第十三章　模型平均

13.1　简介

13.1 简 介

- 模型平均是另一种集成学习技术, 通过独立训练多个相同类型的学习器,然后对它们的预测结果进行平均或投票, 可以降低过拟合风险, 提高模型的鲁棒性. 相较于随机森林, 模型平均是一种更通用的技术, 可以应用于不同类型的模型.
- 模型平均 (model Averaging) 通过整合多个独立模型的预测结果, 旨在提高整体估计的准确性和鲁棒性. 从统计学角度看, 模型平均可视为对概率分布的加权融合, 以降低估计的方差. 通过简单平均或加权平均多个模型的输出, 可以有效减小由单一模型引入的估计误差. 关键在于确保模型之间的适度独立性, 以最大程度地利用它们的差异性. 模型平均在处理有限数据、降低过拟合风险和提高整体模型鲁棒性方面具有广泛应用.

13.1.1 模型不确定性

- 在使用统计方法进行数据分析时, 研究者通常根据分析的问题假定统计模型, 并假定其为真实模型, 即生成实际数据的模型. 这种所谓的真实模型一般是根据研究者的经验或研究者使用数据里面的一些先验信息来建立的. 但这种假设往往是不正确的, 因为这个事先给定的模型只是真实模型的一个近似, 并且有可能是错误的, 而且在很多实际问题的数据分析研究中真实模型都是未知的, 这就是模型不确定性带来的不良影响. 从建模角度看, 模型不确定性一方面体现在采用的模型形式不确定, 例如函数形式、分布假定、模型结构或使用不同的预测变量等, 另一方面是由于模型选择导致的不确定性,例如不同的模型选择方法对应不同的模型选择结果或同一模型选择方法的选择结果会不稳定.

13.1.2 模型选择与模型平均

- 考虑到模型不确定性时, 一种传统的做法是利用模型选择方法从众多的候选模型中选出最优模型, 并将其看作真实模型进行后续的统计推断. 前面章节已经详细介绍了模型选择, 解决了过度复杂或简单的模型均可能使估计或者预测的方差偏大的问题. 研究者们常采用的模型选择方法和准则包括逐步回归、AIC、BIC、C_p、交叉验证、Lasso 回归、SCAD 或 MCP 方法等, 这些方法通过选择最优模型的过程看似在一定程度上解决了模型不确定性问题.但是模型选择方法导致了人们忽视模型选择过程所带来的不确定性, 即模型选择本身就是不确定性的.

- 模型平均方法起源于 20 世纪 60 年代, 早期最有影响的是 Bates 教授和诺贝尔经济学奖获得者 Granger 所做的工作, 他们通过组合两个无偏的预测来说明组合预测的优越性[155]. 他们把来自不同模型的估计或者预测通过一定的权重平均起来, 有时也称为模型组合, 一般包括组合估计和组合预测. 模型平均方法没有追求最优模型, 而以特定的权重对所有候补模型的统计结果进行平均, 在避免了遗失有用信息的同时也充分考虑了模型不确定性, 并且使得估计更加稳健.

13.1.2 模型选择与模型平均

- 因此, 模型平均解决了模型选择带来的一系列问题:
  - (1) 模型平均使用连续的权重去组合来自不同模型的估计, 在表示形式上, 模型选择可以看作模型平均的特例, 但它的权重只取 0 或者 1, 因而模型平均估计一般更加稳健.
  - (2) 模型选择通过一系列的模型选择方法选出一个最优模型, 这样就可能导致遗失其他模型的信息或者其他变量所特有的影响因素信息. 但是, 模型平均方法不会把某个选定的模型当作真实模型, 因为模型平均不轻易地排除任何模型, 给每一个模型赋予一个权重, 去充分利用每一个模型的信息, 因而减少信息的遗失.
  - (3) 模型选择也会导致模型不确定性, 因为可能选到一个与真实模型相差甚远的模型, 统计推断就可能存在很高的风险. 模型平均提供了一种保障机制, 规避了这种风险, 也可以说是避免了把鸡蛋放在同一个篮子里.
- 总的来说, 模型选择是统计推断的基础, 模型选择旨在选定一个最优模型, 基于该模型进行统计推断. 但是, 模型选择的不确定性会影响到统计推断, 从而使得分析结果会出现问题. 模型平均方法从估计和预测角度来看是模型选择的推广, 相比传统的模型选择方法来确定唯一的最优模型, 模型平均方法通过组合不同的模型进行估计和预测, 解决模型选择带来的模型不确定性问题, 常常能够减小估计风险, 得到更加有效的结论. 模型平均本质上是一类集成学习方法.

