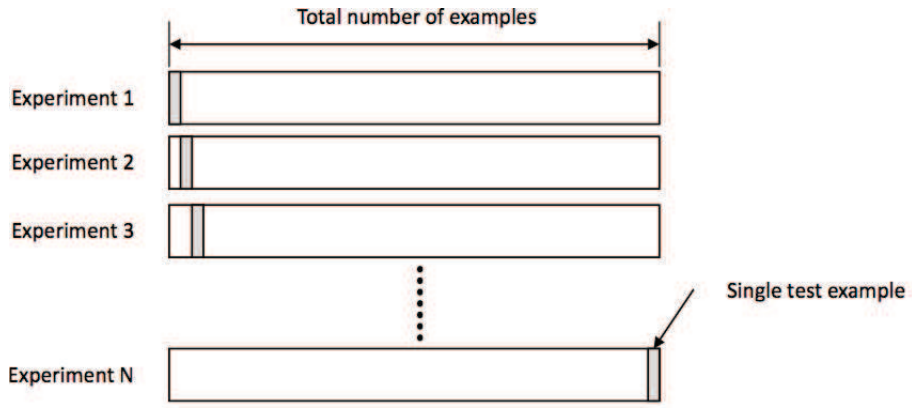


Figure 2.19: 去1交叉

在交叉测试中， 如果划分的数量增大会有什么特征呢？ 1)真是误差的偏差最较小。 2)但是误差的方差会变大。 3) 计算时间会增大。

那么划分数量如何确定呢？ 1）对于大数据量， 由于计算本身就比较大， 所以一般来说3划分就相当准确了。 2）对于稀疏的数据（sparse datasets）， 最好利用LOOCV，来对数据充分测试。 3）中等规模的话，10划分是一个不错的选择。

如果模型选择和性能指标同时计算的话， 数据就要被划分成3大块（图??）。

- 训练集： 用来参数学习。
- 验证集： 用来选择训练模型选择
- 测试集： 用来评估充分训练后的模型。

所以3路数据分割的一般流程如下： 1） 把数据划分成训练，验证和测试集。 2）选择架构和训练参数 3）利用训练集训练模型 4）利用验证集评估模型选择 5）重复2-4，选出最优模型 6）利用测试集评估模型。

2.3.6 后验抽样

2.3.7 马尔科夫蒙特卡洛方法

在线材料UCLA的一个Tutorial

2.4 回归分析

在前面我们讲过误差分析中的期望预测错误(Expected Prediction Error (EPE))， 它的表 达式为$E[(Y - \hat{f}(X))^2]$, 理论上我们要知道数据的概率分

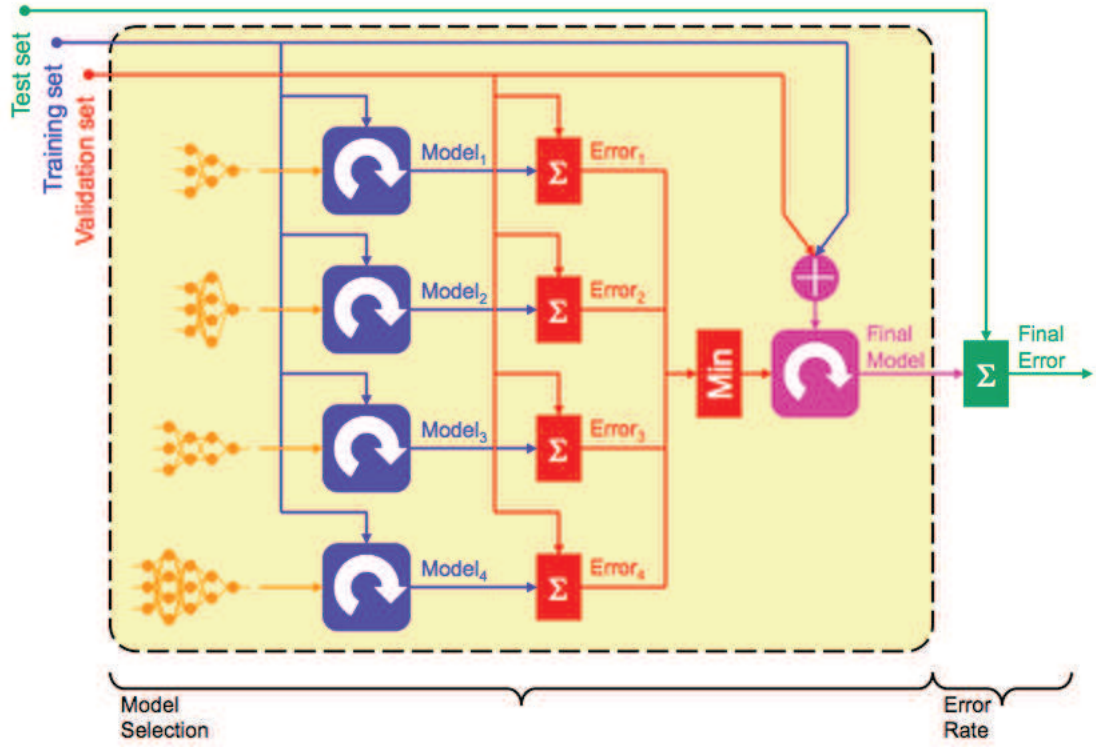


Figure 2.20: 三路（训练， 验证和测试）数据分割

布然后就可以 计算了。但是在实际数据分析中， 我们必须要估算， 一般会用均方差形式 (Mean squared error)。 根据以前的分析，我们希望这个值能够最小， 因此 被称为经验风险最小化（Empirical risk minimization(ERM)）， 我们知道在最小化 中存在偏差和方差平衡的问题。 我们估计的时候会用余留方差和(Residual Sum of Squired RSS)来代替平均值。

$$RSS(f) = \sum_{i=1}^{n} (y_i - f(x_i))^2$$

那么ERM的结果就是要求函数:

$$\hat{f} = \underset{f}{\operatorname{argmin}}\, RSS(f)$$

但是存在过拟合问题(Over-fitting)和函数过于复杂(complexity)。
为了使得函数相对简单平滑(Smooth), 提出粗度惩罚(Roughness Penalty),

