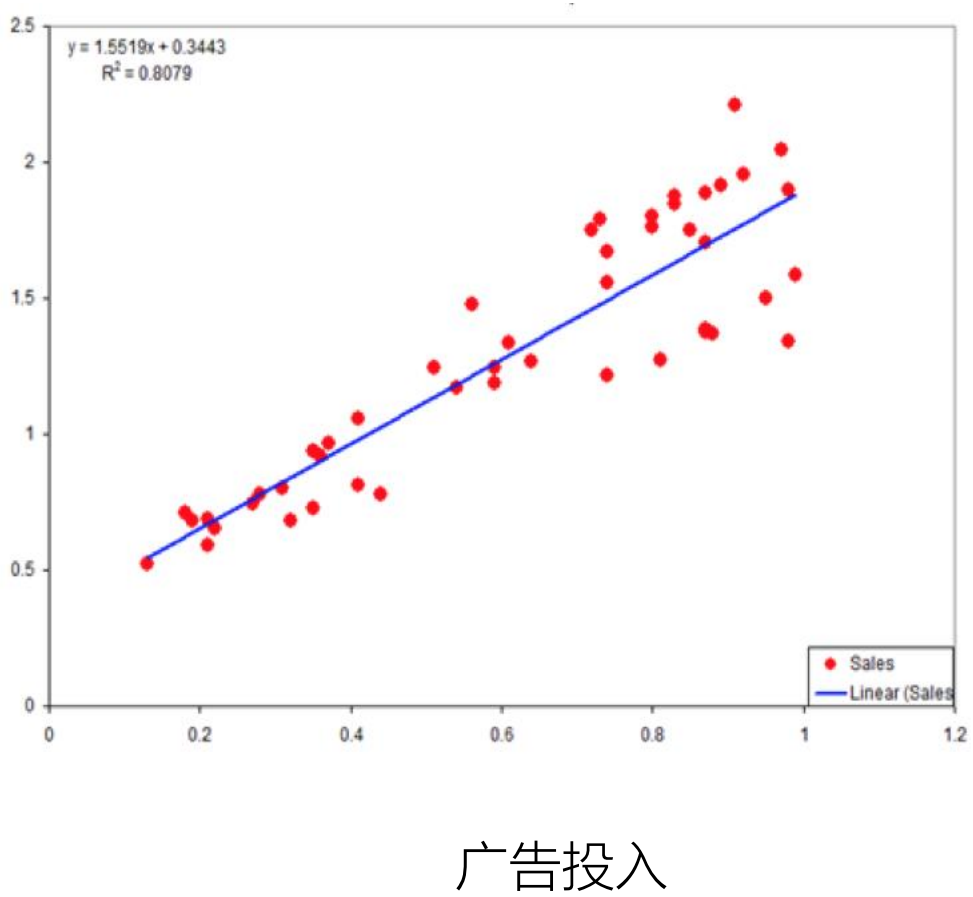
第三部分
回归模型



回归的基本概念

- 回归：目标变量为连续性的预测问题，e.g., 收入预测、销量预测、库存预测



在一个回归模型中：

- 需要预测的变量叫做因变量(目标变量 target variable)，记做 y (连续性的实数值)
- 选取的用来解释因变量变化的变量叫做自变量(解释变量、features)，记为d 维向量$\mathbf{x}$
- 学习目标：找到一个函数 f, 建立$\mathbf{x}$ 到y 的映射关系：

$$y = f(\mathbf{x})$$

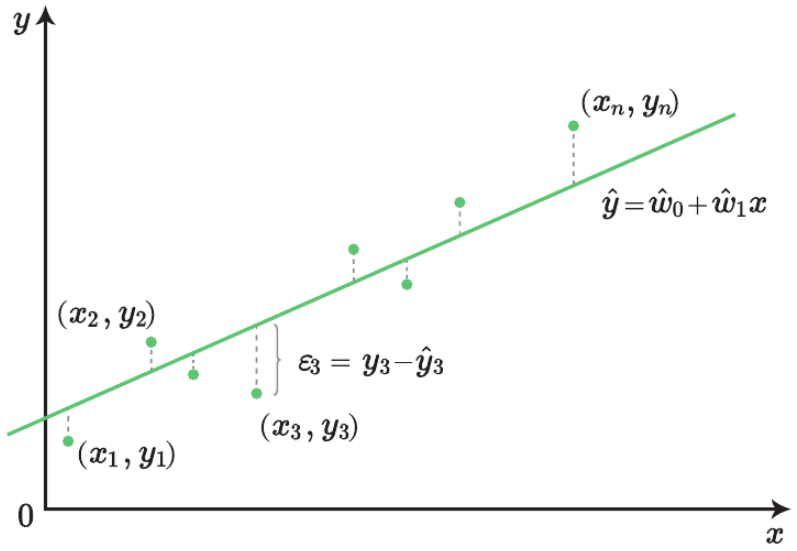
一元线性回归

- 一元线性回归模型，只有一个输入特征：

$$\hat{y}_i = w_1 x_i + w_0$$

其中w_0, w_1为模型参数(回归系数)

- 对于样本 (x_i, y_i)，真实值yi与模型预测值的差成为残差，记为：$\varepsilon_i = y_i - \hat{y}_i$
- 给定样本集合D=$\{(x_1, y_1), \ldots, (x_n, y_n)\}$，我们的目标是找到一条直线$y = w_0 + w_1 x$使得所有样本点尽可能落在它的附近。



一元线性回归 – 优化

- 目标函数：残差平方和（residual sum of squares）

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

将目标函数$\mathrm{RSS}(w_1, w_0)$分别对w_0和w_1求导并令导数等于零：

$$\frac{\partial \mathrm{RSS}}{\partial w_0} = \sum_{i=1}^{n} 2(y_i - w_1 x_i - w_0)(-1) = 0,$$

$$\frac{\partial \mathrm{RSS}}{\partial w_1} = \sum_{i=1}^{n} 2(y_i - w_1 x_i - w_0)(-x_i) = 0.$$

可获得最优解（能够描述目标特征和输入特征之间线性关系的最优直线）：

$$\hat{w}_1 = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^{n} x_i^2 - n(\bar{x})^2}, \quad \hat{w}_0 = \bar{y} - \hat{w}_1 \bar{x},$$

其中均值$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i, \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$. 这种方法叫做 最小二乘法 (Ordinary Least Square, OLS).

