# 统计学习基础

Dai HBG (1800010660@pku.edu.cn)

2020 年 12 月 4 日

## 目录

**1 回归分析与线性模型** **3**
1.1 普通线性回归 . . . 3
1.2 子集选择 . . . 4
1.3 对数几率回归 . . . 6
1.3.1 参数估计 . . . 6
1.4 最大熵模型 . . . 7
1.5 线性判别分析 . . . 8
1.6 正交化回归 . . . 8
1.7 收缩方法 . . . 9
1.7.1 岭回归 . . . 9

**2 核方法** **9**
2.1 一维核光滑方法 . . . 9
2.1.1 局部线性回归 . . . 10
2.1.2 局部多项式回归 . . . 10
2.2 核密度估计和分类 . . . 11

**3 提升算法与集成学习** **11**
3.1 Adaboost 算法 . . . 12
3.2 向前分步算法与指数损失 . . . 13
3.3 梯度提升与树的大小 . . . 13

**4 无监督学习** **14**
4.1 原型聚类 . . . 14
4.1.1 $K$-均值聚类 . . . 14
4.1.2 学习向量量化 . . . 14
4.2 主成分分析 . . . 14
4.2.1 随机向量的主成分 . . . 14
4.2.2 样本主成分与奇异值分解 . . . 15
4.3 谱聚类 . . . 16

**5 神经网络** **17**
5.1 投影寻踪回归 . . . 17
5.2 单隐层神经网络 . . . 17

**6 支持向量机** **17**
6.1 线性可分支持向量机 . . . 17
6.2 软间隔最大化 . . . 18

**7 奇异值分解** **18**
7.1 完全、紧与截断奇异值分解 . . . 18
7.2 奇异值分解与矩阵近似 . . . 19

**8 树模型** **19**
8.1 决策树模型与特征选择 . . . 19
8.2 决策树生成算法 . . . 20
8.2.1 ID3 算法与 C4.5 算法 . . . 20
8.2.2 剪枝 . . . 20
8.2.3 CART . . . 21

**9 贝叶斯分类器** **21**
9.1 朴素贝叶斯分类器 . . . 22

**10 基展开与正则化** **22**
10.1 分段多项式与样条 . . . 22
10.1.1 自然三次样条 . . . 22
10.2 光滑样条 . . . 23
10.3 再生核希尔伯特空间 . . . 23

**11 特征选择与稀疏学习** **24**
11.1 特征选择 . . . 24
11.1.1 过滤式选择与包裹式选择 . . . 24
11.1.2 嵌入式选择与 $L_1$ 正则化 . . . 25
11.2 线性模型中的 LASSO . . . 25
11.3 ADMM 算法 . . . 26
11.4 LASSO 解的性质 . . . 26

**12 模型评估与选择** **28**
12.1 偏差、方差与预测误差 . . . 28
12.2 样本内误差及模型选择准则 . . . 28
12.3 交叉验证 . . . 29

**13 半监督学习** **29**

**A 关于线性代数** **29**
A.1 分块正定矩阵的逆 . . . 29
A.2 二阶优化方法 . . . 29

**B 关于概率论** **30**
B.1 分布的尾估计 . . . 30

# 1 回归分析与线性模型

## 1.1 普通线性回归

考虑概率模型

$$y = f(x_1, \cdots, x_p) + \varepsilon, \tag{1.1.1}$$

其中 $\varepsilon$ 为随机项, 称为*回归模型* (regression model). 当式 (1.1.1) 中的 $f$ 取为因变量的线性函数的时候, 就成为*线性回归模型* (linear regression model).

考虑**高斯-马尔可夫模型** (Gauss-Markov)

$$\mathbb{E}[\boldsymbol{Y}] = \boldsymbol{X}\boldsymbol{\beta}, \quad \mathrm{Var}(\boldsymbol{Y}) = \sigma^2 \boldsymbol{I}_n,$$

也就是假定误差项均值为 0, 且互不相关, 有相同方差 $\sigma^2$. 该模型也称为**独立观测线性模型**, 记为 $(\boldsymbol{Y}, \boldsymbol{X}\boldsymbol{\beta}, \sigma^2\boldsymbol{I}_n)$.

在一元回归模型中, 容易知道系数 $\beta_1$ 的最小二乘估计就是 $\boldsymbol{x}$ 和 $\boldsymbol{y}$ 的样本协方差除以 $\boldsymbol{x}$ 的样本方差. 如果假设误差服从同方差的独立的正态分布, 将 $\beta_0, \beta_1$ 视为参数, 从而对应有 $y_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2)$, 然后采用极大似然估计, 容易发现其等价于 LSE.

假定有训练数据集 $T = \{(x_1, y_1), \cdots, (x_N, y_N)\}$, 每个 $y_i$ 是标量,$X_i$ 有 $p$ 个分量. 按照一般的处理方式,$x$ 加多一个分量 1, 于是令 $\boldsymbol{X}$ 为 $N \times (p+1)$ 的列满秩矩阵, 每一行代表着训练集的输入, 然后基于最小二乘估计, 进行基本的求导运算可得系数估计

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}, \tag{1.1.2}$$

于是有 $\hat{\boldsymbol{y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}}$. 据此可知基于 LSE 的估计是 $\boldsymbol{\beta}$ 的无偏估计.

对于加权的最小二乘估计, 优化问题变为

$$(\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})^T\boldsymbol{W}(\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}) = \min_{\boldsymbol{\beta}} (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})^T\boldsymbol{W}(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}),$$

其中 $\boldsymbol{W}$ 是对角矩阵. 同样的求导可得加权最小二乘估计

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{y}.$$

现在假设对于固定的 $x_i$,$y_i$ 是不相关的, 且具有方差 $\sigma^2$, 从而由1.1.2可知 $\hat{\boldsymbol{\beta}}$ 的协方差矩阵为

$$\mathrm{Var}(\hat{\boldsymbol{\beta}}) = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\sigma^2. \tag{1.1.3}$$

在 $p = 1$ 的情形, 容易看出当样本方差越大的时候 $\mathrm{Var}(\hat{\beta}_1)$ 越小.

由式 (1.1.2) 可知 $\hat{\boldsymbol{y}} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}$, 令 $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T$, 得到 $\hat{\boldsymbol{y}} = \boldsymbol{H}\boldsymbol{y}$, 因此 $\boldsymbol{H}$ 称为帽矩阵. 容易验证 $\boldsymbol{H}$ 是对称半正定幂等矩阵, 且由矩阵的迹的性质可知 $\mathrm{tr}(\boldsymbol{H}) = p + 1$.

**命题 1.1.1** *假设 $y_i$ 的误差是不相关、零均值且方差相等的, 则*

$$\hat{\sigma}^2 = \frac{1}{N-p-1}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2$$

*是 $\sigma^2$ 的无偏估计, 且 $(N-p-1)\hat{\sigma}^2 \sim \chi^2_{N-p-1}$.*

**证明** 只需考虑误差 $\varepsilon_i = y_i - \overline{y_i}$, 注意到由 $\bar{\boldsymbol{y}} = \boldsymbol{X}\boldsymbol{\beta}$, 则 $\bar{\boldsymbol{y}} = \boldsymbol{H}\bar{\boldsymbol{y}}$, 则减去均值后有

$$\mathbb{E}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2 = \mathbb{E}\boldsymbol{\varepsilon}^T[\boldsymbol{I} - \boldsymbol{H}]^T[\boldsymbol{I} - \boldsymbol{H}]\boldsymbol{\varepsilon} = \mathbb{E}\boldsymbol{\varepsilon}^T\boldsymbol{\varepsilon} - \mathbb{E}\boldsymbol{\varepsilon}^T\boldsymbol{H}\boldsymbol{\varepsilon},$$

由 $\mathrm{tr}(\boldsymbol{H}) = p + 1$ 可知结论成立.

对于第二个命题, 考虑 $QR$ 分解 [1] $\boldsymbol{X} = \boldsymbol{Q}\boldsymbol{R}$, 则 $\boldsymbol{H} = \boldsymbol{Q}\boldsymbol{Q}^T\boldsymbol{y}$

$$(N-p-1)\hat{\sigma}^2 = \mathrm{tr}(((\boldsymbol{T} - \boldsymbol{H})\boldsymbol{y})^T((\boldsymbol{T} - \boldsymbol{H})\boldsymbol{y})),$$

[1] QR 分解见1.6

然后类似于一维的情形, 考虑正交变换 $\boldsymbol{z}=\boldsymbol{T}\boldsymbol{y}$, 使得 $\boldsymbol{T}$ 的前 $p+1$ 行是 $\boldsymbol{Q}^T$, 于是简单计算即得. □

由命题 1.1.1 的证明过程可以很容易验证 $D[e_i]=(1-h_{ii})\sigma^2$, 这里 $h_{ii}=\boldsymbol{H}_{i,i}$.

**定理 1.1.2 (高斯-马尔可夫定理)** 假设线性模型正确, 则对于任意参数 $\theta=a^T\boldsymbol{\beta}$,$\boldsymbol{\beta}$ 的最小二乘估计是最佳线性无偏估计 (best linear unbiased estimation, BLUE).

**证明** 首先有

$$\mathbb{E}[\boldsymbol{a}^T\hat{\boldsymbol{\beta}}]=\mathbb{E}\boldsymbol{a}^T(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\beta}=a^T\boldsymbol{\beta},$$

从而无偏性得到证明. 对于任意满足 $\mathbb{E}[\boldsymbol{c}^T\boldsymbol{y}]=a^T\boldsymbol{\beta}$ 的无偏估计, 由数理统计的知识, 只需要说明

$$\mathbb{E}[\boldsymbol{a}^T(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}][\boldsymbol{a}^T(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}-\boldsymbol{c}^T\boldsymbol{y}]=0.$$

考虑 $\theta$ 是一维的情形, 则由条件只需要证明

$$\mathbb{E}\{[\boldsymbol{a}^T(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T]\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^T[\boldsymbol{c}-\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{a}]\}=0,$$

由假设的误差性质, 只需证明

$$\boldsymbol{a}^T(\boldsymbol{X}^T\boldsymbol{X})^{-1}(\boldsymbol{a}-\boldsymbol{X}^T\boldsymbol{c})=0,$$

而由假设知

$$a^T\boldsymbol{\beta}=\boldsymbol{c}^T\boldsymbol{X}\boldsymbol{\beta},$$

对任意 $\boldsymbol{\beta}$ 都成立, 可知结论成立. □

令 $\sum_{i=1}^N(y_i-\overline{y_i})^2$ 表示总平方和 (sum of squares for total, SST),$\sum_{i=1}^N(\hat{y_i}-\overline{y_i})^2$ 表示回归平方和 (sum of squares for regression, SSR),$\sum_{i=1}^N(y_i-\hat{y_i})^2$ 表示残差平方和 (sum of squares for eotal, SSE), 则有

**命题 1.1.3** 依据以上的定义, 有

$$\text{SST=SSR+SSE.}$$

**证明** 只需说明

$$(\boldsymbol{X}\hat{\boldsymbol{\beta}}-\overline{\boldsymbol{y}})^T(\boldsymbol{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}})=0,$$

注意到 $\overline{\boldsymbol{y}}=\boldsymbol{X}\boldsymbol{\beta}$, 从而

$$(\boldsymbol{X}\hat{\boldsymbol{\beta}}-\overline{\boldsymbol{y}})^T(\boldsymbol{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}})=(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})^T\boldsymbol{X}^T(\boldsymbol{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}})=0.$$

最后一步是拉格朗日乘子法的结果. □

## 1.2 子集选择

记 $\boldsymbol{X}_q$ 为 $\boldsymbol{X}$ 前 $q$ 列构成的矩阵, 则称

$$\tilde{\boldsymbol{\beta}}_q=(\boldsymbol{X}_q^T\boldsymbol{X}_q)^{-1}\boldsymbol{X}_q^T\boldsymbol{y} \tag{1.2.4}$$

为选模型, 对应的式 (1.1.2) 称为全模型. 记 $\boldsymbol{X}_t$ 为为 $\boldsymbol{X}$ 后 $t$ 列构成的矩阵, 这里 $t=p+1-q>0$,$\boldsymbol{\beta}_q,\boldsymbol{\beta}_t$ 等做类似含义的解释. 记 $\hat{\boldsymbol{\beta}}_q$ 为采用全模型估计出的 $\boldsymbol{\beta}$ 的前 $q$ 项, 我们有

**定理 1.2.1** 若全模型正确, 则

$$\mathbb{E}(\tilde{\boldsymbol{\beta}}_q)=\boldsymbol{\beta}_q+\boldsymbol{A}\boldsymbol{\beta}_t,\quad \boldsymbol{A}=(\boldsymbol{X}_q^T\boldsymbol{X}_q)^{-1}\boldsymbol{X}_q^T\boldsymbol{X}_t. \tag{1.2.5}$$

**证明** 由式 (1.2.4) 可知

$$\mathbb{E}(\tilde{\boldsymbol{\beta}}_q)=(\boldsymbol{X}_q^T\boldsymbol{X}_q)^{-1}\boldsymbol{X}_q^T\mathbb{E}[\boldsymbol{y}]=(\boldsymbol{X}_q^T\boldsymbol{X}_q)^{-1}\boldsymbol{X}_q^T(\boldsymbol{X}_q,\boldsymbol{X}_t)(\boldsymbol{\beta}_q;\boldsymbol{\beta}_t)=\boldsymbol{\beta}_q+\boldsymbol{A}\boldsymbol{\beta}_t.$$

□

由定理 1.2.1 可知**一般选模型对系数的估计不是无偏的**, 除非选模型本身就正确, 或者 $\boldsymbol{X}$ 前 $q$ 列和后 $t$ 列正交, 此时后面的样本对估计 $\boldsymbol{\beta}$ 不起作用. 根据附录A.1小节有

$$(\boldsymbol{X}^T\boldsymbol{X})^{-1}=\begin{pmatrix}(\boldsymbol{X}_q^T\boldsymbol{X}_q)^{-1}+\boldsymbol{A}\boldsymbol{D}\boldsymbol{A}^T & \boldsymbol{B}\\ \boldsymbol{B}^T & \boldsymbol{D}\end{pmatrix}, \tag{1.2.6}$$

然后利用式1.1.3可以类似说明 $\mathrm{Var}(\hat{\boldsymbol{\beta}}_q)\geq\mathrm{Var}(\tilde{\boldsymbol{\beta}}_q)$, 这里不等号的意义是 $\mathrm{Var}(\hat{\boldsymbol{\beta}}_q)-\mathrm{Var}(\tilde{\boldsymbol{\beta}}_q)$ 是半正定的. 这说明即使全模型是正确的, **选模型不会使系数估计方差增大**.

定义参数 $\boldsymbol{\theta}$ 的有偏估计 $\tilde{\boldsymbol{\theta}}$ 的*平均平方误差矩阵* (mean square error matrix, MSEM) 为

$$\mathrm{MSEM}(\tilde{\boldsymbol{\theta}}):=\mathbb{E}[(\tilde{\boldsymbol{\theta}}-\boldsymbol{\theta})(\tilde{\boldsymbol{\theta}}-\boldsymbol{\theta})^T],$$

则容易证明

$$\mathrm{MSEM}(\tilde{\boldsymbol{\theta}})=\mathrm{Var}(\tilde{\boldsymbol{\theta}})+(\mathbb{E}[\tilde{\boldsymbol{\theta}}]-\boldsymbol{\theta})(\mathbb{E}[\tilde{\boldsymbol{\theta}}]-\boldsymbol{\theta})^T,$$

从而由式 (1.2.5) 及 (1.2.6) 可知

$$\mathrm{MSEM}(\tilde{\boldsymbol{\beta}}_q)=(\boldsymbol{X}_q^T\boldsymbol{X}_q)^{-1}\sigma^2+\boldsymbol{A}\boldsymbol{\beta}_t\boldsymbol{\beta}_t^T\boldsymbol{A}^T,\quad \mathrm{Var}(\hat{\boldsymbol{\beta}}_q)=(\boldsymbol{X}_q^T\boldsymbol{X}_q)^{-1}\sigma^2+\boldsymbol{A}\boldsymbol{D}\boldsymbol{A}^T\sigma^2,$$

由 $\mathrm{Var}(\hat{\boldsymbol{\beta}}_t)=\boldsymbol{D}\sigma^2$, 从而有

**定理 1.2.2** *若全模型正确, 则当 $\mathrm{Var}(\hat{\boldsymbol{\beta}}_t)\geq\boldsymbol{\beta}_t\boldsymbol{\beta}_t^T$ 时有 $\mathrm{Var}(\hat{\boldsymbol{\beta}}_q)\geq\mathrm{MSEM}(\tilde{\boldsymbol{\beta}}_q)$.*

直观理解, 定理 1.2.2 说的是, 当 $\boldsymbol{\beta}_t$ 比较难估计的时候, 不考虑它们可以提高 $\boldsymbol{\beta}_q$ 估计的平均平方误差.