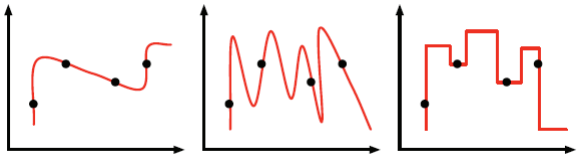


Figure 2.21: 同样的数据可以用二次样条， 高次样条， 以及阶梯来拟合

计算PRSS(Penalized Residual Sum of Squired)

$$PRSS(f,\lambda)=\sum_{1}^{n}(y_i-f(x_i))^2+\lambda J(f)$$

其中λ是惩罚系数， $J(f)$是函数平滑测度(functional measuring smoothness)。 如何最小化PRSS， 使得函数既能够满足误差最下， 又能够保 证简单平滑。

这个过程相对经验风险最下化， 又被称为结构风险最小化 (Structural Risk Minimization(SRM))。 例如， 我们要求$f(x)$的二次导数越小越好。

$$J(f)=||f''(x)||^2$$

SRM主要是为了防止选择了过于复杂的函数， 从而使得函数的通用性变弱， 而使得以后测试过程的预测效果不好。 或者说是过拟合了。

但是， 有时候我们也希望我们估算的参数本身不要过于复杂而不是函数过于复杂。 这时候我们会引入待估参数的惩罚。 一般来说， 这种惩罚(penalty)可以分为L1惩罚和L2惩罚，或者称为L1回归 (regularization)和L2回归。 主要是指函数平滑测度是 L^p空间的表达式。 L^p空间的定义形式:

$$\|x\|_p=(|x_1|^p+|x_2|^p+\cdots+|x_n|^p)^{\frac{1}{p}}$$

其中L^∞空间的定义如下:

$$\|x\|_\infty=\max\{|x_1|,|x_2|,\ldots,|x_n|\}$$

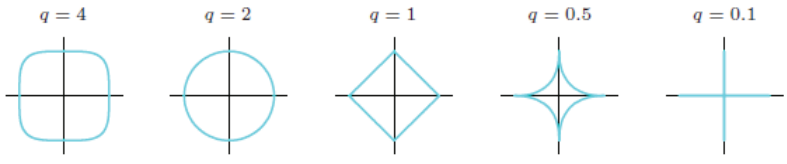


Figure 2.22: Lp空间， 从0到无穷大的过程是从原点到正方形的一个扩张

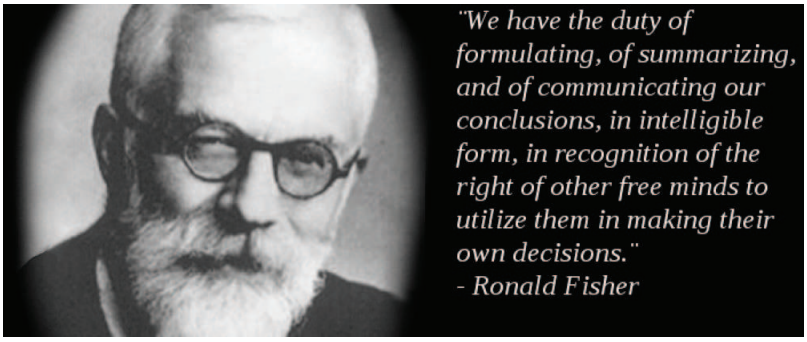
其中最经典的L2算法是Ridge 回归， 而L1算法有LASSO， 另外还有LAR。

- 岭回归 Ridge： $J(f) = \sum_i \beta_i^2$
- 最小绝对值收缩和选择器 LASSO （Least Absolute Shrinkage and Selection Operator）： $J(f) = \sum_i |\beta_i|$
- 最小角度回归 LAR (Least Angular Regression): $J(f) = \sum_i \beta_i$

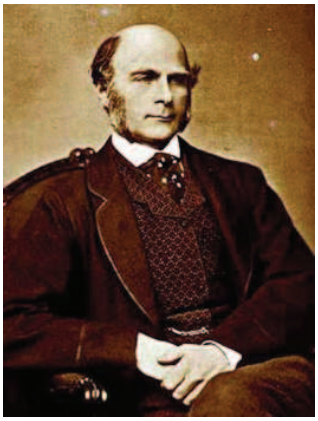
LASSO是Tibshirani在1996年提出的， 在研究LASSO和boost之间关系的时候， Bradley Efron在收到Jackknife工作的启发之后在1979年提出了Bootstrapping。 而Jackknife是Maurice Quenouille最早在1949年的时候开始 研究的。

罗纳德，费希尔(Ronald Fisher)是英国的搞农业出生的， 这家伙视力却很差，特别喜 欢生物统计， 由于不小心读了卡尔皮尔逊(Karl pearson)的论文而中毒， 皮尔 逊这个家伙的确有点魅力，他是剑桥学数学出生的，他提出了矩估计， 发展了 卡方检验， 创立了概率函数簇。 就是他和高尔顿(Francis Galton)一起办了统计的第一份杂志 《生物统计学》。后英国的统计就领先全球了。高尔顿要比皮尔逊大40岁， 这 家伙是个全才，指纹就是他搞出来的， 他开始研究人类学的，热衷并提出了优 生学(eugenics)后，不仅仅吸引了皮尔逊，也深深的吸引了费希尔。他就和凯恩 斯, 达尔文的儿子一起搞了剑桥的优生学会。 他是个天才， 对实验很有理解， 提出了统计方法的一系列优化判断标准， 例如最大似然估计， 费希尔线性判定， 充分性。比皮尔逊小十几岁而又比费希尔大十几岁的还有一位大师是酿啤酒的搞化 学的，他叫威廉戈塞(William Sealy Gosset)， 他就是大名鼎鼎的学生， 因为 他的老板不让他把酿酒的秘密发表出来， 所以他只能用笔名学生氏（Student） 来发表。 他在选酒过程研究了学生分布。据说费希尔特别爱读他的论文。

