基础数理统计

第十三章 线性回归和 Logistic 回归

1. 13.1 简单线性回归
2. 13.2 最小二乘和极大似然
3. 13.3 最小二乘估计的性质
4. 13.4 预测
5. 13.5 多元回归
6. 13.6 模型选择
7. 13.7 Logistic 回归

监督学习 (Supervised learning)

例 1

垃圾邮件分类: 样本 $(x_{i1}, \cdots, x_{ip}, y_i)$, 其中 $y_i \in \{0, 1\}$. 对于这一类问题，给定训练数据集 $(\boldsymbol{x}_i, y_i), i = 1, \cdots, n$, 希望构建合适的 ($p$ 维) 函数 $f(\cdot)$, 使得

$$y_i \approx f(\boldsymbol{x}_i).$$

应用场景：

- 预测：下一次只要知道解释变量 $\boldsymbol{x}$，就可以很好的做出预测 $f(\boldsymbol{x})$;
- 解释：通过构建出的模型，找出解释变量是如何影响被解释变量的;
- 实验设计：通过模型可以反过来对收集哪些特征给出更好的建议。

回归

- 回归：研究响应变量 Y 和协变量 (也称预测变量或特征)$\boldsymbol{X}$ 关系的方法
- 总结 $\boldsymbol{X}$ 和 Y 的关系的一种方法是通过回归函数

$$r(\boldsymbol{x}) = \mathrm{E}(Y|\boldsymbol{X} = \boldsymbol{x}) = \int y f(y|\boldsymbol{x}) dy.$$

- 目标：用形如 $(Y_1, \boldsymbol{X}_1), \ldots, (Y_n, \boldsymbol{X}_n) \sim F_{\boldsymbol{X},Y}$ 的数据估计回归函数 $r(\boldsymbol{x})$。

回归分析

统计学中，“回归” 一词来源于英国科学家 Francis Galton 在研究父辈身高和子女成年身高关系时候最先提出。一般来说，

- 父母身高越高，孩子身高也很高
- 父母身高不高，孩子身高也不高
- 高的没有父母那么高，偏矮的也没有父母那么矮

生物学家称为” 回归” 现象，也是统计中回归分析的来源。

例如，假设父亲身高 X 与儿子身高 Y 满足如下的关系：

$$(Y - 175) \approx 0.7 * (X - 175) \tag{1}$$

那么我们有

- $X = 180,\ Y = 178.5;$
- $X = 170,\ Y = 171.5.$

非监督学习 (Unsupervised learning)

还有一类问题是，训练数据集格式为 $x_i, i = 1, \cdots, n$, 没有特定的被解释变量，例如

- 聚类：从不同的视角对样本进行归类，找到样本背后的结构.
- 特征提取：提取出数据中的核心特征。

13.1 简单线性回归

简单线性回归模型

给定数据 (x_i, y_i)，$i = 1, 2, \ldots, n$。

定义 1 (简单线性回归模型)

对于自变量 x，因变量 y，一元线性回归模型为：

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i. \ (i = 1, 2, \ldots, n)$$

其中，ϵ_i 为误差项，满足

$$E(\epsilon_i | x_i) = 0, \ V(\epsilon_i | x_i) = \sigma^2.$$

- 误差并不是真的错误或差，可以理解为可能影响 y 但未考虑进模型的各种因素随机影响；
- 误差的引入可以让模型具有更好的解释性和泛化能力；
- 误差作为一个随机变量，假定均值为 0，方差的大小表示模型的可解释性大小。

令 $\widehat{\beta}_0$，$\widehat{\beta}_1$ 为 β_0，β_1 的估计，拟合曲线为

$$\widehat{r}(x) = \widehat{\beta}_0 + \widehat{\beta}_1 x,$$

预测值或拟合值为 $\widehat{y}_i = \widehat{r}(x_i)$，残差定义为

$$\widehat{\epsilon}_i = y_i - \widehat{y}_i.$$

对于任意的回归系数 β_0, β_1, 考虑残差平方和 (RSS)

$$\mathrm{RSS} = \sum_{i=1}^{n} \widehat{\epsilon}_i^2 = \sum_{i=1}^{n} (y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_i)^2.$$

定义 2

最小二乘估计是使得 *RSS* 最小的 $\widehat{\beta}_0$, $\widehat{\beta}_1$ 值，即

$$(\widehat{\beta}_0, \widehat{\beta}_1) = \arg\min_{\beta_0, \beta_1} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2.$$