利用命题 1.1.1 同样的手法可证明

**定理 1.2.3** *若全模型是正确的, 则*

$$\mathbb{E}[\tilde{\sigma_q^2}]=\sigma^2+\frac{\boldsymbol{\beta}_t^T\boldsymbol{D}^{-1}\boldsymbol{\beta}_t}{n-q}.$$

定理 1.2.3 是说, 不考虑一些和因变量有关的因子后, **选模型对误差方差的估计也一般不是无偏的**, 而是偏高, 这和直观吻合.

当全模型是正确的时候显然对于给定的 $\boldsymbol{x}=(\boldsymbol{x}_q,\boldsymbol{x}_t)$, 预测是无偏的. 记 $U=y-\boldsymbol{x}\hat{\boldsymbol{\beta}}$, 若假设误差不相关, 则可知

$$D[U]=\sigma^2[1+\boldsymbol{x}^T(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{x}]. \tag{1.2.7}$$

同样记记 $U_q=y-\boldsymbol{x}\tilde{\boldsymbol{\beta}}$, 则由式 (1.2.5) 知

$$\mathbb{E}[U_q]=\boldsymbol{x}_t^T\boldsymbol{\beta}_t-\boldsymbol{x}_t^T\boldsymbol{A}\boldsymbol{\beta}_t.$$

则选模型的*预测均方误差* (mean square error pf prediction, MSEP) 为

$$\mathrm{MSEP}(\tilde{y})=\mathbb{E}[U_q^2]=D[U_q]+\{\mathbb{E}[U_q]\}^2=\sigma^2[1+\boldsymbol{x}_q^T(\boldsymbol{X}_q^T\boldsymbol{X}_q)^{-1}\boldsymbol{x}_q]+(\boldsymbol{x}_t^T\boldsymbol{\beta}_t-\boldsymbol{x}_t^T\boldsymbol{A}\boldsymbol{\beta}_t)^2. \tag{1.2.8}$$

**定理 1.2.4** *若全模型是正确的, 则*

$$D[U]\geq D[U_q],$$

*且当 $\mathrm{Var}(\hat{\boldsymbol{\beta}}_t)\geq\boldsymbol{\beta}_t\boldsymbol{\beta}_t^T$ 时有*

$$D[U]\geq\mathrm{MSEP}(\tilde{y})=\mathbb{E}[U_q^2].$$

**证明** 第一个不等式由式 (1.2.6)(1.2.7)(1.2.8) 即得

$$D[U]-D[U_q]=\sigma^2(\boldsymbol{A}^T\boldsymbol{x}_q)^T\boldsymbol{D}(\boldsymbol{A}^T\boldsymbol{x}_q)\geq 0.$$

由此可知

$$\mathbb{E}[U_q^2]=(\boldsymbol{A}^T\boldsymbol{x}_q)^T\boldsymbol{\beta}_t\boldsymbol{\beta}(\boldsymbol{A}^T\boldsymbol{x}_q)\leq(\boldsymbol{A}^T\boldsymbol{x}_q)^T\mathrm{Var}(\hat{\boldsymbol{\beta}}_t)(\boldsymbol{A}^T\boldsymbol{x}_q)=\sigma^2(\boldsymbol{A}^T\boldsymbol{x}_q)^T\boldsymbol{D}(\boldsymbol{A}^T\boldsymbol{x}_q)=D[U]-D[U_q],$$

移项即证. □

定理 1.2.4 是说, 即使全模型是正确的, **丢掉一些很难估计的变量可以提高预测精度**, 本质上是**用偏差换方差.**

**定理 1.2.5** *若全模型正确, 则*

$$D[U] \geq D[U_q].$$

**证明** 直接使用分解 (1.2.6) 得

$$\begin{aligned}\boldsymbol{x}^T(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{x} - \boldsymbol{x}_q^T(\boldsymbol{X}_q^T\boldsymbol{X}_q)^{-1}\boldsymbol{x}_q &= \begin{pmatrix}(\boldsymbol{X}_q^T\boldsymbol{X}_q)^{-1} + \boldsymbol{A}\boldsymbol{D}\boldsymbol{A}^T & \boldsymbol{B} \\ \boldsymbol{B}^T & \boldsymbol{D}\end{pmatrix}\boldsymbol{x} - \boldsymbol{x}_q^T(\boldsymbol{X}_q^T\boldsymbol{X}_q)^{-1}\boldsymbol{x}_q \\ &= (\boldsymbol{A}^T\boldsymbol{x}_q - \boldsymbol{x}_q)^T\boldsymbol{D}(\boldsymbol{A}^T\boldsymbol{x}_q - \boldsymbol{x}_q) \geq 0\end{aligned}$$

从而结论成立. □

这个定理是说, 选模型预测的方差不会大于全模型.

## 1.3 对数几率回归

现在考虑一个广义线性模型 (generalized linear model)

$$y = g^{-1}(\boldsymbol{\beta}^T\boldsymbol{x}), \tag{1.3.9}$$

其中 $g$ 是一个单调可微函数. 这样的广义线性模型就相当于对输出进行一个映射, 映射到实际问题需要的空间.

现在考虑在模型 (1.3.9) 中取一个特殊的函数, 将 $\mathbb{R}$ 映射到 $(0,1)$ 区间, 就可以得到概率分布的线性模型. 最常见的二项对数几率回归模型 (binomial logistic regression model) 就是令

$$g^{-1} = \frac{x}{1+x},$$

就得到条件概率分布

$$P(Y=1|\boldsymbol{x}) = \frac{e^{\boldsymbol{\beta}^T\boldsymbol{x}}}{1+e^{\boldsymbol{\beta}^T\boldsymbol{x}}}, \quad P(Y=0|\boldsymbol{x}) = \frac{1}{1+e^{\boldsymbol{\beta}^T\boldsymbol{x}}},$$

其中 $\boldsymbol{\beta}$ 是系数向量, 包含了偏置.

定义一个概率为 $p$ 的事件的**几率** (odds) 为 $p/(1-p)$, 则对于对数几率回归而言, 恰好

$$\log\frac{P(Y=1|\boldsymbol{x})}{1-P(Y=1|\boldsymbol{x})} = \boldsymbol{\beta}^T\boldsymbol{x}, \tag{1.3.10}$$

也就是输出 $Y=1$ 的对数几率是 $\boldsymbol{x}$ 的线性函数.

实际上, 对数几率回归就是通过 $\boldsymbol{x}$ 的线性模型对 $K$ 个类的后验概率进行建模, 也就是将式 (1.3.10) 写成更一般的形式

$$\log\frac{P\{G=k|\boldsymbol{X}=\boldsymbol{x}\}}{P\{G=K|\boldsymbol{X}=\boldsymbol{x}\}} = \boldsymbol{\beta}^T\boldsymbol{x},$$

从而多项逻辑斯蒂回归模型 (multi-nomianl logistic regression model) 可以写为

$$P(Y=k|x) = \frac{e^{\boldsymbol{\beta}_k^T x}}{1+\sum_{k=1}^{K-1}e^{\boldsymbol{\beta}_k^T x}}, \quad P(Y=K|x) = \frac{1}{1+\sum_{k=1}^{K-1}e^{\boldsymbol{\beta}_k^T x}},$$

其中 $\boldsymbol{\beta}_k, k=1,\cdots,K-1$ 是系数向量, 包含了偏置.

### 1.3.1 参数估计

下面考虑采用极大似然估计的方法来对模型进行参数估计. 假设给定数量为 $N$ 的训练集, 且设

$$P_{\boldsymbol{\beta}}(Y=1|x) = \pi_{\boldsymbol{\beta}}(x), \quad P_{\boldsymbol{\beta}}(Y=0|x) = 1-\pi_{\boldsymbol{\beta}}(x),$$

则似然函数为

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{N} [\pi(\boldsymbol{x}_i)]^{y_i} [1 - \pi(\boldsymbol{x}_i)]^{1-y_i},$$

取对数似然即可将问题转化为最优化问题. 注意这里的似然函数的上标只不过是恰好可以写成这种简洁的形式而已.

取对数似然为

$$\begin{aligned} l(\boldsymbol{\beta}) &= \sum_{i=1}^{N} [y_i \log \pi_{\boldsymbol{\beta}}(\boldsymbol{x}_i) + (1 - y_i) \log(1 - \pi_{\boldsymbol{\beta}}(\boldsymbol{x}_i))] \\ &= \sum_{i=1}^{N} \left[ y_i \boldsymbol{\beta}^T \boldsymbol{x}_i - \log(1 + e^{\boldsymbol{\beta}^T \boldsymbol{x}_i}) \right], \end{aligned}$$

对数似然对参数 $\boldsymbol{\beta}$ 求一阶导, 并令其为 0 得到

$$\frac{\partial l(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = -\sum_{i=1}^{N} \boldsymbol{x}_i (y_i - \pi_{\boldsymbol{\beta}}(\boldsymbol{x}_i)) = 0,$$

然后当然可以得到二阶导数 (Hessian 矩阵) 为

$$\frac{\partial^2 l(\boldsymbol{\beta})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T} = -\boldsymbol{x}_i \boldsymbol{x}_i^T \pi_{\boldsymbol{\beta}}(\boldsymbol{x}_i)(1 - \pi_{\boldsymbol{\beta}}(\boldsymbol{x}_i)),$$

基于此对参数 $\boldsymbol{\beta}$ 进行参数更新

$$\beta^{\text{new}} = \beta^{\text{old}} - \left( \frac{\partial^2 l(\boldsymbol{\beta})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T} \right)^{-1} \frac{\partial l(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}},$$

这是所谓的 Newton-Raphson 算法 [2].

现在不妨设 $\boldsymbol{y}$ 是样本的输出,$\boldsymbol{p}$ 是当前模型输出的概率向量,$\boldsymbol{W}$ 是 $N \times N$ 的对角矩阵, 第 $i$ 个对角元为 $\pi_{\beta^{\text{old}}}(\boldsymbol{x}_i)(1 - \pi_{\beta^{\text{old}}a}(\boldsymbol{x}_i))$, 则有

$$\begin{aligned} \beta^{\text{new}} &= \beta^{\text{old}} + (\boldsymbol{X}^T \boldsymbol{W} \boldsymbol{X})^{-1} \boldsymbol{X}^T (\boldsymbol{y} - \boldsymbol{p}) \\ &= (\boldsymbol{X}^T \boldsymbol{W} \boldsymbol{X})^{-1} \boldsymbol{X}^T \boldsymbol{W} (\boldsymbol{X} \beta^{\text{old}} + \boldsymbol{W}^{-1} (\boldsymbol{y} - \boldsymbol{p})) \\ &= (\boldsymbol{X}^T \boldsymbol{W} \boldsymbol{X})^{-1} \boldsymbol{X}^T \boldsymbol{W} \boldsymbol{z}, \end{aligned}$$

其中 $\boldsymbol{z} = \boldsymbol{X} \beta^{\text{old}} + \boldsymbol{W}^{-1} (\boldsymbol{y} - \boldsymbol{p})$. 从而可见这是一个迭代的最小二乘, 即是

$$\beta^{\text{new}} \leftarrow \arg\min_{\boldsymbol{\beta}} (\boldsymbol{z} - \boldsymbol{X}\boldsymbol{\beta})^T \boldsymbol{W} (\boldsymbol{z} - \boldsymbol{X}\boldsymbol{\beta}).$$

这个算法确实被称为**迭代加权最小二乘** (iterative weighted least square, IRLS).

## 1.4 最大熵模型

最大熵模型 (maximum extropy model) 是一个输出满足最大熵原理的条件概率分布 $P(Y|X)$ 的模型. 对于给定的训练数据集 $T = \{(x_1, y_1), \cdots, (x_N, y_N)\}$, 则可以确定经验分布

$$\tilde{P}(X = x, Y = y) = \frac{v(X = x, Y = y)}{N}, \quad \tilde{P}(X = x) = \frac{v(X = x)}{N},$$

其中 $v$ 代表频数. 现在给定一个事件集 $Q$, 定义特征函数

$$f(x, y) := \begin{cases} 1, (x, y) \in Q \\ 0, (x, y) \notin Q \end{cases}$$

然后分别定义样本期望和模型期望为

$$\mathbb{E}_{\tilde{P}}(f) := \sum_{x,y} \tilde{P}(x, y) f(x, y), \quad E_P(f) := \sum_{x,y} \tilde{P}(x) P(y|x) f(x, y),$$

定义模型关于样本的条件熵为

$$H(P) := -\sum_{x,y} \tilde{P}(x) P(y|x) \log P(y|x),$$

则满足条件 $E_{\tilde{P}}(f) = E_P(f)$ 且最大化 $H(P)$ 的模型就是最大熵模型.

[2] 二阶方法见附录A.2.

## 1.5 线性判别分析

线性判别分析 (Linear Discriminant Analysis, LDA) 的想法很简单, 就是将样例投影到一条直线上, 使得同样例的投影点尽可能接近, 而异类的投影点尽可能远离.

考虑二分类问题, 将数据投影到直线 $\boldsymbol{w}$ 上 [3], 第 $i$ 类的样本的均值为 $\boldsymbol{\mu}_i$, 协方差矩阵定义为

$$\boldsymbol{\Sigma}_i := \sum_{\boldsymbol{x}\in X_0} (\boldsymbol{x}-\boldsymbol{\mu}_i)(\boldsymbol{x}-\boldsymbol{\mu}_i)^T,$$

由于样本协方差矩阵度量了样本间的差异程度, 所以我们希望投影的同类样本的差异小, 也就是使得 $\boldsymbol{w}^T\boldsymbol{\Sigma}_0\boldsymbol{w}+\boldsymbol{w}^T\boldsymbol{\Sigma}_1\boldsymbol{w}$ 尽可能小, 而使得 $||\boldsymbol{w}^T\boldsymbol{\mu}_0-\boldsymbol{w}^T\boldsymbol{\mu}_1||_2^2$ 尽可能大. 同时考虑二者, 可得优化问题

$$\max_{\boldsymbol{w}} J = \frac{||\boldsymbol{w}^T\boldsymbol{\mu}_0-\boldsymbol{w}^T\boldsymbol{\mu}_1||_2^2}{\boldsymbol{w}^T\boldsymbol{\Sigma}_0\boldsymbol{w}+\boldsymbol{w}^T\boldsymbol{\Sigma}_1\boldsymbol{w}} = \frac{w^T(\boldsymbol{\mu}_0-\boldsymbol{\mu}_1)(\boldsymbol{\mu}_0-\boldsymbol{\mu}_1)^T\boldsymbol{w}}{\boldsymbol{w}^T\boldsymbol{\Sigma}_0\boldsymbol{w}+\boldsymbol{w}^T\boldsymbol{\Sigma}_1\boldsymbol{w}}. \tag{1.5.11}$$

定义类内散度矩阵 (within-class scatter matrix) 为 $\boldsymbol{S}_w := \boldsymbol{\Sigma}_0+\boldsymbol{\Sigma}_1$, 定义类间散度矩阵 (between-class scatter matrix) 为 $\boldsymbol{S}_b := (\boldsymbol{\mu}_0-\boldsymbol{\mu}_1)(\boldsymbol{\mu}_0-\boldsymbol{\mu}_1)^T$, 从而优化问题 (1.5.11) 重写为

$$\max_{\boldsymbol{w}} J = \frac{\boldsymbol{w}^T\boldsymbol{S}_b\boldsymbol{w}}{\boldsymbol{w}^T\boldsymbol{S}_w\boldsymbol{w}}, \tag{1.5.12}$$

进一步等价于

$$\begin{aligned} \min_{\boldsymbol{w}} \quad & -\boldsymbol{w}\boldsymbol{S}_b\boldsymbol{w} \\ \text{s.t.} \quad & \boldsymbol{w}\boldsymbol{S}_w\boldsymbol{w}=1, \end{aligned}$$

利用拉格朗日乘子法以及一些矩阵求导, 这等价于

$$\boldsymbol{S}_b\boldsymbol{w} = \lambda\boldsymbol{S}_w\boldsymbol{w}, \tag{1.5.13}$$

显然 $\boldsymbol{S}_b\boldsymbol{w}$ 的方向就是 $\boldsymbol{\mu}_0-\boldsymbol{\mu}_1$, 因此可以不妨 $\boldsymbol{S}_b\boldsymbol{w}=\lambda(\boldsymbol{\mu}_0-\boldsymbol{\mu}_1)$, 于是代入 (1.5.13) 可得

$$\boldsymbol{w} = \boldsymbol{S}_w^{-1}(\boldsymbol{\mu}_0-\boldsymbol{\mu}_1).$$

## 1.6 正交化回归

对于**多元线性回归模型** (meltiple linear regression), 自然可以依照前面讨论的直接计算回归系数, 但在这一节我们考虑样本矩阵的列是正交的特殊情况.

