天善智能
TIANSHAN SOFT

从统计学到机器学习的算法基础1
-参数估计与模型调优

《Python数据科学：全栈技术详解》

讲师：Ben

自我介绍

- 天善商业智能和大数据社区　　讲师 –Ben
- 天善社区 ID - Ben_Chang
- https://www.hellobi.com – 学习过程中有任何相关的问题都可以提到技术社区数据挖掘版块。

天善智能
TIANSHAN SOFT

主要内容

- 随机变量参数的点估计
  - 矩估计
  - 极大似然估计
- 统计学习的极大似然估计
  - 线性回归的极大似然估计
  - 逻辑回归的极大似然估计
- 模型调优
  - 选取最优模型的标准
  - 模型的内部有效与外部有效
  - 机器学习模型调优的方案

随机变量参数的点估计

统计学推断的基本问题

- 统计推断的基本问题
  - 参数估计问题
    - 点估计
    - 区间估计
  - 假设检验问题

参考资料：经济数学基础(第3分册概率统计)龚德恩,2006,重点阅读第5章，从第115页开始

什么是参数估计

参数是刻画总体某方面的概率特性的数量。

当这个数量是未知的时候，从总体抽出一个样本，用某种方法对这个未知参数进行估计就是参数估计。

例如，$X \sim N(\mu, \sigma^2)$,

若μ, σ^2未知，通过构造样本的函数, 给出它们的估计值或取值范围就是参数估计的内容.

点估计　区间估计

天善智能 TIANSHAN SOFT

参数估计的类型

点估计 —— 估计未知参数的值

- 矩估计
- 极大似然估计
- 最小二乘估计法
- 贝叶斯估计

区间估计—— 估计未知参数的取值范围，使得这个范围包含未知参数真值的概率为给定的值。

矩估计

用**样本的 k 阶矩作为总体的 k 阶矩**的估计量, **建立含待估计参数的方程**，从而可解出待估计参数

一般地，不论总体服从什么分布，总体期望 μ 与方差 σ^2 存在，则根据矩估计法它们的**矩估计量**分别为：

一阶矩： $\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} X_i = \bar{X}$

r阶矩： $B_r = \frac{1}{n}\sum_{i=1}^{n} X_i^r$

二阶矩： $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2 = S_n^2$

$\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2 = S^2$ 是无偏矩估计

矩估计

基本原理：总体矩是反映总体分布的最简单的数字特征，当总体含有待估计参数时，总体矩是待估计参数的函数。

样本取自总体，　$A_k \xrightarrow{P} E(X^k)$　k=1,2,⋯.

样本矩在一定程度上可以逼近总体矩，

故用样本矩来估计总体矩

矩估计(conj.)

设总体X的分布函数为 $F(x;\theta_1,\ldots,\theta_k)$ 其中 $\theta_1,\ldots,\theta_k$ 是待估参数。

$X_1,\ldots,X_k$ 为X的样本，其服从F分布。 设总体的k阶矩

$E(X^k)=\mu_k$ k=1,2,⋯. 存在，则样本的k阶矩

$$\boxed{A_k=\frac{1}{n}\sum_{i=1}^{n}X_i^k} \xrightarrow{P} \boxed{\mu_k}$$ (由大数定理)

令 $A_l=\mu_l$ l=1,2,⋯k. $\Longrightarrow$ k个方程组，

从中解得$\hat{\theta}_1,\ldots,\hat{\theta}_k$ 即为矩估计量。

矩估计量的观察值称为矩估计值。

天善智能 TIANSHAN SOFT

例1 设总体X的均值μ，方差σ都存在，且$\sigma^2>0$，

但μ，σ^2未知，又设X_1,Λ,X_n是一个样本；

求：μ,σ^2的矩估计量。

解: $\mu_1=EX=\mu, \mu_2=E(X^2)=DX+(EX)^2=\sigma^2+\mu^2$

令 $\mu_1=A_1$， $\mu_2=A_2$， 即 $\mu=A_1$， $\sigma^2+\mu^2=A_2$，

所以 $\hat{\mu}=A_1=\bar{X}$，

$$\hat{\sigma}^2=A_2-A_1^2=\frac{1}{n}\sum_{i=1}^{n}X_i^2-\bar{X}^2=\frac{1}{n}\sum_{i=1}^{n}(X_i-\bar{X})^2=\frac{n-1}{n}S^2$$

特别，若 $X\sim N(\mu,\sigma^2)$，μ,σ^2未知；

则 $\hat{\mu}=\bar{X}$， $\hat{\sigma}^2=\frac{1}{n}\sum_{i=1}^{n}(X_i-\bar{X})^2$

矩估计(conj.)-正态分布的矩估计示例

例2 有一批零件，其长度$X\sim N(\mu,\sigma^2)$，现从中任取4件，测的长度(单位：mm)为12.6, 13.4, 12.8, 13.2。试估计μ和σ^2的值。

解： 由

$$\bar{x}=\frac{1}{4}(12.6+13.4+12.8+13.2)=13$$

$$s^2=\frac{1}{4-1}[(12.6-13)^2+(13.4-13)^2+(12.8-13)^2$$

$$+(13.2-13)^2]=0.133$$

得μ和σ^2的估计值分别为13(mm)和0.133(mm)^2

例3 不合格品率 p 的矩法估计

设某车间生产一批产品，为估计该批产品不合格品率，抽取了 n 件产品进行检查.

分析 设总体 X 为抽的不合格产品数，相当于抽取了一组样本 X_1，X_2，…，X_n，且

$$X_i=\begin{cases}1, & \text{第}i\text{次取到不合格品；}\\ 0, & \text{第}i\text{次取到合格品.}\end{cases} \quad i=1,2,\cdots n.$$

解 因 $p=EX$，故 p 的矩估计量为

$$\hat{p}=\overline{X}=\frac{1}{n}\sum_{i=1}^{n}X_i=f_n(A)$$

(即出现不合格产品的频率).

矩估计(conj.)-0-1分布的矩估计示例

例4 作了一次营销活动，营销了1000人 。事后统计结果，120人购买，其余人没有购买。请分别用矩估计法、极大似然估计法计算这个随机事件分布的参数（提示：该随机事件服从伯努利分布。

分析：令伯努利分布的参数为营销后响应的概率(p)，其分布为B(1000,p)

解： 由

$$\hat{p} = \overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i = f_n(A)$$

则P的估计值为120/1000=0.12

矩估计的优缺点

矩估计的优点

- 不依赖总体的分布，简便易行
- 只要n充分大，精确度也很高。

矩估计的缺点

- 矩估计的精度较差；
- **要求总体的某个k阶矩存在；**
- 要求未知参数能写为总体的原点矩的函数形式

例 考虑Cauchy分布，其密度函数为

$$f(x,\theta)=\frac{1}{\pi(1+(x-\theta)^2)},-\infty<x<+\infty,$$

其各阶矩均不存在。

极大似然估计

定义:(1)设随机变量X的概率密度函数为f(x, θ)， 其中θ为未知参数(f为已知函数).

$x_1, x_2, \cdots, x_n$ 为样本 $X_1, X_2, \cdots, X_n$ 的样本观察值,

$$L(\theta) = L(x_1, x_2, \cdots, x_n; \theta) = \prod_{i=1}^{n} f(x_i; \theta)$$

称 $L(x_1, x_2, \cdots, x_n; \theta)$ 为 X关于样本观察值 $x_1, x_2, \cdots, x_n$ 的**似然函数**。

(2)若X是离散型随机变量，似然函数定义为

$$L(x_1, x_2, \cdots, x_n; \theta) = \prod_{i=1}^{n} P(X_i = x_i)$$

极大似然估计(conj.)

