第六章　保形预测

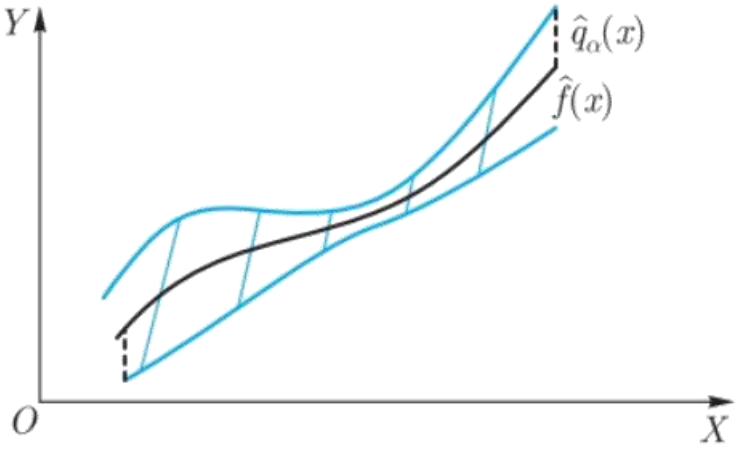
6.1　简介

6.1 简 介

- 保形预测提供了一种量化机器学习模型预测的不确定性的方法. 它不是像决策树、回归分析或支持向量机那样的特定算法, 而是一个可以与各种机器学习模型结合使用的框架. 保形预测输出的不仅仅是预测值, 还包括一个可信区间或概率置信水平, 表示预测的不确定性.
- 在进行预测的时候经常会思考这样的问题: 我们的预测有多好? 如果我们对一个新对象的类别进行预测, 有多少信心认为预测结果 $\hat{Y}$ 确实等于这个未知的真实的类别 Y 呢? 如果类别是一个数字, 我们得到的 $\hat{Y}$ 与真实的类别 Y 有多接近? 在机器学习中, 这些问题通常以过去的经验相当粗略地回答.
- 因此, 在使用机器学习进行预测时, 需要注意到预测值也是随机变量, 存在不确定性. 例如, 使用一个股票自动交易系统中的机器学习方法预测股票价格. 由于股票市场的高度不确定性, 机器学习的点预测可能与实际值有很大不同. 人工智能系统如果以高概率估计出覆盖目标真实值的范围, 交易系统就可以计算出最好和最差的情况, 并做出更明智的决定. 所以, 对实际值预测一个可信范围会更有意义.

6.1 简 介

- 保形预测 (conformal prediction) 是目前机器学习中热门的且非常灵活的预测技术. 即当我们处理回归或分类问题时, 给定输入, 保形预测可以输出一个预测区间或者一个预测集.
- 关于保形预测的理解, 跟传统的区间估计是类似的. 给定一种回归问题的预测方法, 保形预测产生一个 95% 的预测区间 $I^{0.05}$, 即该区间包含 Y 的概率至少为 95%, 通常 $I^{0.05}$ 也包含预测值 $\hat{Y}$. 我们称 $\hat{Y}$ 为点预测, 称 $I^{0.05}$为区间预测. 在分类的情况下, 类别 Y 有有限个可能值, $I^{0.05}$ 可能由这些值中的几个值组成, 或者在理想的情况下 $I^{0.05}$ 只是这些值中的一个值.
- 与传统区间估计所使用的方法不同的是, 保形预测并没有研究预测值的分布或渐近分布, 这也体现了保形预测的优势, 可以适用于任何模型, 因为研究渐近分布往往需要模型的假定. 保形预测在处理连续因变量的回归预测问题时, 对于单因变量的预测输出的是一个预测区间, 对于多因变量的预测输出的是一个预测域. 图 6.1 展示的是对单个连续因变量的区间预测.



6.1 简 介

- 保形预测在处理离散因变量的分类预测问题时, 输出的是预测集. 集合中的元素是由预测类别及对应的概率构成. 图 6.2 中展示了三种不同的输入值下, 对松鼠类型预测的保形预测集. 上述两图表明, 预测区间和预测集都可以保证以较高的概率覆盖真实值.

{ 黑松鼠 0.99 }

{ 黑松鼠, 0.82 灰狐狸, 0.03 桶, 0.02 雨桶, 0.02 }

{ 土拔鼠, 0.30 黑松鼠, 0.22 水貂, 0.18 黄鼠狼, 0.16 河狸, 0.03 臭鼬 0.01 }

6.1.1 简单运用

- 在介绍保形预测的流程之前, 我们要先知道两个定义: 一个是随机变量序列的可交换性 (exchangeability), 另一个是不符合度量 (nonconformity mea_x0002_sure). 在后面章节中, 保形预测的预测集既可代表回归问题的预测区间也可表示分类问题的预测集.

1. 可交换性

- 一个有限的随机变量序列是可交换的, 是指随机变量的联合概率分布对随机变量的排列不变.

$$P(X_1, X_2, \cdots, X_n) = P(X_{\pi(1)}, X_{\pi(2)}, \cdots, X_{\pi(n)})$$

  - 这里 $\pi(1), \pi(2), \cdots, \pi(n)$ 代表自然数 $1, 2, \cdots, n$ 的任意一个排列. 一个无限的随机变量序列是无限可交换 (infinitely exchangeable) 的, 是指它的任意一个有限子序列都是可交换的.

- 如果一个无限随机变量序列 $X_1, X_2, \cdots, X_n, \cdots$ 是独立同分布的, 那么它们是无限可交换的. 反之不然.