首先考虑没有截距的一元回归

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

自然根据公式1.1.2可以知道回归系数估计和残差分别为

$$\hat{\beta} = \frac{\sum_{i=1}^N x_iy_i}{\sum_{i=1}^N x_i^2} = \frac{\boldsymbol{X}^T\boldsymbol{Y}}{\boldsymbol{X}^T\boldsymbol{X}}, \quad r_i = y_i-x_i\hat{\beta}.$$

直接从公式1.1.2可以看出当 $\boldsymbol{X}$ 的列是正交时, 系数 $\boldsymbol{\beta}$ 的每一个分量的估计互不影响, 且由公式

$$\hat{\beta}_i = \frac{\boldsymbol{x}_i^T\boldsymbol{Y}}{\boldsymbol{x}_i^T\boldsymbol{x}_i}$$

给出. 接下来考虑依次添加进 $\boldsymbol{X}$ 的列进行回归, 对 $X_i$ 依次施行关于前 $i-1$ 列的**施密特正交化**, 得到 $\boldsymbol{z}_i$, 则新添加的列的回归不受前面回归的系数影响, 且由于标准施密特正交化的可以表示为

$$\boldsymbol{z}_i = \boldsymbol{x}_i - \sum_{j=1}^{i-1} \frac{\boldsymbol{x}_i^T\boldsymbol{X}_j}{\boldsymbol{x}_i^T\boldsymbol{X}_j}\boldsymbol{X}_j,$$

[3]应该说 $\boldsymbol{w}$ 是这个直线的方向, 显然投影的类间距离只和直线的方向有关, 从而可以假设投影直线过原点, 然后显然若 $|\boldsymbol{w}|=1$, 则 $\boldsymbol{w}^Tx$ 表示投影沿着投影直线方向到原点的距离

也就是只有 $\boldsymbol{z}_p$ 含有 $\boldsymbol{x}_p$, 且系数是 1, 因此根据 $\boldsymbol{Z}$ 进行回归计算得到的系数 $\hat{\beta}_p$ 就是原本用 $\boldsymbol{X}$ 对应的第 $p$ 个系数, 然后再一步步倒回去就可以解出原本的系数, 这实际上和使用 QR 分解后解方程是一样的.

整理以上过程得到：

1. 初始化 $\boldsymbol{z}_0 = \boldsymbol{x}_0 = \mathbf{1}$；
2. 对于 $j = 1, 2, \cdots, p$, 在 $\boldsymbol{z}_0, \cdots, \boldsymbol{z}_{j-1}$ 上对 $x_j$ 进行回归, 并得到残差向量 $\boldsymbol{z}_j$(实际上就是施密特正交化的结果)；
3. 在残差 $\boldsymbol{z}_j$ 上对 $\boldsymbol{Y}$ 进行回归得到系数 $\hat{\beta}_j$.

直观的说线性回归实际上就是一个正交化的过程, 所以我们不停的用新加入的 $\boldsymbol{x}_i$ 对上一回归留下的残差进行回归, 于是便得到了正交的分量.

现在将这个正交化写成矩阵的形式：

$$\boldsymbol{X} = \boldsymbol{Z}\boldsymbol{\Gamma} = \boldsymbol{Z}\boldsymbol{D}^{-1}\boldsymbol{D}\boldsymbol{\Gamma} = \boldsymbol{Q}\boldsymbol{R}, \tag{1.6.14}$$

这里的 $\boldsymbol{D}$ 为对角矩阵, 使得 $||D_{jj}|| = ||\boldsymbol{z}_j||$, 也就是使得 $\boldsymbol{Q}$ 为标准正交矩阵. 式 (1.6.14) 称为矩阵 $\boldsymbol{X}$ 的 **$\boldsymbol{QR}$ 分解**, 其中 $\boldsymbol{Q}$ 为 $N \times (p+1)$ 的正交矩阵, 满足 $\boldsymbol{Q}^T\boldsymbol{Q} = \boldsymbol{I}$,$\boldsymbol{R}$ 是 $(p+1) \times (p+1)$ 的上三角矩阵. 据此可得

$$\hat{\beta} = \boldsymbol{R}^{-1}\boldsymbol{Q}^T\boldsymbol{y}, \quad \hat{\boldsymbol{y}} = \boldsymbol{Q}\boldsymbol{Q}^T\boldsymbol{y}.$$

注意到 $\boldsymbol{Q}$ 实际上是 $\boldsymbol{X}$ 的列向量的线性组合, 因而张成同一个子空间, 而 $\boldsymbol{Q}\boldsymbol{Q}^T$ 实际上就是一个正交投影因子.[4]

在前面提到的正交化算法中, 实际上若想求得原本的回归系数, 需要先执行正交化到最后一步, 然后一步一步往回走, 本质上是通过矩阵的 QR 分解来进行解方程. 换句话说就是正交化的最后一步的回归系数就是原来的回归系数.

## 1.7 收缩方法

### 1.7.1 岭回归

考虑11.1.1中描述的岭回归模型, 除了在那里直接给出的显示公式之外, 我们进一步考虑7节中讨论的 $\boldsymbol{X}$ 的奇异值分解

$$\boldsymbol{X} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^T,$$

这里令 $\boldsymbol{D}$ 是 $p \times p$ 半正定矩阵,[5] 且对角元为 $\boldsymbol{X}$ 的递减的奇异值 $d_i$, 于是可得

$$\boldsymbol{X}\boldsymbol{w} = \boldsymbol{U}\boldsymbol{D}(\boldsymbol{D}^2 + \lambda\boldsymbol{I})^{-1}\boldsymbol{D}\boldsymbol{U}^T\boldsymbol{Y} = \sum_{j=1}^{p} \boldsymbol{u}_j \frac{d_j^2}{d_j^2 + \lambda}\boldsymbol{u}_j^T\boldsymbol{Y}.$$

容易看到 $\boldsymbol{U}$ 是 $\boldsymbol{X}$ 的列的一个正交化, 从而 $\boldsymbol{U}^T\boldsymbol{Y}$ 是 $\boldsymbol{Y}$ 在 $\boldsymbol{U}$ 下的“坐标”. 显然岭回归实际上的作用就是收缩每一个主成分 [6] 的系数, 且越重要的主成分收缩越小, 越不重要的主成分收缩越多.

# 2 核方法

类似于分段拟合的思想, 仅使用靠近目标点 $x_0$ 的观测来拟合简单的模型, 是一种回归技术. 如果希望拟合的函数是光滑的, 可以通过给近邻添加不同的权值来实现, 其中核函数 $K_\lambda(x_0, x_i)$ 就是一个实现途径.

## 2.1 一维核光滑方法

对于回归模型, 考虑一个 $k$-最近邻平均

$$\hat{f}(x) = \mathrm{Ave}(y_i | x_i \in N_k(x)),$$

[4] 可以这样理解, 因为 $\boldsymbol{Q}$ 的列向量是一组正交基, 从而 $\boldsymbol{Q}^T\boldsymbol{y}$ 得到的是 $\boldsymbol{y}$ 在子空间上这组正交基下的坐标上的投影值, 从而再左乘 $\boldsymbol{Q}$ 就得到正交投影.

[5] 注意岭回归是不惩罚截距项的系数的, 因为总可以做总体平移来得到无截距的回归, 这一点可以从上一小节的正交化回归过程中看到.

[6] 主成分分析参考4.2节.

在输入是一维的情形, 得到的回归曲线不是平滑的, 原因在于当 $x$ 移动时, 边缘的元素会发生跳变. 一种平滑的方式是考虑给最近邻中的点加上权重, 这个权重随着到目标点的距离递减, 从而得到**核加权平均** (kernel-weighted average)

$$\hat{f}(x_0) = \frac{\sum_{i=1}^N K_\lambda(x_0, x_i) y_i}{\sum_{i=1}^N K_\lambda(x_0, x_i)}.$$

特别地, 取

$$K_\lambda(x_0, x) = D\left(\frac{|x - x_0|}{\lambda}\right), \quad D(t) = \begin{cases} \frac{3}{4}(1 - t^2), & |t| \le 1 \\ 0, & |t| > 1 \end{cases}$$

称为 *Epanechnikov kernel*.

### 2.1.1 局部线性回归

采用近邻加权平均的回归技术在靠近边界的地方可能会出现较大的偏差, 因为边界处的最近邻集中在一侧. 一种自然的想法是在靠近边界的地方采用局部线性回归的方式拟合数据. 在 $x_0$ 处考虑最小二乘问题

$$\min_{\alpha(x_0), \beta(x_0)} \sum_{i=1}^N K_\lambda(x_0, x_i)[y_i - \alpha(x_0) - \beta(x_0) x_i]^2,$$

令 $b(x)^T = (1, x)$,$\boldsymbol{B}$ 是 $N \times 2$ 回归矩阵, 第 $i$ 行为 $b(x_i)^T$,$\boldsymbol{W}(x_0)$ 是对角矩阵, 第 $i$ 个对角元为 $K_\lambda(x_0, x_i)$, 则由1.1中的解法可以得到局部线性回归

$$\hat{f}(x_0) = b(x_0)^T (\boldsymbol{B}^T \boldsymbol{W}(x_0) \boldsymbol{B})^{-1} \boldsymbol{B}^T \boldsymbol{W}(x_0) \boldsymbol{Y} = \sum_{i=1}^N l_i(x_0) y_i. \tag{2.1.1}$$

这样得出的系数 $l_i(x_0)$ 称为**等价核** (equivalent kernel), 因为已经包含了权重 $K_\lambda(x_0, x_i)$ 的调整.

**命题 2.1.1** *对于局部线性回归 (2.1.1), 有*

$$\sum_{i=1}^N l_i(x_0) = 1, \quad \sum_{i=1}^N (x_i - x_0) l_i(x_0) = 0.$$

**证明** 由于 $b(x_0)^T = (1, x_0)$, 从而由式 (2.1.1) 可得

$$b(x_0)^T (\boldsymbol{B}^T \boldsymbol{W}(x_0) \boldsymbol{B})^{-1} \boldsymbol{B}^T \boldsymbol{W}(x_0) \boldsymbol{B} = (1, x_0) = \left( \sum_{i=1}^N l_i(x_0), \sum_{i=1}^N x_i l_i(x_0) \right),$$

整理可得结论. □

现在由命题 2.1.1 以及模型线性的假设, 考虑展开式

$$\begin{aligned} \mathbb{E}\hat{f}(x_0) &= \sum_{i=1}^N l_i(x_0) f(x_i) \\ &= f(x_0) \sum_{i=1}^N l_i(x_0) + f'(x_0) \sum_{i=1}^N (x_i - x_0) l_i(x_0) + \frac{f''(x_0)}{2} \sum_{i=1}^N (x_i - x_0)^2 l_i(x_0) + R \\ &= f(x_0) + \frac{f''(x_0)}{2} \sum_{i=1}^N (x_i - x_0)^2 l_i(x_0) + R, \end{aligned} \tag{2.1.2}$$

从而误差被高阶项控制.

### 2.1.2 局部多项式回归

从展开式 (2.1.2) 可知使用局部线性回归可以将偏差控制在二次以上. 同理如果我们希望偏差的阶至少是 $d + 1$, 则可以使用局部 $d$ 次多项式回归, 也就是

$$\min_{\alpha(x_0), \beta_j(x_0), j = 1, \cdots, d} \sum_{i=i}^N K_\lambda(x_0, x_1) \left[ y_i - \alpha(x_0) - \sum_{j=1}^d \beta_j(x_0) x_i^j \right]^2.$$

这显然就是一个使用基函数展开的线性回归, 同样类似局部线性回归, 令 $b(x)^T=(1,\cdots,x^d)$, 记号同 (2.1.1), 可直接得到和 (2.1.1) 完全一样的表达式.

**命题 2.1.2** *以 $\boldsymbol{l}_d$ 表示使用 $d$ 次局部多项式回归得到的等价核向量, 则 $||\boldsymbol{l}_d||^2$ 是 $d$ 的单调上升函数.*

**证明** 由于 $\boldsymbol{W}(x_0)$ 对角元都是正的, 从而考虑 $\boldsymbol{B}\sqrt{\boldsymbol{W}(x_0)}$, 然后就化为了 $\boldsymbol{W}(x_0)=\boldsymbol{I}_N$ 的情形, 然后

$$\boldsymbol{l}_d^T\boldsymbol{l}_d=b(x_0)^T(\boldsymbol{B}^T\boldsymbol{B})^{-1}b(x_0),$$

然后根据1.2小节关于子集选择的结论可知原命题成立. □

## 2.2 核密度估计和分类

假设样本 $\{x_1,\cdots,x_N\}$ 是从一个分布中抽取的, 我们希望估计这个分布的密度函数 $f_X(x)$, 一种最直接的做法就是令

$$\hat{f}_X(x_0)=\frac{\mathrm{NUM}\{x_i|x_i\in\mathscr{N}(x_0)\}}{N\lambda},$$

这里 $\mathscr{N}(x_0)$ 是 $x_0$ 的宽度为 $\lambda$ 的邻域.

这种估计方式的问题就是不连续, 为此, 采用核估计

$$\hat{f}_X(x_0)=\frac{1}{N\lambda}\sum_{i=1}^{N}K_\lambda(x_0,x_i),$$

通常采用高斯核 $K_\lambda(x_0,x_i)=\phi(|x_0-x_i|/\lambda)$.

