Chapter 9

Generalized linear models and the exponential family

曾宏生

目录

9.1 Introduction

- 很多分布都是指数家族分布
- 很多模型都是广义线性模型
- 为什么要构建通用的分布和模型？即这章的目的是?

9.2 The exponential family

指数家族为什么重要?

- 在一定正则条件下，指数家族是唯一具有有限数量充分统计量的分布家族，意味着可以将数据压缩成有限大小的表示，但不会损失信息。

充分统计量：将样本加工为统计量时，信息无损失，则称此统计量为充分统计量。数学化表达:

- 随机样本 $X = \{X_1, \ldots, X_n\}$ 是从分布 $f(x|\theta)$ 产生的,这个分布依赖于未知参数 θ, 而我们需要从这随机样本中估计参数 θ
- 统计量：样本观测的任意实值函数 $T(x) = r(X_1, \ldots, X_n)$，例如 $E(X), max(X), median(X)$
- 充分统计量：吸收了样本关于 θ 的所有有效信息，即 $p(\theta|X) = p(\theta|T(x))$
- 例如，对于高斯分布 $\mathcal{N}(\mu, \sigma^2)$ 所生成的样本 X ，其充分统计量是 $(E(X), S^2)$
  其实，在很多问题上，$\hat{\theta}$ 的 MLE 就是一个充分统计量

9.2 The exponential family

指数家族为什么重要?

- 指数家族是一个共轭先验存在的家族（原文不严谨，说是唯一一个，但有一些非指数家族其实也存在共轭先验）
- 在满足用户给定的约束条件下，指数家族是做出最少假设的分布。
- 指数家族是广义线性模型的核心。
- 指数家族是变分推断的核心。

9.2 The exponential family

指数家族定义

对于 $x = (x_1, \ldots, x_m) \in \mathcal{X}^m$ 和 $\theta \in \Theta \subseteq \mathbb{R}^d$，分布的 pdf 或 pmf 满足以下形式的即为指数家族：

$$p(x|\theta) = \frac{1}{Z(\theta)} h(x) exp[\theta^T \phi(x)]$$
$$= h(x) exp[\theta^T \phi(x) - A(\theta)]$$

其中，$Z(\theta) = \int_{\mathcal{X}_m} h(x) exp[\theta^T \phi(x)] dx$ 是归一化因子，$A(\theta) = \log Z(\theta)$

- θ 被称为自然参数 (natural parameters) 或正则参数 (canonical parameters)
- $\phi(x) \in \mathbb{R}^d$ 被称为充分统计量的一个向量
- $Z(\theta)$ 被称为归一化函数(partition function)
- $A(\theta)$ 被称为对数归一化函数(log partition function)，或累积量函数 (cumulant function)
- h(x)是一个缩放常量，通常为1

9.2 The exponential family

指数家族定义

更一般化的写法:

$$p(x|\theta) = h(x)exp[\eta(\theta)^T \phi(x) - A(\eta(\theta))]$$

其中 η 是将参数 θ 映射成自然参数 $\eta = \eta(\theta)$

- 如果$dim(\theta) < dim(\eta(\theta))$，意味着充分统计量个数比参数多，这被称为曲指数家族(curved exponential family)
- 如果$dim(\theta) = dim(\eta(\theta))$，称为正则形式(canonical form), 在这本书里默认模型都是正则形式。

9.2 The exponential family

指数家族例子

- 伯努利分布

  对于 $x \in \{0, 1\}$,可以写成如下指数家族形式:

$$\begin{aligned}\mathrm{Ber}(x|\mu) &= \mu^x(1-\mu)^{(1-x)} = \exp[x\log(\mu) + (1-x)\log(1-\mu)] \\ &= \exp[\phi(x)^T\theta]\end{aligned}$$

  其中$\phi(x) = [\mathbb{I}(x=0), \mathbb{I}(x=1)]$和$\theta = [\log(\mu), \log(1-\mu)]$

但是这种表示方式是过于完备的(over-complete)，因为特征之间是线性相关的： $\mathbb{I}(x=0) + \mathbb{I}(x=1) = 1$

9.2 The exponential family

指数家族例子

- 伯努利分布

所以可以将伯努利分布写成最小形式:

$$\mathrm{Ber}(x|\mu) = (1-\mu)\exp\left[x\log\left(\frac{\mu}{1-\mu}\right)\right]$$

所以有$\phi(x) = x, \theta = \log(\frac{\mu}{1-\mu})$ 和 $Z = 1/(1-\mu)$

$A(\theta) = \log Z(\theta) = \log(1+e^{\theta})$

可以从自然参数 θ 还原均值参数 μ:

$$\mu = sigm(\theta) = \frac{1}{1+e^{-\theta}}$$

9.2 The exponential family

怎么得到最小的指数家族表达式?

