考虑这样的一个问题: 给定一系列样本: , 其中, 是一未知函数, 参数可能具有无限维. 求的最佳估计.

这是非参数回归问题的一般形式. 本项目将介绍两种非参数回归的常用方法: 局部多项式回归和样条回归, 并在最后分析它们与参数回归的基本方法最小二乘法之间的差异.

**1.局部多项式回归介绍**

顾名思义, 局部多项式回归的主要思想是利用与当前点邻近的点所对应的, 对进行估计. 具体做法如下:

写出在点处的泰勒级数: . 取级数的前项, . 一般取0,1,2,3.

现在要做的就是写出损失函数. 局部多项式回归在这一步上引入了权重的概念: 对周围每个点的损失指派一个权重, 权重分布满足, 并且距离越近的点, 权重越大; 反之权重越小. 基于此, 局部多项式回归的损失函数可以写为: .

式中, ,, .

对求偏导来最小化损失函数, 可得. 于是对每个样本点的估计值.

由上述局部多项式回归的过程, 可以发现其有两个超参数: 拟合多项式的最高次数和权重. 下面分别对这两个参数的选取进行讨论:

**i)权重w**

权函数显然是一个关于距离的函数, 理论上任何满足之前所述性质的函数都能作为权函数. 在局部线性回归中, 权重通常采用核函数来度量. 这里的权函数概念不同于支持向量机中的核函数, 它可以理解为是一种概率密度函数. 一般将核函数记为. 常见的核函数有: Gaussian核, Epanechnikov核, tricube核等.

将核函数用于局部线性回归, 则权重. 式中称为窗宽, 也是一个超参数.

窗宽度量了局部多项式回归模型的复杂度. 越大的窗宽, 会有越多的样本点用于估计, 指派到每个样本点的权重也会较小, 导致近样本点与远样本点的权重差距不那么大, 因此这样得到的拟合曲线也会越光滑, 但同样如前所述, 估计的偏差会相对变大, 因为利用了大量无关局域性质的样本点的信息; 而较小的窗宽与之相反, 拟合的方差会因偏差的减小而增大, 模型泛化误差较大.

因此一个好的窗宽必须达到经验误差与泛化误差之间的平衡. 在优化调参的时候, 可以采用交叉验证或施加一定的惩罚来提高模型的泛化能力. 比如可以证明, 在使用高斯核函数的情况下最好的窗宽为, 其中是样本标准差. 另外, 如果在不同区域拟合的时候, 采用不同的窗宽, 也不失为一种选择, 特别是当样本点的密集程度在不同区间有明显差异时, 这种可变窗宽的优势会更明显. 下图是大窗宽与小窗宽的拟合效果对比:

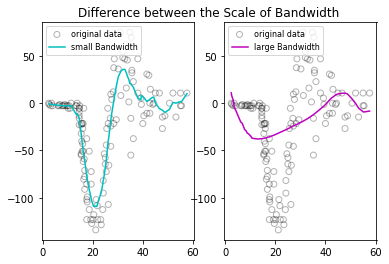


图1 窗宽大小对回归的影响

核函数决定了权重指派的具体方式. 不同的核函数会改变较远样本点与较近样本点权重指派的相对差异, 对回归结果的影响并不明显. 经过几次尝试, 发现较为复杂的核函数可能会使回归曲线变得更为粗糙, 而简单的核函数比如, 回归曲线则相对平滑. 以下是采用高斯核函数和采用简单核函数的拟合效果对比:

图表, 散点图

描述已自动生成

图2 核函数对回归效果的影响

可以看到, 两者虽有区别, 但是显然没有窗宽对模型的影响大.

**ii)最高次数d**

当时, 设计矩阵退化为维全1向量, 因此. 可见在这种情形下, 只是用附近的点的值来估计. 这也是此时多项式回归被成为核平滑的原因.

