高维数据过拟合问题与正则化

经济学系　15320180155372　史九领

2019 年 12 月 14 日

在机器学习中，高维数据越来越普遍。高维数据指数据的维度很高，甚至远大于样本量的一类数据。高维数据的明显表现是：在空间中数据是非常稀疏的，与空间的维数相比样本量总是显得非常少。在分析高维数据过程中碰到最大的问题就是维数的膨胀，也就是通常所说的“Curse of Dimensionality”问题，并由此产生一些列的估计问题。

1　高维数据与过拟合问题

随着维数的增加，分析数据所需的空间样本数会呈指数增长，因而数据变得更稀疏。在高维数据空间，预测将变得不再容易，容易导致模型过拟合问题。如图 1 所示，随着数据维度的增加，模型的分类效果呈现出倒“V”的形状。

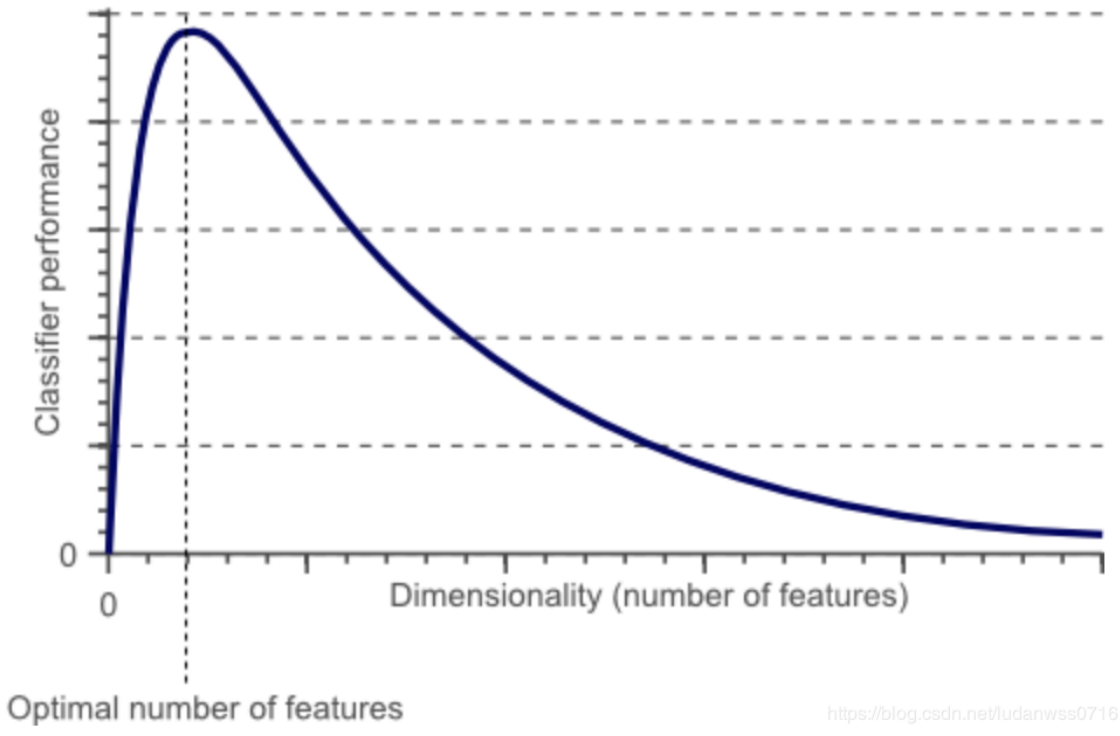

图 1: 数据维度与模型分类效果的关系

过拟合问题的基本解决思路包括两个方面：（1）增加样本量；（2）减少样本特征。主成分分析作为一种数据降维方法，通过整合原本的单一变量可以得到一组新的综合变量，综合变量所代表的意义丰富且变量间互不相关，综合变量包含了原变量大部分的信息，这些综合变量称为主成分。主成分分析是在保留所有原变量的基础上，通过原变量的线性组合得到主成分，然后选取少数主成分就可保留原变量的绝大部分信息，这样就可用这几个主成分来代替原变量，从而达到降维的目的。但现实中经常存在所能获取到的样本数据量有限，甚至远小于数据维度的情况，即：$k >> n$, 主成分分析法的局限性在于它只适用于数据空间维度小于样本量的情况，当数据空间维度很高时，主成分分析法将不再适用。