扩展:

一般指数家族的表达式中的充分统计量向量 $\phi(x)$ 是最小充分统计量向量时，即可得到最小的指数家族表达形式。

最小充分统计量: 对于最小充分统计量 $S = T(X)$, 对于其他充分统计量 $T'(X)$，$T(X)$ 是 $T'(X)$ 的一个函数。

一般，参数的MLE估计就是最小充分统计量

9.2 The exponential family

指数家族例子

- 单变量高斯分布

  单变量高斯分布可以写成如下指数家族形式:

$$\begin{aligned}\mathcal{N}(x|\mu,\sigma^2) &= \frac{1}{(2\pi\sigma^2)^{\frac{1}{2}}}\exp[-\frac{1}{2\sigma^2}(x-\mu)^2] \\ &= \frac{1}{(2\pi\sigma^2)^{\frac{1}{2}}}\exp[-\frac{1}{2\sigma^2}x^2+\frac{\mu}{\sigma^2}x-\frac{1}{2\sigma^2}\mu^2] \\ &= \frac{1}{Z(\theta)}\exp(\theta^T\phi(x))\end{aligned}$$

  其中:

$\theta = [\frac{\mu}{\sigma^2}, \frac{-1}{2\sigma^2}]^T$

$\phi(x) = [x, x^2]^T$

$Z(\mu,\sigma^2) = \sqrt{2\pi}\sigma\exp[\frac{\mu^2}{2\sigma^2}]$

$A(\theta) = \frac{-\theta_1^2}{4\theta_2} - \frac{1}{2}\log(-2\theta_2) - \frac{1}{2}\log(2\pi)$

9.2 The exponential family

对数归一化函数

指数家族的一个重要性质就是对数归一化函数 $A(\theta)$ 的导数可以用来生成充分统计量的累积量 (cumulants)，所以 $A(\theta)$ 也被称为累积量函数。

补充说明

一个分布的第一和第二累计量分布是充分统计量的均值 $\mathbb{E}[X]$ 和方差 $\mathbf{var}[X]$，而第一和第二矩 (moments) 是他们的均值 $\mathbb{E}[X]$ 和 $\mathbb{E}[X^2]$

9.2 The exponential family

对数归一化函数

- $A(\theta)$ 的一阶导

$$
\begin{aligned}
\frac{dA}{d\theta} &= \frac{d}{d\theta}\left(log \int \exp(\theta\phi(x))h(x)dx\right) \\
&= \frac{\frac{d}{d\theta}\int \exp(\theta\phi(x))h(x)dx}{\int \exp(\theta\phi(x))h(x)dx} \quad \text{\#链式求导, 先对log求导} \\
&= \frac{\int \phi(x)\exp(\theta\phi(x))h(x)dx}{\exp(A(\theta))} \quad \text{\#分子对}\theta\text{求导} \\
&= \int \phi(x)\exp(\theta\phi(x) - A(\theta))h(x)dx \\
&= \int \phi(x)p(x)dx \\
&= \mathbb{E}[\phi(x)]
\end{aligned}
$$

9.2 The exponential family

对数归一化函数

- $A(\theta)$ 的二阶导

$$\begin{aligned}\frac{d^2A}{d\theta^2} &= \frac{d}{d\theta}\int \phi(x)\exp(\theta\phi(x)-A(\theta))h(x)dx \\ &= \int \phi(x)\exp(\theta\phi(x)-A(\theta))h(x)(\phi(x)-A^{'}(\theta))dx \\ &= \int \phi(x)p(x)(\phi(x)-A^{'}(\theta))dx \\ &= \int \phi^2(x)p(x)dx - A^{'}(\theta)\int \phi(x)p(x)dx \\ &= \mathbb{E}[\phi^2(x)]-\mathbb{E}[\phi(x)]^2 \\ &= \mathrm{var}[\phi(x)]\end{aligned}$$

- 多变量情况下
  可以推导出 $\nabla^2 A(\theta) = \mathrm{cov}[\phi(x)]$
  因为协方差矩阵是正定矩阵，所以 $A(\theta)$ 是凸函数