令

$$\boldsymbol{y}=\begin{pmatrix} y_1 \\ y_2 \\ \cdot \\ \cdot \\ \cdot \\ y_n \end{pmatrix},\ X=\begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \cdot & \cdot \\ \cdot & \cdot \\ \cdot & \cdot \\ 1 & x_n \end{pmatrix},\ \boldsymbol{\beta}=\begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}.$$

则 $\sum_{i=1}^{n}(y_i-\beta_0-\beta_1 x_i)^2=\|\boldsymbol{y}-X\boldsymbol{\beta}\|_2^2$，故当 $X^\top X$ 可逆时，

$$\begin{pmatrix} \widehat{\beta}_0 \\ \widehat{\beta}_1 \end{pmatrix}=(X^\top X)^{-1}X^\top \boldsymbol{y}=\begin{pmatrix} n & \sum\limits_{i=1}^{n} x_i \\ \sum\limits_{i=1}^{n} x_i & \sum\limits_{i=1}^{n} x_i^2 \end{pmatrix}^{-1}\begin{pmatrix} \sum\limits_{i=1}^{n} y_i \\ \sum\limits_{i=1}^{n} x_i y_i \end{pmatrix}.$$

此时

$$\mathrm{RSS}=\boldsymbol{y}^\top[I_n-X(X^\top X)^{-1}X^\top]\boldsymbol{y}.$$

记 $\bar{x} = \frac{1}{n}\sum_{i=1}^n x_i$, $\bar{y} = \frac{1}{n}\sum_{i=1}^n y_i$ 以及

$$\begin{aligned}\ell_{xy} &= \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^n x_i y_i - n\bar{x}\bar{y},\\ \ell_{xx} &= \sum_{i=1}^n (x_i - \bar{x})^2 = \sum_{i=1}^n x_i^2 - n\bar{x}^2.\end{aligned}$$

定理 1

最小二乘估计为

$$\widehat{\beta}_1 = \frac{\ell_{xy}}{\ell_{xx}},\ \widehat{\beta}_0 = \bar{y} - \widehat{\beta}_1 \bar{x}.$$

σ^2 的无偏估计为

$$\widehat{\sigma}^2 = \frac{1}{n-2}\sum_{i=1}^n \widehat{\epsilon}_i^2 = \frac{1}{n-2} RSS.$$

注意到，$A = I_n - X(X^\top X)^{-1}X^\top$ 满足 $A^2 = A$ 且 A 的秩为 $n-2$，故存在正交矩阵 Γ 使得

$$A = \Gamma \begin{pmatrix} I_{n-2} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix} \Gamma^\top,$$

令 $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_n)^\top$ 且 $\Gamma^\top \boldsymbol{\epsilon} = (\eta_1, \ldots, \eta_n)^\top = \boldsymbol{\eta}^\top$，则 $\mathsf{E}(\boldsymbol{\eta}) = \mathbf{0}$ 且 $\mathsf{V}(\eta) = \sigma^2 \boldsymbol{I}_n$.

$$\mathrm{RSS} = \boldsymbol{y}^\top [I_n - X(X^\top X)^{-1}X^\top]\boldsymbol{y} = \sum_{i=1}^{n-2} \eta_i^2.$$

故 $\mathsf{E}(\mathrm{RSS}) = (n-2)\sigma^2$.

13.2 最小二乘和极大似然

13.1 简单线性回归
13.2 最小二乘和极大似然
13.3 最小二乘估计的性质
13.4 预测
13.5 多元回归
13.6 模型选择
13.7 Logistic 回归

若假定 $\epsilon_i | x_i \sim N(0, \sigma^2)$，则似然函数为

$$
\begin{aligned}
L(\beta_0, \beta_1, \sigma^2) &= \prod_{i=1}^{n} f(x_i, y_i) = \prod_{i=1}^{n} f_x(x_i) \prod_{i=1}^{n} f_{y|x}(y_i|x_i), \\
\prod_{i=1}^{n} f_{y|x}(y_i|x_i) &\propto \sigma^{-n} \exp\left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^{n} [y_i - (\beta_0 + \beta_1 x_i)]^2 \right\}.
\end{aligned}
$$

条件对数似然函数为

$$
\ell(\beta_0, \beta_1, \sigma^2) = -n \log \sigma - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - (\beta_0 + \beta_1 x_i))^2.
$$

定理 2

在正态性的假设下，β_0, β_1 的最大似然估计即为最小二乘估计。σ^2 的最大似然估计为

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \widehat{y}_i)^2 = \frac{1}{n}\sum_{i=1}^{n}\widehat{\epsilon}_i^2.$$

13.3 最小二乘估计的性质

13.1 简单线性回归
13.2 最小二乘和极大似然
13.3 最小二乘估计的性质
13.4 预测
13.5 多元回归
13.6 模型选择
13.7 Logistic 回归