现在我们希望给出核密度估计的误差界限, 从而考虑偏差-方差分解

$$\mathrm{MSE}(x_0)=\mathbb{E}_f\left[\left\{\hat{f}_n(x_0)-f(x_0)\right\}^2\right]=\left\{\mathbb{E}_f\left[\hat{f}_n(x_0)\right]\right\}^2+\mathrm{Var}\left[\hat{f}_n(x_0)\right],$$

将偏差记为 $b(x_0)$, 方差记为 $\sigma^2(x_0)$, 我们有

**命题 2.2.1** *若 $f$ 和 $K$ 分别满足条件*

$$\sup_{x\in\mathbb{R}}f(x)\le f_{\max}<\infty,\quad \int K^2(u)du<\infty,$$

则

$$\sigma^2(x_0)\le\frac{f_{\max}}{n\lambda}\int K^2(u)du.$$

**证明** 直接进行估计

$$\sigma^2(x_0)\le\frac{1}{n\lambda^2}\mathbb{E}_fK^2\left(\frac{x_1-x_0}{\lambda}\right)\le\frac{f_{\max}}{n\lambda}\int K^2(u)du.$$

□

对密度进行估计后, 对于分类问题我们就可以使用贝叶斯公式进行处理. 例如我们给每一个类 $j$ 估计一个密度 $\hat{f}_j$, 然后给每一个类估计一个先验分布 $\hat{\pi}_j$, 于是由贝叶斯公式有

$$P\{G=j|X=x_0\}=\frac{\hat{\pi}_j\hat{f}_j(x_0)}{\sum_{k=1}^{J}\hat{\pi}_k\hat{f}_k(x_0)}.$$

当特征空间维数很高的时候, 给每个类进行密度函数估计会出现困难, 这时 9.1小节讨论的朴素贝叶斯分类方法有较好的表现.

# 3 提升算法与集成学习

提升 (boosting) 方法的思想是训练多个模型进行线性组合以达到一个更强的模型.

## 3.1 Adaboost 算法

对于分类问题,Adaboost 算法的思想是在每一轮改变训练数据的权值, 从而使得新训练的分类器着重将之前分类器分类错误的数据分类对. 假定给定二分类问题训练数据集, 其中输出取值于 $\{-1,1\}$, 首先初始化权值分布为

$$D_1 = (w_{11}, \cdots, w_{1N}) = (\frac{1}{n}, \cdots, \frac{1}{n}),$$

使用 $D_m$ 作为权值训练分类器 $G_m$, 然后计算分类误差率

$$e_m = \sum_{i=1}^{N} P(G_m(x_i) \neq y_i),$$

注意这是用了新的权值 $D_m$, 然后令 $G_m$ 的系数为

$$\alpha_m = \frac{1}{2}\log\frac{1-e_m}{e_m},$$

然后更新权值分布为 $D_{m+1}$, 令

$$w_{m+1,i} = \frac{w_{mi}}{Z_m}\exp\{-\alpha_m y_i G_m(x_i)\},$$

归一化因子为

$$Z_m = \sum_{i=1}^{N} w_{mi}\exp\{-\alpha_m y_i G_m(x_i)\}.$$

. 最后构建基本分类器的线性组合得到最终分类器

$$f(x) = \sum_{m=1}^{M}\alpha_m G_m(x), \quad G(x) = \operatorname{sign}\{f(x)\}.$$

接下来为了分析 Adaboost 算法的分类误差.

**定理 3.2.1** Adaboost 算法最终分类器的误差有上界

$$\frac{1}{N}\sum_{i=1}^{N} I(G(x_i)\neq y_i) \leq \frac{1}{N}\sum_{i=1}^{N}\exp\{-y_i f(x_i)\} = \prod_{m=1}^{M} Z_m.$$

**证明** 不等号只需注意到 $G(x_i)\neq y_i$ 时总有 $-y_i f(x_i) > 0$, 从而成立. 对于第二个等式, 注意到

$$w_{mi}\exp\{-\alpha_m y_i G_m(x_i)\} = Z_m w_{m+1,i},$$

从而注意到 $w_{1i} = 1/N$, 有

$$\frac{1}{N}\sum_{i=1}^{N}\exp\{-y_i f(x_i)\} = \frac{1}{N}\sum_{i=1}^{N}\exp\left\{-\sum_{m=1}^{M}\alpha_m y_i G_m(x_i)\right\} = Z_1\sum_{i=1}^{N} w_{1i}\prod_{m=2}^{M}\exp\{-\alpha_m y_i G_m(x_i)\} = \prod_{m=1}^{M} Z_m.$$

□

由定理 3.2.1 可知如果能让 $Z_m$ 尽可能小, 就可以获得较好的分类效果. 特别地, 对于二分类问题有

**定理 3.2.2** 对于二分类问题 Adaboost 算法有

$$\prod_{m=1}^{M} Z_m = \prod_{m=1}^{M}[2\sqrt{e_m(1-e_m)}] \leq \exp\left\{-2\sum_{m=1}^{M}\gamma_m^2\right\},$$

这里 $\gamma_m = 1/2 - e_m$.

**证明** 直接由 $\alpha_m$ 和 $e_m$ 的定义以及简单的泰勒展开即得. □

定理 3.2.2 告诉我们 Adaboost 二分类的训练误差是指数下降的.

## 3.2 向前分步算法与指数损失

Adaboost 是一般的**加法模型** (additive model) 的一个例子, 所谓加法模型, 指的是形如

$$f(x) = \sum_{m=1}^{M} \beta_m b(x; \gamma_m) \tag{3.2.1}$$

的模型, 其中 $b(x;\gamma_m)$ 是基函数,$\beta_m$ 是系数,$\gamma_m$ 是参数. 对于模型 (3.3.1) 通常采用**向前分步算法** (forward stagewise algorithm) 进行优化, 记 $f_{m-1}(x)$ 为前 $m-1$ 步已优化的, 然后通过极小化损失函数

$$(\beta_m, \gamma_m) = \arg\min_{\beta,\gamma} \sum_{i=1}^{N} L(y_i, f_{m-1}(x_i) + \beta b(x_i; \gamma))$$

得到新的参数, 进而更新得到 $f_m(x)$.

**定理 3.3.1** Adaboost 模型就是由基本分类器组成, 损失函数是指数函数

$$L(y, f(x)) = \exp\{-yf(x)\}$$

的采用向前分布算法的加法模型.

**证明** 令 $f_0(x) = 0$, 首先注意到

$$(\alpha_m, G_m) = \arg\min_{\alpha,G} \sum_{i=1}^{N} \exp\{-y_i f_{m-1}(x_i) - \alpha y_i G(x_i)\},$$

而 $\exp\{-y_i f_{m-1}(x_i)\}$ 恰好是更新前的权重, 然后易见优化问题就和 Adaboost 模型一样了. 然后求解 $\alpha$, 结果也是一样的, 然后容易验证更新后的权重也是一样的. □

实际上 Adaboost 算法提出时并不是基于加法模型进行构建的, 其与采用指数损失的加法模型的等价性是后来才被证明的. 对于采用 $\{-1,1\}$ 作为响应的二分类模型, 容易证明

$$f^*(x \arg\min_{f(x)} \mathbb{E}_{Y|x}[e^{-Yf(x)}] = \frac{1}{2}\log\frac{P\{Y=1|x\}}{P\{Y=-1|x\}},$$

也就是说 Adaboost 算法是对 $P\{Y=1|x\}$ 的对数几率的一半进行估. 另一种等价的损失函数是**二项式对数似然** (binomial log-likelihood) 或称为**交叉熵** (cross entropy)

$$l(Y, p(x)) = Y\log P\{Y=1|x\} + (1-Y)\log(1-P\{Y=1|x\}),$$

这里 $Y \in \{0,1\}$.

现在我们来考虑采用 $\{-1,1\}$ 作为响应的二分类模型, 损失函数将对 $yf(x)$ 进行惩罚, 显然错误分类对应着 $yf(x) \le 0$, 指数损失和交叉熵对于 $yf(x)$ 的惩罚是单调下降的, 也就是对于误分类更多的分类, 惩罚更多, 而对于越有把握的正确分类则给予更少的惩罚, 基于此种观点, 平方损失对于二分类问题来说是不合适的, 因为它也会惩罚非常有把握的正确分类. 另一个问题在于, 指数损失是凸函数, 因此和平方损失一样, 很容易被异常点影响, 所以在信噪比很低的数据中表现不佳.

另外, 使用指数损失的 Adaboost 算法即使在训练集上误分类已经降为 0, 其损失函数还是可以继续下降, 原因就在于指数损失的性质. 指数损失继续下降意味着模型对训练集的分类把握更大.

## 3.3 梯度提升与树的大小

现在我们考虑一个**提升树** (boosting tree) 的模型 [7]

$$f_M(x) = \sum_{m=1}^{M} T(x; \Theta_m),$$

[7] 树模型的详述见8

求解的一般方法就是向前分步算法, 问题在于对于一般的损失函数, 最佳的解可能没有显式表达. 为此采用类似于梯度下降的思想, 每一步加入的树都拟合使得损失下降最快的方向. 对于使用平方损失的情况, 梯度的方向恰好就是误差的方向, 这是因为

$$\frac{\partial(y_i - f(x_i))^2}{\partial f(x_i)} = 2(y_i - f(x_i)),$$

因此对于使用平方损失的提升树模型, 每一步新增的树都是拟合当前模型的残差, 类似于正交线性回归, 因此正交线性回归也可以看成是一种提升算法. 梯度提升算法对于其他的损失函数一样试用, 只需将拟合的方向替换成对应的损失函数的梯度方向即可. 对于回归问题, 这个算法又称为**多重加法回归树** (multiple additive regression tree).

在单一的树模型中, 我们往往首先得到一个很大的树, 然后再通过剪枝的方式来获得适当大小的树. 然而在提升模型中, 过于复杂的树会增加计算开销, 并且可能对于模型没有提升. 一种简单的想法是限制每个加入的树的大小, 例如只允许有两层的树, 这样隐含了不允许任何输入变量之间有任何联系. 只允许有三层的树, 这样隐含所有变量之间最多有二阶的相互联系.

# 4 无监督学习

聚类方法作为一种无监督学习的方法, 通过定义一个样本间的距离度量, 将样本划分为一些类别. 如果规定一个样本只能属于一个类, 就称为*硬聚类* (hard clustering) 方法, 否则称为*软聚类* (soft clustering) 方法. 通常采用归到一类的定义是一个类中任意两个样本之间的距离 $d_{ij} \le T$.

常用的聚类方法为所谓的聚合聚类方法. 设有 $n$ 个样本, 初始时将每个样本归于一类, 然后每一步合并类间距离最小的两个类, 直到终止条件.

## 4.1 原型聚类

### 4.1.1 $K$-均值聚类

$K$-均值聚类给定 $K$ 个类别, 初始给定 $k$ 个中心, 记为 $(m_{1,1}, \cdots, m_{1,k})$, 在第 $i$ 轮迭代将每一个样本归入距离最近的中心, 得到一个分类, 然后重新计算每一类的中心, 更新为 $(m_{i+1,1}, \cdots, m_{i+1,k})$, 重复直到分类不再更新, 或者到达终止条件.

### 4.1.2 学习向量量化

学习向量量化 (Learning Vector Quantization, LVQ) 算法和 $K$-均值聚类类似, 希望用一些样本空间的向量来作为原型, 然后新的样本就根据和这些向量的距离中最近的那个来归类. 另一个不同点是,LVQ 假设数据样本带有类别标记, 利用这些标记作为监督类辅助聚类.

给定数量为 $m$ 的样本 $D = \{(\boldsymbol{x}_i, y_i)\}$, 初始化选择 $q$ 个向量 $\{\boldsymbol{p}_j\}$ 作为初始原型向量, 对应的簇类为 $t_j \in \mathscr{Y}$, 每次随机选取一个样本点 $(\boldsymbol{x}_i, y_i)$, 找到与其距离最近的原型向量 $\boldsymbol{p}_j$, 如果 $t_j = y_i$, 则 $\boldsymbol{p}_j$ 向着 $\boldsymbol{x}_i$ 的方向靠近一个给定的距离, 否则远离一个给定的距离. 重复以上过程直到满足终止条件, 得到的就是最终的原型向量.

## 4.2 主成分分析

### 4.2.1 随机向量的主成分

所谓*主成分分析* (principal component analysis, PCA) 就是对观测数据进行正交变换为线性无关的变量.

考虑 $\boldsymbol{x} = (x_1, \cdots, x_p)^T$ 为 $p$ 维随机变量, 其均值向量记为 $\boldsymbol{\mu}$, 协方差矩阵记为 $\Sigma$, 则我们希望施行正交变换 $\boldsymbol{y} = A^T\boldsymbol{x}$, 其中 $A$ 是 $p \times p$ 的正交矩阵, 且满足 $A^T\Sigma A = \Lambda$, 这里 $\Lambda$ 是对角阵, 意味着变换后 $\boldsymbol{y}$ 的各分量之间是线性无关的. 不妨记 $\lambda_i = \Lambda_{i,i}$, 则存在这样的 $A$, 使得其最大化 $\lambda_1$, 然后在此基础上最大化 $\lambda_2$, 以此类推. 这样得到的 $y_i$ 称为*$\boldsymbol{x}$ 的第 $i$ 个主成分*.

**定理 4.2.1** *$\boldsymbol{x}$ 的第 $i$ 个主成分的方差恰好是 $\Sigma$ 的从大到小排列的第 $i$ 的特征值 $\lambda_i$, 且 $A$ 的第 $i$ 列是对应的特征向量.*

**证明** 考虑优化问题

$$\begin{aligned} \min_{\alpha_1} \quad & -\alpha_1^T \Sigma \alpha_1 \\ \text{s.t.} \quad & \alpha_1^T \alpha_1 = 1 \end{aligned}$$

的拉格朗日函数

$$-\alpha_1^T \Sigma \alpha_1 + \lambda(\alpha_1^T \alpha_1 - 1),$$

求导即可得到 $\lambda_1$ 对应的主成分. 接下来考虑优化问题

$$\begin{aligned} \min_{\alpha_2} \quad & -\alpha_2^T \Sigma \alpha_2 \\ \text{s.t.} \quad & \alpha_2^T \alpha_2 = 1, \quad \alpha_1^T \Sigma \alpha_2 = 0, \quad \alpha_2^T \alpha_1 = 0, \end{aligned}$$

则由于 $\alpha_1^T \Sigma \alpha_2 = \lambda_1 \alpha_2^T \alpha_1$, 从而考虑拉格朗日函数

$$-\alpha_2^T \Sigma \alpha_2 + \lambda(\alpha_2^T \alpha_2 - 1) + \phi \alpha_2^T \alpha_1,$$

对 $\alpha_2$ 求导然后左乘 $\alpha_1$ 可得 $\phi = 0$, 然后类似前面的做法, 以此类推即可. □

### 4.2.2 样本主成分与奇异值分解

现在考虑 $N$ 个维度为 $p$ 的数据集的输入, 考虑对数据的一个秩为 $q \leq p$ 的最佳线性逼近

$$f(\lambda) = \boldsymbol{\mu} + \boldsymbol{V}_q \boldsymbol{\lambda},$$

其中 $\boldsymbol{\mu}$ 和 $\boldsymbol{\lambda}$ 分别是 $p \times 1$ 和 $q \times 1$ 的向量,$\boldsymbol{V}_q$ 是 $p \times q$ 的列正交矩阵, 希望极小化**重构误差** (reconstuction error)

$$\min_{\boldsymbol{\mu}, \boldsymbol{V}_q} ||\boldsymbol{x}_i - \boldsymbol{\mu} - \boldsymbol{V}_q \boldsymbol{\lambda}_i||^2,$$

利用拉格朗日乘子法可得

$$\hat{\boldsymbol{\mu}} = \overline{\boldsymbol{x}}, \quad \hat{\boldsymbol{\lambda}}_i = \boldsymbol{V}_q(\boldsymbol{x}_i - \overline{\boldsymbol{x}}).$$

令输入矩阵 $\boldsymbol{X}$ 的奇异值分解为 $\boldsymbol{X} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^T$, 则极小化重构误差等价于

$$\min_{\boldsymbol{V}_q} ||\boldsymbol{X}^T - \boldsymbol{V}_q \boldsymbol{V}_q^T \boldsymbol{X}^T||_2,$$

根据7.2小节的结论可知当 $\boldsymbol{V}_q$ 为 $\boldsymbol{V}$ 的前 $q$ 列时, 有

$$\boldsymbol{V}_q \boldsymbol{V}_q^T \boldsymbol{V} \boldsymbol{D}^T \boldsymbol{U}^T = \boldsymbol{V}_q \boldsymbol{D}_q^T \boldsymbol{U}_q^T,$$

其中 $\boldsymbol{U}_q$ 是 $\boldsymbol{U}$ 的前 $q$ 列, 且恰好极小化了重构误差.

