Generalized Linear Model

广义线性模型

开始之前

- 目前为止，我们已经介绍了线性回归与logistic回归
- 两个有监督学习方法：前者用于回归，后者用于分类
  - 回归：$y|\boldsymbol{x};\boldsymbol{\beta} \sim N(\mu,\sigma^2),\ \mu = \boldsymbol{\beta}^T\boldsymbol{x}$
  - 分类：$y|\boldsymbol{x};\boldsymbol{\beta} \sim Bernoulli(\phi),\ \phi = \frac{1}{1+e^{-\beta^T x}}$
- 我们将会看到，这些方法都是一类方法的特例：广义线性模型(generalized linear model, GLM)
- 我们还会看到其他的GLM方法是怎么导出并被应用到回归和分类问题当中

Outline

- 指数族分布 (Exponential Family of Distributions)
- 广义线性模型(Generalized Linear Model)

指数族分布

- 我们称一类分布在指数族分布中，如果其概率密度函数可以表示成：

$$p(y;\eta) = \underline{b(y)} \exp\left(\eta^T \underline{T(y)} - \underline{a(\eta)}\right) \rightarrow b(y) \frac{\exp(\eta^T T(y))}{\exp(a(\eta))}$$

  - y：变量。我们将在有监督学习框架下用指数族分布对响应变量建模
  - η: 自然参数（natural parameter）

    $\log(\exp(a(\eta))) = a(\eta)$
  - $T(y)$：充分统计量（sufficient statistic），大多数情况下$T(y) = y$
  - $b(y)$：base measure
  - $a(\eta)$：log-partition function，归一化参数
- 给定$T(y)$，$b(y)$，$a(\eta)$的形式，我们就定义了一族被η参数化的指数族分布
- 判断是否属于指数族分布最常用的做法就是写出一个分布的密度函数，做一些“algebra messaging”使其符合指数族分布的形式

正态分布（固定方差）

不失一般性

- $N(\mu,\sigma^2)$常被用来建模连续数据，假设$\sigma^2=1$：

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

提出二次项　　　y的线性组合

- $\eta=\mu$
- $T(y)=y$
- $b(y)=\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{1}{2}y^2\right)$
- $a(\eta)=\frac{1}{2}\mu^2=\frac{1}{2}\eta^2$

伯努利分布

- Bernoulli(ϕ)常被用来建模二分类数据

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

- $\eta = \log\left(\frac{\phi}{1-\phi}\right) \Rightarrow \phi = \frac{1}{1+e^{-\eta}}$。巧了，sigmoid函数 / logistic 函数
- $T(y) = y$
- $b(y) = 1$
- $a(\eta) = -\log(1-\phi) = \log(1+e^{\eta})$
- 注意上述两个分布自然参数和规范参数（canonical parameter）之间的关系

η $a(\eta)$

练习（15MIN）

- 判断泊松分布和指数分布是否属于指数族分布。若是，给出各个元素的构成

- 泊松分布：$p(y;\lambda) = \frac{\lambda^y}{y!}e^{-\lambda} = \frac{1}{y!}\exp(y\log\lambda - \lambda)$

  很多分布 $T(y)=y$

- $\eta = \log(\lambda)$; $T(y) = y$; $b(y) = \frac{1}{y!}$; $a(\eta) = \lambda = e^{\eta}$

- 指数分布： $p(y;\lambda) = \lambda e^{-\lambda y} = \exp(-\lambda y + \log\lambda)$

- $\eta = -\lambda$; $T(y) = y$; $b(y) = 1$; $a(\eta) = -\log(\lambda) = -\log(-\eta)$

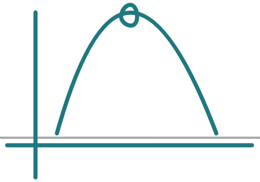
$\frac{\lambda^Y}{y!}e^{-\lambda}$ $\quad \eta = \log\lambda \quad \lambda = e^{\eta} \quad a(\eta) = e^{\lambda}$

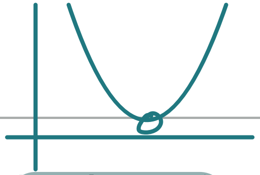
$\searrow \exp(Y \log\lambda - \lambda) \cdot \frac{1}{Y!}$