最小二乘估计的性质

定理 3

令 $\widehat{\beta} = (\widehat{\beta}_0, \widehat{\beta}_1)^\top$ 表示最小二乘估计，则

$$
\begin{aligned}
E(\widehat{\beta}|x_1, x_2, \ldots, x_n) &= \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} \\
V(\widehat{\beta}|x_1, x_2, \ldots, x_n) &= \frac{\sigma^2}{\ell_{xx}} \begin{pmatrix} n^{-1}\sum_{i=1}^n x_i^2 & -\bar{x} \\ -\bar{x} & 1 \end{pmatrix}.
\end{aligned}
$$

记 $\widehat{\text{se}}(\widehat{\beta}_0) = \dfrac{\widehat{\sigma}}{\sqrt{\ell_{xx}}}\sqrt{\dfrac{\sum_{i=1}^{n} x_i^2}{n}}$, $\widehat{\text{se}}(\widehat{\beta}_1) = \dfrac{\widehat{\sigma}}{\sqrt{\ell_{xx}}}$.

定理 4

在适当的条件下，有

1. 相合性：$\widehat{\beta}_0 \xrightarrow{P} \beta_0$，$\widehat{\beta}_1 \xrightarrow{P} \beta_1$。
2. 渐近正态性：

$$\frac{\widehat{\beta}_0 - \beta_0}{\widehat{\text{se}}(\widehat{\beta}_0)} \rightsquigarrow N(0,1), \quad \frac{\widehat{\beta}_1 - \beta_1}{\widehat{\text{se}}(\widehat{\beta}_1)} \rightsquigarrow N(0,1).$$

3. β_0 和 β_1 的 $1-\alpha$ 的渐近置信区间分别为

$$\widehat{\beta}_0 \pm z_{\alpha/2}\widehat{\text{se}}(\widehat{\beta}_0), \quad \widehat{\beta}_1 \pm z_{\alpha/2}\widehat{\text{se}}(\widehat{\beta}_1).$$

4. 检验 $H_0: \beta_1 = 0$ *vs* $H_1: \beta_1 \neq 0$ 的 *Wald* 检验为：
   如果 $|W| > z_{\alpha/2}$，则拒绝 H_0，其中 $W = \widehat{\beta}_1/\widehat{\text{se}}(\widehat{\beta}_1)$。

13.4 预测

预测区间

定理 5 (预测区间)

令

$$\widehat{\xi}_n^2 = \widehat{\sigma}^2 \left(\frac{\sum_{i=1}^{n}(x_i - x_\star)^2}{n\sum_{i=1}^{n}(x_i - \bar{x})^2} + 1 \right) = \widehat{\sigma}^2 \left(1 + \frac{1}{n} + \frac{(\bar{x} - x_\star)^2}{\ell_{xx}} \right),$$

则 $y_\star = \beta_0 + \beta_1 x_\star + \epsilon_\star$ 的 $1-\alpha$ 近似预测区间为

$$\widehat{y}_\star \pm z_{\alpha/2}\widehat{\xi}_n.$$

这里 $\widehat{y}_\star = \widehat{\beta}_0 + \widehat{\beta}_1 x_\star$。

注意到:

$$\widehat{y}_\star - y_\star = (1\ x_\star)\begin{pmatrix}\widehat{\beta}_0 - \beta_0 \\ \widehat{\beta}_1 - \beta_1\end{pmatrix} - \epsilon_\star.$$

故 $\mathrm{E}(\widehat{y}_\star - y_\star) = 0$ 且

$$\begin{aligned}\mathrm{V}(\widehat{y}_\star - y_\star) &= \sigma^2\left\{1 + (1\ x_\star)\frac{\begin{pmatrix}\sum_{i=1}^n x_i^2 & -n\bar{x} \\ -n\bar{x} & n\end{pmatrix}}{n\ell_{xx}}\begin{pmatrix}1 \\ x_\star\end{pmatrix}\right\} \\ &= \sigma^2\left(\frac{\sum_{i=1}^n (x_i - x_\star)^2}{n\sum_{i=1}^n (x_i - \bar{x})^2} + 1\right).\end{aligned}$$

作业:5(必做),10(选做)

13.5 多元回归

13.1 简单线性回归
13.2 最小二乘和极大似然
13.3 最小二乘估计的性质
13.4 预测
13.5 多元回归
13.6 模型选择
13.7 Logistic 回归

回归样本

给定样本

$$(\boldsymbol{X}_1, Y_1), \ \cdots, (\boldsymbol{X}_n, Y_n),$$

其中 $\boldsymbol{X}_i = (X_{i1}, \ldots, X_{ip})^\top, i = 1, 2, \ldots, n$ 为解释变量，$Y_1, \cdots, Y_n \in \mathbb{R}$ 为响应变量.