9.2 The exponential family

对数归一化函数

- 例子：伯努利分布
  从上面可知，$A(\theta) = 1 + e^{\theta}$
  而伯努利分布的充分统计量向量$\phi(x) = x$
  所以均值 $\mathbb{E}(x) = \frac{dA}{d\theta} = \frac{e^{\theta}}{1+e^{\theta}} = \frac{1}{1+e^{-\theta}} = \text{sigm}(\theta) = \mu$
  方差 $\text{var}(x) = \frac{d}{d\theta}(1 + e^{-\theta})^{-1} = (1 + e^{-\theta})^{-2}e^{-\theta} = (1 - \mu)\mu$

9.2 The exponential family

对数归一化函数

- 例子：伯努利分布
  从上面可知，$A(\theta) = 1 + e^{\theta}$
  而伯努利分布的充分统计量向量$\phi(x) = x$
  所以均值 $\mathbb{E}(x) = \frac{dA}{d\theta} = \frac{e^{\theta}}{1+e^{\theta}} = \frac{1}{1+e^{-\theta}} = \text{sigm}(\theta) = \mu$
  方差 $\text{var}(x) = \frac{d}{d\theta}(1 + e^{-\theta})^{-1} = (1 + e^{-\theta})^{-2}e^{-\theta} = (1 - \mu)\mu$

9.2 The exponential family

指数家族的MLE

- 指数家族模型的似然：

$$p(\mathcal{D}|\theta)=\prod_{i=1}^{N}p(x_i|\theta)$$

$$=\left[\prod_{i=1}^{N}h(x_i)\right]g(\theta)^N\exp\left(\eta(\theta)^T[\sum_{i=1}^{N}\phi(x_i)]\right)$$

其中 $g(\theta)=\frac{1}{Z(\theta)}$

可以看出模型的充分统计量是 N 和

$$\phi(\mathcal{D})=[\sum_{i=1}^{N}\phi_1(x_i),\ldots,\sum_{i=1}^{N}\phi_K(x_i)]$$

其中，K 表示分布有 K 个充分统计量

例如：对于伯努利模型，$\phi=[\sum_i \mathbb{I}(x_i=1)]$，对于单变量高斯模型，$\phi=[\sum_i x_i,\sum_i x_i^2]$（同时，我们也需要知道样本大小，$N$）

9.2 The exponential family

指数家族的MLE

- 计算**正则**指数家族的模型的MLE
  对于给定 N 个服从 iid 的数据集 $\mathcal{D} = (x_1, \ldots, x_N)$ ，它的对数似然是:

$$\log p(\mathcal{D}|\theta) = \theta^T \phi(\mathcal{D}) - NA(\theta)$$

因为前面说过 $A(\theta)$ 是凸函数，所以 $-A(\theta)$ 是凹函数，$\theta^T \phi(\mathcal{D})$ 是 θ 的一个线性函数，所以对数似然是凹函数，因此，具有一个唯一的全局最大值。
所以进行求导:

$$\begin{aligned}\nabla_\theta \log p(\mathcal{D}|\theta) &= \phi(\mathcal{D}) - NA'(\theta) \\ &= \phi(\mathcal{D}) - N\mathbb{E}[\phi(X)]\end{aligned}$$

让导数等于零，求解MLE估计:

$$\mathbb{E}[\phi(X)] = \frac{1}{N}\sum_{i=1}^{N} \phi(x_i)$$

例子：伯努利模型

9.2 The exponential family

最大熵的指数家族推导

熵(Entropy): $H(P) = -\sum_x p(x)\log p(x)$

熵，指对不确定性的度量;在信息论中，熵也代表着根据信息的概率分布对信息编码所需要的最短平均编码长度。

最大熵(Maximum Entropy)原理：给定关于一个分布 P 的一些约束（先验信息），我们考虑所有满足约束条件的分布,且信息熵最大的分布 P

9.2 The exponential family

最大熵的指数家族推导

- 假设我们所知道的只是一些特征或函数的期望（先验信息）

$$\sum_x f_k(x)p(x) = F_k$$

- 转化成带约束优化问题，用拉格朗日乘子法求解

$$J(p,\lambda) = -\sum_x p(x)\log p(x) + \lambda_0(1 - \sum_x p(x)) + \sum_k \lambda_k(F_k - \sum_x p(x)f_k(x))$$