当时, , 此时的局部多项式回归被称为局部线性回归. 另外, 当窗宽时, 会用所有的样本点来估计, 因此也可以把局部线性回归看成是线性回归的合理外推. 在某些具有明显线性关系的区域内, 局部线性回归具有极好的性能, 并且相比于的核平滑方式, 在样本点的边界区域, 局部线性回归的偏差较小, 因为它不仅仅利用了响应变量的信息, 还利用了估计点与邻近样本点之间的线性关系, 一定程度上减小了偏差. 下图展示了这两种方式在边界区域的性能（样本由加上高斯噪声产生）, 显然核平滑在边界处产生了明显的偏差（在左边界受到右侧数据影响上偏, 右边界受到左侧数据影响下偏）, 而局部线性回归则几乎没有受到影响.

图表, 散点图

描述已自动生成

图3 两者在边界处的性能对比

局部线性回归的偏差和方差分别为: . 式中是样本容量, 是窗宽. 由此同样可以说明, 当窗宽减小时, 估计值的方差会迅速增加; 而较大的窗宽又会引起较大的经验误差.

当时, 估计方程变为抛物线型, 它在一些具有“峰”或“谷”的位置呈现较好的拟合效果, 而核平滑与局部线性回归在这些位置的估计值均会产生较大偏差.

最高次数与窗宽一样, 也是用来度量模型复杂度的参数. 一般在局部多项式回归中, 最高不会超过3. 即使样本十分复杂, 也可以通过设置节点分割区间, 在每一个小区间内再使用局部多项式回归的方法, 减小模型复杂度.

综上所述, 局部多项式回归是非参数回归中常用的方法, 它综合了多项式回归和线性加权回归的优点, 通过局部加权的方式, 用邻近样本点的信息来对当前数据点进行估计, 对非线性模型具有较好的回归性能.

**2.样条回归介绍**

样条回归与局部多项式回归一样, 都是非参数回归的重要方法. 现在假设样本点有很明显的分段特征, 那么自然想到可以利用分段函数来估计. 样条回归便是采取了分段函数这一策略, 并且加上了一个限制条件: 这族分段函数必须是连续函数.

首先给出最简单的样条回归模型线性样条回归: .

式中, 有序变量被称为控制节点, 为模型待估计的参数, 记号.

可以清楚的看到, 每经过一个控制节点, 估计量中就会加入一个线性项, 从而使模型能拟合控制节点之后的区间. 这就是线性样条的原理. 至于参数的估计, 同样采用最小二乘法最小化损失函数来求得: .

这时的设计矩阵为: .

将它合理外推可以得到阶样条: . 类似于线性样条要求函数连续, 对阶样条有要求阶导数连续. 因此, 高阶样条会更光滑, 具有更好的预测能力. 在这些高阶样条中, 最为常用的是三阶样条模型.

样条回归的一个重要概念是自由度, 可以理解为模型参数的维度, 越高的自由度会带来越高的灵活性. 对于一个具有*K*个控制节点的三阶样条, 其自由度为*K+*4, 也即参数***w***的维度为*K+*4. 通过增加控制节点的个数, 可以很好的降低模型的经验误差, 但模型的解释能力会变差.

为了提高模型对控制节点以外区域的预测能力, 一种可行的办法是使用自然样条. 自然样条是在原有样条的基础上, 限制模型在控制节点以外的高阶（二阶及以上）导数值为0; 具体而言, 自然样条规定模型在控制节点以外区域使用线性模型进行预测, 使得这些区域的预测值符合实际预期. 下面是自然样条的数学推导过程（以自然三阶样条为例）:

三阶样条的模型为. 现在令. 经过计算可以得到, 的系数需要满足. 这就是自然三阶样条的数学模型.

另外一种方法类似于线性回归方法中常用的岭回归和Lasso. 越多的控制节点会带来越低的经验误差, 因此可以首先设定大量的控制节点, 再设置一个罚函数用于平衡节点数量, 从而达到经验误差与泛化误差的平衡. 这种带惩罚项的样条模型被称为平滑样条. 由以上讨论可知, 平滑样条是一个最小化问题:



式中是大于0的常数, 决定了对复杂模型的惩罚力度; 在实际应用时, 可以采用留一交叉验证或者广义交叉验证来选择出最佳惩罚系数.