多元线性回归

在多元线性回归中，输入特征的维度由一维增加到d维度($d > 1$).假设训练集为$D = \{(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_n, y_n)\}$，多元线性回归模型为：

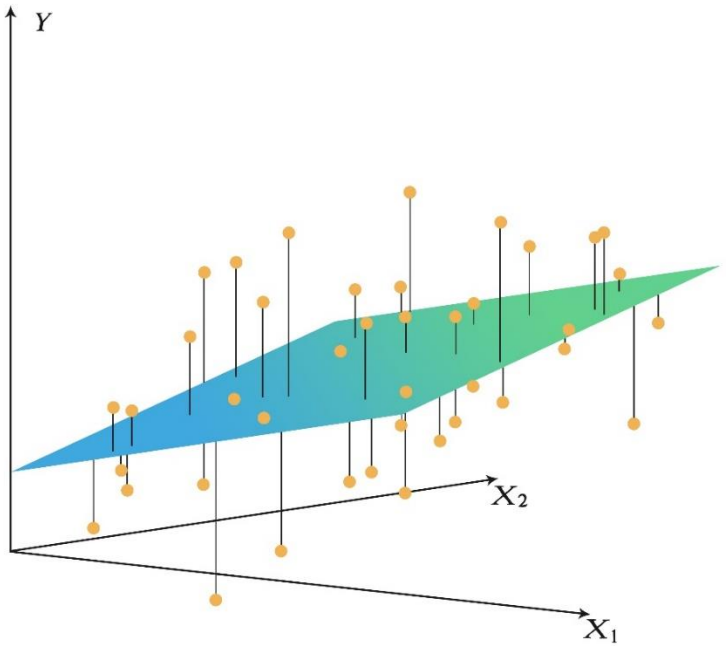
$$y = \boldsymbol{w}^{\mathrm{T}}\boldsymbol{x},$$

其中$\boldsymbol{w} = (w_1, w_2, \ldots, w_d, w_0)^{\mathrm{T}}$为模型参数.

假设我们将训练集中的输入特征部分记为$n \times (d + 1)$维矩阵$\mathbf{X}$，矩阵最后一列值全为1. 训练数据的输出特征部分写成向量形式$\boldsymbol{y} = (y_1, y_2, \ldots, y_n)^{\mathrm{T}}$.

在多元线性模型中，输入$\mathbf{X}$对应的模型输出$\hat{\boldsymbol{y}}$为：

$$\hat{\boldsymbol{y}} = \mathbf{X}\boldsymbol{w},$$



多元线性回归 – 优化

- 目标函数（残差平方和）：

$$\mathrm{RSS}(w) = \|y - \hat{y}\|_2^2 = \|y - \mathbf{X}w\|_2^2,$$

和一元线性回归类似，我们对参数w求导并令导数等于$\mathbf{0}$：

$$\frac{\partial \mathrm{RSS}}{\partial w} = -2\mathbf{X}^{\mathrm{T}}(y - \mathbf{X}w) = \mathbf{0}.$$

因此最优参数为

$$\hat{w} = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}y.$$

- 该方法容易产生过度拟合问题，特别是在特征数量大于训练样本数量时。我们可以通过特征选择和正则化等方法来解决该问题

线性回归的假设

简单线性回归通常对模型作了以下假设：

输入特征是非随机的且互相不相关；

随机误差具有零均值，同方差的特点，且彼此不相关；

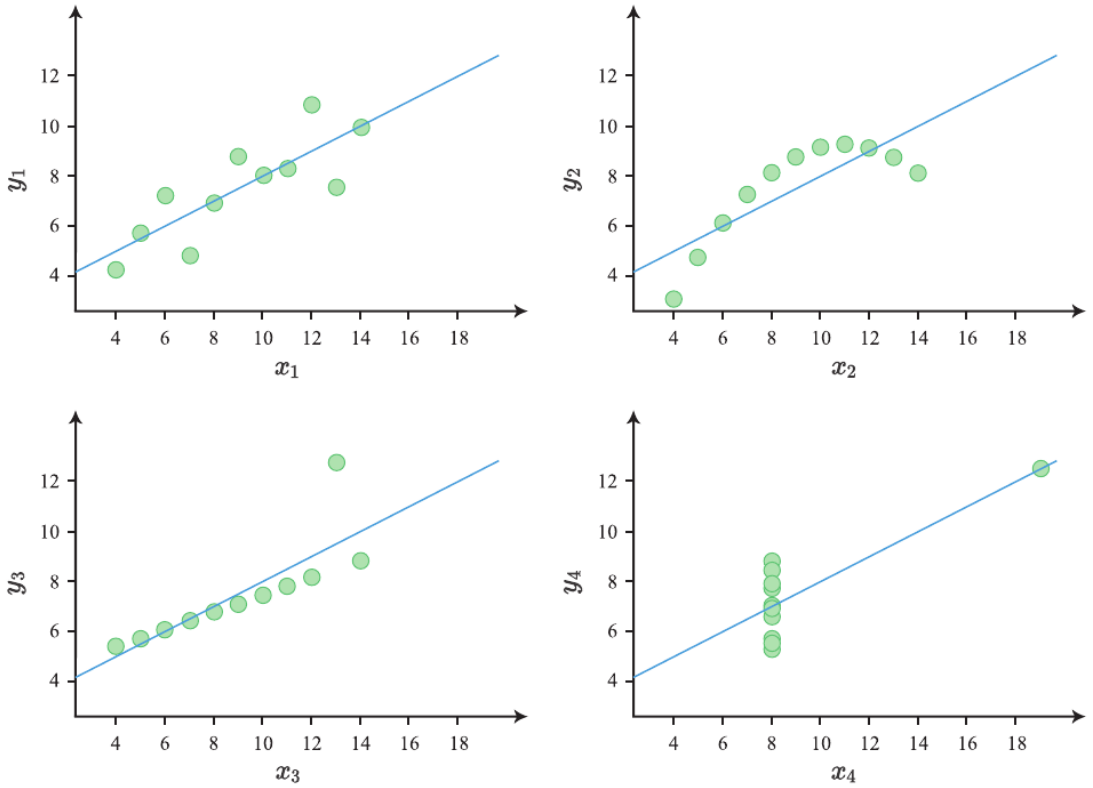
输入特征与随机误差不相关；

随机误差项服从正态分布$\mathcal{N}(0,\sigma^2)$.

因此在得到一个线性回归模型后，通常还要根据上述假设对回归结果进行诊断.