6.1.1 简单运用

2. 不符合度量

- 保形预测使用的不符合度量, 度量了一个新例子与旧例子的不同程度, 也可以说是度量预测值与真实值之间的不符合程度. 因此, 需要寻找一种方法来度量预测值与真实值之间的距离.
- 保形预测要求首先选择一个不符合度量, 而度量预测值和真实值之间距离的方法与数据类型和预测器的选择相关: 在回归问题中, 最常用的不符合度量是残差的绝对值: $R = |\ Y - \hat{Y}\ |$. 在分类问题中, 最常用的不符合度量是:$H = 1 - \hat{Y}$.
- 给定一个不符合度量, 保形算法对每一个误覆盖水平 α 产生一个预测集 $\hat{C}_n$. $\hat{C}_n$ 为 $1-\alpha$ 预测集; 它有至少 $1-\alpha$ 的概率包含真实值 Y.
- 在明确了以上两个概念之后, 我们介绍运用保形预测来构造有效的预测集的几个基本步骤. 假设有一个独立同分布的数据集:

$$\left\{\left(X_1^T, Y_1\right)^T, \cdots, \left(X_n^T, Y_n\right)^T\right\}$$

6.1.1 简单运用

- 其中, $\boldsymbol{X}$ 是输入的特征, Y 是因变量, n 是数据点的数量. 假设一个机器学习模型 f: $\boldsymbol{X} \rightarrow Y$ 已经被训练. 这个模型可以是一个经典的机器学习模型例如线性回归、支持向量机, 或者深度学习技术例如全连接或卷积网络. 目标是去估计模型输出的预测集. 下面我们将描述保形预测的步骤.

2. 保形预测的基本流程

- (1) 选择合适的不符合度量, 计算不符合分数
  - 选择一个适当的不符合度量 $S(\boldsymbol{X},Y) \in R$ 来测量模型输出 $\hat{Y}$ 与真实的 Y 之间的差异, 亦称评分函数, 得到的值称为不符合分数. 这个不符合度量非常重要, 因为它决定了我们能得到什么样的预测集. 例如, 在回归问题中, 可以用 $\left|\hat{Y}_i - Y_i\right|$ 作为评分函数. 通过这种方式, 所得到的保形预测集在预测值 $\hat{Y}_i$ 周围的 L_1 范氏球内; 在分类问题中, 可以用 $1 - \hat{Y}_i$ 作为评分函数, 其中 $\hat{Y}_i$ 是真实类别对应的预测概率.
- (2) 计算不符合分数的 $1-\alpha$ 分位数
  - 不符合分数的 $1-\alpha$ 分位数 $\hat{q}$ 是通过计算 n 个不符合分数 $S_1 = S(\boldsymbol{X}_1, Y_1), \cdots, S_n = S(\boldsymbol{X}_n, Y_n)$ 的 $1-\alpha$ 分位数得到的.

6.1.1 简单运用

- 在完全保形预测方法中, 需要训练m 次来计算评分并构造预测集, 其中 m 为因变量可能的取值的个数, 这样对算力的消耗无疑是巨大的. 为了降低计算复杂度, 可以采用分裂保形预测的方法, 分裂保形预测将整个训练集分割成合适的训练集和校准集 (calibrationset), 然后, 只对训练集进行训练, 在校准集上计算不符合分数.

■ (3) 使用模型预测和不符合分数的 $1-\alpha$ 分位数构造预测集

- 接下来, 使用模型预测值和不符合分数的 $1-\alpha$ 分位数构造保形预测集.假设新的输入为 $\boldsymbol{X}_{n+1}$, 得到的保形预测集可以表示为

$$\hat{C}\left(\boldsymbol{X}_{\mathrm{n}+1}\right)=\left\{Y: S\left(\boldsymbol{X}_{\mathrm{n}+1}, Y\right) \leq \hat{q}\right\}.$$

- 例如, 在使用保形预测进行回归分析时, 不符合度量 $S\left(\boldsymbol{X}_{n+1}, Y\right)=\left|Y-\hat{f}\left(\boldsymbol{X}_{n+1}\right)\right|$,其中 $\hat{f}$ 是用训练数据集得到预测器. 则此时得到的 $1-\alpha$ 预测区间可以表示为

$$\hat{C}\left(\boldsymbol{X}_{\mathrm{n}+1}\right)=\left[\hat{f}\left(\boldsymbol{X}_{\mathrm{n}+1}\right)-\hat{q}, \hat{f}\left(\boldsymbol{X}_{\mathrm{n}+1}\right)+\hat{q}\right].$$

6.1.2 预测集的有效性

- 对于保形预测的有效性讨论需要区分两种有效性: 保守的和精确的. 一般来说, 保形预测是保守有效的: 当它输出一个$1-\alpha$ 集 (也就是说, 一组预测集置信水平为$1-\alpha$时错误的概率不大于α, 并且在预测连续的例子时, 它所犯的错误之间几乎没有相关性. 这意味着, 根据大数定律, 在置信水平$1-\alpha$上的长期错误频率约为 α 或更小. 在实践中, 保守性往往不是很大, 尤其是当n 很大时, 经验结果显示, 长期误差的频率与 α 非常接近. 然而, 从理论的角度来看, 为了获得准确的有效性, 我们必须在预测过程中引入一个随机误差α, 其中 $1-\alpha$ 集出现错误的概率正好是 α, 错误在不同的试验中是独立产生的, 而长期错误的频率收敛到 α.

