第八章　特征筛选

8.1　简介

8.1 简 介

- 模型选择和变量筛选的目标都是提高模型在新数据上的泛化性能, 避免过拟合并提高模型的可解释性. 模型选择侧重在一组不同的模型中选择一个最适合数据的模型, 涉及选择不同的算法、模型结构或超参数. 而变量筛选侧重在一个给定的模型中, 选择对目标变量最具预测能力的特征或变量.
- 尽管变量选择方法可以用来识别重要变量, 但是在维度 p 非常高的情况下, 即维度 p 随着样本数量 n 的增加呈现指数速率增长, $\log(p) = O(n^{\delta}), 0 < \delta < 1$, 例如基因数据, 此时用于优化的算法仍然非常昂贵. 在实践中, 我们可以自然地考虑一个两阶段的方法: 先特征筛选, 然后变量选择. 首先强调一下, 在机器学习领域普遍使用特征表示统计学里的变量, 两个概念本质上是等价的. 具体地说, 我们使用特征筛选方法将超高维 p 降低至中等尺度$d \leqslant n$, 一般情况下, 维度 d 随着样本数量 n 的增加呈现幂增长, $d = O(n^{\zeta}), 0 < \zeta < 1$. 然后再用变量选择方法从剩下的变量中选择真实模型. 如果在每一个降维阶段都保留了所有重要变量, 那么这个两阶段方法要经济得多. 接下来我们介绍诸多特征筛选方法, 其目标是尽可能多地丢弃噪声特征, 同时保留所有的重要特征.

8.1 简 介

- 在本章中, 我们采用以下符号. 设 Y 为响应变量, $\boldsymbol{X}=(X_1, \cdots, X_p)^{\mathrm{T}}$由 p 维预测变量组成, 由此得到 n 个独立的随机样本 $\left\{\boldsymbol{X}_i, Y_i\right\}_{i=1}^n$. $\boldsymbol{Y}=(Y_1, \cdots, Y_n)^{\mathrm{T}}$, $\mathbf{X}=(X_1, \cdots, X_n)^{\mathrm{T}}$ 是 $n \times p$ 的设计矩阵. 设 $\mathbf{X}_j$ 为 $\mathbf{X}$ 的第 j 列, 则 $\mathbf{X}=(\mathbf{X}_1, \cdots, \mathbf{X}_p)$. 我们稍微滥用了符号 $\mathbf{X}$ 和 $\mathbf{X}_j$, 但是在上下文中它们的含义是清楚的. 令 ε 是一个一般随机误差, $\varepsilon=(\varepsilon_1, \cdots, \varepsilon_n)^{\mathrm{T}}$. 设 $\mathcal{M}_*$ 代表一个尺寸为 $s=|\mathcal{M}_*|$ 的真实模型, $\hat{\mathcal{M}}$ 代表尺寸为 $\mathrm{d}=|\hat{\mathcal{M}}|$ 的选择模型. 对于不同的模型和背景, $\mathcal{M}_*$ 和 $\hat{\mathcal{M}}$ 的定义可能有所不同.

8.2　基于边际模型的特征筛选

8.2.1 边际最小二乘

- 对于线性回归模型, 其矩阵形式为

$$\boldsymbol{Y}=\boldsymbol{X}\boldsymbol{\beta}+\varepsilon. \tag{8.2.1}$$

  - 当 $p \gg n$ 时, $\mathbf{X}^{\mathrm{T}}\mathbf{X}$ 是奇异的, 因此 β 的最小二乘估计没有很好的定义. 在这种情况下, 岭回归特别有用. 模型 (8.2.1) 的岭回归估计量由下式给出

$$\hat{\beta}_{\lambda}=\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}+\lambda I_{p}\right)^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{Y},$$

  - 其中 λ 是一个岭参数. 回归分析章节中介绍过岭回归解决了多重共线性的问题, 从正则项看, 岭回归估计量是线性模型的带 L_2 惩罚的惩罚最小二乘的解.当 $\lambda \to 0$ 且 $\mathbf{X}$ 为满秩时, $\hat{\boldsymbol{\beta}}_{\lambda}$ 趋于最小二乘估计量, 而当 $\lambda \to \infty$ 时, λ $\hat{\boldsymbol{\beta}}_{\lambda}$ 趋于 $\mathbf{X}^{\mathrm{T}}\boldsymbol{Y}$. 这意味着当 $\lambda \to \infty$ 时, $\hat{\boldsymbol{\beta}}_{\lambda} \propto \mathbf{X}^{\mathrm{T}}\boldsymbol{Y}$.