13.1.2 模型选择与模型平均

- 在模型平均的理论研究过程中, 如何确定组合权重是最重要的问题. 下面通过组合预测的角度, 介绍模型平均的两类方法：贝叶斯模型平均 (Bayesian model averaging, BMA) 和频率模型平均 (frequentist model averaging, FMA).

13.2　贝叶斯模型平均

13.2 贝叶斯模型平均

- 贝叶斯模型平均是一种基于贝叶斯理论并且将模型本身的不确定性考虑在内的方法. 基本步骤是: 设定待组合模型的先验概率和各个模型中参数的先验分布, 然后用经典的贝叶斯方法进行统计分析.
- 设 Y 为组合预测随机变量, D 为已获得的数据. 因为并不知道哪一个模型是最优模型, 即模型本身存在着不确定性, 我们设定 $\mathcal{M}=\{M_1, M_2, \cdots, M_K\}$ 代表所有可能模型组成的模型空间.
- 根据贝叶斯模型平均方法, 组合预测随机变量 Y 的后验分布为

$$\begin{aligned} P\left(y \middle| D\right) &= \sum_{k=1}^{K} P\left(y, M_k \middle| D\right) \\ &= \sum_{k=1}^{K} P\left(M_k \middle| D\right) P\left(y \middle| M_k, D\right) \left(\text{全概率公式}\right), \end{aligned} \tag{13.2.1}$$

13.2 贝叶斯模型平均

- 其中 $P(M_k|D)$ 为给定数据 D 的条件下模型 M_k 的后验分布, 反映了研究人员对于真实模型的不确定性.根据贝叶斯公式, 其形式为

$$P(M_k|D) = \frac{P(D|M_k)P(M_k)}{P(D)} \text{(贝叶斯公式)}$$

$$= \frac{P(D|M_k)P(M_k)}{\sum_{k=1}^{K} P(D|M_k)P(M_k)} \text{(全概率公式)}, \tag{13.2.2}$$

- 其中 $P(M_k)$ 为候补模型 (candidate model)Mk 的先验概率, 在没有特别先验信息的条件下可取均匀分布, 即 $P(M_k) = 1/K$; $P(D|M_k)$ 为模型M_k的似然函数. 在参数模型假设下, 假定θ_k为模型M_k的参数向量, 例如线性回归模型θ_k包含了回归系数和误差的方差,$P(\theta_k|M_k)$是给定模型M_k下θ_k的先验概率函数,P(D|θ_k, M_k)是给定模型M_k下参数化的似然函数. 可以推出

$$P(D|M_k)\int P(D|\boldsymbol{\theta}_k, M_k)P(\boldsymbol{\theta}_k|M_k)\mathrm{d}\boldsymbol{\theta}_k. \tag{13.2.3}$$

13.2 贝叶斯模型平均

- 从组合预测随机变量的后验分布可以发现, 贝叶斯模型平均方法实际上是以模型的后验分布为权重, 对所有模型的预测后验分布进行加权. 使用贝叶斯模型平均的关键在于确定组合的模型以及各单项模型的后验概率, 即权重. 另外也可以看出, 贝叶斯模型平均有两个困难：第一是 $P(D|M_k)$ 计算中涉及积分运算, 如果模型复杂, 积分也会变得困难; 第二是候补模型的个数也需要科学方法去确定.

