基础数理统计

第十三章 线性回归和 Logistic 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

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

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

简单线性回归模型

定义 1 (简单线性回归模型)

对于自变量 X, 应变量 Y, 一元线性回归模型为:

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i.$$

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)

$$\text{RSS} = \sum_{i=1}^{n} \widehat{\epsilon}_i^2 = \sum_{i=1}^{n} (Y_i - \widehat{\beta}_0 - \widehat{\beta}_1 X_i)^2.$$

定义 2

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

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

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

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

定理 1

最小二乘估计为

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

σ^2 的无偏估计为

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

13.2 最小二乘和极大似然

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

$$L(\beta_0, \beta_1, \sigma^2) = \prod_{i=1}^n f(X_i, Y_i) = \prod_{i=1}^n f_X(X_i) \prod_{i=1}^n f_{Y|X}(Y_i|X_i),$$

$$\prod_{i=1}^n f_{Y|X}(Y_i|X_i) \propto \sigma^{-n} \exp\left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^n [y_i - (\beta_0 + \beta_1 X_i)]^2 \right\}.$$

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

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

13.3 最小二乘估计的性质

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

$$
\begin{aligned}
E(\widehat{\beta}|X_1, X_2, \ldots, X_n) &= \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} \\
V(\widehat{\beta}|X_1, X_2, \ldots, X_n) &= \frac{\sigma^2}{n s_X^2} \begin{pmatrix} \frac{1}{n}\sum_{i=1}^n X_i^2 & -\bar{X}_n \\ -\bar{X}_n & 1 \end{pmatrix}.
\end{aligned}
$$

其中

$$
s_X^2 = \frac{1}{n}\sum_{i=1}^n (X_i - \bar{X}_n)^2, \ \bar{X}_n = \frac{1}{n}\sum_{i=1}^n X_i.
$$

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

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：
$$\frac{\widehat{\beta}_0 - \beta_0}{\widehat{\text{se}}(\widehat{\beta}_0)} \rightsquigarrow N(0,1),\ \frac{\widehat{\beta}_1 - \beta_1}{\widehat{\text{se}}(\widehat{\beta}_1)} \rightsquigarrow N(0,1).$$
3. β_0 和 β_1 的 $1-\alpha$ 的渐近置信区间分别为
$$\widehat{\beta}_0 \pm z_{\alpha/2}\widehat{\text{se}}(\widehat{\beta}_0),\ \ \widehat{\beta}_1 \pm z_{\alpha/2}\widehat{\text{se}}(\widehat{\beta}_1).$$
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

$$\hat{\xi}_n^2 = \hat{\sigma}^2 \left(\frac{\sum_{i=1}^n (X_i - X_\star)^2}{\sum_{i=1}^n (X_i - \bar{X})^2} + 1 \right),$$

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

$$\hat{Y}_\star \pm z_{\alpha/2} \hat{\xi}_n.$$

这里 $\hat{Y}_\star = \hat{\beta}_0 + \hat{\beta}_1 X_\star$。

作业：5，10

13.5 多元回归

回归样本

给定样本

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

其中 $\boldsymbol{X}_1, \cdots, \boldsymbol{X}_n \in \mathbb{R}^p$ 为解释变量， $Y_1, \cdots, Y_n \in \mathbb{R}$ 为响应变量.

最小二乘估计

把样本 (X_i, Y_i) 看成 $\mathbb{R}^{p+1}$ 空间中点，考虑最小二乘估计

$$\operatorname{argmin}_{\beta \in \mathbb{R}^p} \sum_{i=1}^{n} (Y_i - X_i^T \beta)^2. \tag{2}$$

记

$$\mathbb{X} = \begin{pmatrix} \boldsymbol{X}_1^\intercal \\ \vdots \\ \boldsymbol{X}_n^\intercal \end{pmatrix}_{n\times p} = \begin{pmatrix} X_{11} & \dots & X_{1p} \\ \vdots & \vdots & \vdots \\ X_{n1} & \dots & X_{np} \end{pmatrix}, \mathbb{Y} = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix},$$

用矩阵形式优化问题(2)可以表示成

$$\operatorname{argmin}_{\beta\in\mathbb{R}^p}(\mathbb{Y}-\mathbb{X}\beta)^\intercal(\mathbb{Y}-\mathbb{X}\beta).$$

注意

实际使用中，我们会人为的设定数据矩阵的第一列即 $X_{11},\dots,X_{n1}$ 为 1，这样可以自动的把常数项包括进模型.

求导数，

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

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

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

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

思考

$\mathbb{X}^{\intercal}\mathbb{X}$ 不严格正定会发生什么情况?

定理 6

假设 $\mathbb{X} \to \mathbb{X}$ 是可逆的，则

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

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

$$\operatorname{argmin}_\beta \sum_{j=1}^{n} w_i (Y_i - X_i^T \beta)^2.$$

正态噪音

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

$$Y_i \sim N(X_i^T \beta, \sigma^2), \quad i = 1, \ldots, n.$$

我们可以写出似然函数

$$L(\beta, \sigma) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp\{-\frac{1}{2\sigma^2}(Y_i - X_i^T \beta)^2\},$$

- 这里我们把 σ^2 也当未知参数放进了模型.
- 似然函数中逻辑上应该是 $f(x, y, \beta, \sigma) = f(y|x)f(x)$，我们去除了对参数没有影响的 $f(x)$ 部分.

极大似然估计为

$$\arg\max_{\beta\in\mathbb{R}^p,\sigma^2} \frac{1}{(\sqrt{2\pi}\sigma)^n}\exp\{-\frac{1}{2\sigma^2}(\mathbb{Y}-\mathbb{X}\beta)^\top(\mathbb{Y}-\mathbb{X}\beta)\},$$

如果只关注回归系数 β 部分，

$$\widehat{\beta} = \arg\min_{\beta\in\mathbb{R}^p}(\mathbb{Y}-\mathbb{X}\beta)^\top(\mathbb{Y}-\mathbb{X}\beta).$$

当噪音部分为正态分布时候，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：

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

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

$$\arg \max_{\beta \in \mathbb{R}^p, \sigma^2} \frac{1}{(\sqrt{2}\sigma)^n} \exp\{-\frac{\sqrt{2}}{\sigma} \|\mathbb{Y} - \mathbb{X}'\beta\|_1\}$$

以及

$$\text{最小一乘法:} \quad \widehat{\beta} = \text{argmin} \sum_{i=1}^{n} |Y_i - X_i'\beta|.$$

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。定义

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

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

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

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

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 $\widehat{Y} = I(X^{\mathsf{T}}\beta > 0)$,

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

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

$$\widehat{\beta} = \arg\min \frac{1}{n}\sum_{i=1}^{n} I((2Y_i - 1)X_i^{\mathsf{T}}\beta \leq 0).$$

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：

$$r(x) = E\{Y|X = x\} = P(Y = 1|X = x) \tag{3}$$

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

$$r(x) \propto X^{\mathrm{T}}\beta.$$

Logit 模型

定义一个连接函数 (link function)$f : \mathbb{R} \to [0,1]$, 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

$$r(x) = \frac{1}{1+e^{-X^{\mathsf{T}}\beta}} = \frac{e^{X^{\mathsf{T}}\beta}}{1+e^{X^{\mathsf{T}}\beta}}.$$

逻辑回归-极大似然估计

把 Y_i 生成机制设定为概率为 $r(X_i)$ 的二项分布，写出似然函数

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

逻辑回归估计为

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

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

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

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。