最小二乘估计

考虑最小二乘估计

$$\operatorname{argmin}_{\beta_0, \beta_j \in \mathbb{R}} \sum_{i=1}^{n} (Y_i - \beta_0 - \sum_{j=1}^{p} X_{ij}\beta_j)^2. \tag{2}$$

记

$$\mathbb{X} = \begin{pmatrix} 1 & X_{11} & \dots & X_{1p} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & X_{n1} & \dots & X_{np} \end{pmatrix}, \mathbb{Y} = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}, \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ . \\ . \\ . \\ \beta_p \end{pmatrix}$$

优化问题(2)用矩阵形式可以表示成

$$\operatorname{argmin}_{\boldsymbol{\beta} \in \mathbb{R}^{p+1}} (\mathbb{Y} - \mathbb{X}\boldsymbol{\beta})^\top (\mathbb{Y} - \mathbb{X}\boldsymbol{\beta}).$$

求导数，

$$\frac{\partial(\mathbb{Y}-\mathbb{X}\boldsymbol{\beta})^\top(\mathbb{Y}-\mathbb{X}\boldsymbol{\beta})}{\partial\boldsymbol{\beta}} = 2\mathbb{X}^\top(\mathbb{X}\boldsymbol{\beta}-\mathbb{Y}),$$

$$\frac{\partial^2(\mathbb{Y}-\mathbb{X}\boldsymbol{\beta})^\top(\mathbb{Y}-\mathbb{X}\boldsymbol{\beta})}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^\top} = 2\mathbb{X}^\top\mathbb{X} \geq 0.$$

当矩阵 $\mathbb{X}^\top\mathbb{X}$ 严格正定的时候，我们有最小二乘估计

$$\widehat{\boldsymbol{\beta}} = (\mathbb{X}^\top\mathbb{X})^{-1}\mathbb{X}^\top\mathbb{Y}.$$

定理 6

假设 $\mathbb{X}^\top\mathbb{X}$ 是可逆的，则

$$\begin{aligned}
\widehat{\beta} &= (\mathbb{X}^\top\mathbb{X})^{-1}\mathbb{X}^\top\mathbb{Y} \\
V(\widehat{\beta}|\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_n) &= \sigma^2(\mathbb{X}^\top\mathbb{X})^{-1} \\
\widehat{\beta} &\approx N(\boldsymbol{\beta}, \sigma^2(\mathbb{X}^\top\mathbb{X})^{-1}).
\end{aligned}$$

平方和分解公式

恒有公式

$$\sum_{i=1}^{n}(Y_i - \bar{Y})^2 = \sum_{i=1}^{n}(\widehat{Y}_i - \bar{Y})^2 + \sum_{i=1}^{n}(Y_i - \widehat{Y}_i)^2, \qquad (3)$$

其中

$$\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i, \widehat{Y}_i = \hat{\beta}_0 + \sum_{j=1}^{p} \hat{\beta}_j X_{ij}$$

这里 $\widehat{\boldsymbol{\beta}} = (\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \cdots, \hat{\beta}_p)^\top = (\mathbb{X}^\top\mathbb{X})^{-1}\mathbb{X}^\top\mathbb{Y}.$

证明.

注意到，令

$$\begin{aligned} \widehat{\mathbb{Y}} &= (\widehat{Y}_1, \ldots, \widehat{Y}_n)^\top, \\ H &= \mathbb{X}(\mathbb{X}^\top \mathbb{X})^{-1} \mathbb{X}^\top, \end{aligned}$$

$$\begin{aligned} \sum_{i=1}^n (Y_i - \bar{Y})^2 &= \sum_{i=1}^n (\widehat{Y}_i - \bar{Y})^2 + \sum_{i=1}^n (Y_i - \widehat{Y}_i)^2 \\ &\quad + 2\sum_{i=1}^n (\widehat{Y}_i - \bar{Y})(Y_i - \widehat{Y}_i) \end{aligned}$$

$$\begin{aligned} \sum_{i=1}^n (\widehat{Y}_i - \bar{Y})(Y_i - \widehat{Y}_i) &= (\mathbb{Y} - \widehat{\mathbb{Y}})^\top (\widehat{\mathbb{Y}} - \frac{1}{n} \mathbf{1}_n \mathbf{1}_n^\top \mathbb{Y}) \\ &= \mathbb{Y}^\top (\boldsymbol{I}_n - H)(H - \frac{1}{n} \mathbf{1}_n \mathbf{1}_n^\top) \mathbb{Y} = 0. \end{aligned}$$