费希尔有个印度的学生叫拉奥Calyampudi Radhakrishna Rao， 他和另外一个研 究素数的克拉梅尔Harald Cramer提出了最优无偏估算的下届(Cramer-Rao Lower Bound CRLB)， 和利用期望估算来优化估算器的RB定理(Rao-Blackwell theorem)。 其中Blackwell是布莱克韦尔(David Harold Blackwell)是伯克利的 名誉教授，是美国第一位黑人院士。 也正是从拉奥开始统计的重心从英国移到 了美国。 Maurice H. Quenouille是美国的研究统计推理的一位学者，他写过美 国早期的《统计介绍》(Introductory Statistics) 他提出 了Jackknife的核心思想（1949）， 后来John W. Tukey(1915-2000)后来把Jackknife正 式发表出来了（1958）。 他的工作影响了Jerome H. Friedman， Friedman是斯 坦福的教授（1982-）。当时Trevor Hastie正在那里读PHD， 1994年他又从南非 回到斯坦福和Friedman，还有他的斯坦福同学Rob Tibshirani，一起写了The Elementary of Statistics Learning。 Rob Tibshiran发明了LASSO算法， 他的老板是Bradley Efron 这个家伙提出了 bootstrap方法(1979, 1981(Bayesian extension))，正是受到了Turkey工作的影响。Turkey 自己就是斯坦福的博士（1964），后来受他学生的LASSO影响发明了LAR算法。



(a) Ronald Fisher (1890 - 1962)



(b) Sir Francis Galton (1822 - 1911)

(c) Karl pearson（1857-1936）

(d) William Sealy Gosset（1876-1937）

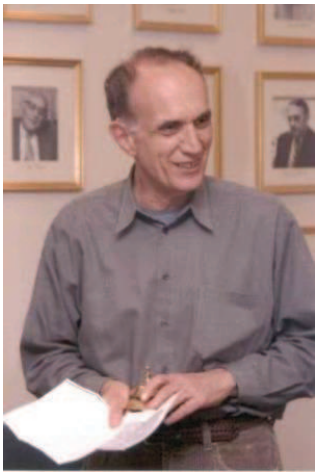
Figure 2.23: 英国统计时代影响费希尔的巨人

另外伯克利的Leo Breiman受到bootstrap的启发，把分类回归树 （classification and regression trees ）应用到bootstrap样本上，提出了 Bootstrap聚合 (Bootstrap aggregation）（1994）。 Breiman是分类器的大牛， 提出了决策树的CART（1984）， 随机森林（Random forest）（2001）。 而boosting的思想是 Robert Schapire回答（1990）Michael Kearns提出的（1988）的关于一组弱学 习器能否生成强学习器的时候提出的思想。 但是一直找不到实现， 直到 AdaBoost的提出 （1995）， 也许受到了Bagging方法的影响。 Robert Schapire 是普林斯顿的教授。 他的合作者Yoav Freund是加州大学的教授。

上面主要提到了目前主要的Boosting和L^p回归两个方向的大牛。 从而引入下面要讲的L^p回归的算法。

(a) Calyampudi Radhakrishna Rao（1920-）

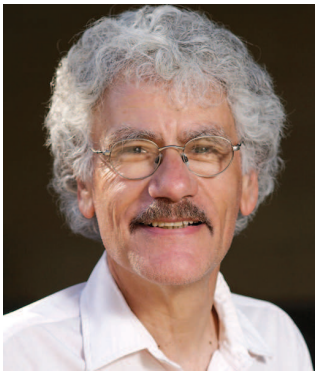
(b) John Tukey (1915-2000)



(c) Bradley Efron（1938-）

(d) Robert Tibshirani（1956-）

(e) Trevor Hastie（1953-）



(f) Jerome Friedman (1939-)

(g) Leo Breiman（1928-2005）

(h) Robert Elias Schapire（196x-）

Figure 2.24: 美国统计学习大牛

2.4.1 LMS

如果ERM条件下， 根据高斯-马尔科夫定理， 在误差为0，同分布， 互不相干的情况下， 最佳线性无偏估计（BLUE）就是最小方差无偏估计（MVUE）， 而BLUE由最小二乘法（OLS）给定的。 （而没有高斯-马尔科夫定理的话， 我们就需要iid条件才能得到这个结论）。

牛顿法

Newton's Method

准牛顿法

Quasi-Newton's Method

牛顿拉福生法

Newton-Raphson Method

LMS 在高维稀疏空间时候， 如何保证待估参数只在尽可能少的维度里面尽可能小，使得 样本选择和计算时候方便高效是一个问题。 和结构风险最小化（SRM）类似。 提出了惩罚回归（Penalized Regulation）。

2.4.2 局部加权线性回归

Locally weighted linear regression(局部加权线性回归)

http://www.cnblogs.com/hust-ghtao/p/3587971.html http://www.dsplog.com/2012/02/05/weigh least-squares-and-locally-weighted-linear-regression/

2.4.3 特征选择

在线性模型情况下， 之前我们已经介绍MVUE和BLUE对参数进行估计。但是在高 维情况下，（可能存在$p > N$的情况）， 这时候应用最小二乘法的时候， 我 们依然可以依据高斯马尔科夫定理来保证无偏估计下的最优估计。 但是， 由 于高维情况下，计算代价太高， 如果牺牲一点偏差来获得更高的效率和更小的 计算量的话， 应该是不错的选择。

最佳子集选择

Allan Miller 最早开始系统的研究最佳子集选择（Best Subset Selection），高维的通常情况下存在一个最佳子集的， 只用这个子集的数据回归的模型是最佳 的。 但是穷尽搜索（Exhaustive Search ）的话，我们会得到一组组合 数$\binom{p}{k+1}$, $\binom{n}{k}$。 这样的效率很低。 因此， 一系列贪心 算法被用来查找最优子集。

