收缩方法相关内容

冯裕祺

2021/6/1

收缩方法相关内容

冯裕祺

收缩的方法

收缩的方法

通过保留一部分预测变量而丢弃剩余的变量，子集选择 (subset selection) 可得到一个可解释的、预测误差可能比全模型低的模型．然而，因为这是一个离散的过程（变量不是保留就是丢弃)，所以经常表现为高方差，因此不会降低全模型的预测误差．而收缩方法 (shrinkage methods) 更加连续，因此不会受高易变性 (high variability) 太大的影响。

岭回归

岭回归

岭回归 (Ridge regression) 根据回归系数的大小加上惩罚因子对它们进行收缩．岭回归的系数使得带惩罚的残差平方和最小：

$$\hat{\beta}^{ridge} = \underset{\beta}{arg\,min}\Big\{\sum_{i=1}^{N}(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda\sum_{j=1}^{p}\beta_j^2\Big\} \quad (1)$$

这里 $\lambda \geq 0$ 是控制收缩程度的参数：λ 值越大，收缩的程度越大．每个系数都向零收缩．通过参数的平方和来惩罚的想法也用在了神经网络，也被称作 **权重衰减** (weight decay)

岭回归问题可以等价地写成：

$$\hat{\beta}^{\text{ridge}} = arg\,min \sum_{\beta}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2$$

$$\text{subject to } \sum_{j=1}^{p} \beta_j^2 \leq t \tag{2}$$

上式用参数显式表达了对回归参数大小的约束。式 (2) 其实是对式 (1) 应用 Lagrange 乘子法得到的。

(1) 中的 λ 和 (2) 中的 t 存在一一对应．当在线性回归模型中有许多相关变量，它们的系数可能很难确定且有高方差．某个变量的较大的正系数可以与相关性强的变量的差不多大的负系数相互抵消．通过对系数加入大小限制，如 (2)，这个问题能得以减轻．

这里说的是，在没有对参数大小进行限制前，会存在一对相关性强的变量，它们系数取值符号相反，但绝对值差不多大，会大大增加方差，这也就是高方差的体现，但其实它们的合作用效果近似为 0，所以考虑引进对参数大小的惩罚。

对输入按比例进行缩放时，岭回归的解不相等，因此求解 (1) 前我们需要对输入进行标准化．另外，注意到惩罚项不包含截距 β_0．对截距的惩罚会使得过程依赖于 Y 的初始选择；也就是，对每个 y_i 加上常数 c 不是简单地导致预测值会偏离同样的量 c．可以证明经过对输入进行中心化（每个 x_{ij} 替换为 $x_{ij}-\bar{x}_j$）后，(1) 的解可以分成两部分．我们用 $\bar{y}=\frac{1}{N}\sum_1^N y_i$ 来估计 β_0．剩余的参数利用中心化的 x_{ij} 通过无截距的岭回归来估计．今后我们假设中心化已经完成，则输入矩阵 X 有 p（不是 $p+1$）列．

将(1)的准则写成矩阵的形式:

$$RSS(\lambda) = (\mathrm{y} - \mathrm{X}\beta)^T(\mathrm{y} - \mathrm{X}\beta) + \lambda\beta^T\beta \tag{3}$$

可以看出岭回归的解为:

$$\hat{\beta}^{\mathrm{ridge}} = \left(\mathrm{X}^{\mathrm{T}}\mathrm{X} + \lambda\mathrm{I}\right)^{-1}\mathrm{X}^T\mathrm{y} \tag{4}$$

其中 I 为 $p \times p$ 的单位矩阵．注意到选择二次函数惩罚 $\beta^T\beta$，岭回归的解仍是 y 的线性函数．解在求逆之前向矩阵 $\mathrm{X}^{\mathrm{T}}\mathrm{X}$ 的对角元上加入正的常数值．即使 $\mathrm{X}^{\mathrm{T}}\mathrm{X}$ 不是满秩，这样会使得问题非奇异，而且这是第一次将岭回归引入统计学中 (Hoerl and Kennard, 1970) 的主要动力。传统的岭回归的描述从定义 (4) 开始．我们选择通过 (1) 和 (2) 来阐述，因为这两式让我们看清楚了它是怎样实现的。