$\lambda e^{-\lambda Y} \rightarrow 1 \cdot \exp(-\lambda Y + \log\lambda)$

$\eta = -\lambda \quad \lambda = -\eta \quad -\log\lambda = -\log(-\eta)$

一些好的性质





- 对η求极大似然估计，一定是一个凹问题；即负对数似然作为损失函数一定是个凸函数
- $E(y;\eta) = \frac{\partial}{\partial \eta} a(\eta)$
- $Var(y;\eta) = \frac{\partial^2}{\partial \eta^2} a(\eta)$，避免了复杂积分

$$p(y;\eta) = b(y)\exp(\eta^T T(y) - a(\eta))$$

$$= b(y)\frac{\exp(\eta^T T(y))}{\exp(a(\eta))}$$

$$\exp(a(\eta)) = \int b(y)\exp(\eta^T T(y))$$

$$\exp(a(\eta))\, a'(\eta) = \frac{d}{d\eta}\int b(y)\exp(\eta^T T(y))$$

$$= \int \frac{d}{d\eta} b(y)\exp(\eta^T T(y))$$

$$= \int b(y)\exp(\eta^T T(y))\, T(y)\, dy$$

$$a'(\eta) = \int \underbrace{b(y)\cdot\exp(\eta^T T(y) - a(\eta))}_{p(y;\eta)}\, T(y)\, dy$$

$$= E(T(y);\eta)$$

$$a''(\eta) = \int \underbrace{b(y)\exp(\eta^T T(y) - a(\eta))}_{p(y;\eta)}\,(T(y) - a'(\eta))\, T(y)\, dy$$

$$= \int p(y;\eta)\cdot T^2(y)\, dy - \int p(y;\eta)\, a'(\eta)\, T(y)\, dy$$

$$= E(T^2(y);\eta) - [E(T(y);\eta)]^2$$

广义线性模型

- 广义线性模型，顾名思义就是经典线性回归模型的一般化。
- 它是指数族分布的自然延伸，在刻画响应变量的分布时引入了协变量
- 在经典线性回归模型中，假定$E(y|\boldsymbol{x})$为$\boldsymbol{x}$的线性组合，而且要求y为连续的且服从正态分布，有固定的方差，也就是对于误差项ϵ的要求
- 很多情况下，y可能是二值或计数。而这显然不能满足正态分布的假设，这给线性回归模型带来很大的问题
- 此时需要对于经典的线性模型进行改进，即推广出应用范围更加广泛的广义线性回归模型
- 而其关键就在于选择恰当的**指数族分布**

GLM三大假设/设计选择

1. $y|\boldsymbol{x};\boldsymbol{\beta} \sim Exponential\ Family(\eta)$

2. $\eta = \boldsymbol{\beta}^T\boldsymbol{x}$, $\boldsymbol{\beta},\boldsymbol{x} \in R^{d+1}$. *如果*$\boldsymbol{\eta}$ *是一个向量，那么*$\eta_j = \boldsymbol{\beta}_j^T\boldsymbol{x}$

3. 给定一个新的$\boldsymbol{x}$，我们的目标是预测$E(T(y)|\boldsymbol{x};\boldsymbol{\beta})$

- 大部分情况下 $T(y) = y$，所以我们的输出是$E(y|\boldsymbol{x};\boldsymbol{\beta})$。这意味着我们的预测函数$f_{\boldsymbol{\beta}}(\boldsymbol{x}) = E(y|\boldsymbol{x};\boldsymbol{\beta})$
- $\boldsymbol{x} \Rightarrow \boldsymbol{\beta}^T\boldsymbol{x} \Rightarrow \eta \Rightarrow Exponential\ Family(\eta) \Rightarrow E(y;\eta) = E(y|\boldsymbol{x};\boldsymbol{\beta}) = f_{\boldsymbol{\beta}}(\boldsymbol{x})$
- 针对不同的y ，我们需要选择不同的指数族分布来构造GLM

GLM Training

- 在训练过程中不需要涉及指数族分布的任何参数，而是针对β进行训练
- 极大似然：$\max_{\beta} \log P\left(y^{(i)} | \boldsymbol{x}^{(i)}; \boldsymbol{\beta}^{T}\right)$，并利用梯度上升