- 保形预测可以提供数学上严格的保证. 设 Y_{n+1} 为真值. Y 可以是分类问题中的一个类别, 也可以是回归问题中的一个实值. 设 $\hat{C}(\boldsymbol{X}_{n+1})$是一个预测集(或区间). 如果 Y_{n+1} 在 $\hat{C}(\boldsymbol{X}_{n+1})$内, 我们将其定义为 $\hat{C}(\boldsymbol{X}_{n+1})$ 覆盖 Y_{n+1}, 即 $Y_{n+1} \in \hat{C}(\boldsymbol{X}_{n+1})$. 然后, 给定一组同独立分布样本 $\left\{(\boldsymbol{X}_1^{\mathrm{T}}, Y_1)^{\mathrm{T}}, \cdots, (\boldsymbol{X}_n^{\mathrm{T}}, Y_n)^{\mathrm{T}}\right\}$,保形预测集满足以下覆盖保证

$$P\left(Y_{n+1} \in \hat{C}(\boldsymbol{X}_{n+1})\right) \geq 1-\alpha.$$

6.1.2 预测集的有效性

- 基于可交换性假设的覆盖保证的证明可以在文章 A Gentle Introductionto Conformal Prediction and Distribution-Free Uncertainty Quantification 的附录中找到. 以上覆盖率保证在有限样本的情况下也是成立的, 且与普通的统计推断通常都对变量的分布有更严格的假定 (比如高斯) 不同, 该保证仅需要非常弱的条件 (可交换性).

6.1.3 效率

- 针对分类问题, 保形预测得到的预测集可能包含一个类别, 也可能包含多个类别. 在处理回归问题时, 预测集通常是包含预测值的一个区间. 在保证一定置信水平的情况下, 保形预测集越小, 效率越高, 即分类预测集的元素越少或者回归预测区间越短, 则越好. 这就是我们通常所说的高效率保形预测集

6.2　保形回归

6.2 保形回归

- 保形预测可以用于回归, 也可以用于分类, 接下来我们首先以回归为例,解释保形预测的思想和原理. 本质上, 保形预测理论背后的基本思想与样本分位数有关.

6.2.1 保形预测基本思想

- 假设有独立同分布的随机变量样本 $U_1, \cdots, U_n$ (事实上, 此处独立同分布的假设可以被更弱的假设——可交换性替代). 对于一个给定的覆盖水平$\alpha \in (0, 1)$ 和另一个独立同分布的样本 U_{n+1}, 基于 $U_1, \cdots, U_n$ 定义样本分位数 $\hat{q}_{1-\alpha}$ 为

$$\hat{q}_{1-\beta} = \begin{cases} U_{(\lceil (n+1)(1-\alpha) \rceil)}, & \lceil (n+1)(1-\alpha) \rceil \leq n \\ \infty, & \text{其他.} \end{cases}$$

  - 可以得到

$$P\left(U_{n+1} \leq \hat{q}_{1-\alpha}\right) \geq 1-\alpha,$$

  - 其中, $U_{(1)} \leqslant \cdots \leqslant U_{(n)}$ 表示 $U_1, \cdots, U_n$ 的次序统计量. 通过可交换性, U_{n+1}在 $U_1, \cdots, U_n, U_{n+1}$ 中的排序是在集合 $\{1, \cdots, n+1\}$ 上均匀分布的, 因此以上有限样本覆盖性质是很容易被验证的.

- 在回归问题中, 观测到独立同分布的样本 $\boldsymbol{Z}_i = \left(\boldsymbol{X}_i^{\mathrm{T}}, Y_i\right)^{\mathrm{T}} \sim P(z), i = 1,2,\cdots,n$, 我们可以考虑以下方法来构造 Y_{n+1} 在新特征值 $\boldsymbol{X}_{n+1}$ 处的预测区间, 其中 $\left(\boldsymbol{X}_{n+1}^{\mathrm{T}}, Y_{n+1}\right)^{\mathrm{T}}$ 是分布 $P(z)$ 中一个独立的随机变量.

6.2.1 保形预测基本思想

- 按照上述思想, 可以构造以下预测区间:

$$\hat{C}_{\text{naive}}\left(\boldsymbol{X}_{n+1}\right)=\left[\hat{f}\left(\boldsymbol{X}_{n+1}\right)-\hat{F}_n^{-1}(1-\alpha), \hat{f}\left(\boldsymbol{X}_{n+1}\right)+\hat{F}_n^{-1}(1-\alpha)\right],$$

  - 其中, $\hat{f}$ 是估计的回归函数预测器, $\hat{F}_n$ 是拟合残差 $\left|Y_i-\hat{f}\left(\boldsymbol{X}_i\right)\right|, i=1,2,\cdots,n$ 的经验分布, $\hat{F}_n^{-1}(1-\alpha)$ 是 $\hat{F}_n$ 的 $(1-\alpha)$ 分位数. 假设估计的回归函数 fb 是准确的, 则该预测区间在大样本情况下是有效的 (即估计的拟合残差分布的$(1-\alpha)$ 分位数 $\hat{F}_n^{-1}(1-\alpha)$ 足够接近总体残差 $\left|Y_i-f\left(\boldsymbol{X}_i\right)\right|, i=1,2,\cdots,n$ 的 $(1-\alpha)$ 分位数). 保证 $\hat{f}$ 的准确性通常需要数据分布 $P(z)$ 和回归预测器 $\hat{f}$ 本身均满足一定的条件, 例如适当地选择预测模型和调优参数.