线性回归的问题

安斯库姆四重奏（每组数据都包括11个样本）



四组数据的基本统计特征一致：

均值、方差和线性回归线

- 实际数据可能不是线性的

使用R^2等指标进行模型诊断

$$\mathrm{SS}_{\mathrm{res}} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$ 残差平方和

$$\mathrm{SS}_{\mathrm{tot}} = \sum_{i=1}^{n}(y_i - \bar{y})^2,$$ 总平方和

$$\mathrm{R}^2(\boldsymbol{y}, \hat{\boldsymbol{y}}) = 1 - \frac{\mathrm{SS}_{\mathrm{res}}}{\mathrm{SS}_{\mathrm{tot}}}$$

越大代表模型越靠谱

线性回归的问题

- 多重共线性

最小二乘的参数估计为$\hat{w} = (X^T X)^{-1} X^T y$，如果变量之间存在较强的共线性，则$X^T X$近似奇异，从而引起$(X^T X)^{-1}$对角线上的值很大，导致参数估计值$\hat{w}$的方差增大，对参数的估计变得不准确

- 奇异矩阵：非满秩的矩阵，i.e., 行列式为0

解决方法：正则化、主成分回归、偏最小二乘回归

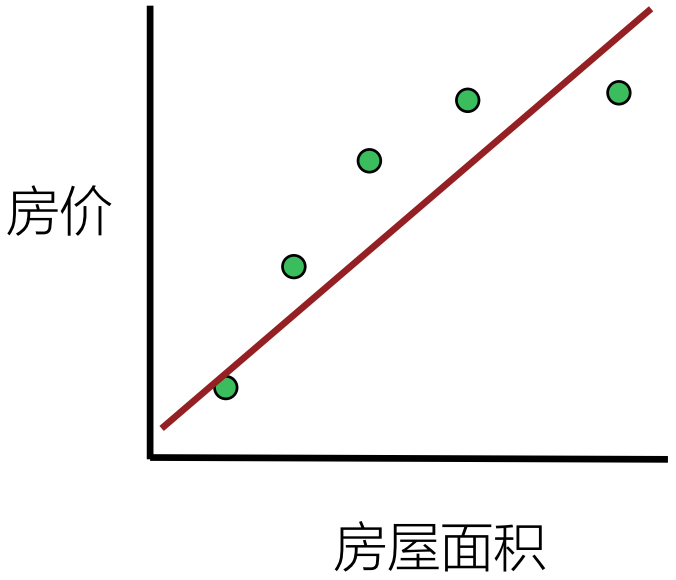
线性回归的问题

- 过度拟合问题：

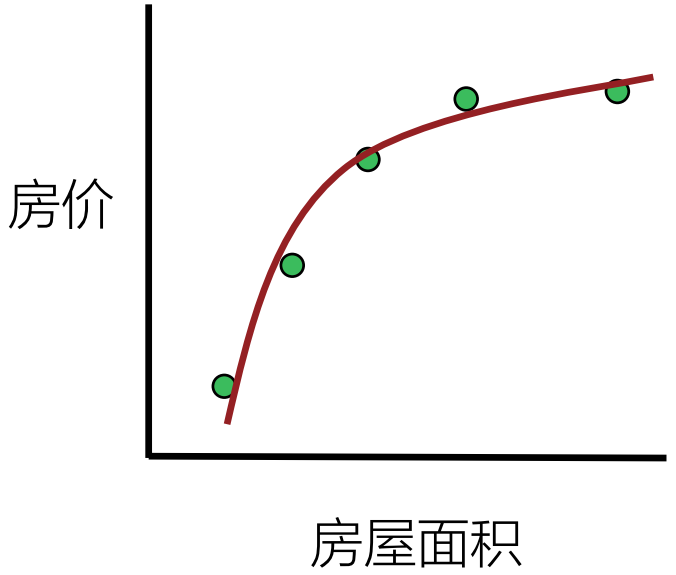
当模型的变量过多时，线性回归可能会出现过度拟合问题

例如在房价预测问题中，假设x表示房屋面积，如果将x^2，x^3等作为独立变量引入可能出现如下情况：

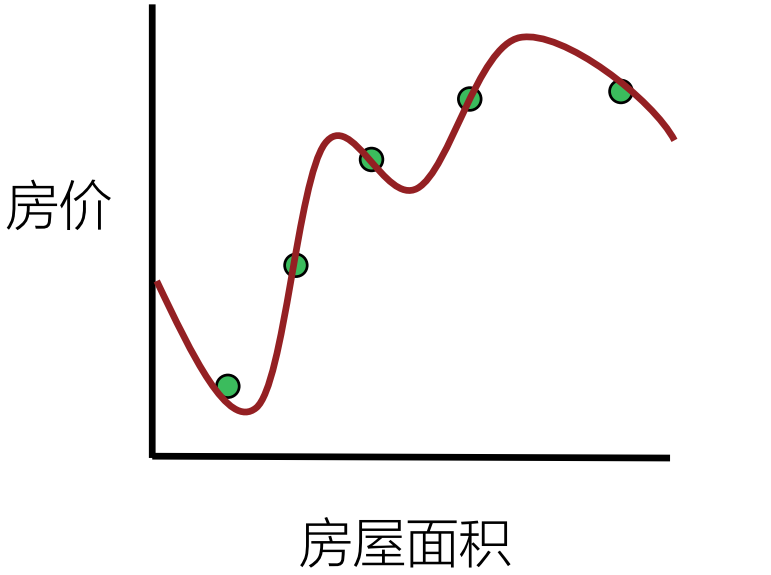
拟合函数$w_0 + w_1x$：
欠拟合 高偏差，训练误差大



拟合函数$w_0 + w_1x + w_2x^2$：
拟合很好



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



错误来自哪里?

If we denote the variable we are trying to predict as Y and our covariates as X, we may assume that there is a relationship relating one to the other such as $Y = f(X) + \epsilon$ where the error term ϵ is normally distributed with a mean of zero like so $\epsilon \sim \mathcal{N}(0, \sigma_\epsilon)$.

We may estimate a model $\hat{f}(X)$ of $f(X)$ using linear regressions or another modeling technique. In this case, the expected squared prediction error at a point x is:

$$Err(x) = E\left[(Y - \hat{f}(x))^2\right]$$

This error may then be decomposed into bias and variance components:

$$Err(x) = \left(E[\hat{f}(x)] - f(x)\right)^2 + E\left[\left(\hat{f}(x) - E[\hat{f}(x)]\right)^2\right] + \sigma_e^2$$

$$Err(x) = \text{Bias}^2 + \text{Variance} + \text{Irreducible Error}$$

That third term, irreducible error, is the noise term in the true relationship that cannot fundamentally be reduced by any model. Given the true model and infinite data to calibrate it, we should be able to reduce both the bias and variance terms to 0. However, in a world with imperfect models and finite data, there is a tradeoff between minimizing the bias and minimizing the variance.