$\square$

说明:

(1). $\sum_{i=1}^{n}(Y_i-\bar{Y})^2$ 为**总偏差平方和**，记为 l_{yy} 或 TSS(Total Sum of Squares)，体现了 Y 的观测值 $Y_1, Y_2, \cdots, Y_n$ 总波动大小。

(2). $\sum_{i=1}^{n}(\widehat{Y}_i-\bar{Y})^2$ 体现了 n 个估计值 $\widehat{Y}_1, \widehat{Y}_2, \cdots, \widehat{Y}_n$ 的波动大小（注意到 $\sum_{i=1}^{n}\widehat{Y}_i=\mathbf{1}_n^\top\widehat{\mathbb{Y}}=\mathbf{1}_n^\top H\mathbb{Y}=\mathbf{1}_n^\top\mathbb{Y}=n\bar{Y}$），它是由于 Y 与自变量之间确定有线性关系并通过自变量的变化而引起，称它为**回归平方和**，记为 U 或 MSS (Model).

(3). $\sum_{i=1}^{n}(Y_i-\widehat{Y}_i)^2=\sum_{i=1}^{n}\widehat{\epsilon}_i^2$ 称为**残差平方和**，记为 Q 或 ESS(SSE). 它是除了自变量对 Y 的线性关系之外的一切其他因素（包括自变量对 Y 的非线性关系及随机误差）引起的，也称为**剩余平方和**。

定理 7

在正态噪声的假设下，有以下结论:

(1) $\widehat{\boldsymbol{\beta}} \sim N_{p+1}(\boldsymbol{\beta}, \sigma^2(\mathbb{X}^\top\mathbb{X})^{-1})$;

(2) $\dfrac{1}{\sigma^2} Q \sim \chi^2_{n-p-1}$;

(3) $\hat{\boldsymbol{\beta}}$ 与 Q 相互独立;

(4) $H_0: \beta_1 = \beta_2 = \cdots = \beta_p = 0$ 成立时，$\dfrac{1}{\sigma^2} U \sim \chi^2_p$;

(5) $H_0: \beta_1 = \beta_2 = \cdots = \beta_p = 0$ 成立时,

$$F = \frac{U/p}{Q/(n-p-1)} = \frac{MMS(\text{模型均方})}{MSE(\text{均方误差})} \sim F(p, n-p-1).$$

其中 $p, n-p-1$ 分别称为模型的自由度和误差的自由度。

回归系数的显著性检验

对回归系数 $\beta_i(i=1,2,\cdots,p)$ 是否为 0 逐个进行检验是很必要的。即检验以下的假设:

$$H_0^{(i)}: \beta_i = 0 \quad (i=1,2,\cdots,p). \tag{4}$$

定义 3

设 U 为 $X_1, X_2, \cdots, X_p$ 对 Y 的回归平方和; $U(i)$ 为去掉 X_i 后余下的 $p-1$ 个自变量对 Y 的回归平方和。则称 $P_i = U - U(i)$ (或 $P_i = Q(i) - Q$) 为变量 x_i 的**偏回归平方和**.

偏回归平方和刻画的是某个自变量对 Y 作用大小的统计量, P_i 越大, 说明 X_i 重要。

回归模型的延伸

- 多项式回归模型:

$$Y = \alpha + \beta_1 X + \cdots + \beta_k X^k + \epsilon;$$

- 变量变换: 例如股票数据

$$Y = \alpha + \beta \log X + \epsilon \quad \text{or} \quad \log Y = \alpha + \beta X + \epsilon,$$

- 局部线性回归和局部多项式回归

$$\operatorname{argmin}_{\beta_0, \boldsymbol{\beta}} \sum_{j=1}^{n} w_i (Y_i - \beta_0 - \boldsymbol{X}_i^\top \boldsymbol{\beta})^2.$$

正态噪音

假定 $\epsilon_1, \cdots, \epsilon_n$, $i.i.d. \sim N(0, \sigma^2)$ 时候，我们有

$$Y_i \sim N(\beta_0 + \sum_{j=1}^{p} X_{ij}\beta_j, \sigma^2), \ i = 1, \ldots, n.$$

我们可以写出似然函数

$$\begin{aligned} L(\boldsymbol{\beta}, \sigma) &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp\{-\frac{1}{2\sigma^2}(Y_i - \beta_0 - \sum_{j=1}^{p} X_{ij}\beta_j)^2\} \\ &= \frac{1}{(\sqrt{2\pi})^n \sigma^n} \exp\left\{-\frac{(\mathbb{Y} - \mathbb{X}\boldsymbol{\beta})^\top (\mathbb{Y} - \mathbb{X}\boldsymbol{\beta})}{2\sigma^2}\right\} \end{aligned}$$

当噪音部分为正态分布时候，$\boldsymbol{\beta}$ 的极大似然估计等价于最小二乘估计.