岭回归的 Bayes 角度

岭回归的 Bayes 角度

当给定一个合适的先验分布，岭回归也可以从后验分布的均值或众数得到. 具体地，假设 $y_i \sim N(\beta_0 + x_i^T\beta, \sigma^2)$，参数 β_j 的分布均为 $N(0, \tau^2)$，每个都相互独立. 则当 τ^2 和 σ^2 值已知时，β 后验分布密度函数的对数值(的负数) 与 (1) 中花括号里面的表达式成比例，且 $\lambda = \sigma^2/\tau^2$。因此岭回归估计是后验分布的众数；又因分布为高斯分布，则也是后验分布的均值。

正态分布均值中位数众数相等

从奇异值分解角度看岭回归

从奇异值分解角度看岭回归

中心化输入矩阵 X 的 **奇异值分解** (SVD) 让我们进一步了解了岭回归的本质．这个分解在许多统计方法分析中非常有用．$N \times p$ 阶矩阵 X 的 SVD 分解有如下形式：

$$\mathrm{X} = \mathrm{UDV}^{\mathrm{T}} \tag{5}$$

这里 U 和 V 分别是 $N \times p$ 和 $p \times p$ 的正交矩阵，U 的列张成 X 的列空间，V 的列张成 X 的行空间．D 为 $p \times p$ 的对角矩阵，对角元 $d_1 \geq d_2 \geq \cdots \geq d_p \geq 0$ 称作 X 的奇异值．如果一个或多个 $d_j = 0$，则 X 为奇异的．

利用奇异值分解，通过化简我们可以把最小二乘拟合向量写成：

$$\begin{aligned}\mathrm{X}\hat{\beta}^{ls} &= \mathrm{X}\left(\mathrm{X}^{\mathrm{T}}\mathrm{X}\right)^{-1}\mathrm{X}^{\mathrm{T}}\mathrm{y} \\ &= \mathrm{U}\mathrm{U}^{\mathrm{T}}\mathrm{y}\end{aligned}$$

注意到 $\mathrm{U}^{T}\mathrm{y}$ 是 y 正交基 U 下的坐标．同时注意其与 (3) 的相似性；

$$\begin{aligned}\hat{\beta} &= \mathrm{R}^{-1}\mathrm{Q}^{\mathrm{T}}\mathrm{y} \\ \hat{\mathrm{y}} &= \mathrm{Q}\mathrm{Q}^{\mathrm{T}}\mathrm{y}\end{aligned}$$

Q 和 U 是 X 列空间的两个不同的正交基。

现在岭回归的解为：

$$\begin{aligned}\mathrm{X}\hat{\beta}^{ridge} &= \mathrm{X}(\mathrm{X}^{\mathrm{T}}\mathrm{X}+\lambda\mathrm{I})^{-1}\mathrm{X}^{\mathrm{T}}\mathrm{y} \\ &= \mathrm{U}\mathrm{D}(\mathrm{D}^2+\lambda\mathrm{I})^{-1}\mathrm{D}\mathrm{U}^{\mathrm{T}}\mathrm{y} \\ &= \sum_{j=1}^{p}\mathrm{u}_j\frac{d_j^2}{d_j^2+\lambda}\mathrm{u_j}^{\mathrm{T}}\mathrm{y}\end{aligned} \tag{7}$$

其中 u_j 是 U 的列向量．注意到因为 $\lambda \geq 0$，我们有 $d_j^2/(d_j^2+\lambda) \leq 1$．类似线性回归，岭回归计算 y 关于正规基 U 的坐标．通过因子 $d_j^2/(d_j^2+\lambda)$ 来收缩这些坐标．这意味着更小的 d_j^2 会在更大程度上收缩基向量的坐标．

d_j^2 值小意味着什么？中心化后的矩阵 X 的奇异值分解是表示 X 中主成分变量的另一种方式．样本协方差矩阵为 $\mathrm{S} = \mathrm{X}^{\mathrm{T}}\mathrm{X}/N$，并且从 (5) 式我们得到

$$\mathrm{X}^{\mathrm{T}}\mathrm{X} = \mathrm{V}\mathrm{D}^2\mathrm{V}^{\mathrm{T}} \tag{8}$$