向前（后）按步选择

向前（后）按步选择 （Forward/backward stepwise selection） 向前和向后的区别是从0到p还是从p到0的过程。

贪心过程是始终选择根据回归计算公式能够贡献最大的特征轴。 特价来说，贡 献最大意味着这个变量有最小的p值对于他的t分布（t测试）或者F分布（F测试））来说。 计算过程是， 假设这个维度特征不选择， 我们估算方差均值， 按原来方差均 值来计算t分布下的p-value。 p-value越小表明这个这个维度特征越重要。 有些算法（软件）允许我们设置一个p-value的上限。 而向后选择的就是每次删 除一个特征， 删除的这个特征会有最大的p-value。

不采用p-value的情况，也会计算最大相关度(correlation)， 假设$\mu = E(Y|X)$， $r = y$。

1. 开始假设$\mu = 0$并且x_j是。
2. 选择和r具有最大绝对相关的X_j^*， 或者说$< X_j^*, r >$最大。
3. 拟合一个线性模型并且更新$\hat{\mu} \leftarrow \hat{\mu} + \hat{\beta}_j X_j$
4. 计算剩余的向量$r = Y - \mu$ ， 把剩余X_k^*正交于X_j投影。
5. 重复选取过程。

可见Stepwise就是不停的相各个方向投影。 但是在这个过程中会选择最相关的。 并且这个过程中，由于余量的作用， 会自动过滤和前面已经选择的相关的样本。 我们知道投影的过程就是最小二乘法的近似的过程。这样就是一种有相关优先的最小二乘法。

向前按阶段选择

向前按阶段选择(Forward stagewise selection)， 算法如下：

1. 设定$\hat{\mu} = 0$
2. 定义当前相关向量$\hat{c} = c(\hat{\mu}) = X^T(Y - \hat{\mu})$
3. 找到使得$j = \text{argmax}\,|\hat{c}_j X_j|$，并且更新$\hat{\mu} \leftarrow \hat{\mu} + \epsilon sign(\hat{c}_j) X_j$。 其中$\epsilon$是一个小常数。

和Stepwise相比， Efron提出的找到最相关的响应后， 不要修正太快，所以引 入了ϵ这个参数。 这样在X_i这个方向上，就不会回归的太快。 这样就能在一个方向上慢慢的前进直到X_j出现成为最相关的。 这样的优点就是 要比Stepwise要谨慎很多。 在这个基础上Efron提出既然可以这样选择路径。 那么何不加快Stagewise的流程， 当多有个X_t同时最相关的时候，就沿着他 们对角线的角度前进。

最小二乘自举

最小二乘自举(Least Square Boosting)

2.4.4 Ridge

岭回归是L^2的回归。 根据图??所示， 参数的L^2使得参数尽量的解决圆形。 所以明显的效果就是参数向量β 的每个分量尽量的小。

假定我们做线性估计$\hat{y} = X\hat{\beta}$， 那么 我们有：

$$PRSS(\beta_i, \lambda) = \sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} \beta_j^2$$

$$\hat{\beta}^{ridge} = \underset{\beta}{\operatorname{argmin}} \left\{ \sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \right\}$$

对此， 随着的$\lambda \geq 0$ 的增大， β_i全部收缩到0.

其实这个式子还有三个等价(equivalent)表达形式：

1. 矩阵形式：$PRSS(\beta, \lambda) = ||\mathrm{X}\beta - \mathrm{Y}||^2 + \lambda||\beta||^2$（假设$\beta_0 = 0$, $E[\hat{y}] = 0$） 。 由于这是二次凸函数， 因此求导。

$$\nabla PRSS(\beta) = 2\mathrm{X}^{\mathrm{T}}(\mathrm{X}\beta - \mathrm{Y}) = 0$$

$$(\mathrm{X}^{\mathrm{T}}\mathrm{X} + \lambda \mathrm{I})\beta = \mathrm{X}^{\mathrm{T}}\mathrm{Y}$$

由于$(XX^T + \lambda I)$始终可逆， 因此$\beta = (\mathrm{X}^{\mathrm{T}}\mathrm{X} + \lambda \mathrm{I})^{-1}\mathrm{X}^{\mathrm{T}}\mathrm{Y}$

2. 限制最值形式：

$$\underset{\beta}{\operatorname{argmin}} \sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2$$

s.t. $||\beta||^2 \leq \Lambda^2$ 我们做一些替换：

$$\underset{\beta}{\operatorname{argmin}} \sum_{i=1}^{n} \epsilon_i$$

s.t. $\epsilon_i = \beta' X_i + \beta_0 - y_i$, $\beta'\beta \leq \Lambda^2$

同样，我们假设Y的期望为零， 做无偏估计的时候， $\beta_0 = 0$ 根据拉格朗日定理，我们得到表达式

$$\mathbb{L}(\epsilon, \beta, \alpha, \lambda) = \sum_{i=1}^{n} \epsilon_i^2 + \sum_{i=1}^{n} \alpha_i(\epsilon_i + y_i - \beta' X_i) + \lambda(\beta'\beta - \Lambda^2)$$

根据KKT条件;

$$\nabla_\beta \mathbb{L} = -\sum_{i=1}^{n} \alpha_i X_i + 2\lambda\beta = 0 \Longleftrightarrow \beta = \frac{1}{2\lambda}\sum_{i=1}^{n} \alpha_i X_i = \frac{1}{2\lambda} X^T \alpha$$

$$\nabla_\epsilon \mathbb{L} = 2\epsilon_i + \alpha_i = 0 \Longleftrightarrow \epsilon_i = -\frac{\alpha_i}{2}$$