- 假设所有的协变量和响应变量标准化, 它们的样本均值和方差分别为0和1. 因此$\frac{1}{n}\mathbf{X}^{\mathrm{T}}\boldsymbol{Y}$ 成为由响应变量和所有协变量之间的皮尔逊相关系数的样本形式组成的向量. 这促使人们使用皮尔逊相关系数作为特征筛选的边际统计量. 具体来说, 首先我们标准化 $\mathbf{X}_j$ 和 $\boldsymbol{Y}$, 然后计算

$$\omega_j=\frac{1}{n}\mathbf{X}_j^{T}\boldsymbol{Y}, \quad j=1,2,\cdots,p, \tag{8.2.2}$$

  - 即第 j 个预测变量和响应变量之间的样本相关系数.

8.2.1 边际最小二乘

- 直观来说, X_j 和 Y 之间的相关性越高, 说明 Xj 越重要. Fan 和 Lv[92] 提出根据 $|\omega_j|$ 对预测变量 X_j 的重要性进行排序 (sure independence screening,SIS), 并开发了一种基于皮尔逊相关系数的特征筛选过程, 也称为确定性筛选, 如下所示. 对于预先指定的比例 $\gamma \in (0, 1)$, 选择排名靠前的 $\lceil \gamma n \rceil$ 个预测变量来获得子模型:

$$\hat{\mathcal{M}}_\gamma = \left\{ 1 \le j \le p : |\omega_j| \text{是排序在前} \lceil \gamma n \rceil \text{中大的} \right\},$$

  - 其中 $\lceil \gamma n \rceil$ 表示 γn 的整数部分. 它将超高维度降至一个相对适中的尺度 $\lceil \gamma n \rceil$, 即 $\hat{\mathcal{M}}_\gamma$ 的大小, 然后对子模型 $\hat{\mathcal{M}}_\gamma$ 再采用变量选择方法. 我们需要设置 γ 值来进行筛选, 一般地, $\lceil \gamma n \rceil$ 的值可以取为 $\lceil n/\log(n) \rceil$.

8.2.2 边际最大似然

- 假设 $\{\boldsymbol{X}_i, Y_i\}_{i=1}^n$. 是前面介绍的广义线性模型的随机样本. 基于第 i 个样本 $\{Y_i, \boldsymbol{X}_i\}$, 用下式$\ell$

$$\ell\left(Y_i, \beta_0 + \boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{\beta}\right) = Y_i\left(\beta_0 + \boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{\beta}\right) - b\left(\beta_0 + \boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{\beta}\right)$$

  - 表示使用正则连接函数 b 的似然函数的负对数 (不失一般性, 离散参数取 ϕ =1). 注意, 当 $p \gg n$ 时, 负对数似然 $\sum_{i=1}^{n} \ell\left(Y_i, \beta_0 + \boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{\beta}\right)$ 的最小值无法给出很好的定义.

- 接下来考虑边际最大似然方法筛选出重要预测变量. 与线性回归模型的边际最小二乘估计相似, 假设每个预测变量进行了标准化, 均值为 0, 标准差为 1, 并定义第 j 个预测变量 X_j 的边际最大似然估计量 (marginal maximumlikelihood estimator, MMLE) $\hat{\boldsymbol{\beta}}_j^M$ 为

$$\hat{\boldsymbol{\beta}}_j^M = \left(\hat{\beta}_{j0}^M, \hat{\beta}_{j1}^M\right) = \arg\min_{\beta_{j0}, \beta_{j1}} \sum_{i=1}^{n} \ell\left(Y_i, \beta_{j0} + \beta_{j1} X_{ij}\right).$$

  - 可以将 $\hat{\beta}_j^M$ 的大小作为边际筛选统计量, 对 X_j 的重要性排序, 并通过给定的阈值 κ_n 选择子模型, 即

$$\hat{\mathcal{M}}_{\mathcal{K}_n} = \left\{1 \le j \le p : \left|\hat{\beta}_{j1}^M\right| \ge \mathcal{K}_n\right\}.$$

  - 阈值 κ_n 的作用与「 γn 」 中的 γ 是一样的, 都是选择一个合适的值确定子模型.

8.2.3 边际非参估计

- 假设有一个随机样本集 $\{\mathbf{X}_i, Y_i\}_{i=1}^n$. 来自可加模型:

$$Y = \sum_{j=1}^{p} f_j\left(X_j\right) + \varepsilon, \tag{8.2.3}$$

  - 其中 $\boldsymbol{X} = (X_1, \cdots, X_p)^{\mathrm{T}}$, ε 是条件均值为 0 的随机误差. 为了快速识别式 (8.2.3) 中的重要变量, 我们考虑以下 p 个边际非参数回归问题:

$$\min_{f_j} E\left[Y - f_j\left(X_j\right)^2\right], \tag{8.2.4}$$

  - 式 (8.2.4) 中的最小值是 $f_j(X_j) = E(Y \mid X_j)$, 即 Y 在 X_j 上的投影. 我们根据 $E\left(f_j^2\left(X_j\right)\right)$ 对式 (8.2.3) 中协变量的统计量进行排序, 并通过阈值选择一小组协变量.