上式是 $\mathrm{X}^{\mathrm{T}}\mathrm{X}$（当忽略因子 N 时，也是 S）的 **特征值分解** $(\mathrm{eigen\ decomposition})$．特征向量 v_j（V 的列向量）也称作 X 的 **主成分** $(\mathrm{principal\ components})$（或 $\mathrm{Karhunen\text{-}Loeve}$）方向．第一主成分方向 v_1 有下面性质：$\mathrm{z}_1 = \mathrm{X}v_1$ 在所有 X 列的标准化线性组合中有最大的样本方差．样本方差很容易看出来是

$$Var\left(\mathrm{z}_1\right) = Var\left(\mathrm{X}v_1\right) = \frac{d_1^2}{N} \tag{9}$$

事实上 $\mathrm{z}_1 = \mathrm{X}v_1 = \mathrm{u}_1 d_1$. 导出变量 z_1 称作 X 的第一主成分，因此 u_1 是标准化的第一主成分. 后面的主成分 z_j 在与前一个保持正交的前提下有最大的方差 d_j^2/N. 所以，最后一个主成分有最小的方差. 因此越小的奇异值 d_j 对应 X 列空间中方差越小的方向，并且岭回归在这些方向上收缩得最厉害。

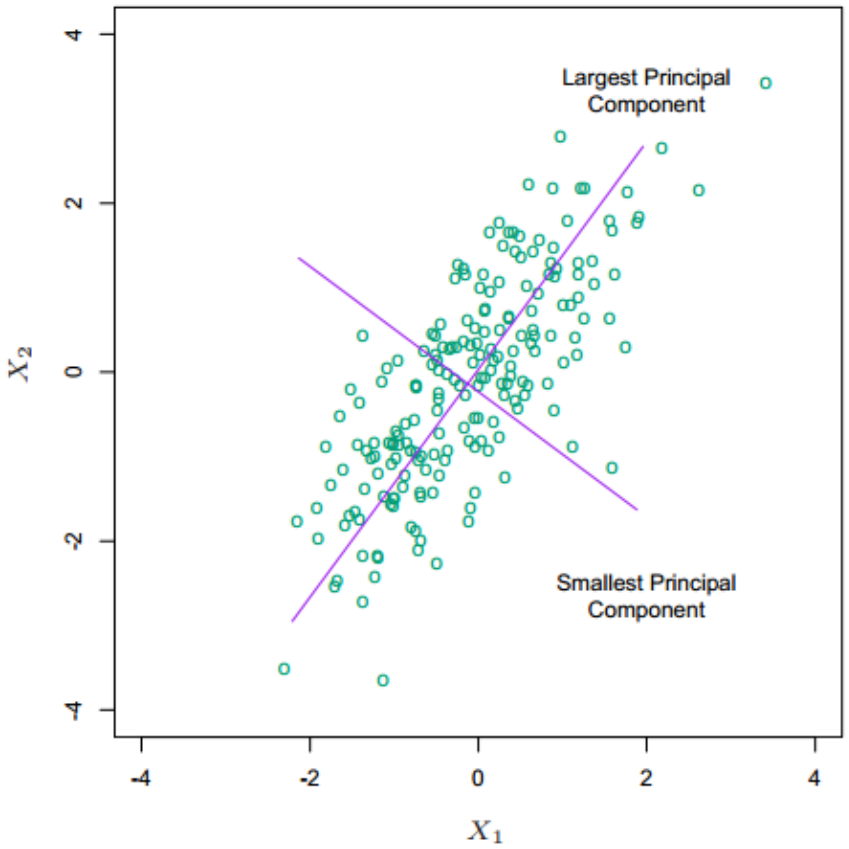

FIGURE 3.9. *Principal components of some input data points. The largest principal component is the direction that maximizes the variance of the projected data, and the smallest principal component minimizes that variance. Ridge regression projects* $\mathbf{y}$ *onto these components, and then shrinks the coefficients of the low–variance components more than the high-variance components.*