但是平滑样条往往趋于选择较少的控制节点. 若将上式做一些简化, 改变一下罚函数的形式:

, 其中

上式的模型就是带惩罚的样条回归, 相比平滑样条, 它的罚函数变为模型参数***w***的2范数的平方, 于是可以像岭回归那样直接求出***w***的最小二乘估计. 它结合了普通样条和平滑样条的优点, 既有较高的灵活性, 又有较强的泛化能力.

**3.样条与最小二乘、局部线性回归之间差异的分析**

就这三者的应用领域而言, 样条和局部线性回归都属于非参数回归, 而最小二乘法则是参数回归中估计待定参数的重要方法.

非参数回归在最开始已经做过介绍, 它最大特点是解释变量与响应变量之间的真实模型未知, 而不同于普通最小二乘回归等参数回归只需要对真实模型的参数做出估计. 先简要介绍一下最小二乘回归:

最小二乘回归的方法为: 给出一系列基函数, 首先假定真实模型为这些基函数的线性组合: . 之后采用最小二乘法最小化残差平方和:

.

式中. 令

, 得待求参数估计.

很多情况下, 基函数通常取为一系列幂函数, 即, 这种回归也被称为多项式回归. 但是当样本点之间的关系十分复杂, 离散程度很不均匀时, 多项式回归的效果并不好. 具体分析如下:

根据预测误差、噪音、偏差、方差之间的关系（预测误差=噪音+偏差+方差）:

,

若采取增加基函数的策略（例如使用六次、七次甚至更高次方的多项式来拟合）来提高拟合水平, 虽然MSE会显著降低, 但是拟合曲线的方差会非常大, 每增加一个样本都会引起曲线的剧烈波动, 也即发生过拟合, 模型的泛化能力非常弱; 而若仍然采用较低次数的多项式来拟合, 则会出现另一种极端——欠拟合, 即估计值虽然有较小的方差, 曲线较为平滑, 但是距离真实响应变量比较远, 这时偏差成为了曲线拟合的主要误差来源. 因此拟合多项式最高次数的选择较为困难.

这便是最小二乘法的一个重要缺陷所在, 即模型泛化能力弱, 如果向样本中添加新样本或是从样本中取出一些样本, 都会导致回归曲线的剧烈波动, 有时甚至需要直接更换模型假设. 这也引出了最小二乘法的另一个不足: 回归函数的模型需要预先假定; 但是单凭经验假定的模型往往缺乏稳健性.

而回归样条、局部线性回归等非参数回归策略, 不仅对样本的分布没有过多的假设, 而且由于并不给出具体回归函数的形状, 具有较强的适应性, 模型完全由样本数据来确定, 因而也具有较高的灵活度. 在非线性、非齐次模型的回归上, 它们相比最小二乘会有明显的优势.

但是, 也正是因为非参数回归的参数灵活多变, 它也容易陷入“维数灾难”, 比如平滑样条中惩罚系数的选取, 直接影响到回归曲线的光滑程度、对样本的拟合程度, 一旦失准就会造成过拟合或欠拟合. 除此之外, 由于样条、局部线性回归等没有给出具体的真实模型, 它们无法像最小二乘那样进行外推运算; 且估计精度很依赖样本的大小, 只有样本容量足够大时才会有较高的精度.

至于样条和局部线性回归之间的差异, 主要是源于两者采用的数学近似不同. 样条使用分段函数进行估计: 在两个控制节点之间使用同一函数得到估计值, 在两个不同的控制区域则使用不同的函数来进行估计. 于是对样条回归效果影响最大的是控制节点数量和位置的选取. 而局部线性回归利用的是真实模型的泰勒级数, 即用预测点邻近的各点来估计真实值. 具体距离多近才算是“邻近”, 则采用窗宽来度量, 它便成为了局部线性回归中影响力最大的超参数.

这些底层逻辑上的差异, 使得样条曲线比局部线性回归更为平滑, 自由度更高; 但是它也是更不稳定的回归方法, 在控制节点以外的区域有时甚至会比最小二乘回归产生更大的泛化误差. 在处理不同问题时, 两种方法应斟酌选用.