参数学习方向不同

- GLM有着统一的迭代准则：

$$\beta_j := \beta_j + \alpha\left(y^{(i)} - f_{\beta}\left(\boldsymbol{x}^{(i)}\right)\right) x_j^{(i)}$$

术语

- η: Natural parameter
- $\mu = E(y;\eta) = g(\eta)$: Canonical response function
- $g(\eta) = \frac{\partial}{\partial \eta} a(\eta)$
- $\eta = g^{-1}(\mu)$: Canonical link function

三种不同的参数化

- Model parameter: $\boldsymbol{\beta}$
- Natural parameter: η
- Canonical parameter:
  - ϕ – Bernulli
  - μ, σ^2 – Normal/Gaussian
  - λ – Poisson, Exponential
- 当我们训练GLM时，只需要学习$\boldsymbol{\beta}$
- $\boldsymbol{\beta}$与η之间的关系是线性的： $\eta = \boldsymbol{\beta}^T \boldsymbol{x}$。这是我们的设计选择
- $\eta \xrightarrow{g(\cdot)} \phi, \mu, \sigma, \lambda$ $\qquad\qquad$ $\phi, \mu, \sigma, \lambda \xrightarrow{g^{-1}(\cdot)} \eta$

线性回归

- $y|\boldsymbol{x} \sim N(\mu, \sigma^2)$，其中$\mu$与$\boldsymbol{x}$有关，不失一般性假设$\sigma^2 = 1$
- 我们令$Exponential\ Family(\eta)$为正态分布$N(\mu, 1)$
- $f_{\boldsymbol{\beta}}(\boldsymbol{x}) = E(y|\boldsymbol{x}; \boldsymbol{\beta}) = \mu = \eta = \boldsymbol{\beta}^T\boldsymbol{x}$

Logistic回归

- $y|\boldsymbol{x} \sim Bernoulli(\phi)$，其中$\phi$与$\boldsymbol{x}$有关
- 我们令$Exponential\ Family(\eta)$为$Bernoulli(\phi)$
- $f_{\boldsymbol{\beta}}(\boldsymbol{x}) = E(y|\boldsymbol{x};\boldsymbol{\beta}) = \phi = \frac{1}{1+e^{-\eta}} = \frac{1}{1+e^{-\boldsymbol{\beta}^T\boldsymbol{x}}}$

  $\eta = \log\left(\frac{\phi}{1-\phi}\right)$
- $\phi = g(\eta) = \frac{1}{1+e^{-\eta}}$
- $\eta = \log\left(\frac{\phi}{1-\phi}\right) = \beta^T x$

Softmax回归

- $y|\boldsymbol{x} \sim Multi(1, \phi_1, \dots, \phi_K)$, $\phi_k \geq 0$, $\sum_{k=1}^{K} \phi_k = 1$
- 我们只利用$\phi_1, \dots, \phi_{K-1}$进行参数化，$\phi_K = 1 - \sum_{k=1}^{K-1} \phi_k$
- 定义$T(y) \in \mathbb{R}^{K-1}$:

$$T(1) = \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix}, T(2) = \begin{pmatrix} 0 \\ 1 \\ \vdots \\ 0 \end{pmatrix}, \dots, T(K-1) = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 1 \end{pmatrix}, T(K) = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix}$$

- $\left(T(y)\right)_i$表示$T(y)$的第i个元素，则$\left(T(y)\right)_i = I(y = i)$

$(T(1))_1 = T(I(1=1)) = 1$

- $E\left[\left(T(y)\right)_i\right] = P(y = i) = \phi_i$

Softmax回归

- 我们现在验证多项分布属于指数族分布

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

- $\eta = \left(\log\left(\frac{\phi_1}{\phi_K}\right), \log\left(\frac{\phi_2}{\phi_K}\right), \ldots, \log\left(\frac{\phi_{K-1}}{\phi_K}\right)\right)^T$

