# Generalized Linear Models

## 引子
- 对P[y|x,theta]进行建模的两种方法
  - 高斯分布函数->Least Squares
  - Bernoulli->Logistic Regression
- 最小二乘和逻辑回归都是广义线性模型得到的特例

## 指数族分布exponential family distributions

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

- 如果一个分布能用此公式表示，则该分布属于指数分布
- 公式参数解释
  - η叫做自然参数(natural parameter)或者典范参数(canonical parameter)
  - T(y)叫做充分统计量(sufficient statistic) — 通常T(y) = y
  - a(η)叫做对数分割函数(log partition function)
  - $e^{-a(\eta)}$ 起到归一化常数的作用(normalization constant)，确保p(y; η)的和为1
- 高斯（正态）分布和伯努利分布都是指数族分布

## Bernoulli(φ) distribution

- y ∈ [0, 1]
  - 伯努利分布，也叫01分布
  - p(y = 1; φ) = φ
  - p(y = 0; φ) = 1−φ
- 换种写法

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

- 与指数族公式做对比，便可得出 η = log（φ / (1 − φ)）以及

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

- 这就说明伯努利分布是指数族分布
- 有趣的是，翻转η = log（φ / (1 − φ)），得到 φ = 1/（1 + e−η），刚好是sigmoid函数

## Gaussian distribution

- 高斯分布，也叫正态分布
- 注：在推导线性回归的时候，σ2 的值对我们最终选择的 θ 和 hθ(x) 都没有影响。所以我们可以给 σ2 取一个任意值。为了简化推导过程，就令σ2 = 1
- 正态分布的公式只要学过概率肯定都会写，如下

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

- 在指数族公式下的取值

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

- 这就说明高斯分布也是指数族分布

## 其他指数族分布
- 多项式分布(multinomial)
- 泊松分布(Poisson) — 用于对计数类问题进行建模
- γ分布(gamma)和指数分布(exponential) — 用于对连续的、非负的随机变量进行建模
- β和狄利克雷分布（the beta and the Dirichlet） — 用于概率

## Constructing GLMs

### 三个假设/设计决策
- $y \mid x;\theta \sim \text{ExponentialFamily}(\eta).$ — 给定x和θ，y服从参数为η的指数族分布
- $T(y) = y.$ — 给定 x，目的是要预测对应这个给定 x 的 T(y) 的期望值。咱们的例子中绝大部分情况都是 T(y) = y，这也就意味着我们的学习得到 h 输出的预测值 h(x) 要满足 h(x) = E[y|x]。(E[y|x]应该就是对给定x时的y值的期望的意思)
- $\eta = \theta^T x.$ — 自然参数 η 和输入值 x 是线性相关的

### Ordinary Least Squares
- 普通最小二乘法，由广义线性模型中的高斯分布推导而来

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

- 以高斯分布建模，得到
- 公式解释
  - 因为高斯分布期望为μ，所以第一步得到第二步
  - 因为高斯公式中μ = η，所以得到第三步
  - 由第三个假设，又得到第四步

### Logistic Regression
- 逻辑回归，二值化问题，由广义线性模型中的伯努利分布推导而来

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

- 以伯努利分布建模，得到
- 公式解释
  - if y|x; θ ∼ Bernoulli(φ), then E[y|x; θ] = φ (伯努利分布的性质)，所以得到第二步
  - φ = 1/(1 + e−η)，在上面证明伯努利分布是指数族分布时证明过的，由假设1得到第三步
  - 由假设3得到第四步
  - 该公式便是逻辑回归的形式
- 两个术语
  - 正则响应函数（canonical response function） — $(g(\eta) = \text{E}[T(y);\eta])$
  - 正则链接函数（canonical link function） — $g^{-1}$