来自：http://scott.fortmann-roe.com/docs/BiasVariance.html

偏差和方差

真实值　预测值

- 泛化误差$E\left[(y - \hat{f}(\boldsymbol{x}))^2\right]$

$$= Var(\boldsymbol{x}) + bias^2(\boldsymbol{x}) + variance(\boldsymbol{x})$$

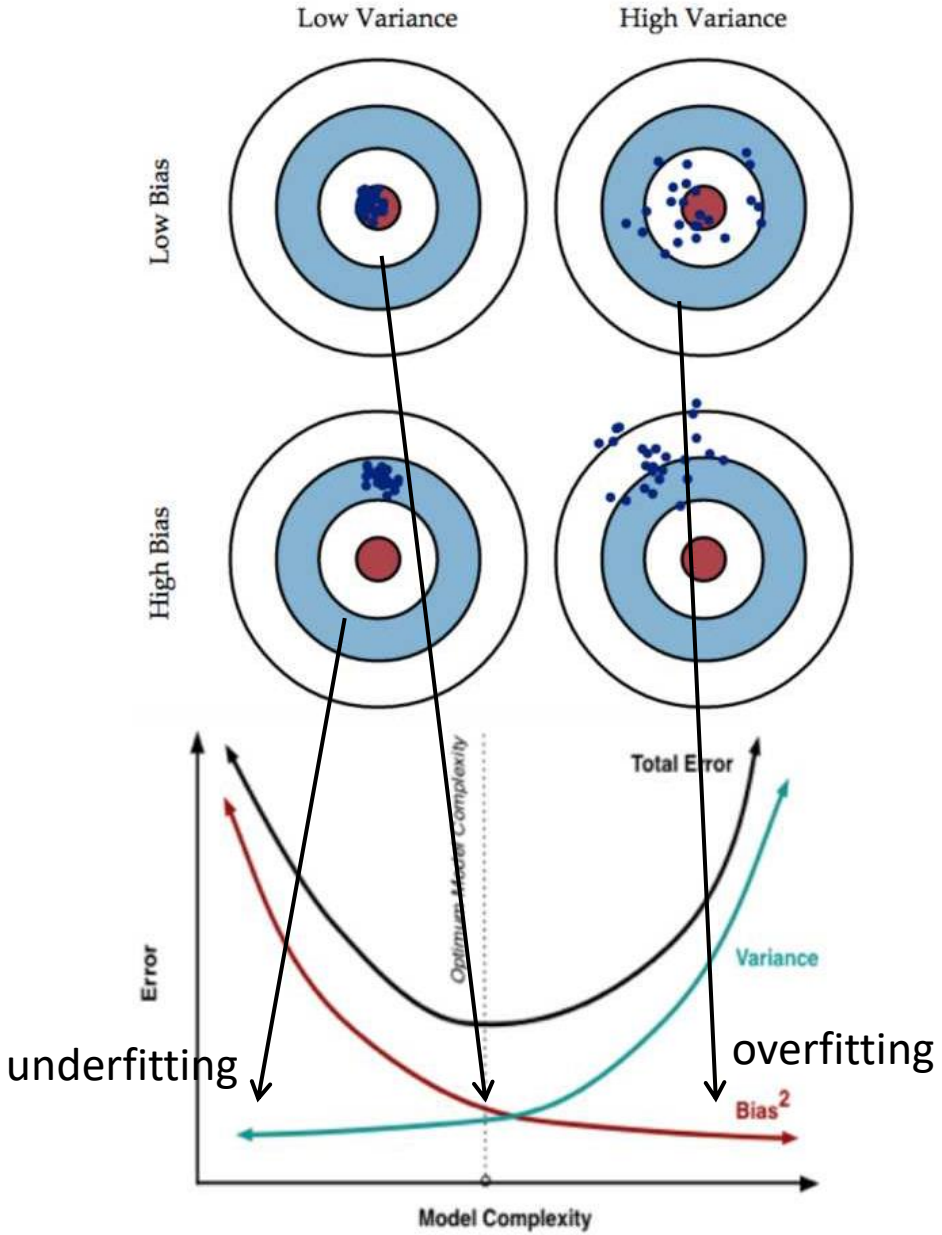
Irreducible Error　偏差平方　方差

真实值与观测值之间的方差

真实值与模型预测值期望之间的方差

模型预测值与预测值期望之间的方差

- 偏差($bias$):模型依靠自身能力进行预测的平均准确程度（准）
- 方差($variance$):模型在不同训练集上表现出来的差异程度（确）



目录

正则化

- 通过在模型中添加惩罚项或约束条件来控制模型复杂度，获得bias-variance trade-off
  - 可以减小线性回归的过度拟合和多重共线性等问题

岭回归和LASSO

具体来讲，岭回归和LASSO分别对应ℓ_2和ℓ_1正则化，对系数向量$\boldsymbol{w}$提出的先验假设分别为$\|\boldsymbol{w}\|_2 \leq C$和$\|\boldsymbol{w}\|_1 \leq C$，$C$为预先取定的常数. 也就是说，我们关注下面带约束的优化问题，对于岭回归

$$\min_{\boldsymbol{w}} \quad \|\boldsymbol{y}-\mathbf{X}\boldsymbol{w}\|_2^2 \qquad \text{s.t.} \quad \|\boldsymbol{w}\|_2 \leq C,$$

利用拉格朗日乘子法，以上约束优化问题等价于无约束的惩罚函数优化问题

$$\min_{\boldsymbol{w}} \quad \|\boldsymbol{y}-\mathbf{X}\boldsymbol{w}\|_2^2+\lambda\|\boldsymbol{w}\|_2^2$$

其中正则化系数$\lambda > 0$是依赖于C的常数.