使用类似随机向量主成分的思想, 假设输入来自于一个分布, 仍然令中心化后的 $N \times p$ 的输入矩阵 [8]$\boldsymbol{X}$ 的奇异值分解为 $\boldsymbol{X} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^T$, 则 $\boldsymbol{X}^T\boldsymbol{X} = \boldsymbol{V}\boldsymbol{D}^T\boldsymbol{D}\boldsymbol{V}^T$ 为协方差矩阵 $\boldsymbol{\Sigma}$ 的一个估计 [9], 且易知

$$\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{V} = \boldsymbol{V}\boldsymbol{D}^T\boldsymbol{D}\boldsymbol{V}^T\boldsymbol{V} = \boldsymbol{V}\boldsymbol{D}^T\boldsymbol{D},$$

可知 $\boldsymbol{V}$ 的每一列都是 $\boldsymbol{X}^T\boldsymbol{X}$ 的特征向量, 且按照特征值从小到大排列, 从而根据定理 4.2.1 可知 $\boldsymbol{X}^T$ 的主成分是 $\boldsymbol{V}\boldsymbol{X}^T = \boldsymbol{D}^T\boldsymbol{U}^T$. 通常我们会记为 $\boldsymbol{U}\boldsymbol{D}$.

[8]注意这里 $\boldsymbol{X}$ 的行是输入, 所以后面的讨论要时刻注意这点.

[9]实际上差一个因子 $1/p$, 只不过不影响讨论

## 4.3 谱聚类

谱聚类的思想是将样本点看作一个图的顶点, 顶点之间的边的权值是定义好的样本点之间的相似度. 谱聚类对图进行切分, 使得子图中的所有边的权值之和尽可能大, 且子图间的所有边的权值之和尽可能小.

现在以 $w_{ij} = w_{ji}$ 表示两个顶点 $i, j$ 之间的距离, 则 $\boldsymbol{W} = \{w_{ij}\}$ 称为相似矩阵 (adjacency matrix). 然后定义

$$g_i := \sum_{j=1}^{N} w_{ij}$$

为顶点 $i$ 的**度** (degree of vertex $i$), 度矩阵是一个对角矩阵, 对角元是 $g_i$, 记为 $\boldsymbol{G}$.

接下来定义拉普拉斯矩阵 (Graph Laplacian) 定义为

$$\boldsymbol{L} := \boldsymbol{G} - \boldsymbol{W},$$

容易验证对于任意向量 $\boldsymbol{f}$, 有

$$\boldsymbol{f}^T \boldsymbol{L} \boldsymbol{f} = \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} w_{ij} (f_i - f_j)^2, \tag{4.3.1}$$

从而可知 $\mathbf{1}^T \boldsymbol{L} \mathbf{1}^T = 0$, 于是拉普拉斯矩阵最小的特征值为 0.

考虑一个图中不交的顶点子集 $A, B$, 定义他们之间的距离为

$$W(A, B) := \sum_{i \in A, j \in B} w_{ij},$$

现在对于将一个图的顶点分为 $k$ 个不交子集的切分, 定义切图的损失函数

$$cut(A_1, \cdots, A_k) = \frac{1}{2} \sum_{i=1}^{k} W(A_i, \overline{A_i}).$$

如果为了极小化切图损失函数, 只需要将拥有最小出边的那个顶点切分出来即可. 因此考虑 *Ratio* **切图**损失函数

$$RatioCut(A_1, \cdots, A_k) = \frac{1}{2} \sum_{i=1}^{k} \frac{W(A_i, \overline{A_i})}{|A_i|},$$

因此最小化 Ratio 切图损失函数相当于同时最小化切图损失函数和最大化每个图中的顶点数.

接下里考虑指示向量 $\boldsymbol{h}_i$, 满足

$$h_{ij} = \begin{cases} 0, & v_j \notin A_i \\ \dfrac{1}{\sqrt{|A_i|}}, & v_j \in A_i \end{cases}$$

容易验证 $\{\boldsymbol{h}_i\}, i = 1, \cdots, k$ 是一组正交单位的向量, 且满足

$$\boldsymbol{h}_i^T \boldsymbol{L} \boldsymbol{h} = \frac{cut(A_i, \overline{A_i})}{|A_i|},$$

令 $\boldsymbol{H}$ 为由 $\{\boldsymbol{h}_i\}$ 组成的矩阵, 则可知

$$RatioCut(A_1, \cdots, A_k) = \mathrm{tr}(\boldsymbol{H}^T \boldsymbol{L} \boldsymbol{H}),$$

于是我们就可以寻找满足 $\boldsymbol{h}_i^T \boldsymbol{h}_i = 1$ 的使得 $\mathrm{tr}(\boldsymbol{H}^T \boldsymbol{L} \boldsymbol{H})$ 最小的若干个 $\boldsymbol{h}_i$. 由线性代数的知识可知其取值范围是 $\boldsymbol{L}$ 的特征值的范围, 且最小的特征值为 0, 因此由 $\{\boldsymbol{h}_i\}$ 的正交性, 可知只需要取 $\boldsymbol{L}$ 的最小的非零特征值对应的特征向量即可.

现在考虑 $\boldsymbol{H}$ 的行向量, 对于前面的定义, 显然其只有 $k$ 种取值, 对应到哪了类, 就只有对应的分量不是 0, 也就是说在一个 $k$ 维空间中, 所有的样本点散落在 $k$ 个点上, 这样就将所有的样本点分成了 $k$ 个类. 现在由特征向量代替 $\boldsymbol{h}_i$, 显然不再具有上述性质, 但是我们可以对这时的 $\boldsymbol{H}$ 的行向量进行 $K$-均值聚类, 得到 $k$ 个类, 于是完成样本聚类.

现在考虑式 (4.3.1), 假设图是连通的, 并且 $\boldsymbol{f}$ 中至少有 $f_i \neq f_j$, 则存在一条通路 $(f_i, f_{k_1}, \cdots, f_{k_n}, f_j)$, 上面所有的权值都大于 0, 从而可知此时 $\boldsymbol{f}^T \boldsymbol{L} \boldsymbol{f} > 0$, 于是 $\boldsymbol{L}$ 的特征值为 0 的特征向量只有 $\mathbf{1}$. 由此可知, 如果原图可以剖分为 $m$ 个不连通的子图, 则 $\boldsymbol{L}$ 可以重排为分块矩阵, 于是有 $m$ 个线性无关的特征值为 0 的特征向量, 于是每个特征向量可以看作是一个指示向量, 只有属于这个连通子图的顶点才能在对应的分量不为 0.

# 5 神经网络

神经网络从本质上说是一个广义的加法模型, 通过非线性函数将特征非线性化, 从而得到在非线性特征上的线性模型.

## 5.1 投影寻踪回归

现在考虑有 $p$ 个分量的输入 $\boldsymbol{X}$ 和目标 $\boldsymbol{Y}$, 则**投影寻踪回归** (projection pursuit regression, PPR) 模型具有形式

$$f(\boldsymbol{X}) = \sum_{m=1}^{M} g_m(\boldsymbol{w}^T \boldsymbol{X}),$$

易见这是一个在导出特征 $V_m$ 上的加法模型. 实际上只要 $M$ 足够大,PPR 模型可以任意逼近任意足够光滑的函数. 这类模型称为**普适逼近** (universal approximator). 换句话说,PPR 模型的拟合能力很强, 缺陷就是可解释性差, 因为特征进入模型的变换可以十分复杂.

现在考虑在给定的数据上极小化误差函数

$$\sum_{i=1}^{N} \left[ y_i - \sum_{m=1}^{M} g_m(\boldsymbol{w}^T \boldsymbol{x}_i) \right]^2,$$

首先给出 $M = 1$ 的算法, 对于 $M \geq 1$ 的情形, 采用向前分步算法的方式每次添加一个 $(\boldsymbol{w}_m, g_m)$. 核心思想是利用泰勒展开, 假设 $\boldsymbol{w}_0$ 是当前估计的参数, 则有

$$g(\boldsymbol{w}^T \boldsymbol{x}_i) \approx g(\boldsymbol{w}_0^T \boldsymbol{x}_i) + g'(\boldsymbol{w}_0^T \boldsymbol{x}_i)(\boldsymbol{w} - \boldsymbol{w}_0)^T \boldsymbol{x}_i,$$

从而有

$$\sum_{i=1}^{N} \left[ y_i - \sum_{m=1}^{M} g_m(\boldsymbol{w}^T \boldsymbol{x}_i) \right]^2 \approx \sum_{i=1}^{N} g'(\boldsymbol{w}_0^T \boldsymbol{x}_i)^2 \left[ \left( \boldsymbol{w}_0^T \boldsymbol{x}_i + \frac{y_i - g(\boldsymbol{w}_0^T \boldsymbol{x}_i)}{g'(\boldsymbol{w}_0^T \boldsymbol{x}_i)} \right) - \boldsymbol{w}^T \boldsymbol{x}_i \right]^2,$$

已知这是一个带权的最小二乘回归问题, 利用1.1的结论即可解决, 得到新的 $\boldsymbol{w}$.

## 5.2 单隐层神经网络

神经网络的基本结构就是单隐层的神经网络, 本质上是一个二阶段回归或分类模型. 不妨单隐层神经网络具有 $K$ 个输出 $Y_k$,$M$ 个隐层 $Z_m$, 以及 $p$ 维输入 $X_i$, 则可以给出

$$\begin{aligned} Z_m &= \sigma(\alpha_{0m} + \boldsymbol{\alpha}_m^T \boldsymbol{X}), \quad m = 1, \cdots, M \\ T_k &= \beta_{0k} + \boldsymbol{\beta}_k^T \boldsymbol{Z}, \quad , k = 1, \ldots, K \\ f_k(\boldsymbol{X}) &= g_k(\boldsymbol{T}), \quad k = 1, \ldots, K. \end{aligned} \tag{5.2.1}$$

直观理解, 通过输入层的线性组合的非线性变换得到隐层, 然后将隐层作为输入层, 其线性组合的非线性变换得到输出层.

# 6 支持向量机

支持向量机 (support vector machine, SVM) 是一种二分类的模型.

## 6.1 线性可分支持向量机

最简单的情形是假定样本是线性可分的. 同样假定给定了特征空间中的 $N$ 个训练集, 分类 $y_i \in \{-1, 1\}$. 我们希望找到一个超平面 $\boldsymbol{\beta}^T \boldsymbol{x} + b$, 它可以完全正确地分类样本, 也就是使得 $y_i(\boldsymbol{\beta}^T \boldsymbol{x}_i + b) > 0$. 定义

$$\gamma_i := y_i \left( \frac{\boldsymbol{\beta}^T \boldsymbol{x}_i + b}{||\boldsymbol{\beta}||} \right)$$

为 $(\boldsymbol{\beta}, b)$ 关于样本点 $(\boldsymbol{x}_i, y_i)$ 的几何间隔 (geometric margin). 我们希望支持向量机能最大化最小间隔, 从而得到优化问题

$$\begin{aligned} \max_{\boldsymbol{\beta},b} \quad & \gamma \\ \text{s.t.} \quad & \gamma_i := y_i \left( \frac{\boldsymbol{\beta}^T \boldsymbol{x}_i + b}{||\boldsymbol{\beta}||} \right) \geq \gamma, \quad i = 1, \cdots, N, \end{aligned}$$

这等价于

$$\begin{aligned} \max_{\boldsymbol{\beta},b} \quad & \frac{\gamma}{||\boldsymbol{\beta}||} \\ \text{s.t.} \quad & \gamma_i = y_i \left( \boldsymbol{\beta}^T \boldsymbol{x}_i + b \right) \geq \gamma, \quad i = 1, \cdots, N, \end{aligned}$$

注意到将 $\boldsymbol{\beta}$ 和 $b$ 同时缩放并不影响结果, 因此优化问题进一步等价于

$$\begin{aligned} \min_{\boldsymbol{\beta},b} \quad & \frac{1}{2}||\boldsymbol{\beta}||^2 \\ \text{s.t.} \quad & \gamma_i = y_i \left( \boldsymbol{\beta}^T \boldsymbol{x}_i + b \right) - 1 \geq 0, \quad i = 1, \cdots, N. \end{aligned} \tag{6.1.1}$$

其拉格朗日对偶函数为

$$L(\boldsymbol{\beta}, b, \boldsymbol{\alpha}) = \frac{1}{2}||\boldsymbol{\beta}||^2 - \sum_{i=1}^{N} \alpha_i [y_i \left( \boldsymbol{\beta}^T \boldsymbol{x}_i + b \right) - 1], \quad \alpha_i \geq 0, \quad i = 1, \cdots, N,$$

由于原问题是凸优化问题, 且满足 slater 条件, 从而强对偶性成立, 该问题满足 KKT 条件, 于是有

$$\alpha_i^* [y_i \left( \boldsymbol{\beta}^T \boldsymbol{x}_i + b \right)] = 0,$$

从而可根据不为 0 的 $\alpha_i^*$ 求出 $\boldsymbol{\beta}^*$.

### 6.2 软间隔最大化

# 7 奇异值分解

## 7.1 完全、紧与截断奇异值分解

给定一个 $m \times n$ 的实矩阵 $A$, 则其完全奇异值分解 (full singular value decomposition, SVD) 为

$$A = U \Sigma V^T, \tag{7.1.1}$$

其中 $U, V$ 都是正交矩阵,$\Sigma$ 是由非负降序排列的 $m \times n$ 矩形对角矩阵 (rectangular diagonal matrix). 由线性代数的知识可知 $\Sigma$ 的对角元 $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_p, p = \min\{m, n\}$ 恰好是 $A^T A$ 的特征值.

现在不妨设 $r = \text{rank}(A) \leq \min\{m, n\}$, 则

$$A = U_r \Sigma_r V_r^T \tag{7.1.2}$$

称为 $A$ 的紧奇异值分解 (compact singular value decomposition), 其中 $U_r$ 是分解 (7.1.1) 中 $U$ 的前 $r$ 列,$\Sigma_r$ 是 $\Sigma$ 的前 $r$ 行 $r$ 列,$V_r$ 是 $V$ 的前 $r$ 列. 紧奇异值分解是最为常用的形式.

又设 $0 < k < r$, 则

$$\tilde{A} = U_k \Sigma_k V_k^T \tag{7.1.3}$$

称为 $A$ 的截断奇异值分解 (truncated singular value decomposition), 其中 $U_k$ 是分解 (7.1.1) 中 $U$ 的前 $k$ 列,$\Sigma_k$ 是 $\Sigma$ 的前 $k$ 行 $k$ 列,$V_k$ 是 $V$ 的前 $k$ 列.

## 7.2 奇异值分解与矩阵近似

类似向量的二范数, 我们可以定义矩阵的二范数为其所有元素的平方和的正平方根, 则我们考虑二范数意义下的矩阵近似.

**命题 7.2.1** 正交变换不改变矩阵的二范数.

**证明** 直接由正交矩阵的性质即得. □

**定理 7.2.2** 设 $A \in \mathbb{R}^{m\times n}$,$0 < k < r = \operatorname{rank}(A)$, 并设 $\mathscr{M}_k$ 为 $\mathbb{R}^{m\times n}$ 中所有秩不超过 $k$ 的矩阵的集合, 则式 (7.1.3) 中的 $\tilde{A}$ 满足

$$\tilde{A} = \arg\min_{X\in\mathscr{M}_k} ||A - X||.$$

其中范数取的是二范数.