如果似然函数 $L(\theta)=L(x_1,x_2,\cdots,x_n;\theta)$ 在 $\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}$ 时达到最大值, 则称 $\hat{\theta}$ 是参数θ的极大似然估计。

伯努利分布的参数计算

伯努利分布

- 从概率的角度来看，“逾期”是一个随机事件。如何刻画它的随机性？
- 伯努利分布：一种离散分布，用于表示0-1型事件发生的概率

例：$P(逾期)=\mathrm{p}$，$P(不逾期)=1-\mathrm{p}$

合并起来，可以是

$$P(\mathrm{Y}=\mathrm{y})=p^{y}(1-p)^{1-y}$$

$$y=\begin{cases}1, & 逾期\\0, & 不逾期\end{cases}$$

伯努利分布的参数计算

伯努利分布(续)

似然函数和对数似然函数

一组申请者在表现期的逾期状态为$\{y_1, y_2, \dots, y_n\}$, $y_i \in \{0,1\}$,似然函数和对数似然函数是

$$L(p) = \prod P(Y = y_i) = \prod p^{y_i}(1-p)^{1-y_i}$$

$$l(p) = \log\big(L(p)\big) = \log\left\{\prod P(Y = y_i) = \prod p^{y_i}(1-p)^{1-y_i}\right\}$$

$$= \sum y_i \log(p) + (1-y_i)\log(1-p) \qquad (1)$$

参数估计

$$\hat{p} = argmax\, l(p)$$

对(1)求关于p的一阶导数并等于0，有

$$\hat{p} = \frac{\sum y_i}{n}$$

极大似然估计的优缺点

- 极大似然估计的优点　利用了分布函数形式, 得到的估计量的精度一般较高。
- 极大似然估计的缺点　要求必须知道总体的分布函数形式

天善智能 TIANSHAN SOFT

统计学习的极大似然估计

线性回归估计方法介绍

矩估计：矩方法的基本思想是将总体矩条件换成样本矩条件，这些条件往往是我们假设的。

最小二乘法：是y服从正态分布情况下，极大似然发的一个特例。同时，是x和ε不相关假设下，矩估计的一个特例。

极大似然估计：我们已经收集到由(x_i,y_i)组成的样本，我们不知道的是描述x和y之间关系的参数，和y分布的参数。虽然我们不知道这些参数，但是我们可以对这些参数的性质进行假设。1）这些参数应该使得我们得到(x_i,y_i)这组样本的可能性最大；2） y要符合某一个给定分布，比如正态分布（线性回归）。

线性回归的方法推导：最小二乘法

基本思想要找到一条直线，使残差平方和最小。

拟合值

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

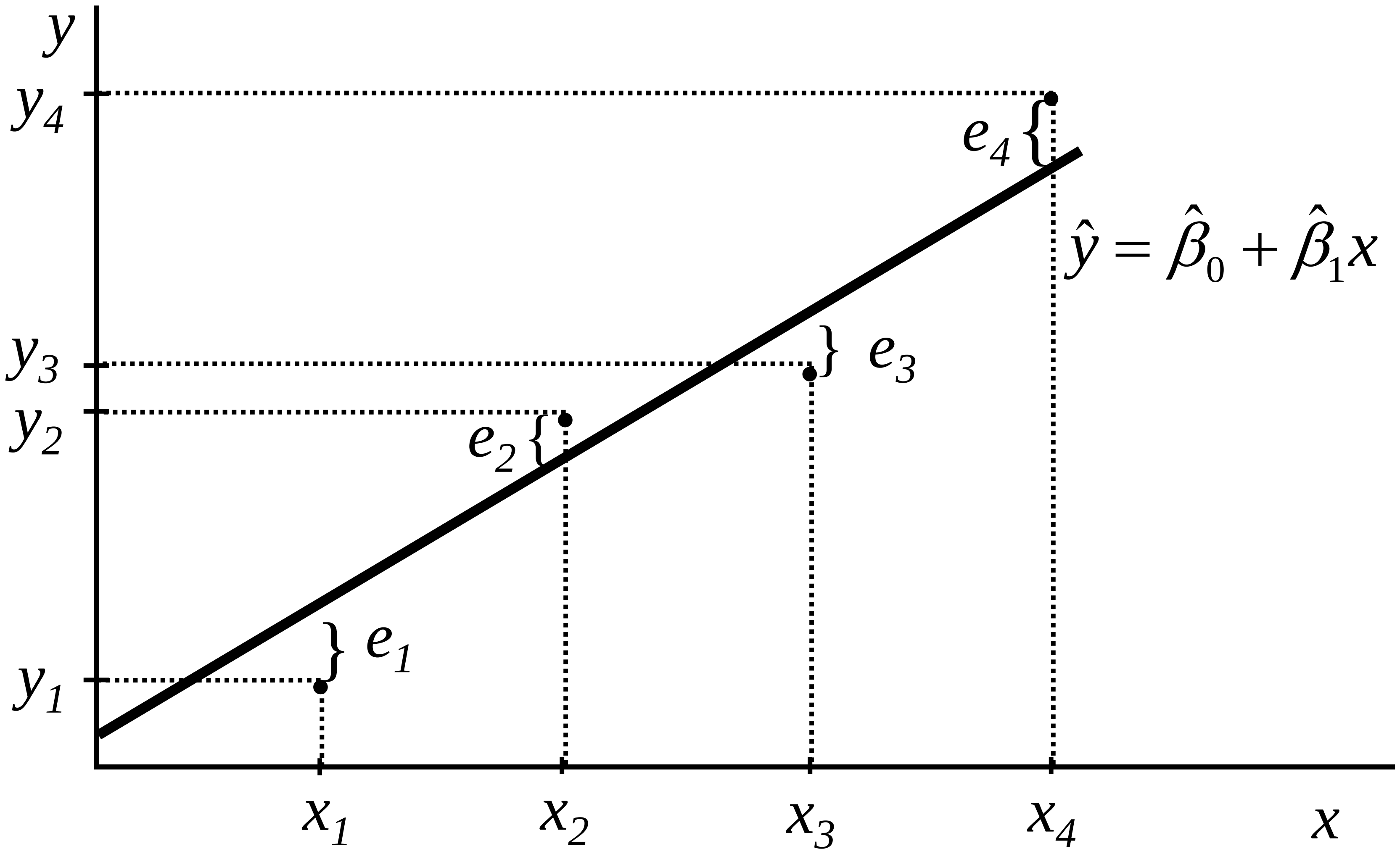
残差是对误差项的估计，因此，它是拟合直线（样本回归函数）和样本点之间的距离。

$$\mathrm{e}_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i$$

天善智能 TIANSHAN SOFT

线性回归的方法推导：最小二乘法

样本回归线，样本数据点和相关的误差估计项



线性回归的方法推导：最小二乘法

- 解一个最小化问题：

$$\sum_{i=1}^{n}(e_i)^2=\sum_{i=1}^{n}\left(y_i-\hat{\beta}_0-\hat{\beta}_1x_i\right)^2$$