- 为了获得边际非参数回归的样本形式, 我们使用B样条基. 令Sn为维度$l \geqslant 1$的多项式样条的空间, $\{\Psi_{jk}, k = 1, 2, \cdots, q_n\}$ 表示具有 $\left\|\Psi_{jk}\right\|_{\infty} \leqslant 1$ 的 B样条基, 其中$\|\cdot\|_{\infty}$ 是 sup 范数

8.2.3 边际非参估计

- 在某些平滑条件下, 非参数投影 $\{f_j\}_{j=1}^{p}$ 可以被 S_n 中的函数 f_{nj} 很好地近似计算, 即 $f_j \approx f_{nj}$. 并且对于任意的 $f_{nj} \in S_n$,存在某些系数 $\{\beta_{jk}\}_{k=1}^{q_n}$, 我们有

$$f_{nj}(x) = \sum_{k=1}^{q_n} \beta_{jk} \Psi_{jk}(x), \qquad 1 \le j \le p.$$

  - 此时, 边际回归问题的样本形式可以表示为

$$\min_{f_{nj} \in S_n} E\left(Y - f_{nj}(X_j)\right)^2 = \min_{\beta_j \in \mathrm{R}^{q^n}} E\left(Y - \boldsymbol{\Psi}_j^{\mathrm{T}} \boldsymbol{\beta}_j\right)^2, \tag{8.2.5}$$

  - 其中 $\boldsymbol{\Psi}_j \equiv \boldsymbol{\Psi}_j(X_j) = (\Psi_1(X_j), \cdots, \Psi_{qn}(X_j))^{\mathrm{T}}$ 表示 q_n 维基函数. 基于上述最小化的目标函数, 我们首先得到回归系数 $\boldsymbol{\beta}_j$ 的总体形式

$$\boldsymbol{\beta}_j = \left(E\left(\boldsymbol{\Psi}_j \boldsymbol{\Psi}_j^{\mathrm{T}}\right)\right)^{-1} E\left(\boldsymbol{\Psi}_j Y\right),$$

8.2.3 边际非参估计

- 其中 E 表示期望. 进而可将 $f_{nj}(X_j)$ 的最小二乘解的总体形式表示出来, 如下所示:

$$\boldsymbol{\beta}_j = \left(E\left(\boldsymbol{\Psi}_j \boldsymbol{\Psi}_j^{\mathrm{T}}\right)\right)^{-1} E\left(\boldsymbol{\Psi}_j Y\right),$$

- 基于观测的随机样本集 $\left\{\mathbf{X}_i, Y_i\right\}_{i=1}^n$ 利用矩估计方法, 即 $E\left(\boldsymbol{\Psi}_j \boldsymbol{\Psi}_j^{\mathrm{T}}\right)$ 和 $E\left(\boldsymbol{\Psi}_j Y\right)$ 用样本矩代替, 就可以得到 $\boldsymbol{\beta}_j$ 的估计了.

$$\boldsymbol{\beta}_j = \left(E\left(\boldsymbol{\Psi}_j \boldsymbol{\Psi}_j^{\mathrm{T}}\right)\right)^{-1} E\left(\boldsymbol{\Psi}_j Y\right),$$

- 接下来, 计算目标 $E\left(f_j^2\left(X_j\right)\right)$ 的统计量的表示形式. 首先, $E\left(f_j^2\left(X_j\right)\right)$ 用 $E\left(f_{nj}^2\left(X_j\right)\right)$ 去代替, $E\left(f_{nj}^2\left(X_j\right)\right)$ 用 $E\left(\boldsymbol{\Psi}_j^{\mathrm{T}}\left(X_j\right)\hat{\boldsymbol{\beta}}_j\right)^2$ 去近似.最后 $E\left(\boldsymbol{\Psi}_j^{\mathrm{T}}\left(X_j\right)\hat{\boldsymbol{\beta}}_j\right)^2$ 所对应的估计量为 $\left\|\hat{f}_{nj}\right\|_n^2$, 具有如下形式:

$$\left\|\hat{f}_{nj}\right\|_n^2 = \frac{1}{n}\hat{f}_{nj}\left(X_{ij}\right)^2 = \frac{1}{n}\sum_{i=1}^n\left\{\boldsymbol{\Psi}_j^T\left(X_{ij}\right)\hat{\boldsymbol{\beta}}_j\right\}^2.$$

