第三章 广义线性模型

- 在前两章我们分别介绍了线性回归与逻辑回归
  - 线性回归问题符合正态分布 $y \mid x;\theta \sim \mathcal{N}\left(\mu, \sigma^2\right)$
  - 逻辑回归问题符合伯努利分布 $y \mid x;\theta \sim \text{Bernoulli}(\phi)$
- 实际上这些模型都是一个更为广泛的模型族的特例，这个模型族被称为**广义线性模型**（Generalized Linear Models）

指数族

- 为了引出广义线性模型，我们首先需要介绍**指数族分布**
- 如果一个分布可以被表示成如下形式，我们就称其属于指数分布族：

$$p(y;\eta) = b(y)\exp\left(\eta^T T(y) - a(\eta)\right) \tag{1}$$

  - η 被称为分布的**自然参数**（或者称为**典范参数**）
  - $T(y)$ 被称为**充分统计量**，通常 $T(y) = y$
  - $a(\eta)$ 被称为**对数分割函数**
  - $e^{-a(\eta)}$ 本质上是一个归一化常数，确保概率 $p(y;\eta)$ 和为 1
- 当选定 T, a, b 时，我们就得到了一种以 η 为参数的分布
- 下面我们来证明伯努利和高斯分布属于指数分布族

伯努利分布的证明

- 伯努利分布可以表示为：

$$\begin{aligned}
p(y;\phi) &= \phi^y(1-\phi)^{1-y} \\
&= \exp(y\log\phi + (1-y)\log(1-\phi)) \\
&= \exp\left(\left(\log\left(\frac{\phi}{1-\phi}\right)\right)y + \log(1-\phi)\right)
\end{aligned}$$

  - 自然参数 $\eta = \log\left(\frac{\phi}{1-\phi}\right)$（这里自然参数不是向量，所以其转置不变）
    - 从该式可以导出 $\phi = \frac{1}{1+e^{-\eta}}$，这正是我们熟悉的 sigmoid 函数！
    - 之后我们推导逻辑回归是广义线性模型时会再提到这个
  - 现在，我们可以得到：

$$
\begin{aligned}
T(y) &= y \\
a(\eta) &= -\log(1-\phi) \\
&= \log(1+e^\eta) \\
b(y) &= 1
\end{aligned}
$$

- 这表明通过设定适当的 T, a, b ，伯努利分布可以写成等式 (1) 的形式，即其属于指数族分布

正态分布的证明

- 之前我们推导线性回归时得出了 σ 的值对 θ 的选择没有影响，所以为了简化推导，这里设定 $\sigma^2=1$，于是我们有：

$$
\begin{aligned}
p(y;\mu) &= \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{1}{2}(y-\mu)^2\right) \\
&= \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{1}{2}y^2\right)\cdot\exp\left(\mu y-\frac{1}{2}\mu^2\right)
\end{aligned}
$$

  - 因此，通过如下选择，我们可以证明高斯分布属于指数族分布：

$$
\begin{aligned}
\eta &= \mu \\
T(y) &= y \\
a(\eta) &= \mu^2/2 \\
&= \eta^2 \\
b(y) &= (1/\sqrt{2\pi})\exp(-y^2/2)
\end{aligned}
$$

- 其实，还有许多其他的分布属于指数族，比如多项式分布、泊松分布、伽马分布等

构建广义线性模型

- 首先，广义线性模型的构建需要基于以下三条假设：
  1. $y \mid x;\theta$ 符合以 η 为参数的指数族分布
  2. 给定 x，我们的目标是预测 $T(y)$ 的理想值，而在大多数的案例中，$T(y)=y$
     - 这意味着我们的假设 h 应该满足 $h(x)=E[y \mid x]$（可以从期望的定义上来进行理解，即反映随机变量平均取值的大小）
  3. 自然参数 η 和 输入 x 满足线性关系 $\eta=\theta^T x$（如果 η 是向量，那么 $\eta_i=\theta_i^T x$ ）
- 基于上面三条假设，我们就可以利用广义线性模型来优雅地解决问题
- 下面，我们将用广义线性模型来推导线性回归和逻辑回归的假设函数，并引出 softmax 回归

线性回归

- 线性回归的目标变量（在 GLM 术语集中也称为**反应变量**）满足**高斯分布**：$y \mid x;\theta \sim \mathcal{N}\left(\mu,\sigma^2\right)$
  - 这里 μ 与 x 相关
- 根据之前推导的结果，我们有：

$$\begin{aligned} h_\theta(x) &= E[y|x;\theta] \\ &= \mu \\ &= \eta \\ &= \theta^T x \end{aligned}$$