- 利用微积分，对两个参数分别求导，得到两个一阶条件

$$\partial\sum_{i=1}^{n}(e_i)^2\Big/\partial\hat{\beta}_0 \Rightarrow \sum_{i=1}^{n}\left(y_i-\hat{\beta}_0-\hat{\beta}_1x_i\right)=0$$

$$\partial\sum_{i=1}^{n}(e_i)^2\Big/\partial\hat{\beta}_1 \Rightarrow \sum_{i=1}^{n}x_i\left(y_i-\hat{\beta}_0-\hat{\beta}_1x_i\right)=0$$

线性回归的方法推导：最小二乘法

解方程组得到：

- 斜率系数：x 和 y 的协方差除以x的方差。

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y}) / (n-1)}{\sum_{i=1}^{n}(x_i - \overline{x})^2 / (n-1)} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2}$$

- 截距估计量：样本回归线穿过样本均值

$$\hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x}$$

线性回归的方法推导：极大似然估计

线性回归中，假设扰动项服从正态分布

$$y_i = \beta x_i + \varepsilon_i,\ \varepsilon_i \sim i.i.d.N\left(0, \sigma^2\right)$$

其中回归系数β和扰动项的方差 σ^2 为参数，这些参数应该使得获得这些(x_i, y_i)组成的样本的可能性最大，因此有：

$$\begin{aligned} L\left(\beta, \sigma^2\right) &= f\left(y_1, y_2, \cdots, y_n | \beta, \sigma^2\right) \\ &= \prod_{i=1}^{n} f\left(y_i | \beta, \sigma^2\right), \end{aligned}$$

—— （1式）

天善智能 TIANSHAN SOFT

线性回归的方法推导：极大似然估计

既然扰动项服从正态分布，那么y_i应该也服从正态分布：

$$y_i \sim i.i.d.N\left(\beta x_i, \sigma^2\right)$$

将这个条件带入（1式），得到似然函数：

$$\begin{aligned} L &= \prod_{i=1}^{n}\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)\exp\left[-\frac{1}{2\sigma^2}\left(y_i-\beta x_i\right)^2\right] \\ &= \left(2\pi\sigma^2\right)^{-\frac{n}{2}}\exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(y_i-\beta x_i\right)^2\right] \end{aligned}$$

——　（2式）

天善智能 TIANSHAN SOFT

线性回归的方法推导：极大似然估计

又到了求极值的时候，但是（2式）显然不好求导，数学家已经证明，一个函数取对数后的极值点和原始的极值点位置相同，因此对（2式）等号两边同时取对数：

$$\ln L = -\frac{n}{2}\ln 2\pi - \frac{n}{2}\ln\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \beta x_i)^2 \quad \text{（3式）}$$

然后分别对 β 和 σ^2求偏导，得到同样的公式：

$$\frac{\partial \ln L}{\partial \beta} = -\frac{1}{\sigma^2}\sum_{i=1}^{n}\left(\beta x_i^2 - x_i y_i\right) = 0,$$

$$\frac{\partial \ln L}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2\left(\sigma^2\right)^2}\sum_{i=1}^{n}(y_i - \beta x_i)^2 = 0$$

$$\Rightarrow \quad \widetilde{\beta} = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2}$$

线性回归的正则方法

岭回归：惩罚项的形式为斜率系数的平方

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

Lasso：惩罚项的形式为斜率系数的绝对值

$$\sum_{i=1}^{n}\left(y_i-\beta_0-\sum_{j=1}^{p}x_{ij}\beta_j\right)^2+\lambda\sum_{j=1}^{p}|\beta_j|$$

逻辑回归的方法推导：极大似然估计

假设我们在推销Ipad，每个消费者都有一个效用函数，消费者对Ipad的需求受一些解释变量的影响，比如阅读的次数、玩游戏的次数等等。我们 y^* 用来代表效用函数，它是x的线性函数。

$$y^* = \mathbf{x}'\boldsymbol{\beta} + \varepsilon.$$

但是我们作为商家，是不知道消费者的效用函数的，我们只能知道他是否购买Ipad，用y来表示观测结果。为了简单起见，我们假设ipad的价格为0.

$$y = \begin{cases} 1 & \text{if } y^* > 0, \\ 0 & \text{if } y^* \le 0. \end{cases}$$

天善智能 TIANSHAN SOFT

逻辑回归的方法推导：极大似然估计

这里面y^*被称为隐变量（latent variable）。接下来我们构造似然函数。

购买Ipad客户的概率为：

$$\begin{aligned}\Pr(y=1|\boldsymbol{\beta},\sigma^2,\mathbf{x}) &= \Pr(y^*>0|\boldsymbol{\beta},\sigma^2,\mathbf{x})\\ &= \Pr(\mathbf{x}'\boldsymbol{\beta}+\varepsilon>0|\boldsymbol{\beta},\sigma^2,\mathbf{x})\\ &= \Pr(\varepsilon>-\mathbf{x}'\boldsymbol{\beta}|\boldsymbol{\beta},\sigma^2,\mathbf{x})\\ &= 1-F(-\mathbf{x}'\boldsymbol{\beta}),\end{aligned}$$

其中F(.)为扰动项ε的累积概率密度函数。

不购买Ipad客户的概率为：

$$\begin{aligned}\Pr\left(y=0|\boldsymbol{\beta},\sigma^2,\mathbf{x}\right) &= 1-\Pr\left(y=1|\boldsymbol{\beta},\sigma^2,\mathbf{x}\right)\\ &= F\left(-\mathbf{x}'\boldsymbol{\beta}\right).\end{aligned}$$

逻辑回归的方法推导：极大似然估计

得到似然函数为：

$$\prod_{y=0} F\left(-\mathbf{x}'\boldsymbol{\beta}\right) \prod_{y=1} \left[1 - F\left(-\mathbf{x}'\boldsymbol{\beta}\right)\right] \quad —— \quad （1式）$$

当假设扰动项ε服从逻辑分布时，则累积概率密度函数为：

$$F\left(-\mathbf{x}'\boldsymbol{\beta}\right) = \frac{\exp\left(-\mathbf{x}'\boldsymbol{\beta}\right)}{1 + \exp\left(-\mathbf{x}'\boldsymbol{\beta}\right)} = \frac{1}{1 + \exp(\mathbf{x}'\boldsymbol{\beta})} \quad —— \quad （2式）$$

和

$$1 - F(-\mathbf{x}'\boldsymbol{\beta}) = \frac{\exp(\mathbf{x}'\boldsymbol{\beta})}{1 + \exp(\mathbf{x}'\boldsymbol{\beta})} \quad —— \quad （3式）$$

将（2式）和（3式）带入（1式），则得到逻辑回归的似然函数。其求解过程和线性回归的极大似然估计完全一样。

天善智能 TIANSHAN SOFT

逻辑回归的方法推导：极大似然估计

得到对数似然函数为：

$$\ln L(Y,\beta)=\sum_{i=1}^{n} y_i \underset{\sim}{x}_i^T\beta-\sum_{i=1}^{n}\ln\left[1+\exp(\underset{\sim}{x}_i^T\beta)\right]$$

对参数求偏导为：

$$\frac{d\ln L(Y,\beta)}{d\beta}=\sum_i y_i \underset{\sim}{x}_i^T-\sum_i\left[\frac{\exp\left(\underset{\sim}{x}_i^T\beta\right)}{1+\exp\left(\underset{\sim}{x}_i^T\beta\right)}\right]\underset{\sim}{x}_i^T=0$$

不过这个上面的这个式子没有解析解，一般使用Newton-Raphson方法进行数值计算。

正则化的逻辑回归

岭回归（L2正则）：

$$\min_{w,c} \frac{1}{2} w^T w + C \sum_{i=1}^{n} \log(\exp(-y_i(X_i^T w + c)) + 1)$$