因此， 我们根据拉格朗日对偶问题把上面的最小值的问题转化为$\mathbb{L}(\epsilon, \beta, \alpha, \lambda)$ 求最大值的问题。 其中λ是变化量， 那么我们把根据KKT条件得到代入到$\mathbb{L}$从而得到关于$\mathbb{L}(\alpha, \lambda)$。

$$\begin{aligned}
&\mathbb{L}(\alpha,\lambda) = \sum^{n} \frac{\alpha_i^2}{4} + \sum^{n} \alpha_i(\frac{-\alpha_i}{2} + y_i - \frac{1}{2\lambda}\sum^{n} \alpha_i X_i' X_i) + \\
&\lambda((\frac{1}{2\lambda}\sum^{n} \alpha_i X_i)'(\frac{1}{2\lambda}\sum^{n} \alpha_i X_i) - \Lambda^2) \\
&= \frac{1}{4}\alpha'\alpha - \frac{1}{2}\alpha'\alpha + \alpha' Y - \frac{1}{2\lambda}\alpha' X^T X\alpha + \frac{1}{4\lambda}\alpha' X^T X\alpha - \lambda\Lambda^2 \\
&= -\frac{1}{4}\alpha'\alpha + \alpha' Y - \frac{1}{4\lambda}\alpha' X^T X\alpha - \lambda\Lambda^2
\end{aligned}$$

由此,

$$\nabla_\alpha \mathbb{L} = -\frac{1}{4}2\alpha + Y - \frac{1}{4\lambda}2XX^T\alpha = 0 \Longleftrightarrow \alpha = 2\lambda(\lambda I + XX^T)^{-1} Y$$

这样,

$$\beta = \frac{1}{2\lambda} X^T \alpha = \frac{1}{2\lambda} X^T 2\lambda(\lambda I + XX^T)^{-1} Y = X^T(\lambda I + XX^T)^{-1} Y$$

前面我们根据矩阵证明了$\beta = (X^T X + \lambda I)^{-1} X^T Y$， 我们如果证明了$(X^T X + \lambda I)^{-1} X^T = X^T(XX^T + \lambda I)^{-1}$ 那么， 这两个结果就是等价的。 我们发现，$(X^T X + \lambda I)X^T = X^T XX^T + \lambda X^T = X^T(XX^T + \lambda I)$ 我们两边同时左乘$(X^T X + \lambda I)^{-1}$和右乘$(XX^T + \lambda I)^{-1}$后， 就得到了我们的证明。

3. 扩展矩阵形式:

$$\begin{aligned} PRSS(\beta_i, \lambda) &= \sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \\ &= \sum_{i=1}^{n}(y_i - \beta_0 - \beta' X_i)^2 + \sum_{j=1}^{p}(0 - \sqrt{\lambda}\beta_j)^2 \end{aligned}$$

同样, 考虑无偏估计$\beta_0 = 0$ 的情况:

$$PRSS(\beta_i, \lambda) = (Y - X\beta)'(Y - X\beta) + (0I - \sqrt{\lambda}I)'(0I - \sqrt{\lambda}I)$$

假设我们有

$$X_\lambda = \begin{pmatrix} X \\ \sqrt{\lambda}I_p \end{pmatrix}_{(n+p)\times p}, Y_\lambda = \begin{pmatrix} Y \\ 0_p \end{pmatrix}_{n+p}$$

这样,

$$PRSS(\beta_i, \lambda) = (Y_\lambda - X_\lambda\beta)'(Y_\lambda - X_\lambda\beta)$$

我们直接利用最小二乘法的结论

$$\beta = (X_\lambda^T X_\lambda)^{-1} X_\lambda^T Y_\lambda$$

我们代入分块矩阵之后, 可以同样计算到

$$\beta = (X^T X + \lambda I_p)^{-1} X^T Y$$

由此, 根据等价表达式, 我们可以求出最小后的参数。可见, λ的引入后, 使得$(X^T X + \lambda I_p)$非奇异 (non-singular)从而逆始终存在。但是 λ如何选择呢? 有个方法叫岭轨迹(ridge traces)

还有一个要注意的是岭估计不是无偏的:

$$\begin{aligned} \hat{\beta}^{ridge} &= (X^T X + \lambda I)^{-1} X^T Y \\ &= (X^T X + \lambda I)^{-1}(X^T X)(X^T X)^{-1} X^T Y \\ &= (X^T X + \lambda I)^{-1}(X^T X)\hat{\beta}^{ls} \\ &= (X^T X + \lambda I)^{-1}(X^T X)\hat{\beta}^{ls} \\ &= (I + \lambda(X^T X)^{-1})^{-1}(X^T X)^{-1}(X^T X)\hat{\beta}^{ls} \\ &= (I + \lambda(X^T X)^{-1})^{-1}\hat{\beta}^{ls} \end{aligned}$$

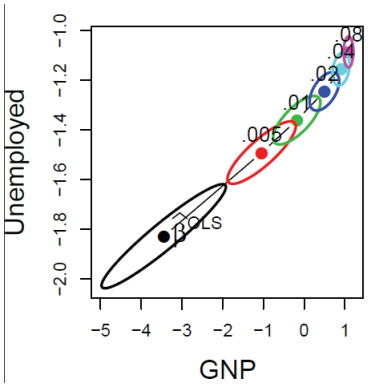


Figure 2.25: 岭轨迹(ridge traces)列子， 每种颜色一个参数连续变化

因为$E(\hat{\beta}^{ridge}) = (I + \lambda(X^TX)^{-1})^{-1}\beta \neq \beta$ 我们令$W_\lambda = (I + \lambda(X^TX)^{-1})$

$$
\begin{aligned}
&Var[\hat{\beta}^{ridge}] = Var[(I + \lambda(X^TX)^{-1})\hat{\beta}^{ls}] \\
&= W_\lambda Var[\hat{\beta}]W_\lambda^T \\
&= \sigma^2 W_\lambda (X^TX)^{-1} W_\lambda^T
\end{aligned}
$$