- 进行求导，并令导数为0

$$\frac{\partial J}{\partial p(x)} = -\log p(x) - 1 - \lambda_0 - \sum_k \lambda_k f_k(x)$$

$$\log p(x) = -1 - \lambda_0 - \sum_k \lambda_k f_k(x)$$

$$p(x) = \frac{1}{Z}\exp(-\sum_k \lambda_k f_k(x))$$

其中Z是归一化常数，可以看出满足最大熵的分布是指数家族的成员。

9.3 GLMs

广义线性模型定义

广义线性模型的定义如下：

- 反应(输出)变量是服从指数家族分布，即 $p(y|x;w) \sim \text{ExpFamily}(\theta)$
- 这个分布的均值参数是输入的一个线性组合，可能通过一个非线性函数，即 $\mu = \mathbb{E}(y) = \eta(w^T x)$

  备注：在cs229课程中，Ng说这里 $\mathbb{E}(y)$ 应该是充分统计量均值 $\mathbb{E}(\phi(y))$，但因为考虑到一般 $\phi(y) = y$

9.3 GLMs

构建表达式

我们先考虑标量输出的、没有限制的 GLMs 表达式:

$$p(y_i|\theta,\sigma^2) = \exp\left[\frac{y_i\theta - A(\theta)}{\sigma^2}\right] + c(y_i,\sigma^2)$$

- σ^2 是离散参数(dispersion parameter), 经常设为1
- θ 是自然参数, 即对应指数家族分布的θ
- A 是归一化函数 (partition function)，即对应指数家族分布的 $A(\theta)$
- c 是一个归一化常量

9.3 GLMs

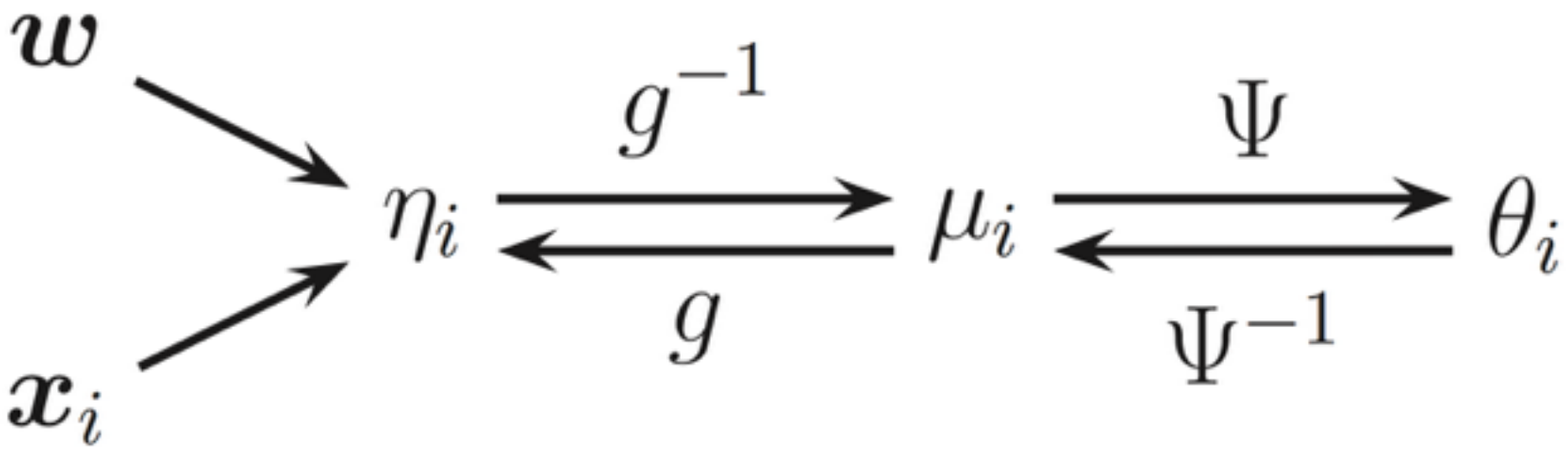
参数定义和转换

1. 用函数 ψ 将均值参数转化成自然参数，即 $\theta = \psi(\mu)$ $\mu = \psi^{-1}(\theta) = A'(\theta)$

2. 考虑输入，定义一个输入的线性函数: $\eta_i = w^T x_i$

3. 现在让分布的均值参数等于上面线性组合的某个可逆单调函数，这个函数被称为均值函数 (mean function)，写成 g^{-1},故: $u_i = g^{-1}(\eta_i) = g^{-1}(w^T x_i)$