- 记 μ_k 和 σ_k^2 为给定数据 D 和候补模型 M_k 的均值和方差, 则组合预测随机变量 Y 的条件期望和条件方差表示为 $E(Y \mid D, M_k) = \mu_k$ 和 $\mathrm{Var}(Y \mid D, M_k) = \sigma_k^2$. 令候补模型 M_k 的后验概率为 $P(M_k \mid D) = \omega_k$, 则组合预测随机变量 Y 的条件期望分解形式为

$$\begin{aligned}\mu(\boldsymbol{\omega}) &= E(Y|D) = E_{\mathcal{M}}\left[E(Y|D,\mathcal{M})\right] \\ &= \sum_{k=1}^{K} P(M_k|D)E(Y|D,M_k) \\ &= \sum_{k=1}^{K} \omega_k \mu_k ,\end{aligned}$$

13.2 贝叶斯模型平均

- $\boldsymbol{\omega}=(\omega_1,\cdots,\omega_K)^{\mathrm{T}}$ 在参数模型假设下, 我们可以使用贝叶斯估计框架, 使用方程 (13.2.3) 得到似然函数 $P(D|M_k)$ 的估计. 接下来使用公式 (13.2.2) 和给定的候补模型 M_k 的先验概率 $P(M_k)$, 估计出所有候补模型的权重 $\hat{\omega}_k$. 另外, 我们基于模型 M_k 的后验概率分布 $\mathrm{P}(y|M_k, D)$ 估计出 $\hat{\mu}_k$, 则贝叶斯模型平均估计的预测值是

$$\hat{\mu}(\hat{\boldsymbol{\omega}})=\sum_{k=1}^{K}\hat{\omega}_k\hat{\mu}_k.$$

- 另外, 条件方差 $\mathrm{Var}(Y\,|\,D)$ 可以分解为

$$\begin{aligned}\mathrm{Var}(Y|D)&=E_{\mathcal{M}}\left[\mathrm{Var}(Y|D,\mathcal{M})\right]+\mathrm{Var}_{\mathcal{M}}\left[E(Y|D,\mathcal{M})\right]\\&=\sum_{k=1}^{K}\omega_k\sigma_k^2+\sum_{k=1}^{K}\omega_k(\mu_k-\mu)^2.\end{aligned}$$

- 研究者认为上述条件方差被作为衡量模型不确定性的基础. $E_{\mathcal{M}}[\mathrm{Var}(Y\,|D,\mathcal{M})]$ 被认为是模型**结构内方差**, $\mathrm{Var}_{\mathcal{M}}[E(Y\,|\,D,\mathcal{M})]$ 则是**结构间方差**, 是由模型结构的不确定性引起的, 原因是当能够确定真实模型时, $\mathrm{Var}_{\mathcal{M}}[E(Y|D,\mathcal{M})]$ 为零.

13.3　频率模型平均

13.3 频率模型平均

- 与贝叶斯模型平均相比较, 频率模型平均方法不需要考虑如何设置候补模型的先验概率, 模型估计和权重估计完全由数据确定. 频率模型平均过程为：假设有 K 个模型 (或者 K 种估计方法), 每个模型 (或方法) 估计出一个预测值, 不妨将其写为

$$\hat{u}_k, \quad k=1,2,\cdots,K,$$

  - 那么频率模型平均方法得到的最终估计的预测值为

$$\hat{u}=\omega_1\hat{u}_1+\omega_2\hat{u}_2+\cdots+\omega_K\hat{u}_K,$$

  - 其中$\boldsymbol{\omega}=(\omega_1,\cdots,\omega_K)^{\mathrm{T}}$ 为权重向量, 通常满足 $0\le\omega_k\le1,\sum_{k=1}^{K}\omega_k=1.$

- 若定义权重系数为示性函数, 即令

$$\omega_k=I\{第k个模型被选到\},$$

  - 则模型平均变为模型选择, 从这里可以看出模型选择是模型平均的一个特例,模型平均是模型选择的推广. 和贝叶斯模型平均方法一样, 频率模型平均法也需要确定权重值, 那么怎么求取候补模型对应的权重呢？下一节介绍几种权重选择方法.