Lasso（L1正则）：

$$\min_{w,c} \|w\|_1 + C \sum_{i=1}^{n} \log(\exp(-y_i(X_i^T w + c)) + 1)$$

模型调优

数值预测模型评估指标

模型建立后，一定要经过检验才能判定其是否合理，只有通过检验的模型才能用来作预测

绝对指标：

1．预测误差 e $$e_i = y_i - \hat{y}_i$$

2．平均绝对误差 MAE $$MAE = \frac{1}{n}\sum |e_i|$$

3．均方差 MSE $$MSE = \frac{1}{n-1}\sum e_i^2$$

相对指标：

1．百分误差 PE $$pe_i = \frac{e_i}{y_i}$$

2．平均绝对百分对误差 MAPE $$MAPE = \frac{1}{n}\sum |pe_i|$$

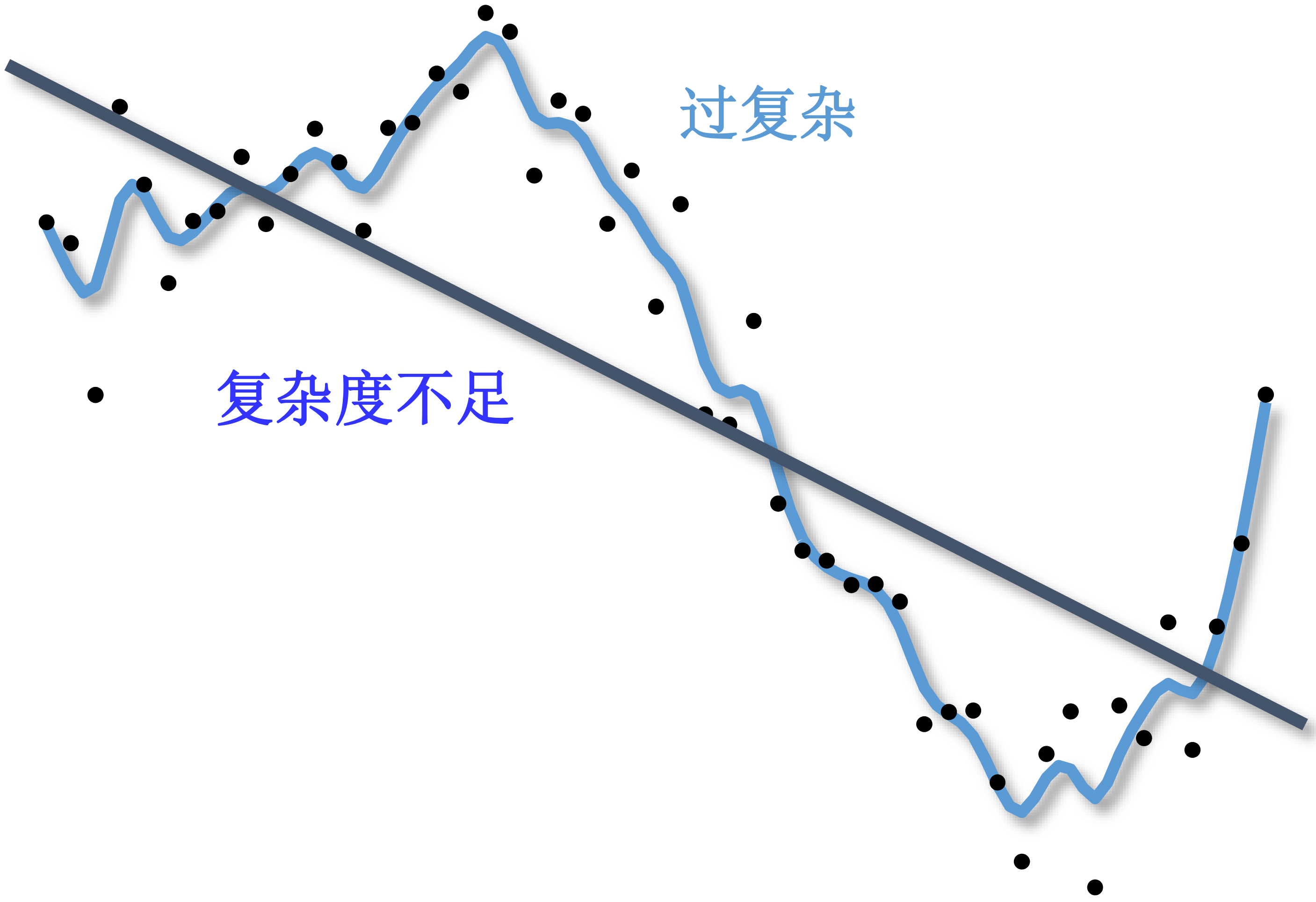
天善智能 TIANSHAN SOFT

二分类模型评估指标

预测类型	统计量
决策（Decisions ）	正确率、召回率 精确度、F1分数
排序（Rankings ）	ROC 指标 (一致性) Gini 指数 K-S统计量 提升度