2 偏误和方差的权衡

当模型做出与实际情况不符的假设时就会引起错误，这种错误称为偏差。如果选择的模型过于简单，这时模型与预测变量和因变量之间的关系就会因为差别太大而产生偏差，即模型“偏差”向了一个错误的方程。方差是一种由于训练数据集的波动引起的错误，在理想情况下，无论使用哪个数据集针对同样的问题使用相同来源的数据建立模型时，所建立的模型应当是相同的。如果算法过于敏感，有可能会找到了一种只存在于当前使用的数据集中的随机模式，当学习算法随着训练数据集的不同而呈现出太大的波动时，所引起的错误就称为方差。如公式 1 所示，对于线性预测模型，模型预测误差是由偏差、方差和系统误差三部分组成。

$$E[(\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})^2] = Var(\boldsymbol{X}'\hat{\boldsymbol{\beta}}) + [bias(\boldsymbol{X}'\hat{\boldsymbol{\beta}})]^2 + Var(\boldsymbol{e}) \tag{1}$$

，

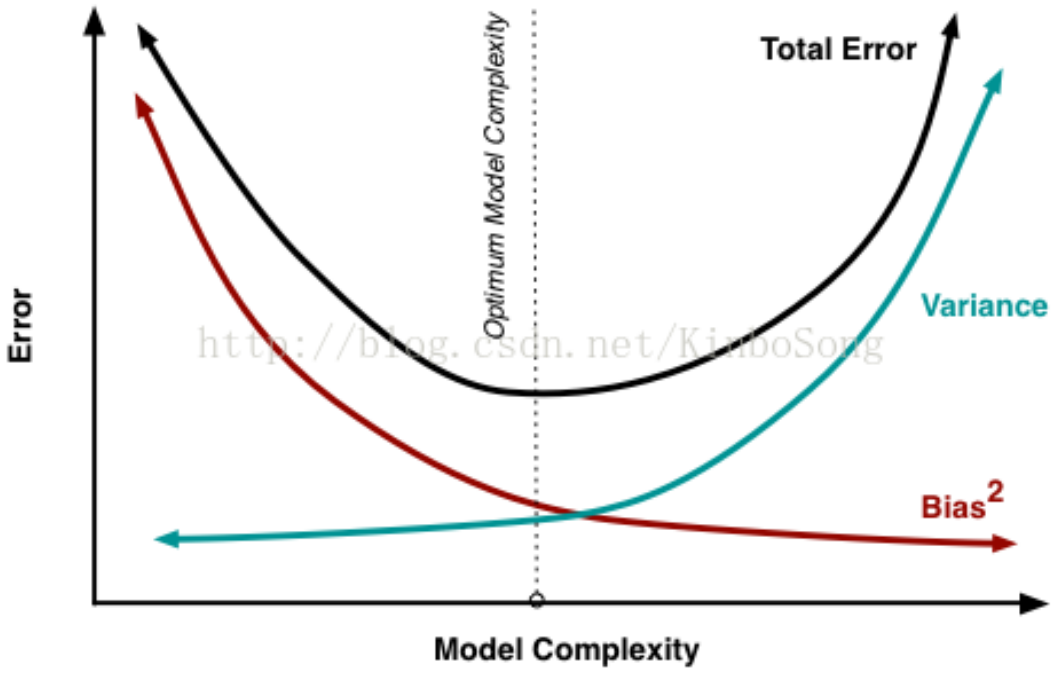

图 2: 模型复杂程度与估计误差

如图 2 所示，当算法的复杂度较低时，偏差错误就会很高，因为模型不能够完全掌握输入与输出之间的关系。当复杂度增加，模型开始能够更好的学习，偏差错误开始下降。当然，与此同时方差错误开始增加，因为模型开始变得越来越敏感。不过这时整体的样本外误差还是呈现下降趋势，因为这时偏差错误中的下降程度相比于方差错误的增加还是高出许多的。但当算法的复杂度越过某个点后，如果继续增加复杂度，方差错误的增加相比与偏差错误的下降就会更加明显，从而开始造成过拟合。