指数噪音

假定噪音是对称的指数分布，即 Laplace 分布：

$$\epsilon_i \sim \mathrm{Laplace}(0, \frac{\sigma}{\sqrt{2}}) : f(x) = \frac{1}{\sqrt{2}\sigma} \exp\{-\frac{\sqrt{2}|x|}{\sigma}\},$$

这里 $\mathrm{E}\epsilon_i = 0, \mathrm{var}(\epsilon_i) = \sigma^2$, 可得极大似然估计：

$$\operatorname*{argmax}_{\boldsymbol{\beta}\in\mathbb{R}^{p+1},\sigma^2} \frac{1}{(\sqrt{2}\sigma)^n} \exp\{-\frac{\sqrt{2}}{\sigma}\|\mathbb{Y} - \mathbb{X}^\top \boldsymbol{\beta}\|_1\}$$

以及

$$\text{最小一乘法:}\ \ \widehat{\beta} = \operatorname{argmin} \sum_{i=1}^{n} |Y_i - \beta_0 - \sum_{j=1}^{p} X_{ij}\beta_j|.$$

13.6 模型选择

一般来说,

- 协变量太少导致偏差很高，称为拟合不足;
- 协变量太多导致方差很高，称为过拟合。

模型选择中有两个问题:

(i) 给每个模型指定一个 “得分”，它在某种意义上衡量模型的好坏;

(ii) 在所有模型中找出得分最好的一个。

Mallows C_p

C_p 统计量达到最小。假设 $\widehat{y}_{ip}, \mathrm{SSE}_p$ 为含 p 个变量的模型预测和 SSE。定义

$$
\begin{aligned}
J_p &= \frac{1}{\sigma^2}\sum_{i=1}^{n}(\widehat{y}_{ip} - \mathrm{E}(y_i))^2 \\
\mathrm{E}(J_p) &= \frac{\mathrm{E}(\mathrm{SSE}_p)}{\sigma^2} - n + 2(p+1) \\
C_p &= \frac{\mathrm{SSE}_p}{\hat{\sigma}^2} - n + 2p \\
&= (n-m-1)\frac{\mathrm{SSE}_p}{\mathrm{SSE}_m} - n + 2p.
\end{aligned}
$$

这里 $\hat{\sigma}^2 = \mathrm{SSE}_m/(n-m-1)$，为全模型中 σ^2 的无偏估计。

AIC 与 BIC 准则

- AIC (Akaike information criterion): 设模型的似然函数为 $L(\boldsymbol{\theta}, \boldsymbol{x})$，$\boldsymbol{\theta}$ 的维数为 p，$\boldsymbol{x}$ 为随机样本，则 AIC 定义为
$$\text{AIC} = -2\ln L(\widehat{\boldsymbol{\theta}}_L, \boldsymbol{x}) + 2p.$$
- BIC (SBC: Schwartz's Bayesian criterion): BIC 定义为
$$\text{BIC} = -2\ln L(\widehat{\boldsymbol{\theta}}_L, \boldsymbol{x}) + p\ln n.$$

交叉验证

$$\widehat{R}_{\mathrm{CV}} = \sum_{i=1}^{n} (Y_i - \widehat{Y}_{(i)})^2$$

$\widehat{Y}_{(i)}$ 是把 Y_i 删去后拟合的模型对 Y_i 的预测值。

13.7 Logistic 回归

1 13.1 简单线性回归
2 13.2 最小二乘和极大似然
3 13.3 最小二乘估计的性质
4 13.4 预测
5 13.5 多元回归
6 13.6 模型选择
7 13.7 Logistic 回归

分类

- **classification**(分类);
- **supervised learning**(监督学习);
- **discrimination**(判别分析);
- **pattern recognition**(模式识别).