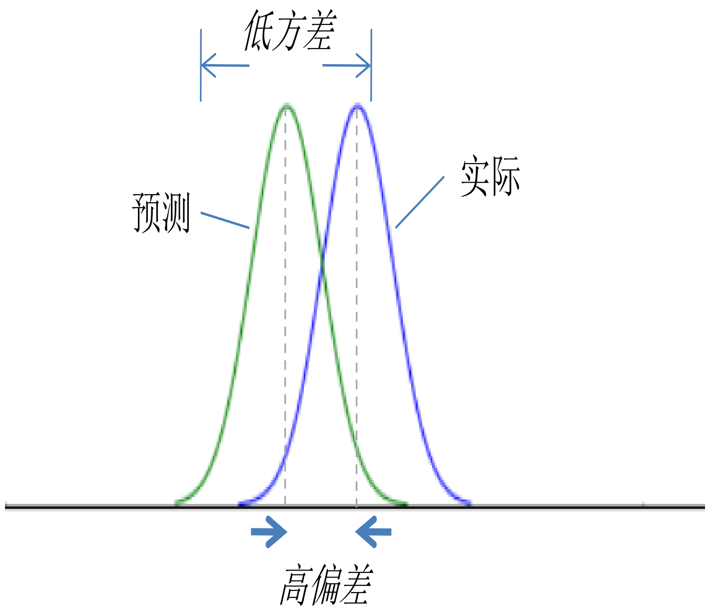
模型的内部有效与外部有效

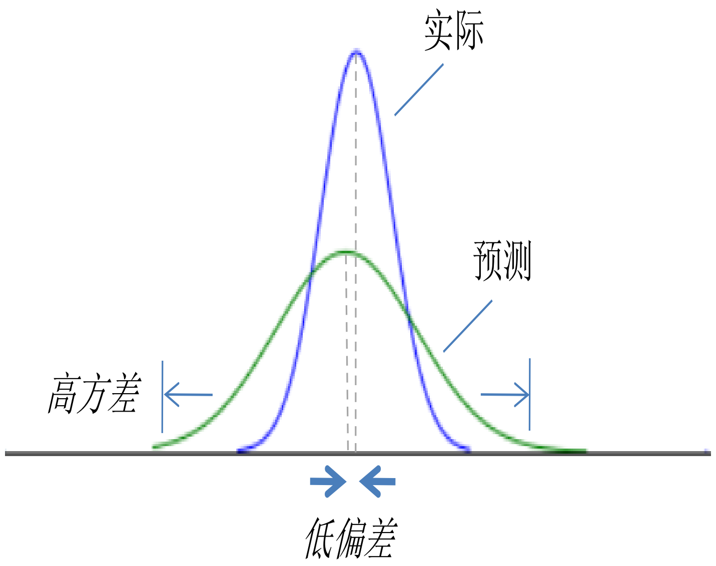
模型复杂度



模型的内部有效与外部有效

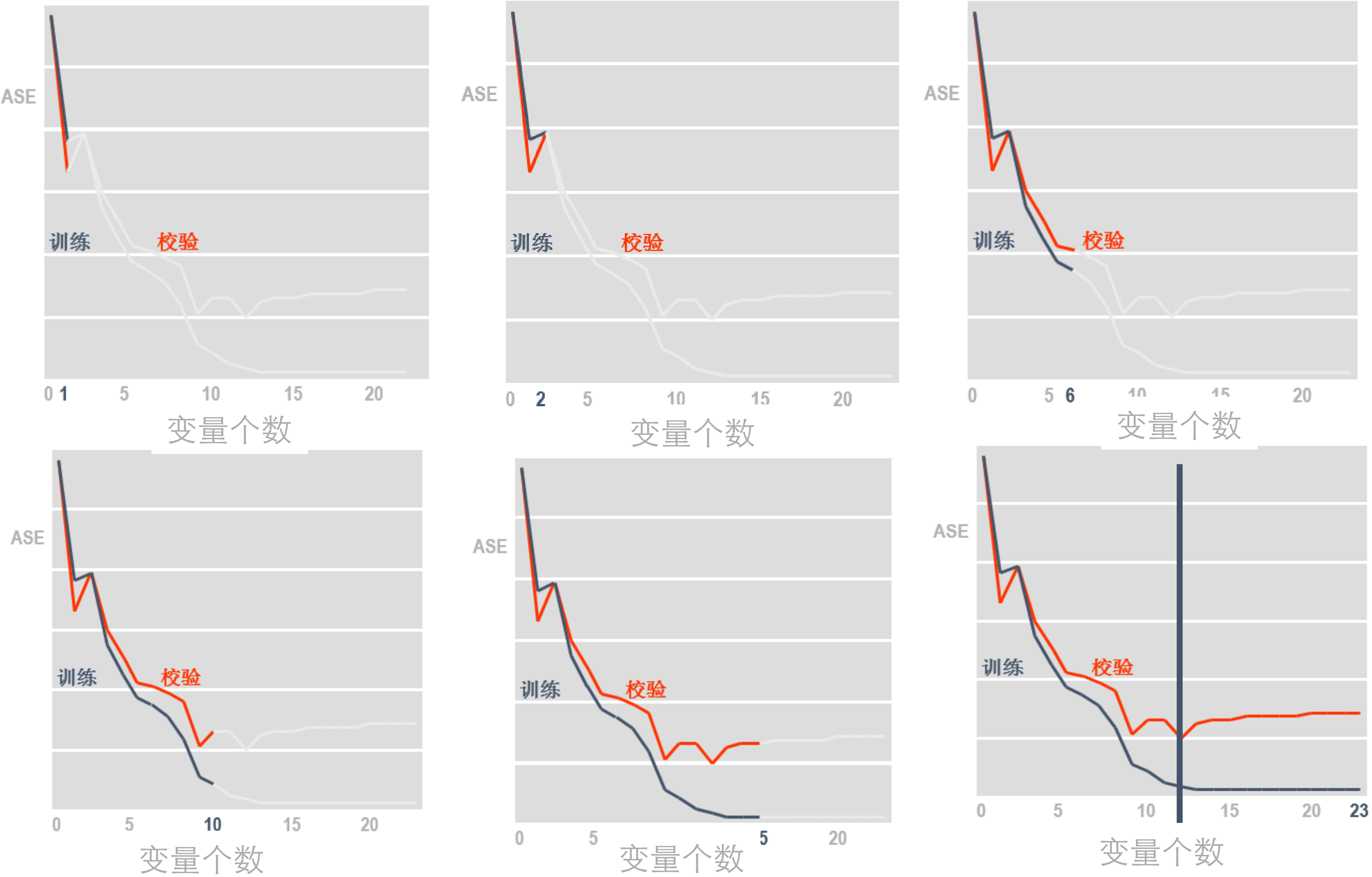
偏差-方差权衡





天善智能
TIANSHAN SOFT

机器学习模型调优的方案



调整模型的超参数，使得模型由简单至复杂。根据模型在校验数据集上的表现，确定最合适的超参数。

Lasso

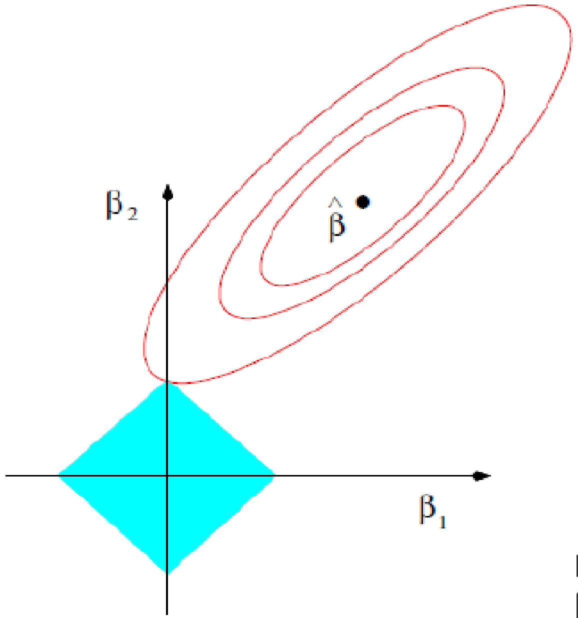
在线性模型中，人们必须选择合适的变量；比如常用的逐步回归法就是选择显著的变量而抛弃那些不显著的。Tibshirani(1996)提出了一个新的方法来处理变量选择的问题。该方法在模型系数绝对值的和小于某常数的条件下，谋求残差平方和最小。该方法既提供了如子集选择方法那样的可以解释的模型，也具有岭回归那样的稳定性。它不删除变量，但使得一些回归系数收缩、变小，甚至为0。因而，该方法被称为lasso(least absolute shrinkage and selection operator，最小绝对值收缩和选择算子。