8.2.3 边际非参估计

- 接下来, 引入阈值 $\mathcal{K}_n$, 根据边际非参数估计量 $\left\|\hat{f}_{nj}\right\|_n^2$ 对重要变量进行排序 (nonparametric independence screening, NIS).

$$\hat{\mathcal{M}}_{\mathcal{K}_n}=\left\{1 \leq j \leq p:\left\|\hat{f}_{nj}\right\|_n^2 \geq \mathcal{K}_n\right\}.$$

8.3　基于边际相关系数的特征筛选

8.3.1 广义和秩相关系数

- 上一节中的 SIS 方法对于具有超高维预测变量的线性回归模型表现良好. 众所周知, 皮尔逊相关系数是用来衡量线性相关性的. 然而对于超高维数据要确定一个回归结构非常困难. 如果线性模型指定错误, 那么 SIS 会失败,因为皮尔逊相关系数只能捕获每个预测变量和响应变量之间的线性关系. 因此, SIS 最有可能错过一些非线性的重要预测变量, NIS 方法也需要可加性的条件. 如果存在更加复杂的非线性关系, 例如协变量外面是更加复杂且未知的函数变换等, NIS 的特征筛选效果也会不好. 因此, 我们需要将转换后的协变量和响应变量之间的皮尔逊相关系数计算出来, 视为边际筛选统计量进行特征筛选.

- 为了捕获这种类型的非线性, Hall 和 Miller[93] 定义第 j 个预测变量 X_j和 Y 的广义相关系数为

$$\rho_g\left(X_j, Y\right)=\sup_{h \in \mathcal{H}} \frac{\operatorname{Cov}\left\{h\left(X_j\right), Y\right\}}{\sqrt{\operatorname{Var}\left\{h\left(X_j\right)\right\} \operatorname{Var}(Y)}}, \qquad j=1,2, \cdots, p.$$

  - 其中 $\mathcal{H}$ 是包含所有线性函数的一类函数, 例如, 它是一类给定次数的多项式函数. 请注意, 若 $\mathcal{H}$ 是所有线性函数的类, 则 $\rho_g\left(X_j, Y\right)$ 是 $h(X_j)$ 和 Y 之间皮尔逊相关系数. 因此, $\rho_g\left(X, Y\right)$ 被认为是传统皮尔逊相关系数的推广.

8.3.1 广义和秩相关系数

- 假设 $\{(X_{ij}, Y_i), i = 1, 2, \cdots, \mathrm{n}\}$ 是总体 (X_j, Y) 的随机样本. 广义相关系数$\rho_g(X_j, Y)$ 可以通过下式来估计:

$$\hat{\rho}_g(X_j, Y) = \sup_{h \in \mathcal{H}} \frac{\sum_{i=1}^{n} \left\{ h(X_{ij}) - \bar{h}_j \right\} (Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n} \left\{ h(X_{ij}) - \bar{h}_j \right\}^2 \sum_{i=1}^{n} (Y_i - \bar{Y})^2}},$$

- 其中, $\bar{h}_j = n^{-1} \sum_{i=1}^{n} h(X_{ij}), \bar{Y} = n^{-1} \sum_{i=1}^{n} Y_i$. 若已知函数类 $\mathcal{H}$, 则可以使用 $\left| \hat{\rho}_g(X_j, Y) \right|$ 筛选重要特征.

- 另外, 我们也可以对响应变量进行函数转换, 并定义转换后的响应变量和协变量之间的相关系数. 一般地, 转换回归模型被定义为

$$H(Y_i) = \boldsymbol{X}_i^{\mathrm{T}} \boldsymbol{\beta} + \varepsilon_i. \tag{8.3.1}$$

- Li 等[94] 通过在模型 (8.3.1) 中假定 $H(\cdot)$ 严格单调, 提出使用秩相关系数衡量每个预测变量的重要性. 他们没有使用之前定义的样本皮尔逊相关系数, 而是提出了边际秩相关系数

8.3.1 广义和秩相关系数

$$\hat{\omega}_j = \frac{1}{n(n-1)} \sum_{i \neq l}^{n} I\left(X_{ij} < X_{lj}\right) I\left(Y_i < Y_l\right) - \frac{1}{4}$$

- 来衡量第 j 个预测变量 X_j 对响应变量 Y 的重要性, 并将方法命名为 RCS(rank correlation screening). 注意, 边际秩相关系数等于响应变量与第 j 个预测变量之间的肯德尔 τ 相关系数的四分之一. 因此, 我们可以基于秩相关系数 $\hat{\omega}_j$ 的绝对值筛选重要预测变量, 即

$$\hat{\mathcal{M}}_{\kappa_n} = \left\{1 \leq j \leq p : \left|\hat{\omega}_j\right| > k_n\right\}.$$

- 其中 κ_n 为预设定的阈值.