### Softmax Regression
- 由广义线性模型中的多项式分布推导而来，逻辑回归只能解决二类值问题，而它可以解决k类值问题y ∈ [1, 2, ..., k]
- 用指数族分布来描述多项式分布

**参数化**

- ❶ 确定参数个数
  - 要对k个不同输出值的多项式进行参数化，可以用k 个参数 φ1,...,φk 来对应各自输出值的概率。
  - 因为 $\sum_{i=1}^k \phi_i = 1$，所以我们可以只用k-1个参数对多项式进行参数化
  - $\phi_i = p(y = i;\phi),$ and $p(y = k;\phi) = 1 - \sum_{i=1}^{k-1}\phi_i,$ 即 $\phi_k = 1 - \sum_{i=1}^{k-1}\phi_i$

- ❷ 确定T(y)
  - define $T(y) \in \mathbb{R}^{k-1}$ as follows:
  $$T(1)=\begin{bmatrix}1\\0\\0\\\vdots\\0\end{bmatrix},T(2)=\begin{bmatrix}0\\1\\0\\\vdots\\0\end{bmatrix},T(3)=\begin{bmatrix}0\\0\\1\\\vdots\\0\end{bmatrix},\cdots,T(k-1)=\begin{bmatrix}0\\0\\0\\\vdots\\1\end{bmatrix},T(k)=\begin{bmatrix}0\\0\\0\\\vdots\\0\end{bmatrix},$$
  - 这里T(y)不再是y，而是一个k-1维的向量
  - T(y)中的第i个元素可以写成 $(T(y))_i$
  - 一种记号
    - 指示函数（indicator function）1{·}
    - 如果参数为真，则等于1；反之则等于0（1{True} = 1, 1{False} = 0)，例如 1{2 = 3} = 0, 而1{3 = 5 − 2} = 1.
  - 所以，$(T(y))_i = 1\{y = i\}.$

- ❸ 所以，$\text{E}[(T(y))_i] = P(y = i) = \phi_i.$

- ❹ 准备就绪，把多项式分布写成指数族分布

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

$$\eta = \begin{bmatrix}\log(\phi_1/\phi_k)\\\log(\phi_2/\phi_k)\\\vdots\\\log(\phi_{k-1}/\phi_k)\end{bmatrix},$$
$$a(\eta) = -\log(\phi_k)$$
其中 $b(y) = 1.$

- ❺ 得出链接函数和响应函数
  - 链接函数 $\eta_i = \log\frac{\phi_i}{\phi_k}$ — 补充 $\eta_k = \log(\phi_k/\phi_k) = 0.$
  - 由公式
  $$\begin{aligned}e^{\eta_i} &= \frac{\phi_i}{\phi_k}\\ \phi_k e^{\eta_i} &= \phi_i\\ \phi_k\sum_{i=1}^k e^{\eta_i} &= \sum_{i=1}^k \phi_i = 1\end{aligned}$$
  - 响应函数 $\phi_i = \frac{e^{\eta_i}}{\sum_{j=1}^k e^{\eta_j}}$ — 这个从η 映射到φ的函数，就称为Softmax函数

- ❻ 根据假设3，得出最终softmax回归模型
  $$\begin{aligned}p(y = i|x;\theta) &= \phi_i\\ &= \frac{e^{\eta_i}}{\sum_{j=1}^k e^{\eta_j}}\\ &= \frac{e^{\theta_i^T x}}{\sum_{j=1}^k e^{\theta_j^T x}}\end{aligned}$$

- 根据上述各种假设和推导可得到我们的假设h(x)
  $$\begin{aligned}h_\theta(x) &= \text{E}[T(y)|x;\theta]\\ &= \text{E}\left[\begin{array}{c}1\{y=1\}\\1\{y=2\}\\\vdots\\1\{y=k-1\}\end{array}\Big| x;\theta\right]\\ &= \begin{bmatrix}\phi_1\\\phi_2\\\vdots\\\phi_{k-1}\end{bmatrix}\\ &= \begin{bmatrix}\frac{\exp(\theta_1^T x)}{\sum_{j=1}^k \exp(\theta_j^T x)}\\\frac{\exp(\theta_2^T x)}{\sum_{j=1}^k \exp(\theta_j^T x)}\\\vdots\\\frac{\exp(\theta_{k-1}^T x)}{\sum_{j=1}^k \exp(\theta_j^T x)}\end{bmatrix}.\end{aligned}$$