$$\min_{w,c} \|w\|_1 + C\sum_{i=1}^{n}\log(\exp(-y_i(X_i^T w + c)) + 1)$$

惩罚项

超参数

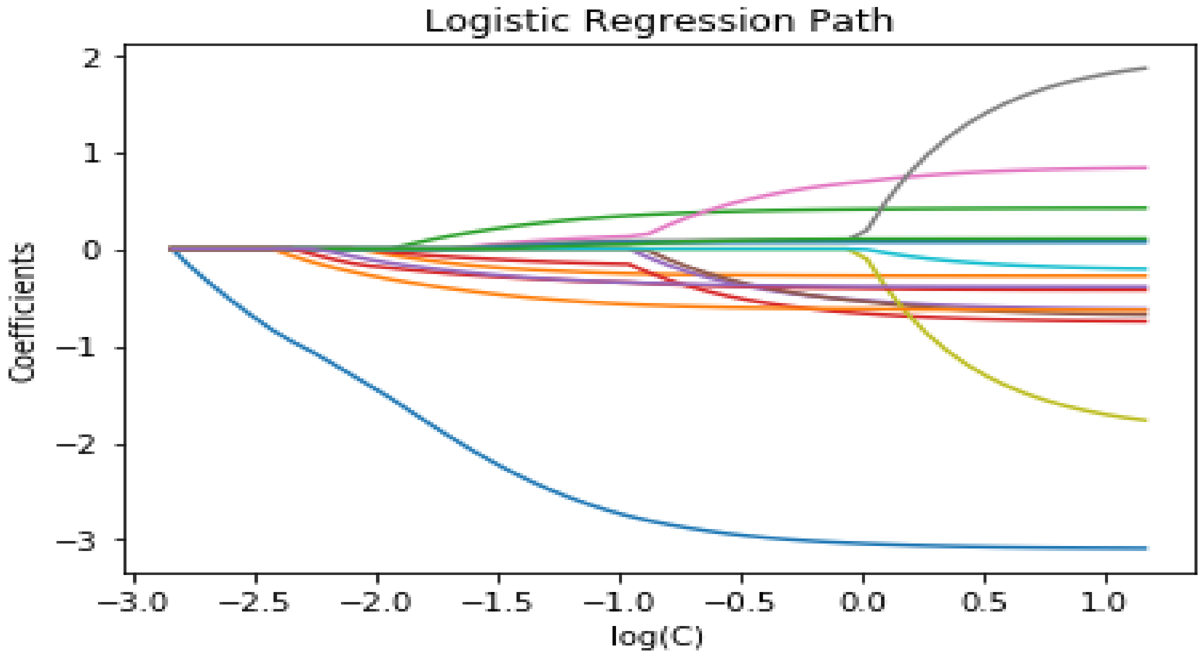
原始的目标函数



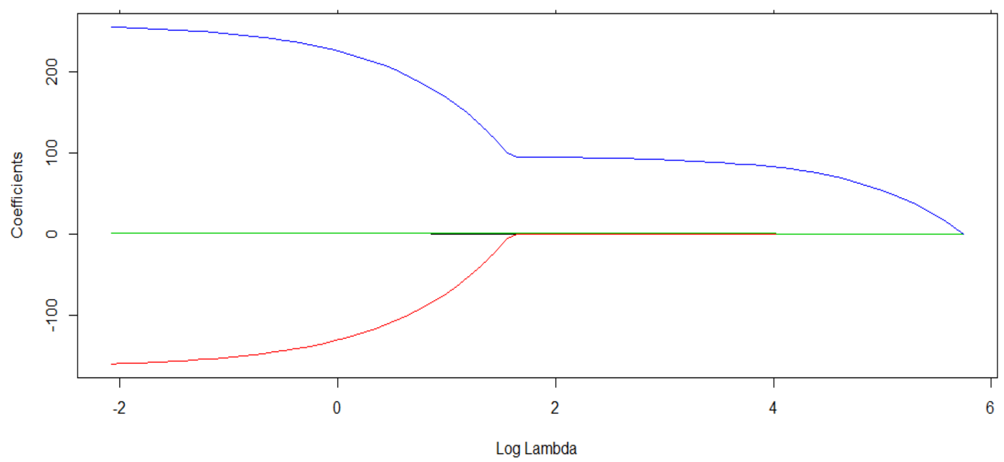
Lasso

随着C的增加，回归系数的值逐一的降为0。
当C很大时，惩罚项的权重很小，因此回归模型可以在所有变量空间上寻找权重组合，得到目标函数的最小值。但是当惩罚项权重提高后，对Y预测弱的变量就被剔除，回归模型只在剩余的变量空间上寻找权重组合，得到目标函数的最小值。

$$\min_{w,c} \|w\|_1 + C\sum_{i=1}^{n} \log(\exp(-y_i(X_i^T w + c)) + 1).$$

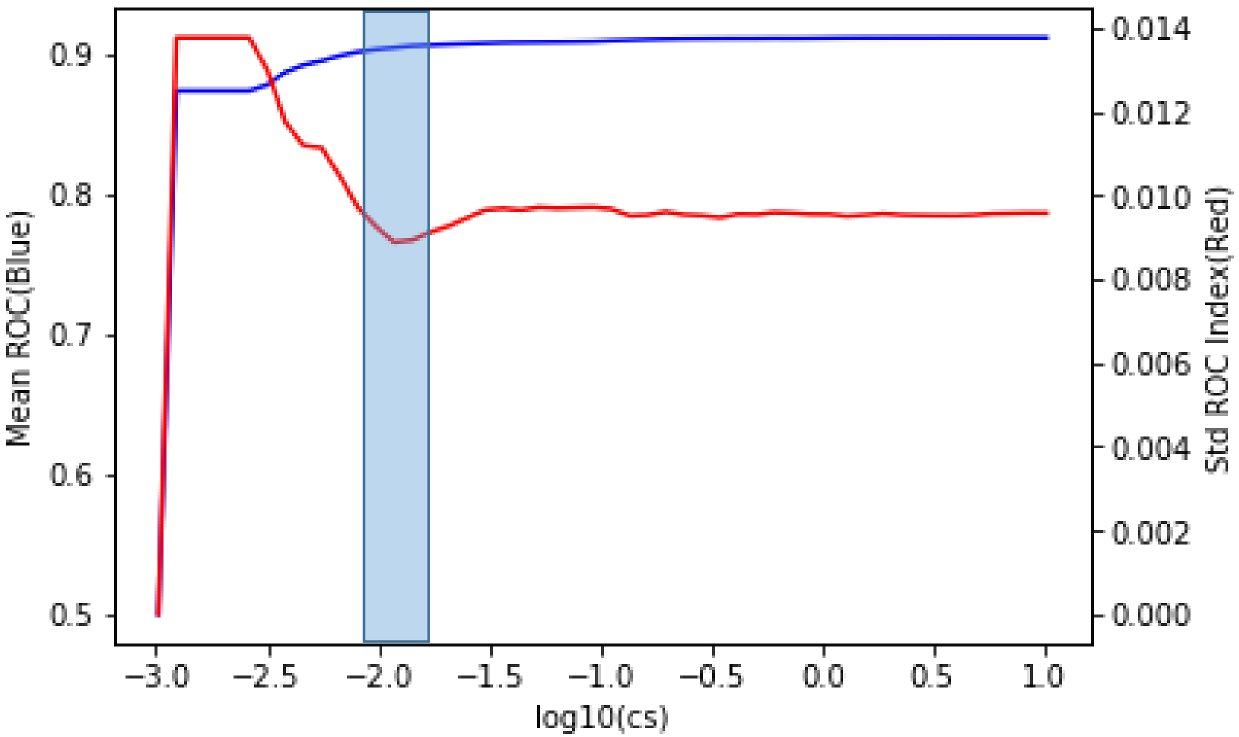


注：阅读其它参考资料时，其参数为Lambda（1/C），如下图所示：



Lasso

```
cs = l1_min_c(X, y, loss='log') * np.logspace(0, 4)
k_scores = []
clf = linear_model.LogisticRegression(penalty='l1')
for c in cs:
    clf.set_params(C=c)
    scores = cross_val_score(clf, X, y, cv=10, scoring='roc_auc')
    k_scores.append([c,scores.mean(),scores.std()])
```



scores.mean()为交叉验证得到的模型ROC曲线下面积的均值；该值越高越好；

scores.std()为交叉验证得到的模型ROC曲线下面积的标准差；该值越低越好。

还是偏差-方差权衡，蓝色区域都可以

机器学习算法的超参数

逻辑回归：LogisticRegression(penalty=’l2’, dual=False, tol=0.0001, C=1.0, fit_intercept=True, …)