8.3.2 确定独立秩筛选

- 我们知道, X_j 和 Y 之间的皮尔逊相关系数仅仅刻画了 X_j 和 Y 之间的线性相关性. Zhu 等人[95] 提出使用 X_j 和 $I(Y<y)$ 的皮尔逊相关系数来刻画 X_j 和 Y 之间的非线性相关性, 因为示性函数具有单调变换不变性, 即针对单调函数 $g(\cdot)$, 有 $I(Y<y)=I(g(Y)<g(y))$. 当 $j=1,2,\cdots,p$ 时, 假设 $E(X_j)=0$, $\mathrm{Var}(X_j)=1$, 得到随机变量 X_j 和 $I(Y<y)$ 相关系数表示为

$$\Omega_j(y)=\mathrm{Cov}\{X_j,I(Y<y)\}=E\{X_jE[I(Y<y)|X_j]\}.$$

  - 直观地说, 若 X_j 和 Y 是独立的, 则对于任意的 y, 都有 $\Omega_j(y)=0$. 另一方面,若 X_j 和 Y 是相关的, 则存在 y 使得 $\Omega_j(y)=0$. 他们提出使用

- 假设 $\{(\boldsymbol{X}_i,Y_i),i=1,2,\cdots,n\}$ 是来自 $\{\boldsymbol{X},Y\}$ 的随机样本. 为了便于表示, 我们假设样本预测变量都是标准化的. 对于任意给定的 y, $\Omega_j(y)$ 的样本矩估计量为

$$\hat{\Omega}_j(y)=n^{-1}\sum_{i=1}^{n}X_{ij}I(Y_i<y).$$

8.3.2 确定独立秩筛选

- 因此, ω_j 的估计量为

$$\hat{\omega}_j = \frac{1}{n}\sum_{k=1}^{n}\hat{\Omega}_j^2(Y_k) = \frac{1}{n}\sum_{i=1}^{n}\left\{\frac{1}{n}\sum_{i=1}^{n}X_{ij}I(Y_i<Y_k)\right\}^2, \quad j=1,2,\cdots,p.$$

- 接下来, 使用 $\hat{\omega}_j$ 对所有候选预测变量 Xj , j = 1, 2, · · · , p 的重要性进行排序,然后选择最前面的几个作为重要预测变量, 此过程被命名为确定独立秩筛选,简记为 SIRS (sure independent ranking screening).

8.3.3 距离相关系数

- 这一章介绍另一种相关系数来有效地度量预测变量与响应变量之间的线性与非线性相关关系, 进行超高维数据的特征筛选. Li, Zhong 和 Zhu[96] 提出了一种基于距离相关系数筛选方法 (distance correlation screening, DCS).前面介绍的皮尔逊相关系数、秩相关系数和广义相关系数仅仅定义了两个随机变量的相关关系, 而距离相关系数是定义了两个不同维度的随机向量的关系.

- 首先, 定义两个随机向量 $\boldsymbol{U} \in \mathbf{R}^{q1}$ 和 $\boldsymbol{V} \in \mathbf{R}^{q2}$ 之间的距离协方差为

$$\mathrm{d}\mathbf{cov}^2\left(\boldsymbol{U},\boldsymbol{V}\right)=\int_{\mathbf{R}^{q1+q2}}\left\|\phi_{\boldsymbol{U},\boldsymbol{V}}\left(\boldsymbol{t},\boldsymbol{s}\right)-\phi_{\boldsymbol{U}}\left(\boldsymbol{t}\right)\phi_{\boldsymbol{V}}\left(\boldsymbol{s}\right)\right\|^2\omega\left(\boldsymbol{t},\boldsymbol{s}\right)\mathrm{d}\boldsymbol{t}\mathrm{d}\boldsymbol{s}, \tag{8.3.2}$$

  - 其中 $\phi_{\boldsymbol{U}}(\boldsymbol{t})$ 和 $\phi_{\boldsymbol{V}}(\boldsymbol{s})$ 是 $\boldsymbol{U}$ 和 $\boldsymbol{V}$ 的边际特征函数, $\phi_{\boldsymbol{U},\boldsymbol{V}}(\boldsymbol{t},\boldsymbol{s})$ 是 $\boldsymbol{U}$ 和 $\boldsymbol{V}$ 的联合特征函数, 并且

$$\omega\left(\boldsymbol{t},\boldsymbol{s}\right)=\left\{c_{q1}c_{q2}\left\|\boldsymbol{t}\right\|_{q1}^{1+q1}\left\|\boldsymbol{s}\right\|_{q2}^{1+q2}\right\}^{-1},$$

  - $c_d=\pi^{(1+d)/2}/\Gamma\{(1+d)/2\}$ (这个选择是为了方便微分的计算).