3 多重共线性问题与惩罚方法

随着数据维度的增加，各种维度之间的相关性也会增加，对于高维数据而言，不同维度之间非常容易产生多重共线性问题。对于传统的 OLS 回归模型,

$$\boldsymbol{Y} = \boldsymbol{X}'\boldsymbol{\beta} + \boldsymbol{\epsilon} \tag{2}$$

其中，响应变量 $\boldsymbol{Y} = (y_1, \cdots, y_n)^T$，协变量为 $\boldsymbol{X} = (\boldsymbol{X_1}, \cdots, \boldsymbol{X_k})$，对于每一个 $\boldsymbol{X_i}$，有 $\boldsymbol{X_i} = (x_{i1}, \cdots, x_{in})^T$。假设每个 $x_{ij}(i = 1, \cdots, k; j = 1, \cdots, n)$ 均中心化和标准化，随机误差项 $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_n)^T$，且 $\epsilon_i \sim N(0, \sigma^2)$。

当 $\boldsymbol{X}$ 为列满秩矩阵时，回归系数 $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_n)^T$ 可以通过普通最小二乘法求得，即

$$\hat{\boldsymbol{\beta}}_{OLS} = \underset{\boldsymbol{\beta} \in R^k}{argmin}(\boldsymbol{Y} - \boldsymbol{X}'\boldsymbol{\beta})^2 = (\boldsymbol{X}'\boldsymbol{X})^{-1}(\boldsymbol{X}'\boldsymbol{Y}) \tag{3}$$

但对于高维数据而言，由于维度之间多重共线性的存在，$\boldsymbol{X}$ 矩阵再不满足列满秩条件，此时 $(\boldsymbol{X}'\boldsymbol{X})^{-1}$ 不存在，普通最小二乘法将不再适用于求解回归系数 $\boldsymbol{\beta}$。这时就需要引入惩罚方法，惩罚方法可以同时实现变量选择和参数估计问题，在参数估计问题中，通过将部分参数压缩为 0 来达到选取变量的效果。惩罚方法是取惩罚似然函数最小时的值作为回归系数的估计值，即

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta} \in R^k}{argmin}[(\boldsymbol{Y} - \boldsymbol{X}'\boldsymbol{\beta})^2 + P_\lambda(|\boldsymbol{\beta}|)] \tag{4}$$

其中，惩罚项 $P_\lambda(|\boldsymbol{\beta}|) = \lambda \sum_{i=1}^{k}(|\boldsymbol{\beta}|^m)$，且 $m \geqslant 0$，λ 成为调节参数，也叫正则参数，λ 要做的就是控制在两个不同的目标中的平衡关系。当 $m = 1$ 时，惩罚项被称为 $L1$ 惩罚项，即 Lasso 惩罚；当 $m = 2$ 时，惩罚项被称为 $L2$ 惩罚项，即 Ridge 惩罚。

4 Ridge 回归

当协变量 $\boldsymbol{X} = (\boldsymbol{X_1}, \cdots, \boldsymbol{X_k})$ 之间相互独立时，采用 OLS 估计参数 $\boldsymbol{\beta}$ 具有很好的性质，$\hat{\boldsymbol{\beta}}_{OLS}$ 不仅是无偏估计，而且在所有的线性无偏估计中方差最小。即使在高度多重相关的情况下，OLS 回归系数估计量依然是线性无偏的，且具有最小的方差，也就是说多重共线性并不影响最小二乘估计量的无偏性和最小方差性。因此在所有线性无偏估计中，OLS 估计具有较小的方差。随着数据维度不断增加，$\boldsymbol{X_j}$ 之间线性相关的概率就会增加，此时设计矩阵 $\boldsymbol{X}$ 不再满足列满秩条件，我们称这种设计矩阵为病态的。当 $\boldsymbol{X}$ 为病态时，$\boldsymbol{X}^T\boldsymbol{X}$ 接近奇异，此时 $\hat{\boldsymbol{\beta}}_{OLS}$ 虽