决策树：DecisionTreeClassifier(criterion=’gini’, splitter=’best’, max_depth=None, min_samples_split=2, min_samples_leaf=1, min_weight_fraction_leaf=0.0,…)

BP神经网络：MLPClassifier(hidden_layer_sizes=(100,), activation=’relu’, solver=’adam’, …)

SVM：SVC(C=1.0, kernel=’rbf’, degree=3, gamma=’auto’, coef0=0.0, shrinking=True,…)

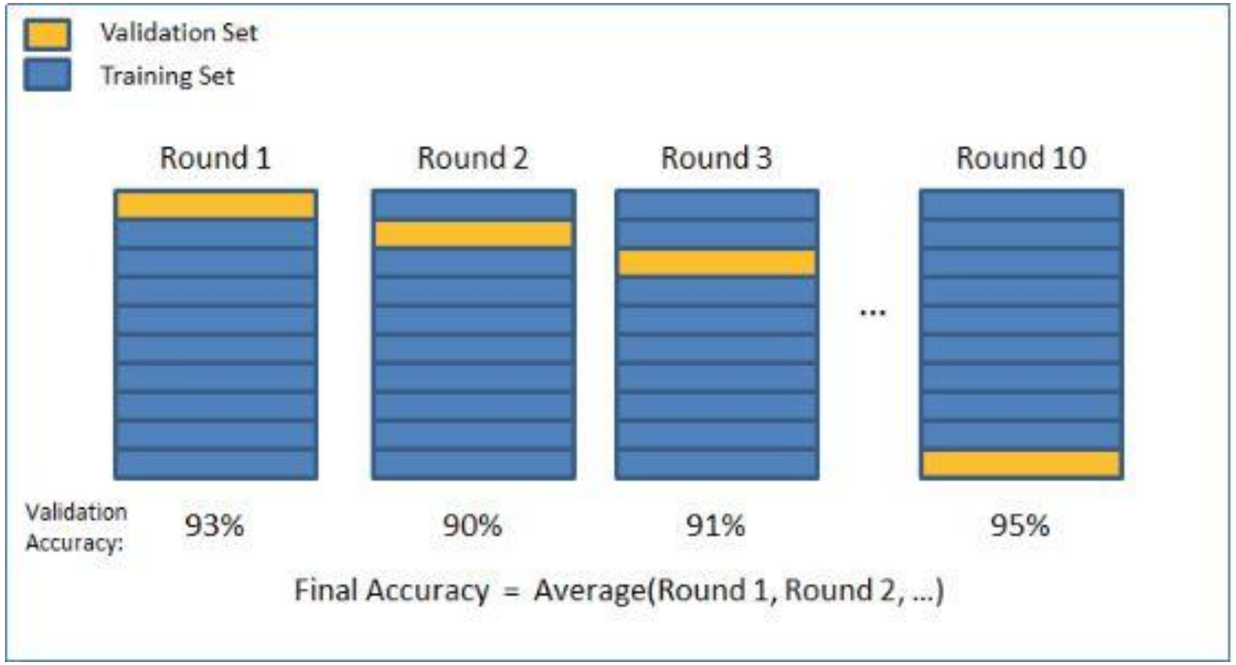
KNN：NearestNeighbors(n_neighbors=5, radius=1.0, algorithm=’auto’,…)