  - 第一个等式来源于假设 2
  - 第二个等式是高斯分布的性质
  - 第三个等式是之前推导过高斯分布属于指数族分布的条件
  - 最后一个等式则来源于假设 3

逻辑回归

- 逻辑回归的反映变量满足**伯努利分布**：$y \mid x;\theta \sim \text{Bernoulli}(\phi)$
  - 之前我们在证明伯努利分布属于指数族分布时已经推导出了$\phi = \frac{1}{1+e^{-\eta}}$
  - 因此，与线性回归类似，我们有：

$$\begin{aligned} h_\theta(x) &= E[y|x;\theta] \\ &= \phi \\ &= \frac{1}{1+e^{-\eta}} \\ &= \frac{1}{1+e^{-\theta^T x}} \end{aligned}$$

    - 上式证明了为什么逻辑回归的假设函数是 sigmod 函数
      - 当反应变量满足伯努利分布时，这是广义线性模型的定义导出的结果
- 此外，我们将表示分布均值（期望）与自然参数 η 关系的函数 $g(\eta) = E[T(y);\eta]$ 称为**正则响应函数**（canonical response function）
  - 将其反函数称为**正则关联函数**（canonical link function）
  - 因此，高斯分布的正则响应函数即为其本身，伯努利分布的正则响应函数即为逻辑函数

softmax 回归

- 如果对于分类问题，y 可以取 k 个值（$k > 2$），那么这就是一个多元分类问题
  - 此时反应变量的条件概率分布模型为**多项分布**
- 下面让我们推导出多项分布数据的广义线性模型
  - 在这之前，需要首先将多项式分布表示为指数族分布
- 假设多项式分布有 k 个输出，一般我们应该定义 k 个参数 $\phi_1, \ldots \phi_k$ 来表示每个输出的概率，但这其实存在冗余，因为第 k 个输出的概率可以用其他 $k-1$ 个输出的概率来表示（概率之和必定为 1 ）
  - 因此，我们只定义 $k-1$ 个参数 $\phi_1, \ldots \phi_{k-1}$ ，其中 $\phi_i = p(y=i;\phi)$，则 $\phi_k = 1 - \sum_{i=1}^{k-1} \phi_i$，
    - 注意其并不是一个参数，而是由 $\phi_1, \ldots \phi_{k-1}$ 确定的
- 为了将多项分布表示为指数族分布，我们首先定义 $T(y) \in \mathbb{R}^{k-1}$ 如下：

$$T(1)=\begin{bmatrix}1\\0\\0\\\vdots\\0\end{bmatrix},T(2)=\begin{bmatrix}0\\1\\0\\\vdots\\0\end{bmatrix},T(3)=\begin{bmatrix}0\\0\\1\\\vdots\\0\end{bmatrix},\cdots T(k-1)=\begin{bmatrix}0\\0\\0\\\vdots\\1\end{bmatrix},T(k)=\begin{bmatrix}0\\0\\0\\\vdots\\0\end{bmatrix}$$

  - 与之前不同，$T(y)$ 与 y 并不相等，$T(y)$ 是一个 $k-1$ 维的向量而非一个实数
  - 我们将用 $(T(y))_i$ 来表示向量 $T(y)$ 的第 i 个元素
- 下面我们将再介绍一个有用的操作符：$1\{\cdot\}$，其运算法则为：

$$1\{\text{True}\}=1,1\{\text{False}\}=0$$

  - 例如：$1\{2=3\}=0,1\{3=5-2\}=1$
- 因此，我们可以得到如下等式：

$$(T(y))_i=1\{y=i\}$$

  - 即只有当 $y=i$ 时，第 i 个元素才为 1，其他都为 0
  - 进一步可以得到：

$$\mathrm{E}[(T(y))_i]=P(y=i)=\phi_i$$

    - 因为求期望时，只有当 $y=i$ 时，乘积不为0
- 基于上述结论，我们可以将多项分布表示为指数族分布：

$$\begin{aligned}
p(y;\phi)&=\phi_1^{1\{y=1\}}\phi_2^{1\{y=2\}}\cdots\phi_k^{1\{y=k\}}\\
&=\phi_1^{1\{y=1\}}\phi_2^{1\{y=2\}}\cdots\phi_k^{1-\sum_{i=1}^{k-1}1\{y=i\}}\\
&=\phi_1^{(T(y))_1}\phi_2^{(T(y))_2}\cdots\phi_k^{1-\sum_{i=1}^{k-1}(T(y))_i}\\
&=\exp\left((T(y))_1\log(\phi_1)+(T(y))_2\log(\phi_2)+\cdots+\left(1-\sum_{i=1}^{k-1}(T(y))_i\right)\log(\phi_k)\right)\\
&=\exp\left((T(y))_1\log(\phi_1/\phi_k)+(T(y))_2\log(\phi_2/\phi_k)+\cdots+(T(y))_{k-1}\log(\phi_{k-1}/\phi_k)+\log(\phi_k)\right)\\
&=b(y)\exp\left(\eta^TT(y)-a(\eta)\right)
\end{aligned}$$