8.3.3 距离相关系数

- 这里 $\|\phi\|^2 = \phi\bar{\phi}$, ϕ 表示复值函数, $\bar{\phi}$ 是 ϕ 的共轭. 由定义 (8.3.2) 可知, 当且仅当 $\boldsymbol{U}$ 和 $\boldsymbol{V}$是独立时, $\mathrm{dcov}^2(\boldsymbol{U},\boldsymbol{V}) = 0$. Székely, Rizzo 和 Bakirov[97] 证明了

$$\mathrm{dcov}^2\left(\boldsymbol{U},\boldsymbol{V}\right) = S_1 + S_2 - 2S_3,$$

- 其中

$$S_1 = E\left(\left\|\boldsymbol{U}-\tilde{\boldsymbol{U}}\right\|\left\|\boldsymbol{V}-\tilde{\boldsymbol{V}}\right\|\right), \qquad S_2 = E\left(\left\|\boldsymbol{U}-\tilde{\boldsymbol{U}}\right\|\right)E\left(\left\|\boldsymbol{V}-\tilde{\boldsymbol{V}}\right\|\right),$$

$$S_3 = E\left\{\left(\left\|\boldsymbol{U}-\tilde{\boldsymbol{U}}\right\|\middle|\boldsymbol{U}\right)E\left(\left\|\boldsymbol{V}-\tilde{\boldsymbol{V}}\right\|\middle|\boldsymbol{V}\right)\right\}.$$

- 并且 $\left(\tilde{\boldsymbol{U}},\tilde{\boldsymbol{V}}\right)$ 是 $(\boldsymbol{U},\boldsymbol{V})$ 的独立副本. 因此, $\boldsymbol{U}$ 和 $\boldsymbol{V}$ 之间的距离协方差可以通过代入对应的样本来估计. 具体来说, 基于总体 $(\boldsymbol{U},\boldsymbol{V})$ 中的随机样本$\{(\boldsymbol{U}_i,\boldsymbol{V}_i), i = 1, 2, \cdots, n\}$, 我们有

$$\widehat{\mathrm{dcov}}^2\left(\boldsymbol{U},\boldsymbol{V}\right) = \hat{S}_1 + \hat{S}_2 - 2\hat{S}_3.$$

8.3.3 距离相关系数

- 其中,

$$\hat{S}_1 = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\|\boldsymbol{U}_i - \boldsymbol{U}_j\|\|\boldsymbol{V}_i - \boldsymbol{V}_j\|,$$

$$\hat{S}_2 = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\|\boldsymbol{U}_i - \boldsymbol{U}_j\|\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\|\boldsymbol{V}_i - \boldsymbol{V}_j\|,$$

$$\hat{S}_3 = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{l=1}^{n}\|\boldsymbol{U}_i - \boldsymbol{U}_l\|\|\boldsymbol{V}_i - \boldsymbol{V}_l\|.$$

- 根据 $\boldsymbol{U}$ 和 $\boldsymbol{V}$ 的距离相关系数定义

$$\mathrm{dcorr}(\boldsymbol{U},\boldsymbol{V}) = \frac{\mathrm{dcov}(\boldsymbol{U},\boldsymbol{V})}{\sqrt{\mathrm{dcov}(\boldsymbol{U},\boldsymbol{U})\,\mathrm{dcov}(\boldsymbol{V},\boldsymbol{V})}},$$

- 根据以上距离协方差矩阵的估计过程, 可以得到距离相关系数估计量

$$\frac{\widehat{\mathrm{dcov}}(\boldsymbol{U},\boldsymbol{V})}{\sqrt{\widehat{\mathrm{dcov}}(\boldsymbol{U},\boldsymbol{U})\,\widehat{\mathrm{dcov}}(\boldsymbol{V},\boldsymbol{V})}}.$$

8.3.3 距离相关系数

- 假设响应变量 $\boldsymbol{Y}$ 是多维度的, 若从 p 个预测变量 X_j 中筛选出重要的预测变量, 则使用如下边际距离相关系数估计量:

$$\hat{\omega}_j = \frac{\widehat{\mathrm{dcov}}\left(\boldsymbol{Y}, X_j\right)}{\sqrt{\widehat{\mathrm{dcov}}\left(\boldsymbol{Y}, \boldsymbol{Y}\right)\widehat{\mathrm{dcov}}\left(X_j, X_j\right)}},$$

- 并且使用 $\hat{\omega}_j^2$ 来对预测变量的重要性进行排序.