13.4　权重选择方法

13.4.1 基于信息准则

- 所谓信息准则 (IC) 的权重选择方法即对于每一个候补模型给出一个基于信息准则的得分, 最后依据得分多少来描述候补模型的重要性. 根据上节介绍的模型选择的知识, 我们可以先计算每一个模型的 AIC 和 BIC 值, 然后通过如下公式计算各个组合权重:

$$\omega_k = \frac{\exp\left(-\frac{\mathrm{IC}_k}{2}\right)}{\sum_{k=1}^{K}\exp\left(-\frac{\mathrm{IC}_k}{2}\right)}, \tag{13.4.1}$$

  - 其中 K 表示候补模型集合中模型的个数, ω_k表示第 k 个候补模型的权重;IC_k 为第 k 个候补模型的 AIC 或 BIC 值. 由上式可看出基于信息准则的组合权重计算方法比较简单, 只需计算每个候补模型下的 AIC 或 BIC 值即可,对应的模型平均方法称为 AIC 模型平均和 BIC 模型平均. 由于计算方便, 基于信息准则的思路是比较常用的权重选择方法.

13.4.2 基于马洛斯准则

- 在模型选择准则中介绍了 C_p 准则, C_p 准则的全称为 Mallows's C_p 准则.我们接下来要介绍的马洛斯准则是对 Mallows's C_p 准则的推广, 基于该准则的权重选择方法是最小化马洛斯准则来得到组合预测的权重. 下面以线性模型为例介绍具体的过程.
- 记响应变量的观测样本 $\boldsymbol{Y} = (Y_1, \cdots, Y_n)^{\mathrm{T}}$, 估计的预测值记为 $\hat{\boldsymbol{\mu}} = (\hat{\mu}_1, \cdots, \hat{\mu}_n)^{\mathrm{T}}$, 协变量样本矩阵记为 $\mathbf{X} = (\boldsymbol{X}_1, \cdots, \boldsymbol{X}_n)^{\mathrm{T}}$, 其中 $\boldsymbol{X_i} = (1, X_{i1}, \cdots, X_{ip})^{\mathrm{T}}$, p 是变量的个数, 误差项 $e = (e_1, \cdots, e_n)^{\mathrm{T}}$. 如果用传统线性模型去描述 p 个协变量与响应变量之间的线性关系表示为

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + e,$$

  - 其中 $\boldsymbol{\beta}$ 为协变量对应的系数向量, 形式为 $\boldsymbol{\beta} = (\beta_0, \beta_1, \cdots, \beta_p)^{\mathrm{T}}$.
- 假设研究者采用 K 个候补模型去获得 $\hat{\boldsymbol{\mu}}$, 其中第 k 个采用的候补模型为