**证明** 见李航《统计学习基础》. □

由命题 7.2.1 及定理 7.2.2 可知此时有

$$||X - \tilde{A}||^2 = \sum_{i=k+1}^{p} \sigma_i^2.$$

# 8 树模型

## 8.1 决策树模型与特征选择

决策树 (decision tree) 模型是一个采用了 if-then 规则的顺序分类模型, 从根开始, 对特征空间进行依次划分, 最后到达一个叶结点, 从而将该实例分入该叶结点对应的类中. 决策树模型的学习本质上是从训练数据集中归纳出一组分类规则. 给定损失函数, 从所有的决策树中选择最优的决策树是一个 NP 完全问题, 从而一般采用递归的方式选择最优特征, 从而我们需要一个**度量分类能力**的指标, 于是可以根据该指标对特征进行选择, 留下具有“足够”分类能力的特征.

**定义 8.1.1** 设离散随机变量 $X$ 具有分布

$$P\{X = x_i\} = p_i, \quad i = 1, 2, \cdots, n, \tag{8.1.1}$$

随机变量 $X$ 的分布的熵 (entropy) 定义为

$$H(X) := -\sum_{i=1}^{n} p_i \log p_i. \tag{8.1.2}$$

当式 (8.1.2) 中的对数以 2 为底时, 熵的单位称为比特 (bit), 以 $e$ 为底时, 熵的单位称为奈特 (nat).

**定义 8.1.2** 离散随机变量 $X$ 给定的条件下离散随机变量 $Y$ 的条件熵 (conditional entropy) 定义为

$$H(Y|X) := \sum_{i=1}^{n} p_i H(Y|X = x_i),$$

其中 $p_i$ 如 (8.1.1) 定义. 当熵和条件熵中的概率由样本估计得到时, 所对应的估计熵分别称为**经验熵** (empirical entropy) 和**经验条件熵** (empirical conditional entropy).

**定义 8.1.3** 特征 $A$ 对训练数据集 $D$ 的信息增益 (information gain) 定义为

$$g(D, A) := H(D) - H(D|A),$$

表示由于特征 $A$ 而使得训练数据集 $D$ 的分类不确定性减少的程度. 设样本 $D$ 具有容量 $|D|$, 以下的容量同样定义, 有 $K$ 个类 $\{C_k\}$, 且设特征 $A$ 具有 $n$ 个不同取值, 根据 $A$ 的取值将样本划分为 $n$ 个不交子集 $\{D_i\}$, 记 $D_{ik} = D_i \bigcap C_k$, 表示在子集 $D_i$ 中属于类 $C_k$ 的样本, 从而定义特征 $A$ 对数据集 $D$ 的经验条件熵为

$$H(D|A) := \sum_{i=1}^{n} \frac{|D_i|}{|D|} H(D_i) = \sum_{i=1}^{n} \frac{|D_i|}{|D|} \sum_{k=1}^{K} \frac{|D_{ik}|}{|D_i|} \log \frac{|D_{ik}|}{|D_i|}.$$

从直观上来讲, 熵越小, 则分布的不确定性越小, 也就是分布中取某一些值的概率比较大. 而 $H(D|A)$ 可以形象地理解为特征 $A$ 分类后样本的“纯度”,“纯度”越高,$H(D_i)$ 就越小, 从而分类效果越好.

同时我们注意到, 分布可能的取值越少, 则其熵倾向于更小, 从而**特征可能的取值越多, 其经验熵倾向于更大**, 从而有可能导致依靠信息增益准则倾向于选择取值较多的特征 (假设一种极端情况, 如果一个特征导致对任意 $i$ 有 $|D_i| = 1$, 那这个特征的信息增益自然最大). 因此有

**定义 8.1.4** 特征 $A$ 对训练数据集 $D$ 的信息增益比 (information gain ratio) 定义为

$$g_R(D, A) := \frac{g(D, A)}{H_A(D)},$$

其中 $H_A(D)$ 定义为

$$H_A(D) := -\sum_{i=1}^{n} \frac{|D_i|}{|D|} \log \frac{|D_i|}{|D|},$$

表示特征 $A$ 的取值的熵.

**定义 8.1.5** 设离散随机变量 $X$ 具有分布

$$P\{X = x_i\} = p_i, \quad i = 1, 2, \cdots, n, \tag{8.1.3}$$

随机变量 $X$ 的分布的基尼指数 (Gini index) 定义为

$$\mathrm{Gini}(X) = \sum_{i=1}^{n} p_i(1 - p_i) = 1 - \sum_{i=1}^{n} p_i^2.$$

样本基尼指数定义为

$$\mathrm{Gini}(D) = 1 - \sum_{k=1}^{K} \left( \frac{|C_k|}{|D|} \right)^2.$$

样本 $D$ 在特征 $A$ 的条件下的基尼指数类似 8.1.3 中定义的条件熵, 不再重述.

## 8.2 决策树生成算法

### 8.2.1 ID3 算法与 C4.5 算法

ID3 算法采用信息增益作为特征选择的基准, 选定一个信息增益阈值 $\varepsilon$, 若不是空树或所有样本属于一个类, 则对每个子节点遍历当前可选特征, 选出信息增益最大的特征 $A_0$, 若其信息增益超过 $\varepsilon$, 则以该特征生成子树, 然后子树成为新的训练集, 且其可选特征中去除 $A_0$. 容易看出 **ID3 算法生成的决策树每个特征最多使用一次**, 也就是每一层的划分都一定要完全分割.

C4.5 算法和 ID3 算法一致, 只不过特征选择依据的准则是信息增益比.

### 8.2.2 剪枝

决策树的剪枝 (pruning) 是一种避免过拟合的手段. 现在假设有一个决策树 $T$, 其叶结点个数为 $T$,$H_t(T)$ 为叶结点 $t$ 上的经验熵, 如同定义 8.1.3 中的特征经验熵. 于是对权值 $\alpha > 0$, 可以定义这个决策树学习的损失函数为

$$C_\alpha(T) = \sum_{t=1}^{T} N_t H_t(T) + \alpha |T|,$$

其中 $N_t$ 表示叶结点 $t$ 的样本点个数.

依据此损失函数, 递归地从叶结点往上搜索, 如果剪掉该叶结点能减小损失函数, 则剪枝.

### 8.2.3 CART

分类与回归树 (classification and regression tree, CART) 在每次划分特征空间时并不是全部划分完全, 而是**假设决策树是二叉树**, 从而每次只是将特征空间划分为两部分, 从而其**特征可以重复使用**.

对于回归树, 其将输入空间划分为 $M$ 个单元 $\{R_m\}$, 在每一个单元 $R_m$ 上固定输出为 $c_m$, 从而可以将回归树模型表示为

$$f(x) = \sum_{m=1}^{M} c_m I(x \in R_m).$$

特别地, 当采用平方损失函数时, 最优的 $c_m$ 的估计值为

$$\hat{c_m} = \text{ave}\{y_i | x_i \in R_m\}.$$

由于回归树每次只划分为两个子集, 从而需要选取切分变量 (splitting variable)$j$ 和切分点 (spliting point)$s$, 这里切分变量是输入的特征向量的分量, 从而得到优化问题

$$\min_{j,s} \left[ \min_{c_1} \sum_{x_i \in R_1(j,s)} (y_i - c_1)^2 + \min_{c_2} \sum_{x_i \in R_2(j,s)} (y_i - c_2)^2 \right],$$

这里

$$R_1(j,s) = \{x | x^j \le s\}, \quad R_2(j,s) = \{x | x^j > s\}.$$

这样得到的回归树称为**最小二乘回归树** (least squares regression tree).

对于分类树, 其选择特征的依据是使得基尼指数最小, 需要遍历所有的特征以及每个特征可能的切分点.

CART 的剪枝如同 8.2.2 小节中所述. 将 $\alpha$ 从 0 开始递增, 得到的树会越来越简单, 且是嵌套的, 然后通过交叉验证选择出最优的树.

# 9 贝叶斯分类器

贝叶斯决策是一种在概率框架下实施决策的基本方法, 基于决策损失函数来进行决策. 对于有 $N$ 种类别标记 $\{c_i\}$ 的分类任务, 记 $\lambda_{ij}$ 为将真实样本为 $c_j$ 的类的损失, 从而基于后验概率可得将样本 $\boldsymbol{x}$ 分类为 $c_i$ 的期望损失 (或称条件风险) 为

$$R(c_i | \boldsymbol{x}) = \sum_{j=1}^{N} \lambda_{ij} P\{c_j | \boldsymbol{x}\},$$

我们希望找到的分类准则 $h$ 能最小化总体风险

$$R(h) = \mathbb{E}_{\boldsymbol{x}}[R(h(\boldsymbol{x}) | \boldsymbol{x})].$$

**贝叶斯判定准则** (Bayes decision rule) 说的就是为了最小化总体风险, 只需要对每个样本, 都最小化其条件风险, 即

$$h^*(\boldsymbol{x}) = \arg\min_c R(c | \boldsymbol{x}),$$

此时的 $h^*$ 称为**贝叶斯最优分类器** (Bayes optimal classifier), 对应的风险 $R(h^*)$ 称为**贝叶斯风险** (Bayes risk).[10] 容易看出, 对于 0/1 损失函数, 贝叶斯最优分类器为

$$h^*(\boldsymbol{x}) = \arg\max_c P\{c | \boldsymbol{x}\},$$

也就是选择使得后验概率最大的类别标记.

综上所述, 贝叶斯分类器的学习目标就是从样本中尽可能准确地估计后验概率. **生成式模型** (generative models) 首先对联合概率分布 $P(\boldsymbol{x}, c)$ 进行建模, 然后得到后验概率. **判别式模型** (discriminative models) 则是根据样本直接对后验概率建模.

由贝叶斯公式可知, 为了学习一个贝叶斯分类器, 需要估计先验分布 $P(c)$ 与条件分布 $P(\boldsymbol{x}|c)$. 对于先验分布, 可以从样本的频数进行估计. 但是对于后验分布 $P(\boldsymbol{x}|c)$, 由于其涉及到所有的类别属性, 所以当类别及其属性较多时很困难.

[10] 关于贝叶斯决策理论可参看另一部分的总结.

## 9.1 朴素贝叶斯分类器

为了解决上述困难, **朴素贝叶斯分类器** (naive Bayes classifier) 采用**属性条件独立性假设** (attribute conditional), 也就是对于已知的类别, 假设所有的属性相互独立, 从而由贝叶斯公式得

$$P\{c|\boldsymbol{x}\} = \frac{P(c)}{P(\boldsymbol{x})} = \prod_{i=1}^{d} P\{x_i|c\},$$

其中 $x_i$ 是样本的第 $i$ 个分量, 于是我们就可以为每一个分量, 利用样本频数对其条件分布进行估计.

为了避免出现那些没有在样本中出现的属性值被频数估计置为 0, 可以通过一些平滑的方式来改进概率估计, 例如将频数估计中的分子和分母都分别加上一个正数.

# 10 基展开与正则化

线性模型虽然简单, 但是很多时候具有较好的性质, 例如说可解释性较好, 以及当特征维数很大的时候, 线性模型是能够避免过拟合的唯一方法.

对于输入向量 $\boldsymbol{X} \in \mathbb{R}^p$, 考虑变换 $h_m(\boldsymbol{X})$, 然后考虑**线性基展开** (linear basis expansion) 模型

$$f(\boldsymbol{X}) = \sum_{m=1}^{M} \beta_m h_m(\boldsymbol{X}). \tag{10.0.1}$$

线性基展开模型的优点是只要确定了基函数 $h_m$, 余下的工作和一般的线性回归一样.

## 10.1 分段多项式与样条

在模型 (10.0.1) 中, 考虑所有的基函数都是幂函数的情形. 假定输入 $X$ 是一维的, 考虑将 $X$ 的定义域分为 $K+1$ 段, 由 $K$ 个结点 $\{\varepsilon_l\}, l = 1, \cdots, K$ 间隔, 我们希望在每段都拟合一个 $M$ 次多项式, 并且要求在整个定义域上有连续的 $M-1$ 阶导数. 最直接的想法是给拟合加上在结点处的约束. 另一种更好的方法是采用基函数

$$\begin{aligned} h_j(X) &= X^{j-1}, \quad j = 1, \cdots, M+1, \\ h_{M+l}(X) &= (X - \varepsilon_l)_+^M, \quad l = 1, \cdots, K, \end{aligned} \tag{10.1.2}$$

又称为**截断幂基集** (truncated-power basis), 其中包含 $K + M + 1$ 个元素.

这么做的理由非常直接, 对于第一段的 $M$ 次函数, 显然可以由前 $M+1$ 个基函数的线性组合给出, 然后因为在第一个结点处和第二段的分段多项式具有相等的前 $M-1$ 阶导数, 从而第二段的多项式表达式减去第一段的多项式表达式得到的至多 $M$ 次多项式在 $\varepsilon_1$ 处的前 $M-1$ 阶导数都是 0, 从而只能是 $\lambda(X - \varepsilon_1)_+^M$ 的形式. 以此类推即可得到整个截断幂基集.

根据截断幂基集拟合出的分段 $M$ 次多项式又称为 $M$ **次样条** (order-M spline), 也称**回归样条** (regression spline). 根据 (10.1.2) 给出的的截断幂基集, 根据训练集进行最小二乘回归即可估计出系数.

### 10.1.1 自然三次样条

多项式拟合在靠近边界的地方不稳定, 为此我们可以要求在边界的两个结点之外函数是线性的, 在其余地方是三次函数, 从而得到**自然三次样条** (natural cubic splines). 注意到当取 $M+3$ 时对应于 (10.1.2) 的基函数集有 $K+4$ 个元素, 现在我们尝试给出自然三次样条类似的基函数集.

首先在第一段是线性函数, 从而得到前两个基函数 $N_1(X) = 1, N_2(X) = X$, 然后从第二段开始, 我们希望新加入的函数满足在 $\varepsilon_1$ 处前两阶导数为 0, 同时当 $X \geq \varepsilon_K$ 时, 其二阶导数和三阶导数都是 0. 从 (10.1.2) 给出的截断幂基集出发, 我们希望用那里给出的基函数出发组合得到我们需要的基函数. 注意到 $(X - \varepsilon_1)_+^3 - (X - \varepsilon_K)_+^3$ 在 $X \geq \varepsilon_K$ 时变为二次函数了, 并且具有连续的前两阶导数. 于是我们只需要再这样操作一次就可以得到一次函数了. 具体地, 令

$$d_k(X) = \frac{(X - \varepsilon_1)_+^3 - (X - \varepsilon_K)_+^3}{\varepsilon_K - \varepsilon_k},$$

然后令 $N_{k+2}(X) = d_k(X) - d_{K-1}(X)$, 从而得到一个自然三次样条的基函数集

$$N_1(X) = 1, \quad N_2(X) = X, \quad N_{k+2}(X) = d_k(X) - d_{K-1}(X), \quad k = 1, \cdots, K-2,$$

从而此时的基函数集只有 $K$ 个元素.