因此， $\hat{\beta}^{ridge} \leq \hat{\beta}$

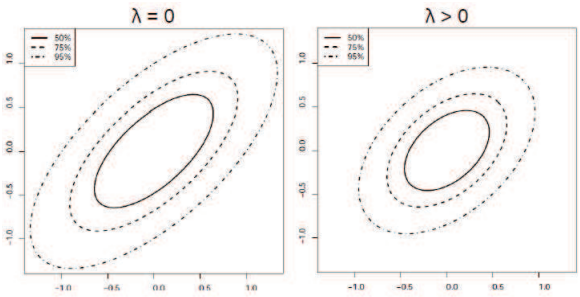


Figure 2.26: 对比最小二乘法岭轨迹是有偏估计，但是方差会更小些

奇异值分解

奇异值分解是把1个矩阵变成3个矩阵的乘积:

$$X = UDV^T$$

其中D是奇异值的对角阵， 并且$UU^T = I$和$VV^T = I$。

$$\begin{aligned}\hat{\beta} &= (X^T X)^{-1} X^T Y \\ &= (VDU^T UDV^T)^{-1} VDU^T Y \\ &= (VD^2 V^T)^{-1} VDU^T Y \\ &= VD^{-2} V^T VDU^T Y \\ &= VD^{-2} DU^T Y\end{aligned}$$

而岭估计的奇异值分解

$$\begin{aligned}\hat{\beta}^{ridge} &= (X^T X + \lambda I)^{-1} X^T Y \\ &= (VD^2 V^T + \lambda VV^T)^{-1} VDU^T Y \\ &= V(D^2 + \lambda I)^{-1} V^T VDU^T Y \\ &= V(D^2 + \lambda I)^{-1} DU^T Y\end{aligned}$$

如果令$(D)_{jj} = d_{jj}$ 那么比较得到:

$$d_{jj}^{-1} \geq fracd_{jj}d_{jj}^2 + \lambda$$

贝叶斯理解

岭回归可以理解为引入一个高斯先验后的贝叶斯估计。 假设满足高斯分布的$\mathrm{N}(0, \sigma^2)$。并且已知待估参数β满足先验概率:

$$\beta|\sigma^2 \sim \mathrm{N}(0, \frac{\sigma^2}{\lambda} I)$$

那么关于β和σ^2的后验概率,

$$\begin{aligned}f_{\beta,\sigma^2}(\beta, \sigma^2|X, Y) &\propto f_Y(Y|X, \beta, \sigma^2) f_\beta(\beta|\sigma^2) f_{\sigma^2}(\sigma^2) \\ &\propto \sigma^{-n} exp[-\frac{1}{2\sigma^2}(Y - X\beta)^T (Y - X\beta)] \\ &\times \sigma^{-p} exp[-\frac{\lambda}{2\sigma^2}\beta'\beta] \\ &\times [\sigma^2]^{-\alpha_0 - 1} exp[-\frac{\beta_0}{2\sigma^2}]\end{aligned}$$

假定是$\beta_0 = 0$的情况下， 那么根据最大似然估计:

$$\hat{\beta}_{MAP} = \underset{\beta}{\operatorname{argmax}}\, Ln(\beta, X, Y)$$

$$\begin{aligned} Ln(\beta, X, Y) &= log\left(\sigma^{-n} exp[-\frac{1}{2\sigma^2}(Y - X\beta)^T(Y - X\beta)] \times \sigma^{-p} exp[-\frac{\lambda}{2\sigma^2}\beta'\beta]\right) \\ &= -(n+p)log(\sigma) - \frac{1}{2\sigma^2}[(Y - X\beta)^T(Y - X\beta) + \lambda\beta'\beta] \end{aligned}$$

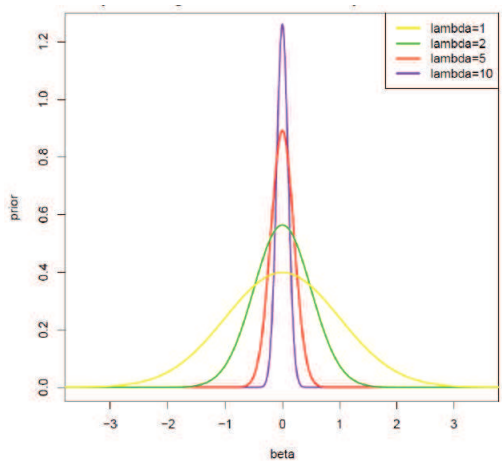


Figure 2.27: 惩罚系数越大， 对应先验的方差越小

更进一步， 后验概率也可以写成另外一个形式:

$$f_{\beta,\sigma^2}(\lambda, \sigma^2|X, Y) \propto f_\beta(\beta|\sigma^2, X, Y) \times f_{\sigma^2}(\sigma^2|X, Y)$$

其中β的分布为

$$f_\beta(\beta|\sigma^2, X, Y) = \sigma^k exp\left\{-\frac{1}{2\sigma^2}[\beta - \hat{\beta}^{ridge}]^T(X^TX + \lambda I)[\beta - \hat{\beta}^{ridge}]\right\}$$

其中$\hat{\beta}^{ridge} = (X^TX + \lambda I)^{-1}X^TY$

自由度分析

自由度(Degrees of Freedom)定义为估计的的迹(trace)。 由于:

$$\hat{\beta}^{ridge} = (X^TX + \lambda I)^{-1}X^TY$$

那么对结果的估算为:

$$\hat{Y} = X\hat{\beta}^{ridge} = X(X^TX + \lambda I)^{-1}X^TY$$

那么

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

自由度为(根据以前的奇异值分解):

$$tr[H(\lambda)] = tr[X(X^TX + \lambda I)^{-1}X^T] = \sum_{j=1}^{p} \frac{d_{jj}^2}{d_{jj}^2 + \lambda}$$

因此自由度是关于λ的单调减函数。 当$\lambda \to \infty$的时候， 自由度为0:

$$\lim_{\lambda\to\infty} tr[H(\lambda)] = \lim_{\lambda\to\infty} \sum_{j=1}^{p} \frac{d_{jj}^2}{d_{jj}^2 + \lambda} = 0$$