$$M_k : \boldsymbol{Y} = \boldsymbol{X}_k \boldsymbol{\beta}_k + e,$$

13.4.2 基于马洛斯准则

- 其中 $\mathbf{X}_k = (\boldsymbol{X}_{k1}, \cdots, \boldsymbol{X}_{kn})^{\mathrm{T}}$ 是 $n \times p_k$ 的第一列为 1 的矩阵, 即它是由 X除去第一列外的任意 $p_k - 1$ 列组成的, $\boldsymbol{\beta}_k$ 是其相应的p_k维的系数向量. 一般情况下$\mathbf{X}_k^{\mathrm{T}}\mathbf{X}_k$是可逆的, 故 $\boldsymbol{\beta}_k$ 基于第 k 个候补模型的最小二乘估计为 $\hat{\boldsymbol{\beta}}_k = (\mathbf{X}_k^{\mathrm{T}}\mathbf{X}_k)^{-1}\ \mathbf{X}_k^{\mathrm{T}}\ \boldsymbol{Y}$. 相应地, $\boldsymbol{\mu}_k$ 的估计为 $\hat{\boldsymbol{\mu}}_k = \mathbf{X}_k(\mathbf{X}_k^{\mathrm{T}}\ \mathbf{X}_k)^{-1}\ \mathbf{X}_k^{\mathrm{T}}\ \boldsymbol{Y} = \boldsymbol{H}_k\boldsymbol{Y}$.其中 $\boldsymbol{H}_k = \mathbf{X}_k(\ \mathbf{X}_k^{\mathrm{T}}\ \mathbf{X}_k)^{-1}\ \mathbf{X}_k^{\mathrm{T}}$ 是帽子矩阵. 记权重向量 $\boldsymbol{\omega} = (\omega_1, \cdots, \omega_K)^{\mathrm{T}}$, 且满足

$$\mathrm{H} = \left\{ \boldsymbol{\omega} \in [0,1]^K : \sum_{k=1}^{K} \omega_k = 1 \right\},$$

- 那么 $\boldsymbol{\mu}$ 的模型平均估计为

$$\hat{\boldsymbol{\mu}}(\omega) = \sum_{k=1}^{K} \omega_k \hat{\boldsymbol{\mu}}_k . \qquad (13.4.2)$$

- 接下来, 采用以下**马洛斯准则**估计权重:

$$\mathbf{C}_n(\boldsymbol{\omega}) = \boldsymbol{\omega}^{\mathrm{T}}\ \hat{\mathbf{E}}^{\mathrm{T}}\ \hat{\mathbf{E}}\,\omega + 2\hat{\sigma}^2\boldsymbol{\omega}^{\mathrm{T}}\phi,$$

13.4.2 基于马洛斯准则

- 其中 $\hat{\mathbf{E}}=(\hat{e}_1,\cdots,\hat{e}_K)^{\mathrm{T}},\hat{e}_k=\boldsymbol{Y}-\hat{\boldsymbol{\mu}}_k$ 为基于第 k 个候补模型估计的残差向量, $\phi=(p_1,\cdots,p_K)^{\mathrm{T}}$, $\hat{\sigma}^2=\dfrac{\hat{e}_{\tilde{k}}^T\hat{e}_{\tilde{k}}}{n-\phi_{\tilde{k}}}$, $\tilde{k}$ 满足 $\phi_{\tilde{k}}=\max\{p_1,\cdots,p_K\}$. 通过极小化马洛斯准则所得到的权重为

$$\hat{\boldsymbol{\omega}}_{\mathrm{M}}=\arg\min_{\boldsymbol{\omega}\in\mathrm{H}}\mathrm{C}_n(\boldsymbol{\omega}). \tag{13.4.3}$$

- 该权重对应的模型平均估计成为马洛斯模型平均 (Mallows model average, MMA) 估计, 将其代入式 (13.4.2) 可得到 MMA 估计的观测值 $\hat{\boldsymbol{\mu}}(\hat{\boldsymbol{\omega}})$.

13.4.3 基于刀切法准则

- 基于刀切法 (Jackknife) 准则的模型平均方法是通过最小化刀切法准则得到组合预测的权重. 该方法适用于随机误差项异方差的情形, 即当$\mathrm{Cov}(e|X) = \mathrm{diag}(\ \sigma_1^2, \cdots, \sigma_n^2\ )$ 时可用基于刀切法准则的模型平均方法去选择权重.
- 假设研究者依然以线性模型为例, 并且采用 K 个候补模型, 其中第 k 个采用的候补模型使用的协变量观测矩阵是 $\mathbf{X}_k = (\boldsymbol{X}_{k1}, \cdots, \boldsymbol{X}_{kn})^{\mathrm{T}}$. 使用刀切法对第 k 个候补模型估计预测值的具体流程是：对于 $i = 1, 2, \cdots, n$,
- (1) 删除第 i 个样本, 使用最小二乘得到回归参数的估计

$$\hat{\boldsymbol{\beta}}_{(-i)} = \left(\mathbf{X}_{k(-i)}^{\mathrm{T}} \mathbf{X}_{k(-i)}\right)^{-1} \mathbf{X}_{k(-i)}^{\mathrm{T}} \boldsymbol{Y}_{(-i)},$$