类似的，如果我们采用L1 正则化，则可获得LASSO （Least Absolute Shrinkage and Section Operator）：

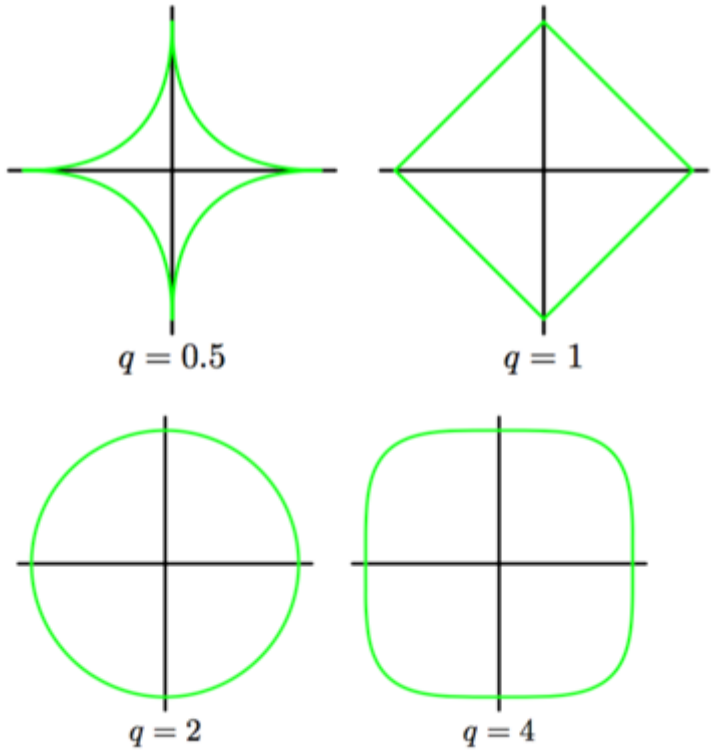
$$\begin{aligned}&\min_{\boldsymbol{w}} \quad \|\boldsymbol{y}-\mathbf{X}\boldsymbol{w}\|_2^2\\ &\text{s.t.} \quad \|\boldsymbol{w}\|_1 \leq C.\end{aligned} \quad \Rightarrow \quad \min_{\boldsymbol{w}} \quad \frac{1}{2}\|\boldsymbol{y}-\mathbf{X}\boldsymbol{w}\|_2^2+\lambda\|\boldsymbol{w}\|_1,$$

正则化

- L_q 正则化的通用形式：

$$\sum_{i=1}^{n}\left(y_1 - w^T x_i\right)^2 + \lambda \|w\|^q$$

- q=2：岭回归(Ridge Regression) ⇔ loss + L_2 惩罚项
- q=1：LASSO ⇔ loss + L_1惩罚项



$||w||^q$ =1示意图

岭回归

对于岭回归问题，对目标函数直接求导并令梯度等于零易得

$$\boldsymbol{w}^{\text{ridge}} = \arg\min_{\boldsymbol{w}} \left(\|\boldsymbol{y} - \mathbf{X}\boldsymbol{w}\|_2^2 + \lambda \|\boldsymbol{w}\|_2^2 \right) = (\mathbf{X}^{\mathrm{T}}\mathbf{X} + \lambda \mathbf{I})^{-1} \mathbf{X}^{\mathrm{T}} \boldsymbol{y}.$$

与式线性回归的解比较可以得知，岭回归得到的估计只多了一个正则项$\lambda\mathbf{I}$. 这一项的存在使得$(\mathbf{X}^{\mathrm{T}}\mathbf{X} + \lambda\mathbf{I})^{-1}$在数值计算上表现更加稳定. 尤其是当多重共线性(multi-collinearity)情况发生时，$\mathbf{X}^{\mathrm{T}}\mathbf{X}$接近奇异，岭回归还是能得到稳定的结果. 另外，在岭回归中还可以通过观察岭迹来剔除多重共线性的特征.

正则化路径分析

在线性回归中，正则化路径是指回归系数的估计值$\hat{w}$随着正则化系数增大而变化的曲线. 它能够帮助我们分析特征之间是否存在相关性以及进行特征选择.

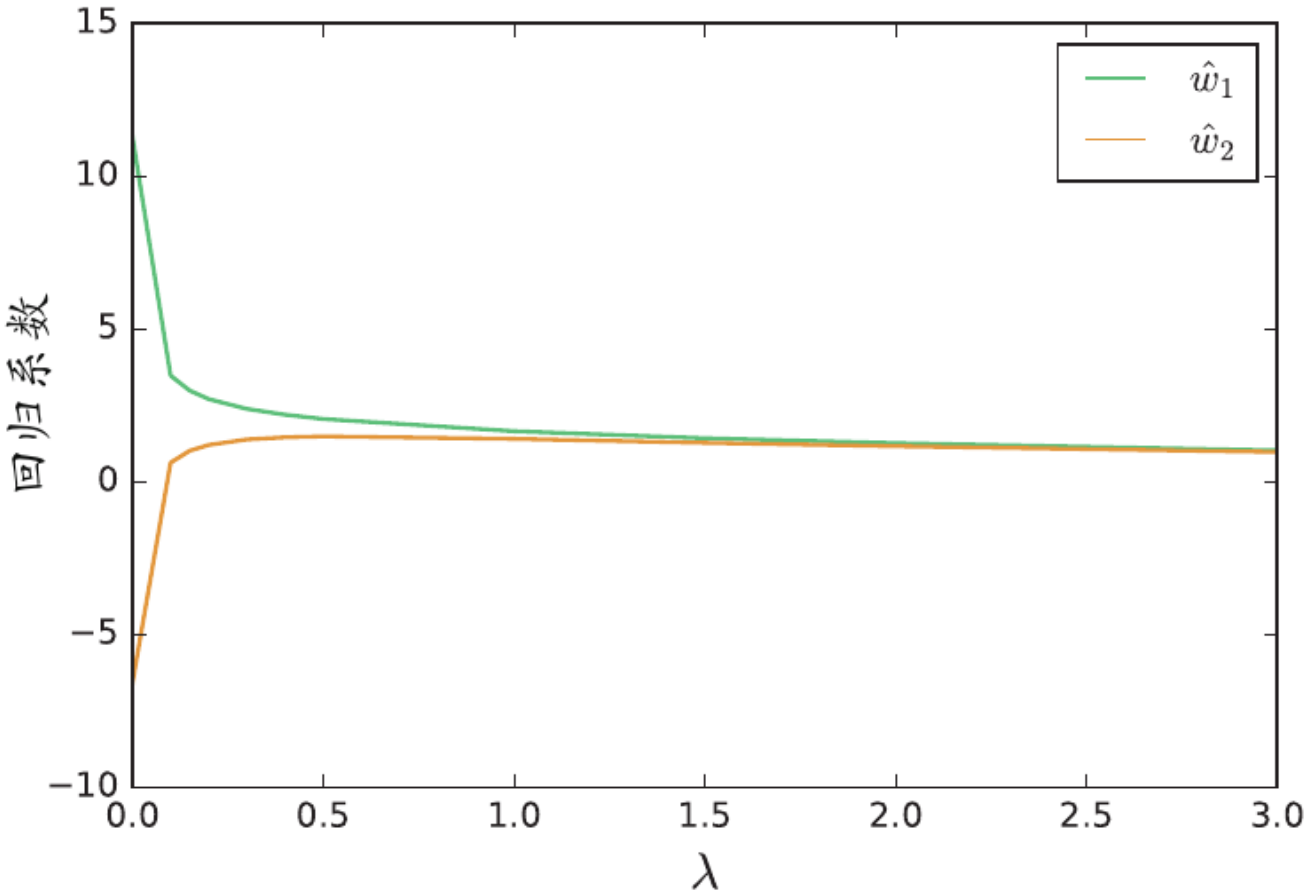
岭回归的正则化路径也被称为岭迹. 如果岭迹波动很大，说明该特征与其他特征有相关性. 假设我们得到在不同的正则化系数λ条件下的回归系数，如表1所示.