6.2.2 完全保形预测

- 一般来说, 上述方法得到的预测区间可以粗略地覆盖真实值, 因为拟合残差分布往往是向下倾斜的. 保形预测区间[73–76] 克服了上述原始方法预测区间的缺陷, 并且, 值得注意的是, 某种程度上保形预测可以保证提供适当的有限样本覆盖, 而无须对 $P(z)$ 和回归预测器 $\hat{f}$ 进行任何假设 (除了 f 是数据点的对称函数).
- 具体算法如下: 对于一个新的输入值 $\boldsymbol{X}_{n+1}$, 其对应的 Y_{n+1} 是未知的.因此我们为 Y_{n+1} 考虑一个试验集合 $\boldsymbol{Y}_{\text{trial}}$. 从中选取某一值 $Y \in \boldsymbol{Y}_{\text{trial}}$, 基于增广数据集 $\boldsymbol{Z}_1, \cdots, \boldsymbol{Z}_n, (\boldsymbol{X}_{n+1}^{\mathrm{T}}, Y)^{\mathrm{T}}$ 进行训练, 构造一个增广回归预测器 $\hat{f}_Y$. 接下来计算不符合分数, 即残差的绝对值.

$$R_{Y,i} = \left| Y_i - \hat{f}_Y(\boldsymbol{X}_i) \right|, \quad i = 1,2,\cdots,n,$$

$$R_{Y,i+1} = \left| Y - \hat{f}_Y(\boldsymbol{X}_{n+1}) \right|,$$

  - 并且基于 $n+1$ 个不符合分数对 $R_{Y,n+1}$ 进行排序, 得到

$$\pi(Y) = \frac{1}{n+1} \sum_{i=1}^{n+1} I\left(R_{Y,i} \leq R_{Y,n+1}\right) = \frac{1}{n+1} + \frac{1}{n+1} \sum_{i=1}^{n} I\left(R_{Y,i} \leq R_{Y,n+1}\right),$$

6.2.2 完全保形预测

- 即 $\pi(Y)$ 是增广样本中残差绝对值小于最后一个样本点的残差绝对值 $R_{Y,n+1}$的样本所占的比例, 其中 $I(\cdot)$ 表示示性函数. 由于数据点的可交换性和 $\hat{f}_Y$ 的对称性, 当我们在估计 Y_{n+1} 时, 发现构造的次序统计量 $\pi(Y_{n+1})$ 在集合$\{1/(n+1), 2/(n+1), \cdots, 1\}$ 上是均匀分布的, 这意味着

$$P\left\{(n+1)\pi(Y_{n+1}) \leq \lceil (1-\alpha)(n+1) \rceil\right\} \geq 1-\alpha.$$

- 上述推理可以通过如下假设检验过程理解, 即对应原假设 $H_0: Y_{n+1} = Y$. 在给定显著性水平 α 下, 我们有如下结论:
  - (1) 若事件 $\{(n+1)\pi(Y) > \lceil (1-\alpha)(n+1) \rceil\}$ 成立, 则拒绝原假设;
  - (2) $1-\pi(Y)$ 相当于 p 值, 若 $1-\pi(Y) < \alpha$, 则拒绝原假设.
- 通过在 $\boldsymbol{Y}_{\text{trial}}$ 遍历所有 Y 的取值, 可以得到在 $\boldsymbol{X}_{n+1}$ 处的保形预测区间,即

$$\hat{C}_{\text{conf}}(\boldsymbol{X}_{n+1}) = \left\{Y \in \boldsymbol{Y}_{\text{trial}} : (n+1)\pi(Y) \leq \lceil (1-\alpha)(n+1) \rceil\right\}.$$

6.2.2 完全保形预测

- 针对每一个新的自变量, 都必须重复以上步骤产生一个预测区间. 为了完整起见, 我们总结为如下**算法 1**.

算法 1: 完全保形预测

输入: 数据 $(\boldsymbol{X}_i^{\mathrm{T}}, Y_i)^{\mathrm{T}}, i=1,2,\cdots,n$, 显著性水平 $\alpha \in (0,1)$, 回归算法 $\mathcal{A}$, 用于构造预测区间的新的点 $\boldsymbol{X}_{\text{new}}=\{\boldsymbol{X}_{n+1}, \boldsymbol{X}_{n+2}, \cdots\}$, 和作为试验集合 $\boldsymbol{Y}_{\text{trial}}=\{Y_1, Y_2, \cdots\}$

输出: $\boldsymbol{X}_{\text{new}}$ 中每一个元素对应的预测区间

for $\boldsymbol{x} \in \boldsymbol{X}_{\text{new}}$ **do**

  for $Y \in \boldsymbol{Y}_{\text{trial}}$ **do**

$$\widehat{f}_Y = \mathcal{A}\left(\left\{\left(\boldsymbol{X}_1, Y_1\right), \cdots, \left(\boldsymbol{X}_n, Y_n\right), (\boldsymbol{x}, Y)\right\}\right)$$

$$R_{Y,i} = \left|Y_i - \widehat{f}_Y\left(\boldsymbol{X}_i\right)\right|, \quad i=1,2,\cdots,n,$$

以及 $R_{Y,n+1} = \left|Y - \widehat{f}_Y(\boldsymbol{x})\right|$

$$\pi(Y) = \left(1 + \sum_{i=1}^{n} I\left(R_{Y,i} \leqslant R_{Y,n+1}\right)\right) /(n+1)$$

  end for

$$\widehat{C}_{\text{conf}}(\boldsymbol{x}) = \{Y \in \boldsymbol{Y}_{\text{trial}} : (n+1)\pi(Y) \leqslant \lceil (1-\alpha)(n+1) \rceil\}$$

end for

返回 $\widehat{C}_{\text{conf}}(\boldsymbol{x})$, 对于所有 $\boldsymbol{x} \in \boldsymbol{X}_{\text{new}}$