例子

例 2

Iris Data Set(鸢尾属植物数据集) 是历史最悠久的数据集，它首次出现在著名的英国统计学家和生物学家 *Ronald Fisher 1936* 年的论文中。在这个数据集中，包括了三类不同的鸢尾属植物：*Setosa*，*Versicolour*，*Virginica*。每类收集了 *50* 个样本，整个数据集一共包含了 *150* 个样本。该数据集测量了所有 *150* 个样本的 *4* 个特征，分别是：

- *sepal length*（花萼长度）
- *sepal width*（花萼宽度）
- *petal length*（花瓣长度）
- *petal width*（花瓣宽度）

线性回归做分类

对于分类问题，如果预测值也是 0 或者 1，那么

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \widehat{Y}_i)^2 = \frac{1}{n}\sum_{i=1}^{n} I(Y_i \neq \widehat{Y}_i),$$

恰好就是分类问题中错分的比例，即错分率。

对于每个 β, 我们设定预测为 $\widehat{Y} = I(X^\top \beta > 0)$,

$$
\begin{aligned}
MSE &= \frac{1}{n}\sum_{i=1}^{n} I(Y_i \neq I(X_i^\top \beta > 0)) \\
&= \frac{1}{n}\sum_{i=1}^{n} I((2Y_i - 1)X_i^\top \beta \leq 0),
\end{aligned}
$$

这里 $2Y_i - 1$ 把原来的 $\{0, 1\}$ 转化为了 $\{+1, -1\}$. 类似于最小二乘法，可以考虑

$$
\widehat{\beta} = \text{argmin}\frac{1}{n}\sum_{i=1}^{n} I((2Y_i - 1)X_i^\top \beta \leq 0).
$$

缺点

- 这里的损失函数不连续，优化问题很难求解.
- 因为示性函数的特点，优化问题可能具有多个解，不易解释.
- 统计性质也很难分析。

回归函数

假定 $Y \in \{0,1\}$, 即有两个类别；在均方损失下，最优的回归函数为:

$$r(\boldsymbol{x}) = E\{Y|X=\boldsymbol{x}\} = P(Y=1|X=\boldsymbol{x}) \tag{5}$$

分类问题的回归函数对应的是一个 0-1 之间的数，反映了因变量取 1 的概率大小。如果我们考虑线性函数，可以假设存在一个 β, 使得

$$r(\boldsymbol{x}) \propto \boldsymbol{x}^\top \boldsymbol{\beta}.$$

Logit 模型

定义一个连接函数 (link function)$f\colon \mathbb{R} \to [0,1]$, 从解释性以及计算角度，还期待：

- 函数是单调增的;
- 函数是光滑连续的;
- 函数是常用的;
- $\cdots$

逻辑回归采用的是

$$r(\boldsymbol{X}) = \frac{1}{1+e^{-\boldsymbol{X}^\top \beta}} = \frac{e^{\boldsymbol{X}^\top \beta}}{1+e^{\boldsymbol{X}^\top \beta}}.$$

逻辑回归-极大似然估计

把 Y_i 生成机制设定为概率为 $r(\boldsymbol{X}_i)$ 的二项分布，写出似然函数

$$L(\beta) = \prod_{i=1}^{n} r(\boldsymbol{X}_i)^{Y_i}(1 - r(\boldsymbol{X}_i))^{1-Y_i} = \prod_{i=1}^{n} \frac{e^{Y_i \boldsymbol{X}_i^\top \beta}}{1 + e^{\boldsymbol{X}_i^\top \beta}},$$

逻辑回归估计为

$$\begin{aligned}
\widehat{\beta} &= \underset{\beta}{\operatorname{argmax}} \prod_{i=1}^{n} \frac{e^{Y_i \boldsymbol{X}_i^\top \beta}}{1 + e^{\boldsymbol{X}_i^\top \beta}} \\
&= \underset{\beta}{\operatorname{argmax}} \frac{1}{n} \sum_{i=1}^{n} \{Y_i \boldsymbol{X}_i^\top \beta - \log(1 + e^{\boldsymbol{X}_i^\top \beta})\} \\
&= \operatorname{argmin}_{\beta} \frac{1}{n} \sum_{i=1}^{n} \{\log(1 + e^{\boldsymbol{X}_i^\top \beta}) - Y_i \boldsymbol{X}_i^\top \beta\}.
\end{aligned}$$

逻辑回归的算法

逻辑回归的求解过程没有显示解，一般通过优化算法迭代的过程来完成。对优化问题

$$\widehat{\beta} = \mathrm{argmin} f(\beta).$$

如果 $f(\cdot)$ 是连续可导的, 常用的优化算法有

- Gradient descent
- Newton's method

Gradient descent

对于优化问题 $\mathrm{argmin} f(\boldsymbol{x})$, 给定一个初始点 $\boldsymbol{x}_s$, Gradient descent 对函数做一个二阶逼近:

$$f(\boldsymbol{x}) \approx f(\boldsymbol{x}_s) + \nabla f(\boldsymbol{x}_s)(\boldsymbol{x} - \boldsymbol{x}_s) + \frac{1}{2\gamma}\|\boldsymbol{x} - \boldsymbol{x}_s\|_2^2,$$