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

模型参数是关于正则化系数的函数

岭迹分析

- 在λ~(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λ=1
- 根据岭迹图做超参数λ的选择
- 除了这种定性的超参选择方法，我们还可以使用交叉验证方法



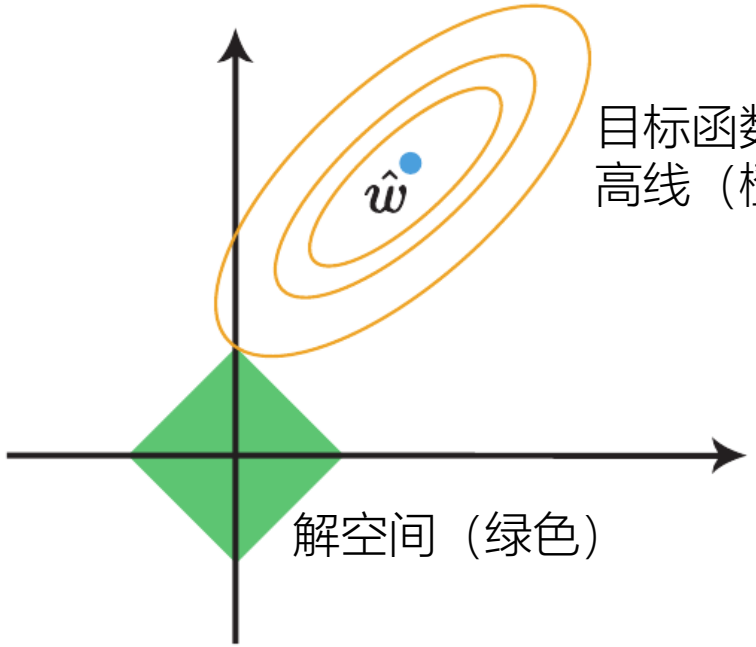
LASSO回归

- LASSO 是一种系数压缩估计方法，它的基本思想是通过追求稀疏性自动选择重要的变量
- LASSO 的目标函数：$\sum_{i=1}^{n}\left(y_i - w^T x_i\right)^2 + \lambda||w||_1$
- LASSO的解 $\hat{w}^{LASSO}$没有解析表达式，常用的求解算法包括坐标下降法、LARS算法和ISTA算法等

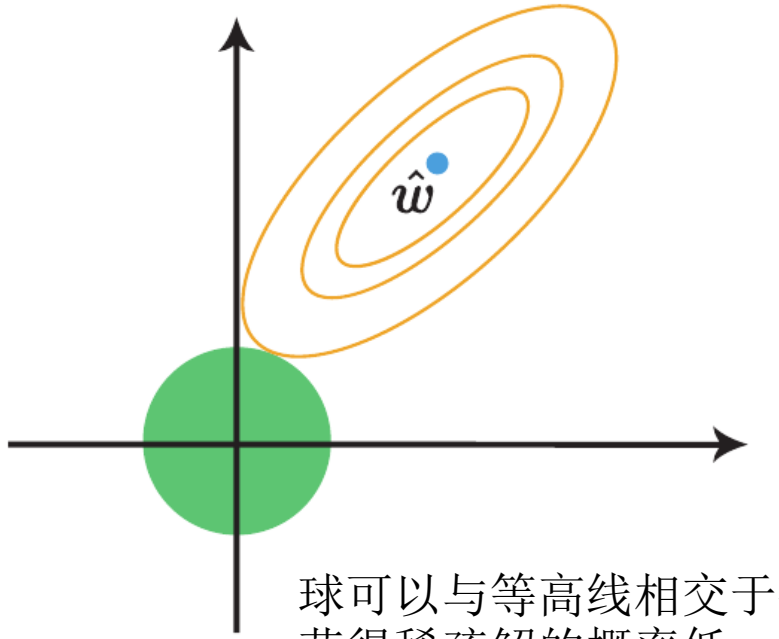
- 为什么LASSO可以产生稀疏解?

最优解：发生在目标函数的等高线和可行区域的交集处.