8.4　高维分类数据的特征筛选

8.4.1 柯尔莫哥洛夫-斯米尔诺夫统计量

- Fan 和 Fan[98] 提出在高维二值分类中使用双样本 t 检验统计量作为特征筛选的边际统计量. 尽管基于双样本 t 检验统计量的特征筛选在高维分类问题中表现很好, 但它可能会被重尾分布或者具有异常值的数据破坏, 而且它是基于模型的, 当数据结构不满足假定模型时, 筛选结果也是失效的.
- 为了克服这些缺点, Mai 和 Zou[99] 提出了一种新的基于柯尔莫哥洛夫–斯米尔诺夫 (Kolmogorov-Smirnov) 统计量的二值分类特征筛选方法. 为了便于标记, 重新标记 $Y=+1,-1$ 为类标签. 令 $F_{1j}(x)=P(X_j \leqslant x \mid Y=1)$ 和F2j (x) = $P(X_j \leqslant x \mid Y=-1)$ 分别为给定 $Y=1,-1$ 时 X_j 的条件分布函数.因此, 若 X_j 和 Y 独立, 则 $F_{1j}(x)=F_{2j}(x)$. 基于这一观察, 上述条件分布函数的差可用于构建特征筛选的方法.
- 因此 Mai 和 Zou[99] 提出柯尔莫哥洛夫–斯米尔诺夫边际效用:

$$\omega_j=\sup_{x\in\mathbf{R}},\left|F_{1j}(x)-F_{2j}(x)\right|,$$

  - 并且将

$$\hat{\omega}_j=\sup_{x\in\mathbf{R}},\left|\hat{F}_{1j}(x)-\hat{F}_{2j}(x)\right|$$

8.4.1 柯尔莫哥洛夫–斯米尔诺夫统计量

- 作为特征筛选的边际统计量, Mai 和 Zou[99] 将这种特征筛选方法命名为柯尔莫哥洛夫–斯米尔诺夫统计量 (KS), 其中 $\hat{F}_{1j}(x)$ 和 $\hat{F}_{2j}(x)$ 为对应的经验条件分布函数, 即

$$\hat{F}_{1j}(x)=\frac{1}{n_1}\sum_{i:Y_i=1} I\left(X_{ij}\le x\right), \qquad \hat{F}_{2j}(x)=\frac{1}{n_2}\sum_{i:Y_i=-1} I\left(X_{ij}\le x\right),$$

- 其中 $n_1=\sum_{i=1}^{n} I(Y_i=1)$ 和 $n_2=\sum_{i=1}^{n} I(Y_i=-1)$,最后使用 $\hat{\omega}_j$ 来对预测变量的重要性进行排序.

8.4.2 均值–方差统计量

- Cui, Li 和 Zhong[100] 提出了一种利用均值–方差指数进行超高维分类问题的确定独立筛选方法, 形式如下:

$$\mathrm{MV}\left(X_i, Y\right)=E_{X_j}\left[\mathrm{Var}_Y\left(F\left(X_i \middle| Y\right)\right)\right]. \tag{8.4.1}$$

  - 它不仅保留了柯尔莫哥洛夫滤波器的优点, 而且允许分类响应变量具有 $O\left(n^k\right)$个发散的类别, 其中 $k \geqslant 0$. 假设分类响应变量 Y 有 K 个类别 $\{y_1, \cdots, y_k\}$.令 $F_j(x)=P\left(X_j \leqslant x\right)$ 表示第 j 个特征 X_j 的边际分布函数, $F_{jk}(x)=P\left(X_j \leqslant x \mid Y=y_k\right)$ 表示给定 $Y=y_k$ 时 X_j 的条件分布函数. 由于 Y 是离散响应变量, Cui, Li 和 Zhong[100] 推导出

$$\mathrm{MV}\left(X_i, Y\right)=\sum_{k=1}^{K} p_k \int\left[F_{jk}(x)-F_j(x)\right]^2 \mathrm{d}F_j(x).$$

  - 如果 Xj 和 Y 在统计上是独立的, 那么对于任意的 k 和 x, 理论上都有$F_{jk}(x)=F_j(x)$. 通过上式推导, 均值–方差是给定 $Y=y_k$ 时 X_j 的条件分布函数与 X_j 的无条件分布函数之间的克拉默–沃姆米塞斯 (Cramér-vonMises) 距离的加权平均, 其中 $p_k=P(Y=y_k)$. 他们进一步表明当 X_j 和 Y在统计上独立时, MV $(X_j \mid \mathrm{Y})=0$.