迭代过程为:

$$\begin{aligned} \boldsymbol{x}_{s+1} &= \mathrm{argmin}_{\boldsymbol{x}}\{f(\boldsymbol{x}_s) + \nabla f(\boldsymbol{x}_s)(\boldsymbol{x} - \boldsymbol{x}_s) + \frac{1}{2\gamma}\|\boldsymbol{x} - \boldsymbol{x}_s\|_2^2\} \\ &= \boldsymbol{x}_s - \gamma \nabla f(\boldsymbol{x}_s). \end{aligned}$$

Newton's method

给定一个初始点 $\boldsymbol{x}_s$, Newton's method 对函数做一个二阶 Taylor 展开:

$$f(\boldsymbol{x}) \approx f(\boldsymbol{x}_s) + \nabla f(\boldsymbol{x}_s)(\boldsymbol{x} - \boldsymbol{x}_s) + \frac{1}{2}(\boldsymbol{x} - \boldsymbol{x}_s)^\top \nabla^2 f(\boldsymbol{x}_s)(\boldsymbol{x} - \boldsymbol{x}_s),$$

迭代过程为:

$$\begin{aligned}\boldsymbol{x}_{s+1} &= \operatorname{argmin}_{\boldsymbol{x}} \{ f(\boldsymbol{x}_s) + \nabla f(\boldsymbol{x}_s)(\boldsymbol{x} - \boldsymbol{x}_s) + \frac{1}{2}(\boldsymbol{x} - \boldsymbol{x}_s)^\top \nabla^2 f(\boldsymbol{x}_s)(\boldsymbol{x} - \boldsymbol{x}_s) \} \\ &= \boldsymbol{x}_s - (\nabla^2 f(\boldsymbol{x}_s))^{-1} \nabla f(\boldsymbol{x}_s).\end{aligned}$$

数学直观上，Newton 法是二阶方法，梯度下降是一阶方法。前者有更好的逼近，算法收敛上会较快，缺点是需要求解二阶 Hessian 矩阵相关线性方程组，计算量上要大一点。

Newton's method

Newton's method 常见形式是求解方程 $g(\boldsymbol{x}) = 0$，对于优化问题

$$\mathrm{argmin} f(\boldsymbol{x}),$$

Newton's method 考虑的问题为

$$\nabla f(\boldsymbol{x}) = 0$$

迭代过程为:

$$\boldsymbol{x}_{s+1} = \boldsymbol{x}_s - (\nabla^2 f(\boldsymbol{x}_s))^{-1} \nabla f(\boldsymbol{x}_s),$$

其中 $\nabla^2 f(\boldsymbol{x})$ 为函数的 Hessian 矩阵, 记 $\boldsymbol{x} = (X_1, \ldots, X_p)'$

$$\nabla^2 f(\boldsymbol{x}) = (\frac{\partial^2 f(\boldsymbol{x})}{\partial X_i \partial X_j})_{p \times p}.$$

令

$$p_i = \frac{e^{\boldsymbol{X}_i^\top \boldsymbol{\beta}}}{1 + e^{\boldsymbol{X}_i^\top \boldsymbol{\beta}}}, \quad i = 1, 2, \ldots, n.$$

对于 Logistic 回归，Hessian 矩阵为

$$H = \mathbb{X}^\top W \mathbb{X},$$

这里

$$\mathbb{X} = \begin{pmatrix} \boldsymbol{X}_1^\top \\ \boldsymbol{X}_2^\top \\ \vdots \\ \boldsymbol{X}_n^\top \end{pmatrix},$$

W 是一个对角矩阵，它的 (i, i) 对角元素为 $p_i(1 - p_i)$.

重加权最小二乘算法

选择初始值 $\widehat{\beta}^0$，计算 p_i^0。令 $s=0$ 并循环迭代下面的步骤直至收敛。

- 令 $\boldsymbol{Z}=(Z_1,Z_2,\ldots,Z_n)^\top$，这里
$$Z_i=\log\left(\frac{p_i^s}{1-p_i^s}\right)+\frac{Y_i-p_i^s}{p_i^s(1-p_i^s)},\quad i=1,2,\ldots,n.$$
- 令对角矩阵 W 的 (i,i) 对角元素为 $p_i^s(1-p_i^s)$.
$$\widehat{\beta}^s=(\mathbb{X}^\top W\mathbb{X})^{-1}\mathbb{X}^\top W\boldsymbol{Z},$$
- 令 $s=s+1$ 并回到第一步。