LASSO



岭回归

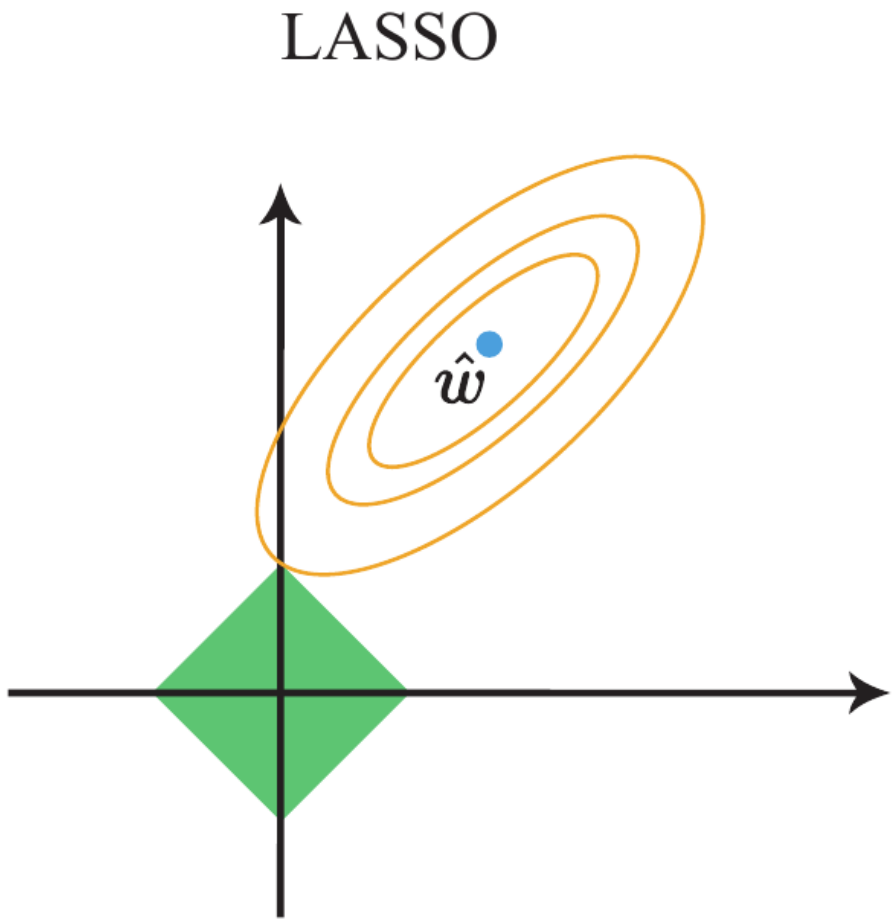


球可以与等高线相交于任何一点，
获得稀疏解的概率低

角比边更有可能与等高线相交，这个现象在高维的情况下尤其明显，因为高维的角更加“凸出”=> 产生稀疏解

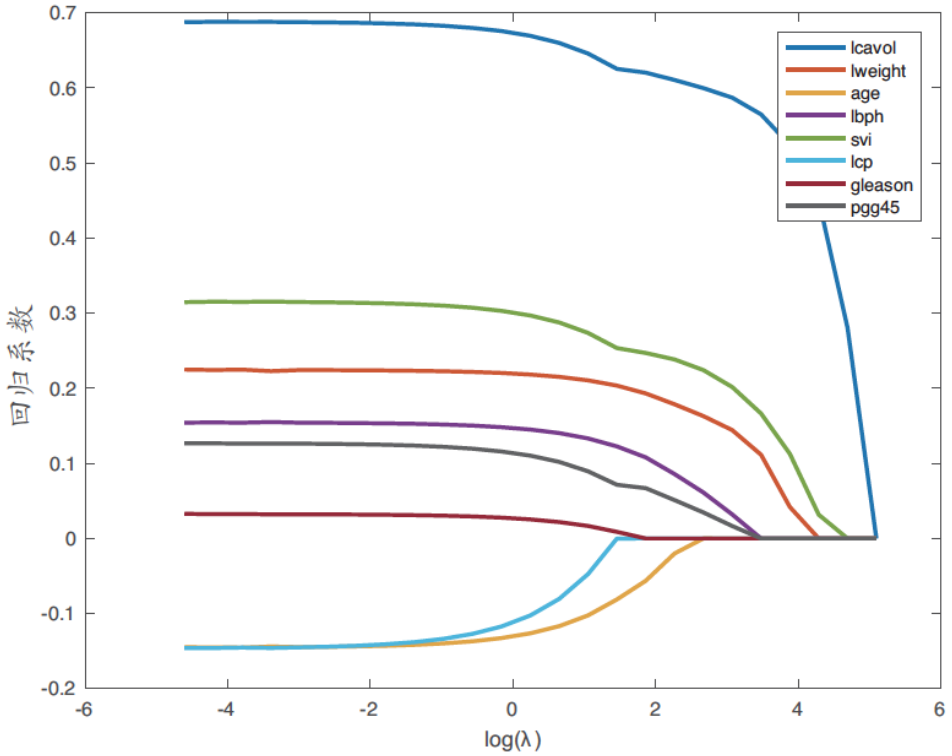
L_1进行特征自动选择

- 如右图所示：图中绿色区域表示约束区域，黄色线为残差平方和函数的等高线
- 通过添加 $\mathbf{L_1}$ 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 $\hat{\mathbf{w}}_2 = 0$，我们可以将对应的变量从模型中删除。



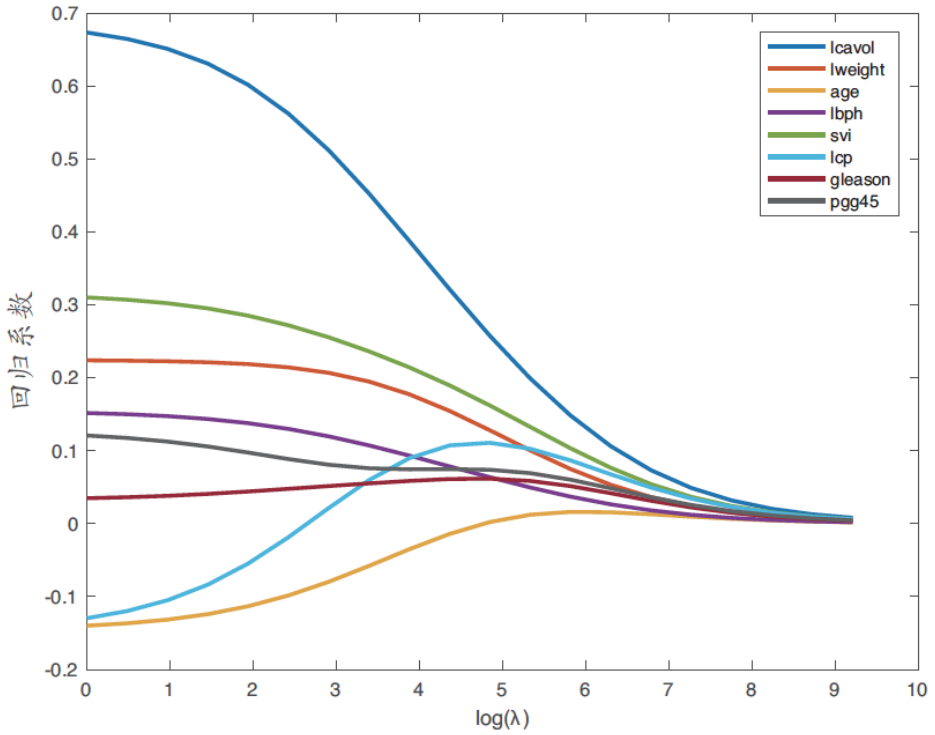
正则化路径分析

前列腺抗原(目标特征)和其他9个临床测量数据(9维输入特征)的关系.