然具有最小方差，但 $\hat{\boldsymbol{\beta}}_{OLS}$ 的值却非常大，从而导致精度非常差，表现很不稳定。

对于这一问题，常见的解决方法是 Ridge 回归，Ridge 回归是一种专用于共线性数据分析的有偏估计回归方法，实质上是一种改良的最小二乘估计法，通过放弃最小二乘法的无偏性，以损失部分信息、降低精度为代价，获得回归系数更为符合实际、更可靠的回归方法，对病态数据的耐受性远远强于最小二乘法。Ridge 回归通过找到一个有偏估计量，这个估计量虽然有较小的偏差，但它的精度却能够大大高于无偏的估计量。Ridge 回归分析就是根据这个原理，通过在正规方程中引入有偏常数项求得回归估计量的。

$$\hat{\boldsymbol{\beta}}_{Ridge} = \underset{\boldsymbol{\beta}\in R^k}{argmin}(\boldsymbol{Y} - \boldsymbol{X}'\boldsymbol{\beta})^2, \quad s.t. \sum_{j=1}^{k} |\hat{\beta}_j|^2 \leqslant t, \quad t \geqslant 0 \tag{5}$$

利用拉格朗日方法，可以得

$$\hat{\boldsymbol{\beta}}_{Ridge} = \underset{\boldsymbol{\beta}\in R^k}{argmin}[(\boldsymbol{Y} - \boldsymbol{X}'\boldsymbol{\beta})^2 + \lambda \sum_{j=1}^{k} |\hat{\beta}_j|^2] \tag{6}$$

进一步求解，可得 Ridge 估计系数

$$\hat{\boldsymbol{\beta}}_{Ridge} = (\boldsymbol{X}'\boldsymbol{X} + \lambda \boldsymbol{I})^{-1}\boldsymbol{X}'\boldsymbol{Y} \tag{7}$$

其中 t 与 λ 一一对应，为调节系数。这里假设每个 $x_{ij}(i = 1, \cdots, k; j = 1, \cdots, n)$ 均中心化和标准化，此时有 $\boldsymbol{X}'\boldsymbol{X} = \boldsymbol{I}$, $\boldsymbol{I}$ 是 $k \times k$ 单位矩阵，则可求得

$$\hat{\boldsymbol{\beta}}_{Ridge} = [(1 + \lambda)\boldsymbol{X}'\boldsymbol{X}]^{-1}\boldsymbol{X}'\boldsymbol{Y} = \frac{\hat{\boldsymbol{\beta}}_{OLS}}{1 + \lambda} \tag{8}$$

此时可以很好地克服了 OLS 在多重共线性线而造成的不稳定问题。在多重共线性情况下，由于 $\boldsymbol{X}^T\boldsymbol{X} \approx 0$, 从而使得 OLS 估计值 $\hat{\boldsymbol{\beta}}_{OLS}$ 偏大，且非常不稳定。而 Ridge 回归通过给 $\boldsymbol{X}^T\boldsymbol{X}$ 加上一个正常数矩阵 $\lambda\boldsymbol{I}$，使得 $\boldsymbol{X}'\boldsymbol{X} + \lambda\boldsymbol{I}$ 等于 0 的可能性大大降低，从而使得估计结果变小，且更加稳定。

5 Lasso 回归

Lasso 是另一种数据降维方法，该方法不仅适用于线性情况，也适用于非线性情况。Lasso 是基于惩罚方法对样本数据进行变量选择，通过对原本的系数进行压缩，将原本很小的系数直接压缩至 0，从而将这部分系数所对应的变量视为非显著性变量，将不显著的变量直接舍弃。

$$\hat{\boldsymbol{\beta}}_{Lasso} = \underset{\boldsymbol{\beta}\in R^k}{argmin}(\boldsymbol{Y} - \boldsymbol{X}'\boldsymbol{\beta})^2, \quad s.t. \sum_{j=1}^{k} |\hat{\beta}_j| \leqslant t, \quad t \geqslant 0 \tag{9}$$

利用拉格朗日方法，可以得

$$\hat{\boldsymbol{\beta}}_{Lasso} = \underset{\boldsymbol{\beta} \in R^k}{argmin}[(\boldsymbol{Y} - \boldsymbol{X}'\boldsymbol{\beta})^2 + \lambda \sum_{j=1}^{k} |\hat{\beta}_j|] \tag{10}$$