- $a(\eta) = -\log(\phi_K)$

- $b(y) = 1$

Softmax回归

- Link function: $\eta_k = \log\left(\frac{\phi_k}{\phi_K}\right), k = 1, \ldots, K$. 这里为了简便定义$\eta_K = \log\left(\frac{\phi_K}{\phi_K}\right) = 0$
- $e^{\eta_k} = \frac{\phi_k}{\phi_K} \implies \phi_K e^{\eta_k} = \phi_k \implies \phi_K \sum_{k=1}^{K} e^{\eta_k} = \sum_{k=1}^{K} \phi_k = 1 \implies \phi_K = \frac{1}{\sum_{k=1}^{K} e^{\eta_k}}$
- Response function: $\phi_k = \frac{e^{\eta_k}}{\sum_{k=1}^{K} e^{\eta_k}}$ the softmax function
- $\eta_k = \boldsymbol{\beta}_k^T \boldsymbol{x}, k = 1, \ldots, K$. $\beta_k \in \mathbb{R}^{d+1}, k = 1, \ldots, K-1; \beta_K = \mathbf{0}$
- $P(y = k|\boldsymbol{x}; \boldsymbol{\beta}) = \phi_k = \frac{e^{\eta_k}}{\sum_{k=1}^{K} e^{\eta_k}} = \frac{e^{\beta_k^T x}}{\sum_{k=1}^{K} e^{\beta_k^T x}}$

以上就是在GLM框架下推导出的Softmax回归

β_1

$p(y=1 \mid x; \beta) = \frac{e^{\beta_1^T x}}{e^{\beta_1^T x} + \underset{1}{\underset{\shortparallel}{e^{\beta_K^T x}}}}$

logistic 回归为其特例

Softmax回归

- $f_{\beta}(\boldsymbol{x}) = E(T(y)|\boldsymbol{x};\boldsymbol{\beta}) = E\left(\begin{matrix} I(y=1) \\ I(y=2) \\ \cdots \\ I(y=K-1) \end{matrix} |\boldsymbol{x};\boldsymbol{\beta}\right) = \left(\begin{matrix} \phi_1 \\ \phi_2 \\ \cdots \\ \phi_{K-1} \end{matrix}\right) = \left(\begin{matrix} \frac{e^{\boldsymbol{\beta}_1^T \boldsymbol{x}}}{\sum_{k=1}^{K} e^{\boldsymbol{\beta}_k^T \boldsymbol{x}}} \\ \frac{e^{\boldsymbol{\beta}_2^T \boldsymbol{x}}}{\sum_{k=1}^{K} e^{\boldsymbol{\beta}_k^T \boldsymbol{x}}} \\ \cdots \\ \frac{e^{\boldsymbol{\beta}_{K-1}^T \boldsymbol{x}}}{\sum_{k=1}^{K} e^{\boldsymbol{\beta}_k^T \boldsymbol{x}}} \end{matrix}\right)$

- $P(y = K|\boldsymbol{x};\boldsymbol{\beta}) = 1 - \sum_{k=1}^{K-1} \phi_k$

练习（15MIN）

- 已知泊松分布属于指数族分布，试推导泊松回归，写出预测函数、对数似然函数、梯度上升法则（以随机梯度上升为例）
- $y|\boldsymbol{x} \sim Poisson(\lambda)$, λ与$\boldsymbol{x}$有关
- $f_{\boldsymbol{\beta}}(\boldsymbol{x}) = E(y|\boldsymbol{x};\boldsymbol{\beta}) = \lambda = e^{\eta} = e^{\boldsymbol{\beta}^T\boldsymbol{x}}$
- 对数似然$l(\boldsymbol{\beta}) = \sum_{i=1}^{n}\left(y^{(i)}\log e^{\boldsymbol{\beta}^T\boldsymbol{x}^{(i)}} - e^{\boldsymbol{\beta}^T\boldsymbol{x}^{(i)}}\right)$
- 梯度上升：$\beta_j \coloneqq \beta_j + \alpha\left(y^{(i)} - e^{\boldsymbol{\beta}^T\boldsymbol{x}^{(i)}}\right)x_j^{(i)}$

$P(y;\eta) = \frac{1}{y!}\exp(y\log\lambda - \lambda)$

预测函数

$a(\eta) = e^{\eta}$

$f_\beta(x) = E(y|x;\beta) = e^{\eta} = e^{\beta^T x}$

对数似然

似然 $p(y^{(i)}|x^{(i)};\beta^T) =$

$\ln p(y^{(i)}|x^{(i)};\beta^T)$

梯度上升 $\beta_j := \beta_j + \alpha(y^{(i)} - e^{\beta^T x})\, x_j^{(i)}$

glm(参数)