6.2.2 完全保形预测

- 如果 $(\boldsymbol{X}_{n+1}^{\mathrm{T}}, Y_i)^{\mathrm{T}}, i=1,2,\cdots,n$ 为独立同分布的, 那么对于一个新的独立同分布的点 $(\boldsymbol{X}_{n+1}^{\mathrm{T}}, Y_{n+1})^{\mathrm{T}}$, 保形预测建立的预测区间 $\hat{C}_{\mathrm{conf}}(\boldsymbol{X}_{n+1})$有以下覆盖保证:

$$P\left(Y_{n+1} \in \hat{C}_{\mathrm{conf}}\left(\boldsymbol{X}_{n+1}\right)\right) \geq 1-\alpha,$$

  - 如果我们另外假设对所有 $Y \in \boldsymbol{R}$, 拟合的绝对残差 $R_{Y,i}=\left|Y_i-\hat{f}_Y\left(\boldsymbol{X}_i\right)\right|, i=1,2,\cdots,n$ 有一个连续的联合分布, 那么下式也成立:

$$P\left(Y_{n+1} \in \hat{C}_{\mathrm{conf}}\left(\boldsymbol{X}_{n+1}\right)\right) \leq 1-\alpha+\frac{1}{n+1}.$$

  - 该结论的证明详见文献 [77].

6.3　保形方法

6.3.1 分裂保形预测

- 上一节中讲述的完全保形预测方法计算量较大, 因为它要遍历所有 Y 的试验集. 即对于任何 $\boldsymbol{X}_{n+1}$ 和未知的 Y, 为了辨别某一个给定的 Y 是否包含在 $\hat{C}_{\text{conf}}\left(\boldsymbol{X}_{n+1}\right)$ 中, 我们在增广数据集 (其中包括新的点 $(\boldsymbol{X}_{n+1}, Y)$) 上重新训练模型, 并重新计算和排序绝对残差. 这个步骤需要在 $\boldsymbol{Y}_{\text{trial}}$ 集合里面的所有元素上重复进行, 计算成本很高. 在非线性回归问题中, 核密度估计或样条方法去训练模型的较高复杂度都会影响完全保形预测. 特别是在高维回归中, 可能会使用相对复杂的回归预测器, 如 Lasso 回归或深度学习等, 模型的训练仍会花费大量时间, 因此执行有效的完全保形预测仍然是一个有待解决的问题.

- 幸运的是, 有一种完全通用的方法, 我们称之为分裂保形预测, 其计算成本大大低于完全保形预测方法. 分裂保形方法采用样本分割将拟合步骤和排序步骤分离, 其计算代价与拟合步骤相对简单. **算法 2** 总结了分裂保形预测算法, 其他细节请参考文献 [78]. 在这里, 以及以后讨论分裂保形预测时, 为了简单起见, 我们假设样本量 n 是偶数, 当 n 是奇数时只需要非常小的改变.

6.3.1 分裂保形预测

算法 2：分裂保形预测

输入： 数据 $(\boldsymbol{X}_i^{\mathrm{T}}, Y_i)^{\mathrm{T}}, i = 1, 2, \cdots, n$, 显著性水平 $\alpha \in (0, 1)$, 回归算法 $\mathcal{A}$.

输出： $\boldsymbol{x} \in \mathbf{R}^p$ 上的预测区间

将 $\{1, 2, \cdots, n\}$ 随机分裂为两个大小相等的数据集 $\mathcal{I}_1, \mathcal{I}_2$

$$\widehat{f} = \mathcal{A}\left(\{(\boldsymbol{X}_i^{\mathrm{T}}, Y_i)^{\mathrm{T}} : i \in \mathcal{I}_1\}\right)$$

$$R_i = \left|Y_i - \widehat{f}(\boldsymbol{X}_i)\right|, i \in \mathcal{I}_2$$

$$d = \{R_i : i \in \mathcal{I}_2\} \text{ 中第 } k \text{ 小的值, 其中 } k = \left\lceil (1-\alpha)\left(\frac{n}{2}+1\right)\right\rceil$$

返回 $\widehat{C}_{\mathrm{split}}(\boldsymbol{x}) = [\widehat{f}(\boldsymbol{x}) - d, \widehat{f}(\boldsymbol{x}) + d]$, 对于所有 $\boldsymbol{x} \in \mathbf{R}^p$

- 如果 $(\boldsymbol{X}_i^{\mathrm{T}}, Y_i)^{\mathrm{T}}, i = 1, 2, \cdots, n$ 是独立同分布的, 对于一个新的数据点 $(\boldsymbol{X}_{n+1}^{\mathrm{T}}, Y_{n+1})^{\mathrm{T}}$, 由**算法 2** 建立的分裂保形预测区间 $\hat{C}_{\mathrm{split}}(\boldsymbol{X}_{n+1})$, 满足

$$P\left(Y_{n+1} \in \hat{C}_{\mathrm{split}}(\boldsymbol{X}_{n+1})\right) \geq 1 - \alpha.$$

6.3.1 分裂保形预测

- 另外, 如果我们假设残差 R_i, $i \in \mathcal{I}_2$ 具有连续的联合分布, 那么

$$P\left(Y_{n+1} \in \hat{C}_{\text{split}}\left(\boldsymbol{X}_{n+1}\right)\right) \leq 1-\alpha+\frac{2}{n+2}.$$

- 与完全保形预测方法相比, 分裂保形预测除了具有极高的效率外, 在内存需求方面也具有优势. 例如, 如果回归算法 $\mathcal{A}$ 涉及变量选择, 如 Lasso 或双向逐步回归等, 当在估计新的点 $\boldsymbol{X}_i$, $i \in \mathcal{I}_2$ 时, 只需运用基于样本 $\mathcal{I}_1$ 选择的变量, 来拟合模型计算基于样本 $\mathcal{I}_2$ 的残差. 因此算法 2 可以大大节省内存.
- 分裂保形预测也可以使用不均衡分裂实现. 采用 $\rho \in (0, 1)$, 令 $|\mathcal{I}_1| = \rho n$,则 $|\mathcal{I}_2| = (1-\rho)n$. 例如, 当回归算法很复杂的时候, 可以选择 $\rho > 0.5$ 使训练得到的回归预测器 $\hat{f}$ 更准确.