均值函数的逆函数 $g()$，被称为联结函数 (link function): $\eta_i = g(\mu_i)$



9.3 GLMs

正则联结函数

联结函数的一个相当简单的形式就是 $g = \psi$，这被称为正则联结函数 (canonical link function)，在这种情况下，因为 $\mu = \psi^{-1}(\theta) = g^{-1}(\eta_i)$, 所以 $\theta_i = \eta_i = w^T x_i$，故GLMs 的表达式可以写成：

$$p(y_i|x_i, w, \sigma^2) = \exp\left[\frac{y_i w^T x_i - A(w^T x_i)}{\sigma^2} + c(y_i, \sigma^2)\right]$$

下面列出了一些分布的正则联结函数：

分布	**联结函数** $g(\mu)$	$\theta = \psi(\mu)$	$\mu = \psi^{-1}(\theta) = \mathbb{E}[y]$
$\mathcal{N}(\mu, \sigma^2)$	identity	$\theta = \mu$	$\mu = \theta$
$\text{Bin}(N, \mu)$	logit	$\theta = \log(\frac{\mu}{1-\mu})$	$\mu = \text{sigm}(\theta)$
$\text{Poi}(\mu)$	log	$\theta = \log(\mu)$	$\mu = e^\theta$

9.3 GLMs

输出变量的均值和方差

从指数家族部分可知:

$$\mathbb{E}[y|x_i, w, \sigma^2] = \mu_i = A^{'}(\theta_i)$$
$$\text{var}[y|x_i, w, \sigma^2] = \sigma_i^2 = A^{''}(\theta_i)\sigma^2$$

下面用具体例子进行验证:

- 对于 linear regression, 使用正则联结函数，则 $\theta_i = u_i$，所以GLMs:

$$p(y_i|x_i, w, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{\frac{1}{2}}} \exp[-\frac{1}{2\sigma^2}(y_i - \mu_i)^2]$$
$$= \exp\Big[\frac{-\frac{1}{2}y_i^2 + \mu_i y_i - \frac{1}{2}\mu_i^2}{\sigma^2} + \log\Big(\frac{1}{(2\pi\sigma^2)^{\frac{1}{2}}}\Big)\Big]$$
$$= \exp\Big[\frac{\mu_i y_i - \frac{1}{2}\mu_i^2}{\sigma^2} - \frac{1}{2}\Big(\frac{y_i^2}{\sigma^2} + \log(2\pi\sigma^2)\Big)\Big]$$

其中 $\theta_i = \mu_i = w^T x_i$

因为 $A(\theta) = \frac{1}{2}\theta^2$，所以 $\mathbb{E}[y_i] = A^{'}(\theta_i) = \mu_i$, $\text{var}[y_i] = A^{''}(\theta_i)\sigma^2 = \sigma^2$

9.3 GLMs

MLE 和 MAP 估计

- GLMs 的一个重要的属性就是可以用拟合logistic regression一样的方法去拟合GLMs
- Step1，对数似然形式：

$$\ell(w) = \log p(\mathcal{D}|w) = \frac{1}{\sigma^2}\sum_{i=1}^{N}\ell_i \qquad \ell_i \triangleq \theta_i y_i - A(\theta_i)$$

- Step2，求梯度向量

$$\nabla_w \ell(w) = \frac{1}{\sigma^2}\sum_{i=1}^{N}(y_i - \mu_i)x_i$$

- Step3，求海森矩阵

$$H = -\frac{1}{\sigma^2}\sum_{i=1}^{N}\frac{d\mu_i}{d\theta_i}x_i x_i^T = -\frac{1}{\sigma^2}X^T S X$$

其中 $S = \text{diag}(\frac{d\mu_1}{d\theta_1}, \ldots, \frac{d\mu_N}{d\theta_N})$ 是对角权重矩阵。

9.4 Probit regression

定义

在 logistic regression 中，我们构建了 $p(y=1|x_i,w)=\text{sigm}(w^T x_i)$ 的模型，其实我们可以写成 $p(y=1,x_i,w)=g^{-1}(w^T x_i)$，均值函数 g^{-1} 可以是任意函数，只要它将 $[+\infty,-\infty]$ 映射成 [0,1]，下表列举了一些可能的均值函数。