  - 其中 $\mathbf{X}_{k(-i)}$ 和 $\boldsymbol{Y}_{(-i)}$ 分别是去掉第 i 个样本之后的 $(n-1) \times p_k$ 协变量观测矩阵和 $(n-1)$ 维的响应变量观测向量.
- (2) 估计出第 i 个样本的预测值为 $\tilde{\mu}_{ki} = \mathbf{X}_{ki}^{\mathrm{T}} \hat{\boldsymbol{\beta}}_{(-i)}$.
  - 最后得到基于第 k 个候补模型的对应 n 个样本的预测向量 $\tilde{\boldsymbol{\mu}}_k = \left(\tilde{\mu}_{k1}, \cdots, \tilde{\mu}_{kn}\right)^{\mathrm{T}}$.

13.4.3 基于刀切法准则

- 注意到 $\mathbf{X}_{k(-i)}^{\mathrm{T}}\mathbf{X}_{k(-i)}=\mathbf{X}_{k}^{\mathrm{T}}\mathbf{X}_{k}-\boldsymbol{X}_{ki}\boldsymbol{X}_{ki}^{\mathrm{T}}$, 根据 Sherman–Morrison 公式, 可以推导出

$$\tilde{\mu}_{ki}=\sum_{j\neq i}\frac{H_{k,ij}}{1-H_{k,ii}}Y_{j}, \tag{13.4.3}$$

- 其中 $H_{k,ij}$ 代表第 k 个模型的帽子矩阵第 i 行第 j 列的元素. 上式表明 $\tilde{\mu}_{ki}$ 不依赖于 Y_i. 可以用矩阵表示第 k 个候补模型的预测向量为

$$\tilde{\boldsymbol{\mu}}_{ki}=\left(\boldsymbol{D}_{k}\left(\boldsymbol{H}_{k}-\boldsymbol{I}_{n}\right)+\boldsymbol{I}_{n}\right)\boldsymbol{Y},$$

- 其中 $\boldsymbol{D}_k$ 是对角矩阵, 它的第 i 个对角元素为 $(1-H_{k,ii})^{-1}$, 并且 $\boldsymbol{I}_n$ 是单位矩阵.

- 令 $\tilde{\mathbf{E}}=\left(\tilde{e}_{1},\tilde{e}_{2},\cdots,\tilde{e}_{K}\right)^{\mathrm{T}}$, $\tilde{e}_{k}=\boldsymbol{Y}-\tilde{\boldsymbol{\mu}}_{k}$ 为基于第 k 个模型估计的残差向量, 而且 $\tilde{e}_k$ 是弃一的交叉验证的残差向量, 在统计学中弃一的交叉验证又称为刀切法. **刀切法准则**定义为

$$\mathrm{J}_{n}\left(\boldsymbol{\omega}\right)=\boldsymbol{\omega}^{\mathrm{T}}\,\tilde{\mathbf{E}}^{\mathrm{T}}\,\tilde{\mathbf{E}}\,\boldsymbol{\omega}. \tag{13.4.4}$$

13.4.3 基于刀切法准则

- 并且通过极小化刀切法准则所得到的权重为

$$\hat{\boldsymbol{\omega}}_{\mathrm{J}} = \arg\min_{\omega \in H} \mathrm{J}_n\left(\boldsymbol{\omega}\right),$$

- 并且基于刀切法准则得出的组合模型权重对应的模型平均方法为刀切法模型平均 (Jackknife model average, JMA).

13.5　模型平均实践

实践代码