6.3.2 刀切法保形预测

- 刀切法 (Jackknife) 保形预测是一种计算复杂性介于完全保形法和分裂保形法之间的保形预测方法. 该方法利用留一残差的分位数来定义预测区间,具体内容如**算法 3** 所示.

算法 3: 刀切法保形预测

输入: 数据 $(\boldsymbol{X}_i^{\mathrm{T}}, Y_i)^{\mathrm{T}}, i = 1, 2, \cdots, n$, 显著性水平 $\alpha \in (0, 1)$, 回归算法 $\mathcal{A}$.

输出: $\boldsymbol{x} \in \mathbf{R}^p$ 上的预测区间

for $i \in \{1, 2, \cdots, n\}$ **do**

$$\widehat{f}^{(-i)} = \mathcal{A}(\{(\boldsymbol{X}_\ell, Y_\ell) : \ell \neq i\})$$

$$R_i = \left| Y_i - \widehat{f}^{(-i)}(\boldsymbol{X}_i) \right|$$

end for

$d = \{R_i : i \in \{1, 2, \cdots, n\}\}$ 中第 k 小的值, 其中 $k = \lceil n(1-\alpha) \rceil$

返回 $\widehat{C}_{\mathrm{Jack}}(\boldsymbol{x}) = [\widehat{f}(\boldsymbol{x}) - d, \widehat{f}(\boldsymbol{x}) + d]$, 对于所有 $\boldsymbol{x} \in \mathbf{R}^p$

6.3.2 刀切法保形预测

- 与分裂保形法相比, 刀切法的一个优点是它在构造绝对残差时使用更多的训练数据, 随后再构造分位数. 这意味着它通常可以产生更短的间隔长度.我们注意到, 由于对称性, 刀切法具有有限样本内覆盖性质:

$$P\left(Y_{n+1} \in \hat{C}_{\text{Jack}}\left(\boldsymbol{X}_i\right)\right) \geq 1-\alpha, \quad i=1,2,\cdots,n.$$

  - 但是就样本外覆盖而言 (真正的预测推断), 它的属性比较脆弱.

- 接下来看几篇经典的刀切法文章. Butler 和 Rothman[79] 表明, 在低维线性回归中, 刀切法在足够强的正则条件下产生渐近有效区间, 也意味着需要保证线性回归估计量的一致性. 最近, Steinberger 和 Leeb[80] 在高维回归中建立了刀切法区间的渐近有效性, 它们虽然不需要基本估计量 $\hat{f}$ 的一致性,但需要 $\hat{f}$ 的一致渐近均方误差有界的条件. 此外, Butler 和 Rothman[79] 以及 Steinberger 和 Leeb[80] 的回归分析均基于线性模型, 即回归函数是特征的线性函数, 特征独立于误差, 误差是同方差的. 但是这里介绍的刀切保形预测方法并不需要这些条件.

6.3.3 局部加权保形预测

- 当噪声是异方差的时候, 我们可以用局部加权的思路替换完全保形法或分裂保形法中的绝对残差, 即使用

$$V_i = \frac{\left|Y_i - \hat{f}(\boldsymbol{X}_i)\right|}{\hat{\sigma}(\boldsymbol{X}_i)}$$

  - 替代

$$R_i = \left|Y_i - \hat{f}(\boldsymbol{X}_i)\right|,$$

  - 其中 $\hat{\sigma}^2(\boldsymbol{x})$ 是绝对残差$\mathrm{Var}\left(\left|Y - \hat{f}(\boldsymbol{X})\right| \boldsymbol{X} = \boldsymbol{x}\right)$ 的方差函数的一个估计.(注意: $\hat{f}$ 和 $\hat{\sigma}$ 可以联合计算, 也可以单独计算.)

- 局部加权对预测波段的局部宽度可能有很大变化. 以下是分裂保形预测区间形式:

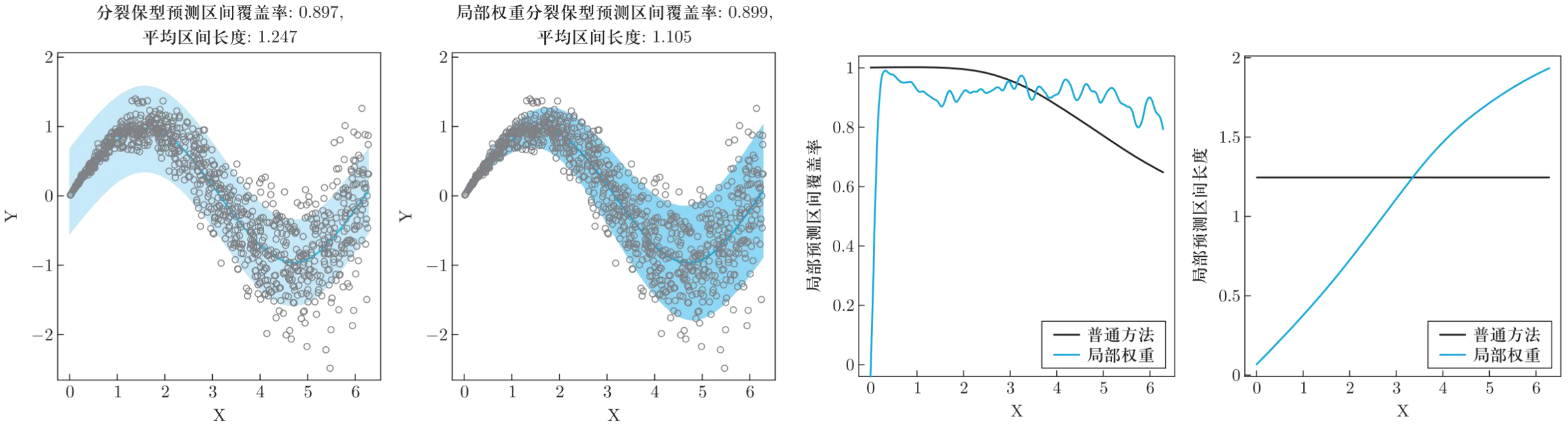
$$\widetilde{C}_{\text{local}}(\boldsymbol{x}) = \left[\hat{f}(\boldsymbol{x}) - \hat{\sigma}(\boldsymbol{x})\widetilde{q}, \hat{f}(\boldsymbol{x}) + \hat{\sigma}(\boldsymbol{x})\widetilde{q}\right],$$

  - 其中 $\widetilde{q}$ 表示 $V_i, i \in \mathcal{I}_2$ 的 $1-\alpha$ 分位数.

6.3.3 局部加权保形预测

异方差噪声示例

- 从图 6.3 可以看出, 通过引入局部权重, 可以获得能适应数据异质性的预测区域, 在保证和之前的方法具有相同有效性的同时, 它具有更小的平均长度, 局部的覆盖率也约等于我们想要的值 (比如说 0.9).



6.4　保形分类

6.4 保形分类

- 选择合适的不符合度量是进行保形预测的重要步骤, 我们在使用保形预测法进行分类时, 应该根据数据类型和预测方法来选择不符合度量. 下面介绍两种不同的不符合度量下保形预测在分类中的应用.

6.4.1 Softmax 法

- 在机器学习尤其是深度学习中, Softmax 是个常用而且比较重要的函数,尤其在多分类的场景中使用广泛. 假设预测目标是 K 个类别, 则通过神经网络多层变换后得到一个 K 维的输入 $z = (z_1, \cdots, z_K)$, 然后采用 Softmax 函数将 z 映射为 K 个 0 到 1 之间的实数, 并且归一化保证和为 1, 因此多分类的概率之和也刚好为 1. Softmax 函数形式如下:

$$\text{Softmax}(z)_k = \frac{e^{z_k}}{\sum_{j=1}^{K} e^{z_j}}, \quad k = 1,2,\cdots,K.$$

  - 在分类保形预测中, 可以将 Softmax 函数作为我们的预测函数, Softmax 方法经常用于图片分类, 下面简单介绍 Softmax 方法下的保形预测.

- 假设我们的预测任务是判断一个 p 维图像 $\boldsymbol{X}$ 属于哪一个类别 $Y \in \{1,2, \cdots, K\}$. 首先使用训练数据和 Softmax 函数得到预测模型, 也可称其为分类预测器 $\hat{f}$.

6.4.1 Softmax 法

- 分类预测器可以对于每个输入 $\boldsymbol{x}$ 输出对应每个类别的 Softmax分数:

$$\hat{f}(\boldsymbol{x}) = \text{Softmax}(z) \in [0,1]^K$$

  - 其中 $\hat{f}(\boldsymbol{x})_k = \text{Softmax}(z)_k$. 然后, 取适当数量的未用于训练的新数据点 $(\boldsymbol{X}_1^{\mathrm{T}}, Y_1)^{\mathrm{T}}, \cdots, (\boldsymbol{X}_n^{\mathrm{T}}, Y_n)^{\mathrm{T}}$ 作为校准数据集. 利用 $\hat{f}$ 和校准数据, 我们试图构建一个可能的类别的置信集 $\mathcal{T}(\boldsymbol{x}) \subset \{1, 2, \cdots, K\}$, 且这个置信集在下面意义下是有效的:

$$1-\alpha \leq P(Y_{n+1} \in \mathcal{T}(\boldsymbol{X}_{n+1})) \leq 1-\alpha+\frac{1}{n+1}.$$

  - 其中, $(\boldsymbol{X}_{n+1}^{\mathrm{T}}, Y_{n+1})^{\mathrm{T}}$ 来自同一分布的新测试点. 换句话说, 置信集包含正确类别的概率几乎正好是 $(1-\alpha)$, 称这种性质为边际覆盖. 为了使用 $\hat{f}$ 和校准数据构造 $\mathcal{T}$, 我们将执行一个简单的校准步骤. 首先, 设置不符合分数为