2.4.5 LASSO

最小绝对值收缩和选择器 LASSO (Least Absolute Shrinkage and Selection Operator)是L^1的惩罚回归。

$$\begin{aligned} L(\beta, \lambda) &= ||Y - X\beta||_2^2 + \lambda||\beta||_1 \\ &= \sum_{i=1}^{n}(y_i - \beta' X_i)^2 + \lambda \sum_{j=1}^{p} |\beta_i| \end{aligned}$$

类似的限制形式如下:

$$L(\beta, \lambda) = ||Y - X\beta||_2^2 + \lambda||\beta||_1$$

s.t. $\sum_{j=1}^{p} |\beta_i| \leq C$

由于$|\beta|$的导数不连续，LASSO 没有闭合公式表示(closed form)， 因此经常用近似计算， 利用岭回归近似估算:

$$|\beta| = |\beta_0| + \frac{1}{2|\beta_0|}(\beta^2 - \beta_0^2)$$

计算中我们用顺序近似的方式计算:

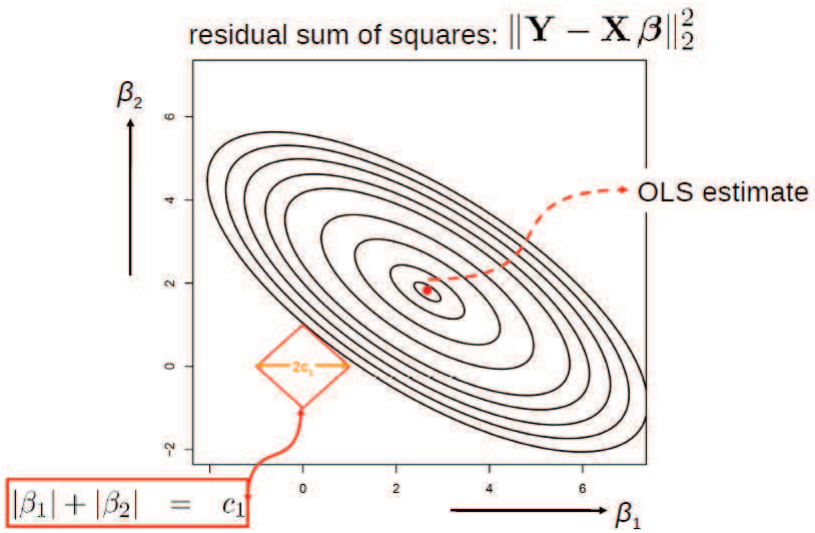


Figure 2.28: 二维的LASSO估算，并且有个估算为0

$$
\begin{aligned}
&||Y - X\beta^{(k+1)}||_2^2 + \lambda||\beta^{(k+1)}|| \\
&\simeq ||Y - X\beta^{(k+1)}||_2^2 + \lambda(||\beta^k|| + \frac{1}{2}\sum_j^p \frac{[\beta_j^{(k+1)}]^2 - [\beta_j^k]^2}{|\beta_j^k|}) \\
&= (Y - X\beta^{(k+1)})'(Y - X\beta^{(k+1)}) + \frac{\lambda}{2}\sum_j^p \frac{[\beta_j^{(k+1)}]^2}{|\beta_j^k|} + \frac{\lambda}{2}\sum_j^p \frac{[\beta_j^k]^2}{|\beta_j^k|} \\
&\propto (Y - X\beta^{(k+1)})'(Y - X\beta^{(k+1)}) + \frac{\lambda}{2}\sum_j^p \frac{1}{|\beta_j^k|}[\beta_j^{(k+1)}]^2
\end{aligned}
$$

类似岭回归的公式， 那么我们有：

$$\beta^{(k+1)} = \{X^T X + \lambda\Psi[\beta^k]\}^{-1} X^T Y$$

其中$diag\{\Psi[\beta^k]\} = (1/|\beta_1^k|, 1/|\beta_2^k|, \ldots, 1/|\beta_p^k|)$

有了这个递推式， 我们就可以迭代计算了。 在迭代计算的时候， 可以利用梯度递减(Gradient Descent)等能加速收敛的方 法。

贝叶斯理解

LASSO回归可以理解为引入一个拉普拉斯先验 （Laplace distribution）后的贝叶斯估计。 假设满足高斯分布的Laplace$(0, 1/\lambda)$。并且已知待估参

数β满足先验概率:

$$\beta_j|\sigma^2 \sim \text{Laplace}(0, \frac{\sigma^2}{\lambda}) = \frac{\lambda}{2\sigma^2}\exp\left(-\frac{\lambda}{\sigma^2}|x|\right)$$

那么关于β和σ^2的后验概率,

$$\begin{aligned} f_{\beta,\sigma^2}(\beta,\sigma^2|X,Y) &\propto f_Y(Y|X,\beta,\sigma^2) f_\beta(\beta|\sigma^2) f_{\sigma^2}(\sigma^2) \\ &\propto \sigma^{-n} exp[-\frac{1}{2\sigma^2}(Y-X\beta)^T(Y-X\beta)] \\ &\times (\frac{\lambda}{2\sigma^2})^p \exp[-\lambda\sum_j \frac{|\beta_j|}{\sigma^2}] \\ &\times [\sigma^2]^{-\alpha_0-1}\exp[-\frac{\beta_0}{2\sigma^2}] \end{aligned}$$

假定是$\beta_0 = 0$的情况下, 那么根据最大似然估计:

$$\hat{\beta}_{MAP} = \underset{\beta}{\text{argmax}}\, Ln(\beta, X, Y)$$

$$\begin{aligned} Ln(\beta,X,Y) &= log\left(\sigma^{-n}exp[-\frac{1}{2\sigma^2}(Y-X\beta)^T(Y-X\beta)] \times (\frac{\lambda}{2\sigma^2})^p \exp[-\frac{\lambda}{\sigma^2}\sum_j|\beta_j|]\right) \\ &= -nlog(\sigma) + plog(\frac{\lambda}{2\sigma^2}) - \frac{1}{2\sigma^2}[(Y-X\beta)^T(Y-X\beta)] - \frac{\lambda}{\sigma^2}\sum_j|\beta_j| \end{aligned}$$