## 10.2 光滑样条

类似于线性回归中防止过拟合的方法, 我们也给多项式拟合添加惩罚项. 具体地, 在所有具有二阶连续导数的函数 $f$ 中, 找一个函数最小化罚残差平方和 (penalized residual sum of squares), 也就是

$$f(x) = \arg\min_f \quad \mathrm{RSS}(f, \lambda) = \sum_{i=1}^{N} [y_i - f(x_i)]^2 + \lambda \int [f''(t)]^2 dt. \tag{10.2.3}$$

**命题 10.2.1** 对 $0 < \lambda < \infty$, 式 (10.2.3) 具有唯一的解, 即自然三次样条, 结点是 $x_i, i = 1, \cdots, N$.

**证明** 不妨设 $g$ 是对应的自然三次样条, 在结点 $x_j$ 上取值 $z_j$, 而 $\tilde{g}$ 是定义域上任意的二阶可导的函数, 也在结点 $x_j$ 上取值 $z_j$. 令 $h(x) = g(x) - \tilde{g}(x)$, 则由自然三次样条的边界条件, 分部积分可得

$$\int g''(x)h''(x)dx = -\int g'''(x)h'(x)dx = \sum_{j=1}^{N} g'''(x_j^+)[h(x_{j+1}) - h(x_j)] = 0,$$

最后一步是因为 $h(x_j) = 0$. 从而

$$\int [\tilde{g}''(t)^2 - g''(t)^2]dt = \int h''(t)[h''(t) + 2g''(t)]dt = \int h''(t)^2 dt \geq 0,$$

且等号成立当且仅当 $h''(x) \equiv 0$, 于是 $\tilde{g}$ 也是一个自然样条. 现在考虑优化问题 (10.2.3), 则对于任意函数 $\tilde{f}$, 找一个在 $x_i$ 上和它取值一样的自然三次样条, 那么可以更优, 于是我们就可以仅仅从自然三次样条中去找最优解即可. □

现在可以将问题 (10.2.3) 写成

$$\min_{\boldsymbol{\theta}} \quad \left[(\boldsymbol{y} - \boldsymbol{N}\boldsymbol{\theta})^T(\boldsymbol{y} - \boldsymbol{N}\boldsymbol{\theta}) + \lambda \boldsymbol{\theta}^T \Omega_{\boldsymbol{N}} \boldsymbol{\theta}\right], \tag{10.2.4}$$

这里 $\{\boldsymbol{N}\}_{ij} = N_j(x_i)$, 如同线性回归, 第 $i$ 行第 $j$ 列表示第 $j$ 个基函数在第 $i$ 个样本点上的取值, 并且

$$\{\Omega_{\boldsymbol{N}}\}_{ij} = \int N_j''(t) N_i''(t) dt,$$

于是这就变成了一个广义的岭回归问题, 容易解得参数估计

$$\hat{\boldsymbol{\theta}} = (\boldsymbol{N}^T\boldsymbol{N} + \lambda\Omega_{\boldsymbol{N}})^{-1}\boldsymbol{N}^T\boldsymbol{y}.$$

接下来考虑所有样本点上的估计

$$\hat{\boldsymbol{f}} = \boldsymbol{N}(\boldsymbol{N}^T\boldsymbol{N} + \lambda\Omega_{\boldsymbol{N}})^{-1}\boldsymbol{N}^T\boldsymbol{y} = \boldsymbol{S}_\lambda \boldsymbol{y},$$

这里 $\boldsymbol{S}_\lambda = \boldsymbol{N}(\boldsymbol{N}^T\boldsymbol{N} + \lambda\Omega_{\boldsymbol{N}})^{-1}\boldsymbol{N}^T$ 称为**光滑子矩阵** (smoother matrix).

## 10.3 再生核希尔伯特空间

设 $\mathcal{H}$ 是一个希尔伯特空间, 二元函数 $K(x, y) : \mathcal{H} \times \mathcal{H} \to \mathcal{F}$ 满足对任意 $f \in \mathcal{H}, t \in \mathcal{F}$ 有

$$f(t) = \langle f, K(\cdot, t)\rangle \tag{10.3.5}$$

成立, 则称 $K$ 是一个**核函数**. 此时可以将 $f$ 和 $K(\cdot, t)$ 视为 $\mathcal{H}$ 中的一个无限维向量.

所谓核的**可再生性**是由核函数的定义有

$$K(x, y) = \langle K(\cdot, x), K(\cdot, y)\rangle.$$

从另一个角度来理解, 类似线性代数, 将 $K(x,y)$ 视为无穷维的矩阵, 则如果对于任意函数有

$$\iint f(x)K(x,y)f(y)dxdy \geq 0$$

且 $K(x,y)=K(y,x)$, 则 $K$ 是一个核函数. 考虑特征函数

$$\int K(x,y)\psi(x)dx = \lambda\psi(y),$$

于是定义内积

$$\langle \psi_1, \psi_2 \rangle = \iint \psi_1(x)\psi_2(y)dxdy,$$

可知特征方程是正交的. 和矩阵类似, 由 Mercer 定理知

$$K(x,y) = \sum_{i=0}^{\infty} \lambda_i \psi_i(x)\psi_i(y).$$

接下来以 $\{\sqrt{\lambda_i}\psi\}_{i=1}^{\infty}$ 作为一组正交基构成一个希尔伯特空间 $\mathcal{H}$, 在此时才能使等式 (10.3.5) 成立.

# 11 特征选择与稀疏学习

## 11.1 特征选择

我们知道很多时候统计学习任务中, 一些特征是无关紧要的, 增加这些特征往往导致模型训练变得困难, 例如说维数灾难. 从所有的特征构成的幂集中搜索所有的子集显然从时间开销上是不可接受的, 因此往往采用其他的一些方法寻找次优的特征子集.

**子集搜索** (subset search) 方法就是一种贪心算法, 一开始选择一个最优的特征, 然后递推地每次加入一个最优的特征, 直到加入的特征没有优化为止. 这种子集搜索方式也称为前向搜索. 后向搜索将这个过程反过来, 首先选择所有的特征, 然后每次去掉一个最无关的特征.

**子集评价** (subset evaluation) 通过计算样本集合 $D$ 关于特征子集 $A$ 的信息增益 [11], 然后以此作为子集选择的评价. 基于此, 我们知道决策树可以作为特征子集选择的一种模型, 决策树每一行都可以作为一个可选择的子集.

### 11.1.1 过滤式选择与包裹式选择

过滤式选择是在训练模型之前就先将特征选好, 而包裹式选择是通过训练模型然后评价模型性能来选择特征.

Relief(Relevant Features) 是一种二分类任务的过滤式特征选择方法, 给定大小为 $m$ 的训练集 $\{(\boldsymbol{x}_i, y_i)\}$, 从 $\boldsymbol{x}_i$ 的同类样本中选择最近邻 $\boldsymbol{x}_{i,nh}$, 称为**猜中近邻** (near-hit), 同样从异类样本中寻找最近邻 $\boldsymbol{x}_{i,nm}$, 称为**猜错近邻** (near-miss), 然后对于属性 $j$, 定义统计量

$$\delta_j = \sum_i -\mathrm{diff}(x_i^j, x_{i,nh}^j)^2 + \mathrm{diff}(x_i^j, x_{i,nm}^j)^2,$$

其中 $x_i^j$ 代表样本 $\boldsymbol{x}_i$ 的第 $j$ 个分量的取值,diff 是一个合适的表示距离的函数. 容易看出统计量 $\delta_j$ 刻画了特征 $j$ 将样本分离的能力, 同类样本距离越近, 异类样本距离越远,$\delta$ 统计量越大, 从而可以通过这种方式选择合适的特征. 显然这个思想类似于1.5中的线性判别分析.

LVW(Las Vegas Wrapper) 是一种包裹式特征选择方法, 其本质是一种启发式算法, 在给定步数内, 每次随机产生特征子集 $A'$, 然后训练模型, 如果效果提升超过阈值, 或者效果不变的情况下特征数目越少, 就更新子集. 其中评价效果的方式可以有很多, 例如说交叉验证.

[11] 信息增益在8.1中定义

### 11.1.2 嵌入式选择与 $L_1$ 正则化

嵌入式选择是特征选择和模型训练同时进行的特征选择方式. 考虑经典线性回归模型, 一种称为**岭回归** (ridge regression) 的避免过拟合的方式是引入**对非截距项系数的** $L_2$ 正则项, 即是

$$\min_{\boldsymbol{w}} \quad (\boldsymbol{y}-\boldsymbol{X}^T\boldsymbol{w})^T(\boldsymbol{y}-\boldsymbol{X}^T\boldsymbol{w})+\lambda\boldsymbol{w}^T\boldsymbol{w}, \tag{11.1.1}$$

从1.1小节我们知道上述优化问题的解是 [12]

$$\boldsymbol{w}=(\boldsymbol{X}^T\boldsymbol{X}+\lambda\boldsymbol{I})^{-1}\boldsymbol{X}^T\boldsymbol{y}.$$

另一种被称为**最小绝对值收缩算子** (Least Absolute SHrinkage and Selection Pperator, LASSO) 的方法, 或者直接称为“LASSO 回归”, 是引入 $L_1$ 正则项, 即是

$$\min_{\boldsymbol{\beta},\beta_0} \quad (\boldsymbol{y}-\beta_0\boldsymbol{1}-\boldsymbol{X}^T\boldsymbol{\beta})^T(\boldsymbol{y}-\beta_0\boldsymbol{1}-\boldsymbol{X}^T\boldsymbol{\beta})+\lambda||\boldsymbol{\beta}||_1. \tag{11.1.2}$$

注意到 $L_1$ 范数是分量绝对值相加, 所以有“尖”, 形象地理解, 这样更容易产生“稀疏”的解, 也就是最优选择使得 $\boldsymbol{w}$ 有一些分量是 0.

类似 LASSO, 考虑低秩回归 (reduce rank regression, RRR), 其形式为

$$\begin{aligned}\min_{\boldsymbol{B}} \quad & \frac{1}{2}||\boldsymbol{y}-\boldsymbol{X}\boldsymbol{B}||_F^2\\ \text{s.t.} \quad & \text{rank}(\boldsymbol{B})\le r\end{aligned}$$

此时的约束类似于 $L_0$ 约束, 性质不好处理, 因此考虑矩阵的**核范数** (nuclear norm), 其等于矩阵所有奇异值的和, 从而得到类似 LASSO 的约束优化问题

$$\min_{\boldsymbol{B}} \quad \frac{1}{2}||\boldsymbol{y}-\boldsymbol{X}\boldsymbol{B}||_F^2+||\boldsymbol{B}||_*,$$

其中 $||\cdot||_*$ 是矩阵核范数.

## 11.2 线性模型中的 LASSO

现在详述线性回归模型中的 LASSO 正则化, 在不引起歧义的情况下, 也成为 $L_1$ 正则化. 使用 $L_1$ 正则化的原因在于普通的线性回归虽然是无偏的或者偏差很小, 但是方差一般较大. 从1.2小节的结果知道减少一些自变量可以获得较小的方差, 从而可能提高预测精度, 而 LASSO 正则化同时具有收缩系数以及选取自变量的作用.

现在重写 (11.1.2) 为

$$\begin{aligned}\min_{\boldsymbol{w},w_0} \quad & \frac{1}{2N}||\boldsymbol{y}-\beta_0\boldsymbol{1}-\boldsymbol{X}\boldsymbol{\beta}||_2^2\\ \text{s.t.} \quad & ||\boldsymbol{\beta}||_1\le t,\end{aligned} \tag{11.2.3}$$

由于问题 (11.1.2) 是凸的, 因此容易使用拉格朗日乘子法说明 (11.2.3) 和 (11.1.2) 之间建立对应关系. 之所以加上了系数 $1/N$, 目的在于有的时候并不是使用所有的训练集进行训练, 因此这样可以保证 (11.2.3) 中的 $t$ 是不随训练样本大小而改变. 在应用中往往采用 (11.2.3) 这种形式, 因为这样可以将 $t$ 从 0 开始增大, 观察进入模型的变量.

由于问题是凸的, 因此可以采用求导的方式来解决. 问题在于 $L_1$ 正则化项在 0 处没有导数, 因此实际解决的时候采用迭代的方法, 最常用的就是**坐标下降** (coordinate descent). 考虑问题 (11.1.2), 其中训练样本进行了中心化, 定义参数为 $\lambda$ 的**软阈值函数** (soft thresholding function) 为

$$S_\lambda(x):=\text{sign}(x)(|x|-\lambda)_+,$$

考虑一维 $\beta$ 的 LASSO 回归

$$\min_{\beta} \quad \frac{1}{2N}\sum_{i=1}^N(y_i-z_i\beta)^2+\lambda|\beta|,$$

[12] 我们知道当样本点很少的时候,$\boldsymbol{X}^T\boldsymbol{X}$ 不一定满秩, 所以一开始引入岭回归就是为了解决这个问题.

容易证明其最优解为

$$\hat{\beta} = S_\lambda\left(\frac{1}{N}\boldsymbol{z}^T\boldsymbol{y}\right).$$

坐标下降就是依次对 $\boldsymbol{\beta}$ 的每一个分量采用这个做法, 固定其余所有的分量, 对其余分量回归的误差进行 LASSO 回归, 每次只优化一个分量.

## 11.3 ADMM 算法

考虑 LASSO 问题

$$\min_{\boldsymbol{\beta}} \quad \frac{1}{2}||\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}||_2^2 + \lambda||\boldsymbol{\beta}||_1,$$

当 $X$ 的列向量是正交的时候, 可以得到一个闭形式的解, 但是一般情况则不行. 现在令

$$f(\boldsymbol{\beta}) = \frac{1}{2}||\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}||_2^2, \quad g(\boldsymbol{\gamma}) = \lambda||\boldsymbol{\gamma}||_1,$$

则 LASSO 问题等价于优化问题

$$\begin{aligned} \min_{\boldsymbol{\beta},\boldsymbol{\gamma}} \quad & f(\boldsymbol{\beta}) + g(\boldsymbol{\gamma}) \\ \text{s.t.} \quad & \boldsymbol{\beta} = \boldsymbol{\gamma} \end{aligned}$$

于是可以考虑增广拉格朗日函数

$$L_\rho(\boldsymbol{\beta},\boldsymbol{\gamma},\boldsymbol{\alpha}) = f(\boldsymbol{\beta}) + g(\boldsymbol{\gamma}) + \boldsymbol{\alpha}^T(\boldsymbol{\beta}-\boldsymbol{\gamma}) + \frac{\rho}{2}||\boldsymbol{\beta}-\boldsymbol{\gamma}||_2^2,$$

所谓交替方向乘子法 (Alternating Direction Method Of Multipliers, ADMM) 就是交替地优化 $\boldsymbol{\beta}$ 和 $\boldsymbol{\gamma}$. 注意到在 $L_\rho(\boldsymbol{\beta},\boldsymbol{\gamma},\boldsymbol{\alpha})$ 中关于 $\boldsymbol{\beta}$ 就是一个岭回归问题, 关于 $\boldsymbol{\gamma}$ 就是一个正交设计下的 LASSO 问题, 其有闭形式的解, 所以在第 $k$ 步对 $\boldsymbol{\beta}$ 和 $\boldsymbol{\gamma}$ 的更新分别为

$$\begin{aligned} \boldsymbol{\beta}^{(k+1)} &\leftarrow \left(\boldsymbol{X}^T\boldsymbol{X} + \rho\boldsymbol{I}\right)^{-1}\left(\boldsymbol{X}^T\boldsymbol{y} + \rho\boldsymbol{\gamma}^{(k)} - \boldsymbol{\alpha}^{(k)}\right) \\ \boldsymbol{\gamma}^{(k+1)} &\leftarrow S_{\lambda/\rho}\left(\boldsymbol{\beta}^{(k+1)} + \frac{\boldsymbol{\alpha}^{(k)}}{\rho}\right) \\ \boldsymbol{\alpha}^{(k+1)} &\leftarrow \boldsymbol{\alpha}^{(k)} + \rho\left(\boldsymbol{\beta}^{(k+1)} - \boldsymbol{\gamma}^{(k+1)}\right) \end{aligned}$$

## 11.4 LASSO 解的性质

使用 LASSO 方法的时候我们希望知道这种方式的系数估计和预测的精度, 因此需要研究 LASSO 问题

$$\min_{\boldsymbol{\beta}} \quad \frac{1}{2n}||\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}||_2^2 + \lambda||\boldsymbol{\beta}||_1$$

的解 $\hat{\boldsymbol{\beta}}$ 的性质.

**命题 11.4.1** LASSO 问题的解 $\hat{\boldsymbol{\beta}}$ 满足

$$\frac{1}{n}||\boldsymbol{X}^T\left(\boldsymbol{y}\boldsymbol{X}\hat{\boldsymbol{\beta}}\right)||_\infty \le \lambda.$$

**证明** 由于原问题是凸优化问题, 因此由一阶条件可知 0 是原问题关于 $\hat{\boldsymbol{\beta}}$ 的次梯度 (sub gradient), 即是

$$0 \in \frac{1}{n}\boldsymbol{X}^T\left(\boldsymbol{y}\boldsymbol{X}\hat{\boldsymbol{\beta}}\right) + \partial||\hat{\boldsymbol{\beta}}||_1,$$

这里 $\partial||\hat{\boldsymbol{\beta}}||_1$ 就是 $L_1$ 项的次导数 (sub differantial), 从而由 1 范数的次梯度的性质可知结论成立. □

现在不妨假设 $\boldsymbol{\beta}^*$ 确实是稀疏的, 记 $S = \text{supp}(\boldsymbol{\beta}^*)$, 也就是其非零的分量, 然后用 $\boldsymbol{\beta}_S^*$ 表示对应的分量构成的向量, 所谓 $\boldsymbol{\beta}^*$ 确实是稀疏的, 我们希望 $S$ 远小于 $n$.