求解可得：

$$\hat{\boldsymbol{\beta}}_{Lasso} == \begin{cases} \hat{\boldsymbol{\beta}}_{OLS} - \frac{\lambda}{2}, & 若\hat{\boldsymbol{\beta}}_{OLS} > \frac{\lambda}{2} \\ \hat{\boldsymbol{\beta}}_{OLS} + \frac{\lambda}{2}, & 若\hat{\boldsymbol{\beta}}_{OLS} < -\frac{\lambda}{2} \\ 0, & 若|\hat{\boldsymbol{\beta}}_{OLS}| \leqslant \frac{\lambda}{2} \end{cases} \tag{11}$$

由求解结果可知，当 $|\hat{\boldsymbol{\beta}}_{OLS}| \leqslant \frac{\lambda}{2}$ 时，一部分系数就会被压缩到 0，从而降低了于 $\boldsymbol{X}$ 的维度，有效缓解了高维数据的多重共线性问题，并达到了缩小模型复杂度的目的。

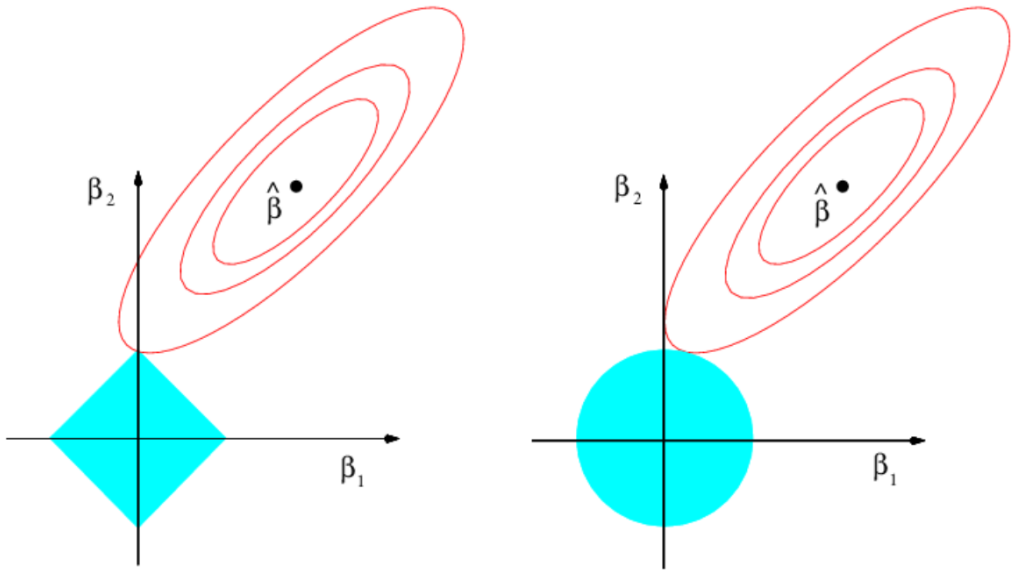

图 3: Lasso 和 Ridge 两种方法的差异

下面以二维数据空间为例，说明 Lasso 和 Ridge 两种方法的差异，左图对应的是 Lasso 方法，右图对应的是 Ridge 方法，如图 3 所示。两个图是对应于两种方法的等高线与约束域。红色的椭圆代表的是随着 λ 的变化所得到的残差平方和，$\hat{\boldsymbol{\beta}}_{OLS}$ 为椭圆的中心点，为对应普通线性模型的 OLS 估计。左右两个图的区别在于约束域，即对应的蓝色区域。等高线和约束域的切点就是目标函数的最优解，Ridge 方法对应的约束域是圆，其切点只会存在于圆周上，不会与坐标轴相切，则在任一维度上的取值都不为 0，因此没有稀疏；对于 Lasso 方法, 其约束域是正方形，会存在与坐标轴的切点，使得部分维度特征权重为 0，因此很容易产生稀疏的结果。所以，Lasso 方法可以达到变量选择的效果，将不显著的变量系数压缩至 0，而 Ridge 方法虽然也对原本的系数进行了一定程度的压缩，但是任一系数都不会压缩至 0，最终模型保留了所有的变量。

References

[1] 周志华:《机器学习，清华大学出版社 2015 版，第 11 章“嵌入式选择与 L1 正则化”。

[2] https://blog.csdn.net/pxhdky.

[3] https://blog.csdn.net/xiaozhu1024.

[3] https://blog.csdn.net/KinboSong.