Ridge VS LASSO

由于优化的中惩罚的L^1和L^2的差别。 那么Ridge在一定的惩罚系数范围内, 优先考虑圆内的点(根据等价形式$||\beta||_2^2 < \Lambda^2$的限制), 然后 优化方差最小。 而LASSO优先考虑菱形范围内的点（根据等价形式$||\beta||_1 < C$）, 再考虑方差最小（图**??**）。 所以LASSO优先考 虑轴上的点, 然后考虑方差最小。

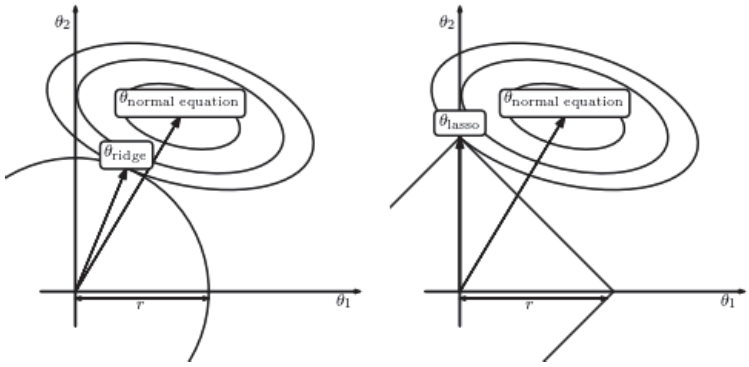


Figure 2.29: LASSO和Ridge对比示意图

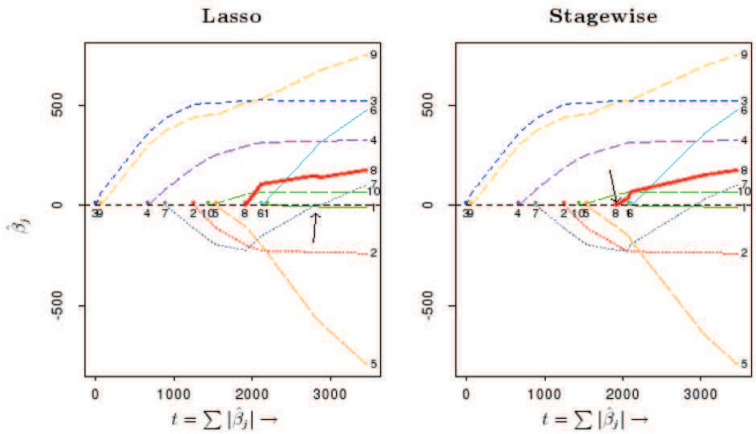


Figure 2.30: 随着样本数的增加的参数变化过程，LASSO和Stagewise接近

2.4.6 LAR

由于Lasso和Stagewise算法在随样本增加不停的修正参数的过程中， 有很类似 的表现（图??）。那么到底什么使得他们如此类似呢。 而LAR的提出使得。 Stagewise和Lasso多成为一种LAR的特例。

LAR的意思是最小角度回归，其实是指在回归路径选择过程中， 到每个样本轴的 角度相等并且最小。 或者说是对角线。

2.5 异常检测

http://dnene.bitbucket.org/docs/mlclass-notes/lecture16.html

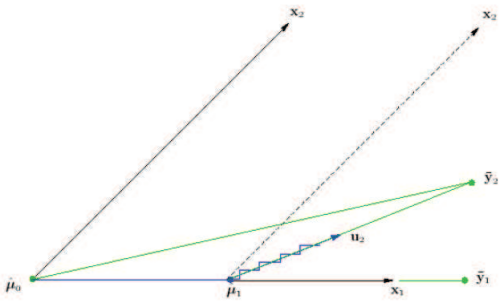


Figure 2.31: 随着逐步的发现多个样本轴有相同的相关度， 路径沿着对角线前进

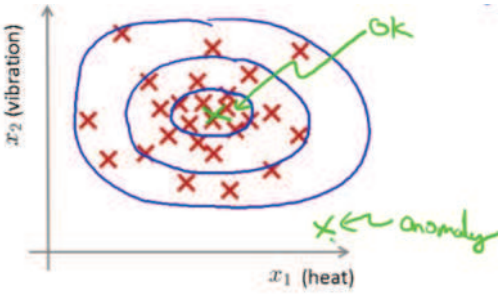


Figure 2.32: 异常检测示意

2.6 机器学习算法

常见数据挖掘算法有CART, C4.5, k-Means, SVM, Apriori, EM, PageRank, AdaBoost, kNN, Naive Bayes.

2.6.1 CART

分类和回归树(Classification And Regression Tree (CART))算法(1984)是加州伯克 利的大牛 Leo Breiman首先提出来的。1938年Jerzy Neyman在44岁的时候， 来 到加州伯克利大学。 当时美国还没有统计系， 他离开伦敦的统计系后以数学系 的教授的身份在伯克利工作， 以后他一直未伯克利的统计系独立而奋斗。 从而 开创了美国统计时代。 当时Breiman的老师Michel Loeve， 他是法国著名统计 学家Paul Levy的学生，在Neyman到美国十年后1948年的时候也从伦敦访问了哥 伦比亚。 当时伯克利已经有了自己培养的6位老师。 Michel Loeve是另外的4名 外来和尚之一。 这十个人是伯克利开统计系的老师。 Breiman就是统计系毕业 的学生。 这个自己毕业的早期学生最后成为了这个系的大牛。

Breiman是1980年加入伯克利教师队伍的， 他喜欢单枪匹马的战斗，