朴素贝叶斯：没有超参数

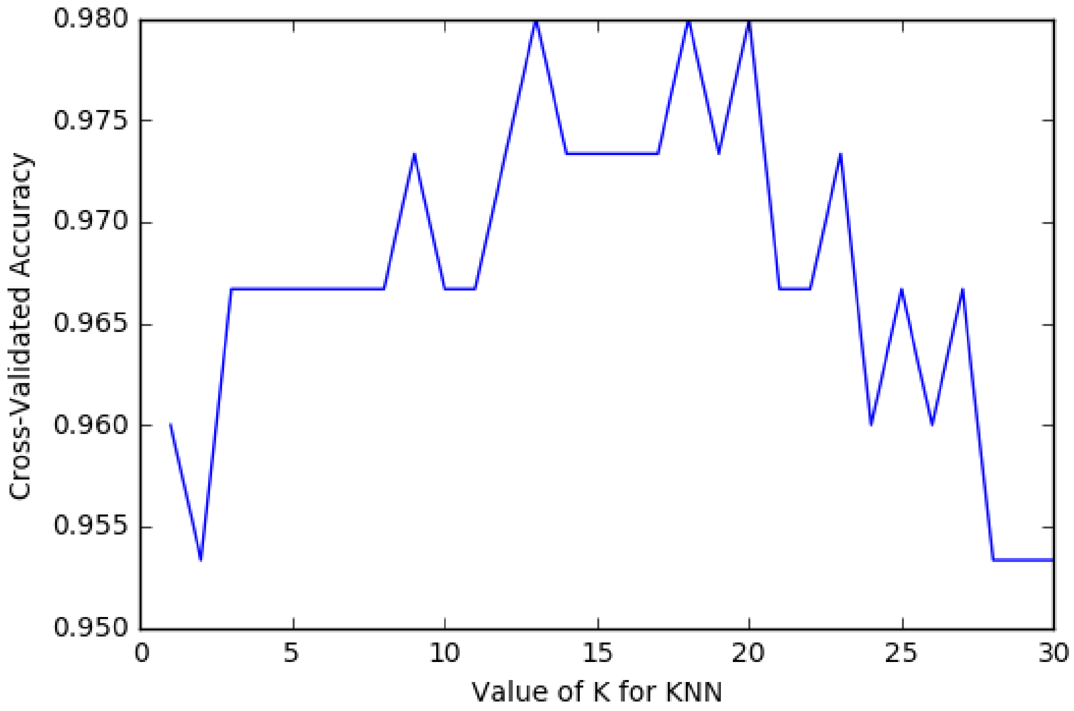
天善智能 TIANSHAN SOFT

交叉验证与网格搜索

交叉验证思路：



确定超参数：



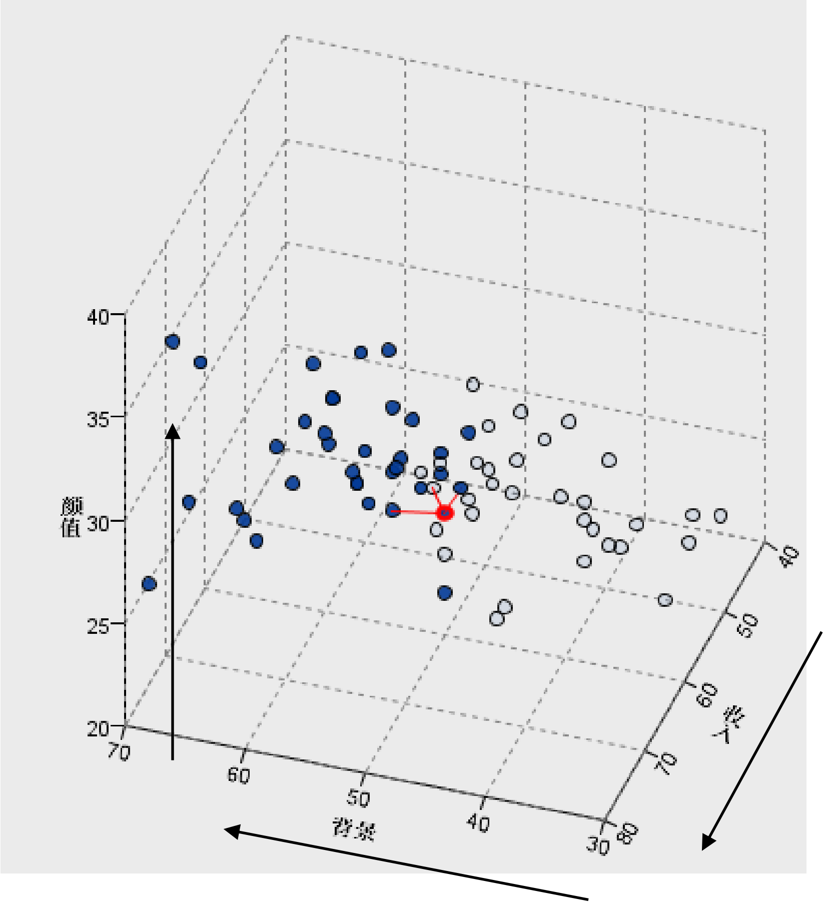
网格搜索实现：

```
for k in k_range:
    knn = KNeighborsClassifier(n_neighbors=k)
    scores = cross_val_score(knn, X, y, cv=10, scoring='accuracy')
    k_scores.append(scores.mean())
```

请参考：
https://blog.csdn.net/yueguizhilin/article/details/77711789

最近邻域（KNN）算法

示例:是否约会成功的KNN法



（蓝色代表约会成功的人）

如何预测一个婚恋网站新注册的男生是否会约会成功呢？这很简单，看看和这个新来的男生条件最接近的男生是否约会成功了。

比如蓝色点代表约会成功的人，红色点代表新来的男生，他和两个蓝色点一个灰色点最近，因此该点约会成功地可能性是2/3。

K邻域法属于惰性算法,其特点是不事先建立全局的判别公式或规则。当新数据需要分类时，根据每个样本和原有样本之间的距离，取最近K个样本点的众数（Y为分类变量的情形）或均值（ Y为连续变量的情形）作为新样本的预测值。这体现了一句老话“近朱者赤，近墨者黑”。

KNN介绍

K邻域法属于惰性算法,其特点是不事先建立全局的判别公式或规则。当新数据需要分类时，根据每个样本和原有样本之间的距离，取最近K个样本点的众数（Y为分类变量的情形）或均值（ Y为连续变量的情形）作为新样本的预测值。这体现了一句老话“近朱者赤，近墨者黑”。

对自变量和因变量的类型没有任何限制,最主要的参数就是K，即取多少个邻近点合适。

KNN算法

1、定义距离$d(x_i,x_j)$，该距离代表两个观测之间的差异程度，常用的距离如下：

欧式距离：

$$d_2(\mathbf{x}_r,\mathbf{x}_s)=\left[(\mathbf{x}_r-\mathbf{x}_s)'(\mathbf{x}_r-\mathbf{x}_s)\right]^{\frac{1}{2}}$$

$$=\left[\sum_{j=1}^{p}\left(x_{rj}-x_{sj}\right)^2\right]^{\frac{1}{2}}$$

Minkowshi距离

$$d(\mathbf{x},\mathbf{y})=\left[\sum_{r=1}^{p}|x_r-y_r|^m\right]^{\frac{1}{m}},$$

默认$m=2$（即默认采用欧式距离）。

当m=1时，为Manhattan 距离（Block距离）

天善智能 TIANSHAN SOFT

kNN之前的数据标准化

- 极差标准化
- 中心标准化（z-score
- 生成哑变量(m-1 princ

$$X_{new} = \frac{X - \min(X)}{\max(X) - \min(X)}$$

$$X_{new} = \frac{X - \mu}{\sigma} = \frac{X - \mathrm{Mean}(X)}{\mathrm{StdDev}(X)}$$

$$\mathrm{male} = \begin{cases} 1 & \text{if } x = \text{male} \\ 0 & \text{otherwise} \end{cases}$$

天善智能 TIANSHAN SOFT

K的选取

n	accuracy	Recall	Precision
1	0.850	1.0000000	0.7600000
2	0.875	1.0000000	0.7916667
3	0.900	1.0000000	0.8260870
4	0.850	1.0000000	0.7600000
5	0.950	1.0000000	0.9047619
6	0.850	0.8947368	0.8095238
7	0.900	0.9473684	0.8571429
8	0.850	0.8421053	0.8421053
9	0.875	0.8947368	0.8500000
10	0.875	0.9473684	0.8181818
11	0.850	0.8421053	0.8421053
12	0.925	0.9473684	0.9000000
13	0.875	0.8421053	0.8888889
14	0.875	0.8421053	0.8888889
15	[illegible]850	[illegible]94737	0.882352[illegible]

K值越小，模型越依赖于最近的样本点的取值，不稳健；K值越大，虽然模型稳健性增强了，但是敏感度下降。因此需要采用遍历的方法，选取最合适的K值。

如左表所示，根据准确度、召回率和精确度综合确定一个合理的K值。为了避免无法决策的麻烦，K一般取奇数。

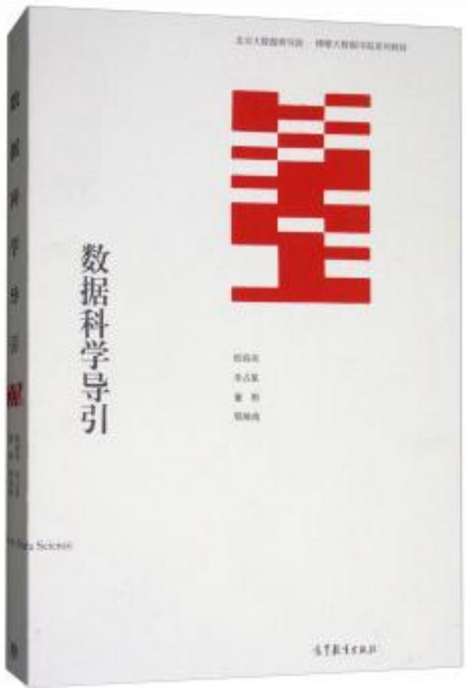
学习方式

天善智能
TIANSHAN SOFT

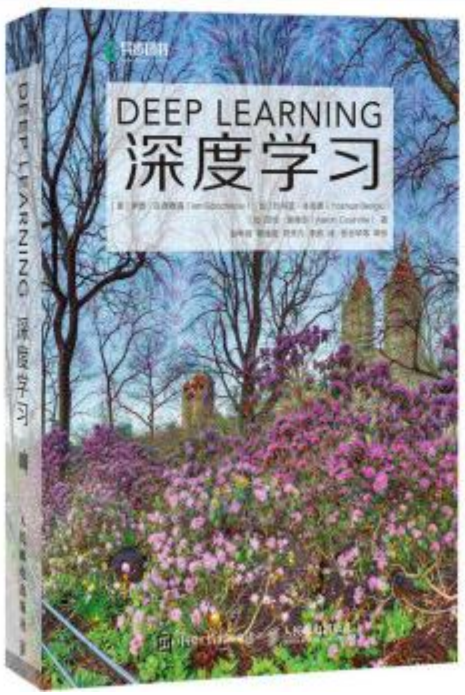
参考书

统计学习方法

机器学习



数据科学导引



深度学习 deep learning