LASSO

随着λ增大，LASSO的变量系数逐个减小为0，可以做特征选择



岭回归

而岭回归变量系数几乎同时减小为0，难以起到特征选择的作用

其他正则化的线性模型

- LASSO在实际应用中的问题：
  - 当数据中存在一组高度相关的特征，LASSO倾向于只选择其中的一个，且选择有较大的任意性
  - 实际应用中希望能把整组的特征都用上，e.g.,在基因数据中，对于表达同一蛋白的多个基因会存在很高的相关性，我们关心的是整组基因，而不仅仅是其中的一个

为了克服以上的缺点，Zou等人于2005年提出了弹性网络(elastic net)正则化. 它综合了岭回归和 LASSO 的正则化方式，最小化如下目标函数

$$J(\boldsymbol{w}) = \|\boldsymbol{y} - \mathbf{X}\boldsymbol{w}\|_2^2 + \lambda_1\|\boldsymbol{w}\|_1 + \lambda_2\|\boldsymbol{w}\|_2^2.$$

弹性网络正则化会呈现一个特征分组的效果，即高度相关的特征对应的回归系数倾向于相等(对于负相关的特征，系数有符号差别). 它克服了LASSO在这方面的缺点，在微阵列分类和基因选择中得到了成功的应用.

其他正则化的线性模型

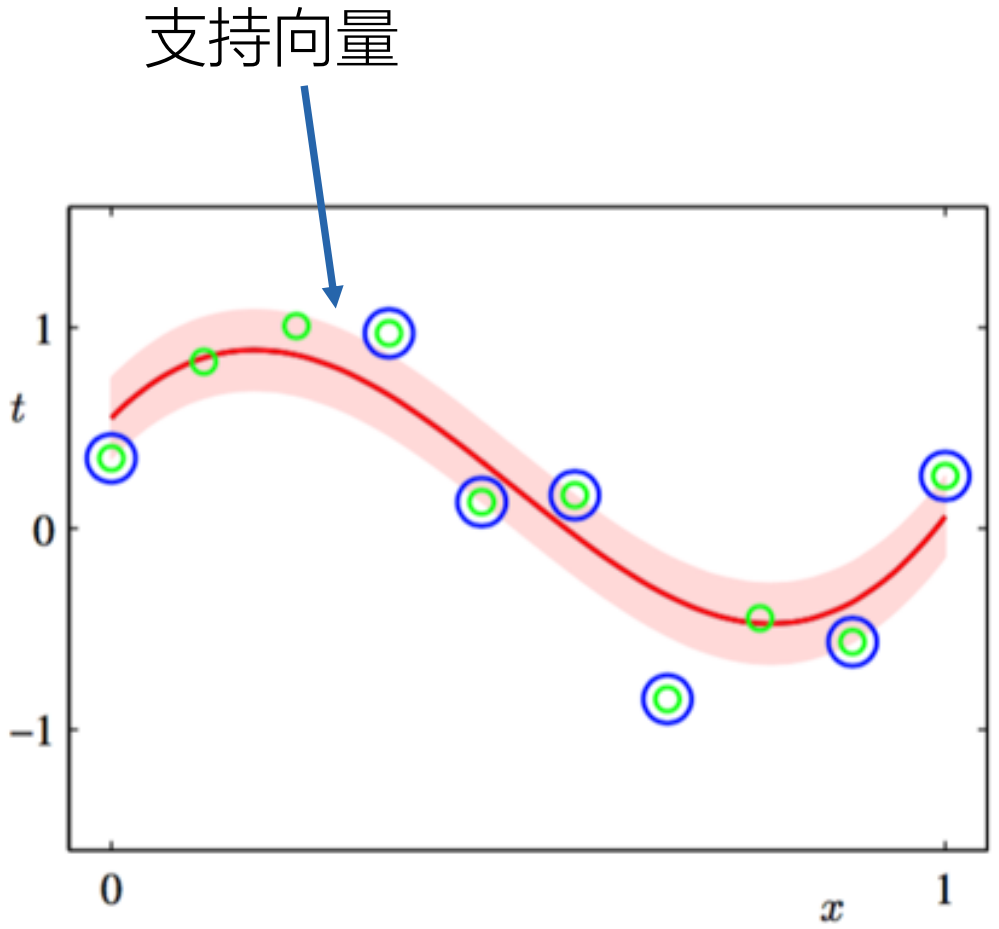
在另外一些应用中，特征不是以单个出现的，而是以事先定义好的组(往往依据具体的问题确定组的划分)的形式呈现，比如属于同一生物过程的一组基因. 比如对特征进行One-Hot编码后的一组特征. 在这些情况下，我们期望能够以组为单位进行特征选择，这就是Group LASSO的基本思想. 假如有事先确定好的G组，$\boldsymbol{w} = (\boldsymbol{w}_1, \boldsymbol{w}_2, \ldots, \boldsymbol{w}_G)$，Group LASSO最小化如下目标函数来进行组特征选择

$$J(\boldsymbol{w}) = \|\boldsymbol{y} - \mathbf{X}\boldsymbol{w}\|_2^2 + \sum_{g=1}^{G} \lambda_g \|\boldsymbol{w}_g\|_2.$$

简单来说，正则化是将数据科学家对于问题的先验理解加入到模型中，进而控制模型解的结构的有效手段.

非线性回归方法

- 样条回归
- 径向基网络
- SVR: SVM for regression
- Regression Tree
- Gaussian process
- ……



总结

- 一元线性回归、多元线性回归、最小二乘估计
- 多重共线性问题、过度拟合问题、偏差与方差
- 正则化、岭回归、LASSO回归

案例分析 – 医疗费用预测

- 精准预测医疗费用对保险公司有重要的价值
- 本案例提供的数据集是从美国人口普查局的人口统计资料中整理得出。数据集共包含1338个样本，具体特征如下：
  - Age: 被保险人年龄
  - Sex: 性别
  - BMI：身体指数
  - Children: 计划中所包含的孩子/受抚养者的数量
  - Smoker: 被保险人是否吸烟
  - Region: 在美国的居住地
  - Charges: 已经结算的医疗费用

- 数据集下载链接：http://staff.ustc.edu.cn/~hexn/data-science/insurance.csv
- 编程平台：Python (+pandas, +numpy, + matplotlib), Jupyter