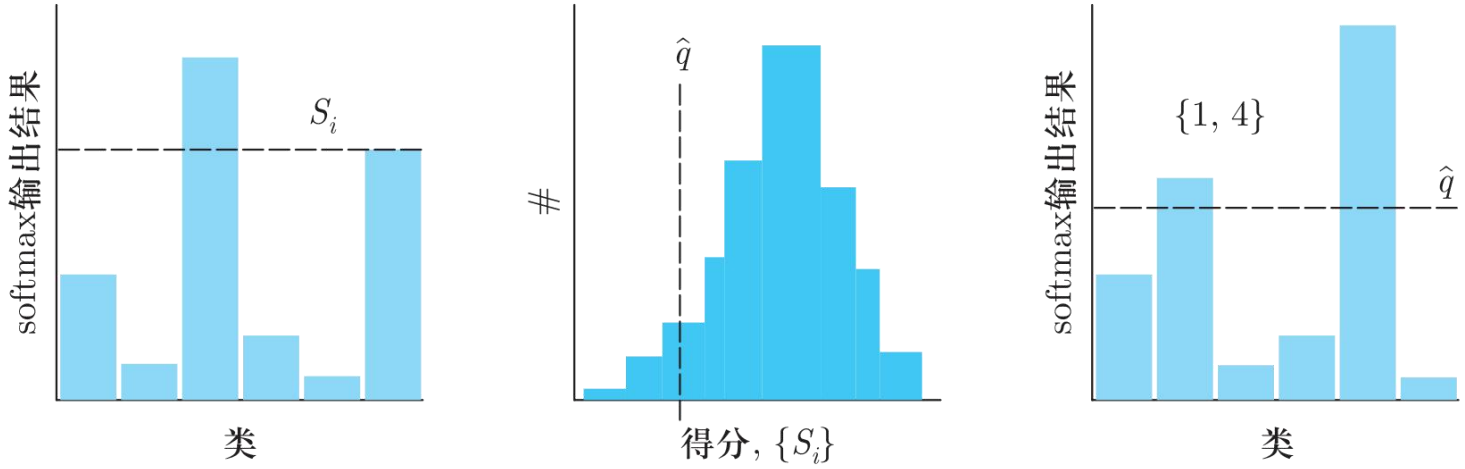
$$S_i = S(\boldsymbol{X}_i, Y_i) = 1 - \hat{f}(\boldsymbol{X}_i)_{Y_i},$$

6.4.1 Softmax 法

- 如果预测器不好, S_i 有可能会很大. 接下来定义 $\hat{q}$ 为 $S_1, \cdots, S_n$ 的 $\lceil (1-\alpha)(n+1) \rceil$ 分位数, 这实际上是经过小小修正的 n 个样本的 $(1-\alpha)$ 分位数.对于一个新的输入 $\boldsymbol{X}_{n+1}$ (Y_{n+1} 未知), 产生一个保形预测集:

$$\mathcal{T}(\boldsymbol{X}_{n+1}) = \{Y : S(\boldsymbol{X}_{n+1}, Y) < \hat{q}\} = \{Y : \hat{f}(\boldsymbol{X}_{n+1}, Y)_Y > 1-\hat{q}\}.$$

- 该预测集包含那些对应 Softmax 分数足够大的类别, 如图 6.4 所示. 值得注意的是, 不论我们应用什么样的预测模型, 选择任何分布的数据集, 这个算法得到的预测集都是具有覆盖保证的.



(1) 计算留出数据集中的得分　(2) 获取分位数　(3) 得出预测集

6.4.2 最近邻法

- 最近邻法也是进行分类时常用的方法之一. 在保形预测中, 我们使用最近邻法分类时, 分类预测器将新的样本分类为与其距离最近的样本所对应的类别.
- 假设有 n 个样本 $\boldsymbol{Z}_1=\left(\boldsymbol{X}_1^{\mathrm{T}}, Y_1\right)^{\mathrm{T}}, \boldsymbol{Z}_2=\left(\boldsymbol{X}_2^{\mathrm{T}}, Y_2\right)^{\mathrm{T}}, \cdots, \boldsymbol{Z}_n=\left(\boldsymbol{X}_n^{\mathrm{T}}, Y_n\right)^{\mathrm{T}}$,每个 $\boldsymbol{Z}_i$ 包含特征向量 $\boldsymbol{X}_i$ 和类别 $Y_i \in\{1, 2, \cdots, K\}$. 对于一个新的样本点 $\boldsymbol{Z}_{n+1}=\left(\boldsymbol{X}_{n+1}^{\mathrm{T}}, Y_{n+1}\right)^{\mathrm{T}}$, 我们只能观察到 $\boldsymbol{X}_{n+1}$ 而不知道 Y_{n+1}. 最近邻法寻找距离 $\boldsymbol{X}_{n+1}$ 最近的 $\boldsymbol{X}_i$, 并使用它的类别 Y_i 作为 Y_{n+1} 的预测值. 如果没有合适的预测器, 我们很难度量这个预测的正确性. 但是可以通过比较 $\boldsymbol{X}$ 到与其有相同类别的旧例子的距离和 $\boldsymbol{X}$ 到与其有不同类别的旧例子的距离, 来得到不符合度量, 因此我们用下式表示不符合分数:

$$S_i=S\left(\boldsymbol{X}_i, Y_i\right)=\frac{\min\left\{\left\|\boldsymbol{X}_j-\boldsymbol{X}_i\right\|: j \neq i, Y_j=Y_i\right\}}{\min\left\{\left\|\boldsymbol{X}_j-\boldsymbol{X}_i\right\|: j \neq i, Y_j=Y_i\right\}}$$

  - 不符合分数 S_i 大小, 也可以代表分类是否正确. 接下来定义 $\hat{q}$ 为 $S_1, \cdots, S_n$ 的 $\lceil(1-\alpha)(n+1)\rceil$ 分位数. 对于一个新的输入 $\boldsymbol{X}_{n+1}$, 我们得到保形预测集:

$$\mathcal{T}\left(\boldsymbol{X}_{n+1}\right)=\left\{Y: S\left(\boldsymbol{X}_{n+1}, Y\right)<\hat{q}\right\}$$

6.5　保形预测实践

实践代码