8.4.2 均值–方差统计量

- 设 $\{(X_{ij}, Y_i): 1 \leqslant i \leqslant n\}$ 是总体 (X_j, Y) 中大小为 n 的随机样本, 则均值方差估计量是

$$\widehat{\mathrm{MV}}(X_j, Y) = \frac{1}{n}\sum_{k=1}^{K}\sum_{i=1}^{n}\hat{p}_k\left[\hat{F}_{jk}(X_{ij}) - \hat{F}_j(X_{ij})\right]^2,$$

  - 其中 $\hat{p}_k = n^{-1}\sum_{i=1}^{n} I(Y_i = y_k)$, $\hat{F}_{jk}(x) = n^{-1}\sum_{i=1}^{n} I(X_{ij} \le x, Y_i = y_k)/\hat{p}_k$, $\hat{F}_j(x) = n^{-1}\sum_{i=1}^{n} I(X_{ij} \le x)$. 因此我们可以使用$\widehat{\mathrm{MV}}(X_j, Y)$筛选出重要的预测变量. 这个过程被称为基于 MV 的确定独立筛选方法.

8.4.3 类别自适应筛选统计量

- 假设观察到具有 K $(K>2)$ 类的分类响应变量 Y, 即 $\{y_1, y_2, \cdots, y_K\}$,对于所有的 $k=1, 2, \cdots, K$ 都有 $p_k = P(Y=y_k)>0$. 在大数据时代, 不同分类响应变量的样本的来源可能不同, 例如在不同的时间段或者在不同的实验方法下收集到. 换句话说, 具有分类响应变量的高维数据通常是异构的. 因此, 我们考虑
  - (1) (异构性) 一组重要的预测变量 $\mathcal{A}_k \equiv \{1 \leqslant j \leqslant p : P(W_k \leqslant \omega_k \mid \mathbf{X})$依赖于 $X_j\}$, 其中对于不同的 $k=1, 2, \cdots, K$, $W_k \equiv I(Y=y_k)$ 重要变量集合 $\mathcal{A}_k$ 可能不同;
  - (2) (稀疏性) 对于某个常数 $\alpha>0$ 的维度 $p=o\{\exp(n^\alpha)\}$, $|\mathcal{A}_k| = s_k = o(n)$, 其中 $|\mathcal{A}_k|$ 是 $\mathcal{A}_k$ 的基数, n 是样本大小.
- 为了寻找不同类别下的重要变量, Xie et al.[101] 考虑了给定 $I(Y=y_k)$ 时 X_j 的条件分布函数, 即 $F_{jk}(x) = P(X_j \leq x \mid Y=y_k) = \dfrac{P(X_j \leq x, Y=y_k)}{P(Y=y_k)}$, 并提出如下边际筛选指数

$$\tau_{jk} = E_{X_j}\left\{F_{jk}(X_j)\right\} - \frac{1}{2}.$$

8.4.3 类别自适应筛选统计量

- 很明显, 若 $I(Y=y_k)$ 与 X_j 独立, 则 $\tau_{jk}=0$. 令 $\{(\boldsymbol{X}_i, Y_i), i=1, 2, \cdots, n\}$是独立同分布随机样本. 定义 $\hat{p}_k=\frac{1}{n}\sum_{i=1}^{n} I\left(Y_i=y_k\right), k=1,2,\cdots,K$. 我们得到$\tau_{jk}, j=1, 2, \cdots, p$ 的样本估计为

$$\hat{\tau}_{jk}=\frac{1}{n+1}\sum_{i=1}^{n}\left\{\frac{1}{n}\sum_{i=1}^{n}\frac{I\left(X_{ij}\le X_{lj}, Y_i=y_k\right)}{\hat{p}_k}\right\}-\frac{1}{2}. \tag{8.4.2}$$

- 接下来, 根据 $\left|\hat{\tau}_{jk}\right|$ 值的大小, 对所有候选预测变量 X_j , j = 1, 2, · · · , p 的重要性进行排序. $\left|\hat{\tau}_{jk}\right|$ 也成为类别自适应筛选统计量, 并简记为 CAS (cate_x0002_gory adaptive screening) 方法.
- 如果研究者对影响一个类集 $\mathcal{G}\subset\{1, 2, \cdots, K\}$ 的重要变量感兴趣, 可以采用如下筛选统计量

$$\hat{\tau}_{j,\mathcal{G}}=\sup_{k\in\mathcal{G}}\left|\hat{\tau}_{jk}\right|.$$

- 特别地, 若筛选出影响到所有类别的重要变量, 则采用

$$\hat{\tau}_{j}=\sup_{k\in\{1,2,\cdots,K\}}\left|\hat{\tau}_{jk}\right|.$$

8.4.3 类别自适应筛选统计量

- 因此, 类别自适应筛选统计量方法非常灵活, 既可以筛选出影响某一类的重要变量, 也可以用于筛选影响某一类集以及所有类的重要变量.

8.5　特征筛选实践

实践代码