**引理 11.4.2 (Basic inequality)** 记

$$\lambda = C\sigma\sqrt{\frac{\ln p}{n}}, \quad C > 2\sqrt{2},$$

则以概率至少 $1 - p^{-C^2/8}$ 成立不等式

$$\frac{1}{n}\left|\left|\boldsymbol{X}\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\right)\right|\right|_2^2 + \lambda||\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*||_1 \leq 4\lambda||\hat{\boldsymbol{\beta}}_S - \boldsymbol{\beta}_S^*||_1. \tag{11.4.4}$$

**证明** 由于 $\hat{\boldsymbol{\beta}}$ 是原 LASSO 问题的解, 从而有不等式

$$\frac{1}{2n}||\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}||_2^2 + \lambda||\hat{\boldsymbol{\beta}}||_1 \leq \frac{1}{2n}||\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}_S^*||_2^2 + \lambda||\boldsymbol{\beta}_S^*||_1,$$

将 $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta}^* + \boldsymbol{\varepsilon}$ 带入上式并整理得到

$$\frac{1}{2n}\left|\left|\boldsymbol{X}\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\right)\right|\right|_2^2 \leq \frac{1}{n}\boldsymbol{\varepsilon}^T\boldsymbol{X}\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\right) + \lambda||\boldsymbol{\beta}^*||_1 - \lambda||\hat{\boldsymbol{\beta}}||_1,$$

于是我们现在要估计 $\boldsymbol{\varepsilon}^T\boldsymbol{X}\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\right)$, 为此根据 $\boldsymbol{\varepsilon}$ 是高斯噪声的假设, 有不等式

$$P\left\{\frac{1}{n}||\boldsymbol{X}^T\boldsymbol{\varepsilon}||_\infty > \frac{\lambda}{2}\right\} \leq \sum_{j=1}^{p} P\left\{\frac{1}{n}|\boldsymbol{x}_j^T\boldsymbol{\varepsilon}| > \frac{\lambda}{2}\right\} \leq pe^{-n\lambda^2/8\sigma^2},$$

其中我们对 $\boldsymbol{x}_j$ 做了正则化, 使得 $||\boldsymbol{x}_j||_2 = \sqrt{n}$. 于是我们得到以高的概率成立不等式

$$\frac{1}{n}\left|\left|\boldsymbol{X}\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\right)\right|\right|_2^2 \leq \lambda||\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*||_1 + 2\lambda(||\boldsymbol{\beta}^*||_1 - ||\hat{\boldsymbol{\beta}}||_1),$$

不等式两边同时加上 $\lambda||\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*||_1$, 然后不等式右边按照 $S$ 和 $S^C$ 分开计算可得结果. □

以上引理给出了预测误差被系数估计误差限制的一个界, 一般地我们想据此给出预测误差的界. 令 $\Delta_S = \hat{\boldsymbol{\beta}}_S - \boldsymbol{\beta}_S^*$, 则当不等式成立时, 由柯西不等式还可进一步得到 $4\lambda||\hat{\boldsymbol{\beta}}_S - \boldsymbol{\beta}_S^*||_1 \leq 4\lambda\sqrt{S}||\hat{\boldsymbol{\beta}}_S - \boldsymbol{\beta}_S^*||_2$. 进一步也可以得到 $\lambda||\Delta_{S^C} + \Delta_S||_1 \leq 4\lambda||\Delta_S||_1$, 于是 $\lambda||\Delta_{S^C}||_1 \leq 3\lambda||\Delta_S||_1$.

为了给出预测误差的界, 需要矩阵 $\boldsymbol{X}$ 在稀疏的方向不能太奇异, 于是需要加一个 restricted eigenvalue 条件, 也就是存在 $\kappa > 0$ 满足

$$\frac{||\boldsymbol{X}\Delta||_2}{\sqrt{n}||\Delta_S||_2} \geq \kappa$$

对任意 $\Delta$ 成立. 若此条件成立, 则可以据此把 $||\Delta_S||_2$ 消掉得到预测误差的一个界.

如果不对 $X$ 提任何条件, 对于另一种形式的 LASSO 问题的写法, 也有类似的结果:

**命题 11.4.3** 假设真实系数 $\boldsymbol{\beta}^*$ 满足 $||\boldsymbol{\beta}^*||_1 \leq K$, 则 LASSO 问题

$$\min_{\boldsymbol{\beta}} \quad ||\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}||_2^2 \quad \text{s.t.} \quad ||\boldsymbol{\beta}||_1 \leq K$$

的解为 $\hat{\boldsymbol{\beta}}$, 则预测误差满足不等式

$$\frac{1}{n}\left|\left|\boldsymbol{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)\right|\right|_2^2 \leq CK\sigma\sqrt{\frac{\log p}{n}}.$$

**证明** 由解的最优性可以得到

$$0 \geq (\hat{\boldsymbol{y}} - \boldsymbol{y}^*)^T(\hat{\boldsymbol{y}} - \boldsymbol{y}) = ||\hat{\boldsymbol{y}} - \boldsymbol{y}^*||_2^2 - (\hat{\boldsymbol{y}} - \boldsymbol{y}^*)^T(\hat{\boldsymbol{y}} - \boldsymbol{y}^*),$$

于是

$$\frac{1}{n}\left|\left|\boldsymbol{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)\right|\right|_2^2 \leq \frac{1}{n}(\hat{\boldsymbol{y}} - \boldsymbol{y}^*)^T(\hat{\boldsymbol{y}} - \boldsymbol{y}^*) = \frac{1}{n}\boldsymbol{\varepsilon}^T\boldsymbol{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*) \leq \frac{1}{n}|\boldsymbol{X}^T\boldsymbol{\varepsilon}||_\infty||\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*||_1,$$

再由引理 11.4.2 可知以高概率有

$$\frac{1}{n}||\boldsymbol{X}^T\boldsymbol{\varepsilon}||_\infty||\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*||_1 \leq \frac{\lambda}{2}(2K) = \lambda K,$$

其中 $\lambda$ 的取值和引理 11.4.2 中一样即可. □

# 12 模型评估与选择

## 12.1 偏差、方差与预测误差

现在我们有一个目标变量 $Y$ 以及一个由训练样本估计的预测模型 $\hat{f}(X)$, 给定一个损失函数 $L(Y,\hat{f}(X))$, 则给定训练集 $T$, 定义*检验误差* (test error) 或*泛化误差* (generalization) 为

$$\mathrm{Err}_T := \mathbb{E}[L(Y,\hat{f}(X)|T)],$$

也就是**给定训练集** $T$ 时模型在独立的测试集上的期望损失. 定义*期望预测误差* (expected prediction error) 为

$$\mathrm{Err} := \mathbb{E}[L(Y,\hat{f}(X))] = \mathbb{E}[\mathrm{Err}_T],$$

也就是对所有可能的变量都取了期望. 一般来说, 我们的目标是希望估计泛化误差, 也就是 $\mathrm{Err}_T$.

接下来定义模型的**训练误差** (training error) 为

$$\overline{\mathrm{err}} = \frac{1}{N}\sum_{i=1}^{N} L(y_i,\hat{f}(x_i)).$$

训练误差就是训练集内的误差, 一般来说训练误差并不是泛化误差的一个好的估计.

对于使用平方损失函数的回归模型, 容易证明所谓的**偏差-方差分解**, 即是

$$\begin{aligned}\mathrm{Err}(x_0) &= \mathbb{E}[(Y-\hat{f}(x_0))^2|X=x_0]\\ &= \mathbb{E}[(Y-f(x_0)+f(x_0)-\mathbb{E}[\hat{f}(x_0)]+\mathbb{E}[\hat{f}(x_0)]-\hat{f}(x_0))^2]\\ &= \sigma^2+\mathrm{Bias}^2(\hat{f}(x_0))+\mathrm{Var}(\hat{f}(x_0)),\end{aligned}$$

这里 $\sigma^2$ 是由模型假设 $Y=f(X)+\varepsilon$ 导出的, 也就是默认存在一个“真实的”模型.

## 12.2 样本内误差及模型选择准则

定义**样本内误差** (in-sample error) 为

$$\mathrm{Err}_{\mathrm{in}} := \frac{1}{n}\sum_{i=1}^{N}\mathbb{E}_{Y_i^0}[L(Y_i^0,\hat{f}(x_i))|T],$$

也就是在观测到的输出有随机误差的假设下, 用固定的模型对所有样本点重新采样的损失的期望. 定义*乐观度* (optimism) 为

$$\mathrm{op} := \mathrm{Err}_{\mathrm{in}} - \overline{\mathrm{err}}.$$

注意到此时用于估计模型的那一组观测值是固定的, 于是对这些观测再取一次期望得到平均乐观度

$$\omega := \mathbb{E}_y[\mathrm{op}].$$

对于使用平方损失函数的模型, 使用类似偏差-方差分解的技术容易证明

$$\omega = \frac{2}{N}\mathrm{Cov}(\hat{y}_i,y_i), \tag{12.2.1}$$

也就是平均乐观度取决于观测值 $y_i$ 如何影响它本身的估计 $\hat{y}_i$.

对于1.1小节中叙述的普通线性回归模型, 则容易知道

$$\mathrm{Cov}(\hat{y}_i,y_i) = (p+1)\sigma^2,$$

也就是

$$\mathbb{E}_y[\mathrm{Err}_{\mathrm{in}}] = \mathbb{E}_y[\overline{\mathrm{err}}] + \frac{2(p+1)}{N}\sigma^2,$$

对于线性回归模型, 显然希望上式最小, 从而变成了选择进入模型的参数的个数 $p+1$. 基于这种思路,$C_p$ **统计量**就定义为

$$C_p := \overline{\mathrm{err}} + \frac{2d}{N}\hat{\sigma}^2,$$

其中 $\hat{\sigma}^2$ 使用具有低偏差的模型估计, 而 $d$ 是模型需要估计的参数个数.

**AIC 信息准则** (Akaike information criterion, AIC) 是这种模型选择思想的另一种形式, 此时假设模型的损失函数是负

$$\mathrm{AIC} := -\frac{2}{N}\cdot \mathrm{loglik} + \frac{2d}{N},$$

其中 loglik 是模型对数似然的最大值. 容易验证对于高斯模型, 当 $\hat{\sigma}^2$ 已知时 $AIC$ 和 $C_p$ 统计量是等价的. 也可以将 AIC 准则写成带参数的形式, 即

$$AIC = \overline{\mathrm{err}}(\alpha) + \frac{2d(\alpha)}{N}\hat{\sigma}^2,$$

其中 $\alpha$ 是控制模型复杂度的参数.

到这一步我们发现, 重要的是需要知道模型的参数个数的确切意思, 因为对于一些稍微复杂的模型, 参数个数 $d$ 并不是显然可以看出的, 因此我们需要在一些场合定义*参数的有效个数* (effective number of parameters).

对于线性模型

$$\hat{\boldsymbol{y}} = \boldsymbol{S}\boldsymbol{y},$$

我们定义其参数的有效个数为 $\mathrm{df}(\boldsymbol{S}) = \mathrm{tr}(\boldsymbol{S})$. 这里 df 是 degrees of freedom 的意思, 也就是有效的自由度.

### 12.3 交叉验证

交叉验证是一种十分简单的直接估计样本外误差 Err 的方法.

# 13 半监督学习

# A 关于线性代数

## A.1 分块正定矩阵的逆

**引理 A.1 (分块正定矩阵的逆)** *分块正定矩阵的逆可以表示为*

$$\begin{pmatrix} \boldsymbol{A} & \boldsymbol{B} \\ \boldsymbol{B}^T & \boldsymbol{D} \end{pmatrix}^{-1} = \begin{pmatrix} \boldsymbol{A}^{-1} + \boldsymbol{A}^{-1}\boldsymbol{B}(\boldsymbol{D}-\boldsymbol{B}^T\boldsymbol{A}^{-1}\boldsymbol{B})^{-1}\boldsymbol{C}\boldsymbol{A}^{-1} & -\boldsymbol{A}^{-1}\boldsymbol{B}(\boldsymbol{D}-\boldsymbol{B}^T\boldsymbol{A}^{-1}\boldsymbol{B})^{-1} \\ -(\boldsymbol{D}-\boldsymbol{B}^T\boldsymbol{A}^{-1}\boldsymbol{B})^{-1}\boldsymbol{B}^T\boldsymbol{A}^{-1} & (\boldsymbol{D}-\boldsymbol{B}^T\boldsymbol{A}^{-1}\boldsymbol{B})^{-1} \end{pmatrix} \tag{1.1.1}$$

**证明** 只需要证明 $(\boldsymbol{D}-\boldsymbol{B}^T\boldsymbol{A}^{-1}\boldsymbol{B})$ 确实可逆即可. 这是因为原矩阵是满秩的, 因此

$$\begin{pmatrix} \boldsymbol{B} \\ \boldsymbol{D} \end{pmatrix} - \begin{pmatrix} \boldsymbol{A} \\ \boldsymbol{B}^T \end{pmatrix} \boldsymbol{A}^{-1}\boldsymbol{B} = \begin{pmatrix} \boldsymbol{0} \\ \boldsymbol{D}-\boldsymbol{B}^T\boldsymbol{A}^{-1}\boldsymbol{B} \end{pmatrix}$$

仍然是列满秩的, 从而 $(\boldsymbol{B}^T\boldsymbol{A}^{-1}\boldsymbol{B})$ 确实可逆. □

## A.2 二阶优化方法

考虑凸的目标函数 $f$, 在 $f(\boldsymbol{\theta}_0)$ 处进行二阶展开

$$f(\boldsymbol{\theta}) = f(\boldsymbol{\theta}_0) + (\boldsymbol{\theta}-\boldsymbol{\theta}_0)^T\nabla_{\boldsymbol{\theta}} f(\boldsymbol{\theta}_0) + \frac{1}{2}(\boldsymbol{\theta}-\boldsymbol{\theta}_0)^T\boldsymbol{H}(\boldsymbol{\theta}-\boldsymbol{\theta}_0),$$

假设海森矩阵 $\boldsymbol{H}$ 正定, 则直接令上式取到最小值, 即是

$$\nabla_{\boldsymbol{\theta}} f(\boldsymbol{\theta}_0) + \boldsymbol{H}(\boldsymbol{\theta}-\boldsymbol{\theta}_0) = 0,$$

从而得到参数更新

$$\boldsymbol{\theta} = \boldsymbol{\theta}_0 - \boldsymbol{H}^{-1}\nabla_{\boldsymbol{\theta}} f(\boldsymbol{\theta}_0).$$

# B 关于概率论

## B.1 分布的尾估计

对于**非负**随机变量 $Y$, 显然有

$$YI_{\{Y\geq t\}} \geq tI_{\{Y\geq t\}},$$

两边取期望就得到所谓的**马尔可夫不等式** (Markov inequality)

$$P\{Y \geq t\} \leq \frac{\mathbb{E}[YI_{\{Y\geq t\}}]}{t} \leq \frac{\mathbb{E}[Y]}{t}. \tag{2.1.1}$$

更一般地, 对任意的在正半轴非负单调非减函数 $\phi(x)$, 则有

$$P\{Y \geq \phi(t)\} \leq \frac{\mathbb{E}[Y]}{\phi(t)},$$

特别地, 取 $\phi(t) = t^2$ 可得**切比雪夫不等式**.