  - 其中：

$$\begin{aligned}
\eta&=\begin{bmatrix}\log(\phi_1/\phi_k)\\\log(\phi_1/\phi_k)\\\vdots\\\log(\phi_{k-1}/\phi_k)\end{bmatrix}\\
a(\eta)&=-\log(\phi_k)\\
b(y)&=1
\end{aligned}$$

- 上述推导表明了多项分布属于指数族分布，并得到了**关联函数**如下（前面已经证明了期望值即为 ϕ_i ）：

$$\eta_i = \log \frac{\phi_i}{\phi_k}$$

- 类似地，我们定义 $\eta_k = \log(\phi_k / \phi_k) = 0$。下面我们将推导出**响应函数**：

$$\begin{aligned} e^{\eta_i} &= \frac{\phi_i}{\phi_k} \\ \phi_k e^{\eta_i} &= \phi_i \\ \phi_k \sum_{i=1}^{k} e^{\eta_i} &= \sum_{i=1}^{k} \phi_i = 1 \end{aligned} \tag{2}$$

  - 这表明 $\phi_k = 1/\sum_{i=1}^{k} e^{\eta_i}$，将其代回（2）式，即可得到响应函数为：

$$\phi_i = \frac{e^{\eta_i}}{\sum_{j=1}^{k} e^{\eta_j}}$$

    - 这个将 η 映射到 ϕ 的函数又被称为 **softmax**（柔性最大值）函数
- 根据之前的假设 3，我们有 $\eta_i = \theta_i^T x \ (i = 1, \ldots, k-1)$，其中 $\phi_1, \ldots, \phi_{k-1} \in \mathbb{R}^{n+1}$
  - 为了方便，我们定义$\theta_k = 0$，这样 $\eta_k = \theta_k^T x = 0$，因此，我们的模型给出 y 的条件分布如下：

$$\begin{aligned} p(y = i \mid x; \theta) &= \phi_i \\ &= \frac{e^{\eta_i}}{\sum_{j=1}^{k} e^{\eta_j}} \\ &= \frac{e^{\theta_i^T x}}{\sum_{j=1}^{k} e^{\theta_j^T x}} \end{aligned}$$

  - 这个模型可以应用于多元分类问题 $y \in \{1, \ldots k\}$，被称为 **softmax 回归**，它是逻辑回归的推广

- 综上，我们的假设函数为：

$$
\begin{aligned}
h_\theta(x) &= E[T(y) \mid x;\theta] \\
&= E\left[\begin{array}{c|}
1\{y=1\} \\
1\{y=2\} \\
\vdots \\
1\{y=k-1\}
\end{array}\; x;\theta\right] \\
&= \begin{bmatrix} \phi_1 \\ \phi_2 \\ \vdots \\ \phi_{k-1} \end{bmatrix} \\
&= \begin{bmatrix} \frac{e^{\theta_1^T x}}{\sum_{j=1}^k e^{\theta_j^T x}} \\ \frac{e^{\theta_2^T x}}{\sum_{j=1}^k e^{\theta_j^T x}} \\ \vdots \\ \frac{e^{\theta_{k-1}^T x}}{\sum_{j=1}^k e^{\theta_j^T x}} \end{bmatrix}
\end{aligned}
$$

- 该假设函数给出了 y 取每个可能的值的条件概率（$i=1,\ldots,k$）
  - 其中 $p(y=k \mid x;\theta)$ 由 $1-\sum_{i=1}^{k-1}\phi_i$ 得到

- 最后，我们来讨论 softmax 回归的参数拟合：
  - 与之前类似，如果我们有一个训练集 $\{(x^{(i)},y^{(i)});i=1,\ldots,m\}$，希望学习出这个模型的参数 θ_i，我们首先会给出其对数似然函数：

$$
\begin{aligned}
\ell(\theta) &= \sum_{i=1}^m \log p(y^{(i)} \mid x^{(i)};\theta) \\
&= \sum_{i=1}^m \log \prod_{l=1}^k \left(\frac{e^{\theta_l^T x^{(i)}}}{\sum_{j=1}^k e^{\theta_j^T x^{(i)}}}\right)^{1\{y^{(i)}=l\}}
\end{aligned}
$$

    - 下面我们就可以通过最大似然分析求出参数 θ，使用梯度上升或牛顿方法