Name	Formula
Logistic	$g^{-1}(\eta)=\text{sigm}(\eta)=\frac{e^\eta}{1+e^\eta}$
Probit	$g^{-1}(\eta)=\Phi(\eta)$
Log-log	$g^{-1}(\eta)=\exp(-\exp(-\eta))$
Complementary log-log	$g^{-1}(\eta)=1-\exp(-\exp(\eta))$

在这一节，我们主要考虑 $g^{-1}(\eta)=\Phi(\eta)$ 的情况，其中 $\Phi(\eta)$ 是正态分布的cdf，这也被称为 Probit regression。probit 函数和logistic 函数很相似，如下图，但这个模型比 logistic 模型有一些更好的优势。

9.4 Probit regression

定义

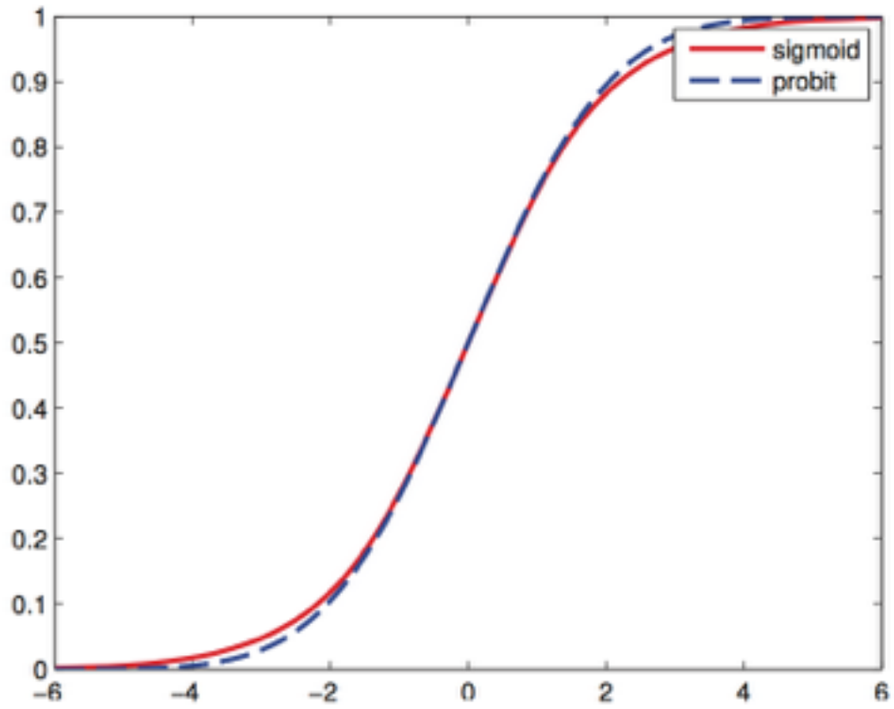
- 构建模型
  符号声明：Φ 是正态分布的cdf；ϕ 是正态分布的pdf，是 Φ 的导数
  令 $\mu_i = w^T x_i$，$\tilde{y_i} \in \{-1, 1\}$，则有：

$$p(\tilde{y} = 1 | w^T x_i) = \Phi(w^T x_i) = \Phi(\mu_i)$$

$$p(\tilde{y} = -1 | w^T x_i) = 1 - \Phi(w^T x_i) = 1 - \Phi(\mu_i) = \Phi(-\mu_i)$$

合并上面两个等式，可得：

$$p(\tilde{y} | w^T x_i) = \Phi(\tilde{y} w^T x_i) = \Phi(\tilde{y} \mu_i)$$



9.4 Probit regression

MLE/MAP 估计

- 对于一个样本的对数似然的梯度表达式是：

$$
\begin{aligned}
g_i &\triangleq \frac{d}{dw}\log p(\tilde{y}_i|w^T x_i) \\
&= \frac{d}{dw}\log\Phi(\tilde{y}_i\mu_i) \\
&= \frac{d\mu_i}{dw}\frac{d}{d\mu_i}\log\Phi(\tilde{y}_i\mu_i) \\
&= x_i\frac{\tilde{y}_i\phi(\tilde{y}_i\mu_i)}{\Phi(\tilde{y}_i\mu_i)} \qquad \#\ \phi(-\mu_i)=\phi(\mu_i) \\
&= x_i\frac{\tilde{y}_i\phi(\mu_i)}{\Phi(\tilde{y}_i\mu_i)}
\end{aligned}
$$

- 类似地，单个样本的海森矩阵如下：

$$
H_i = \frac{d}{dw^2}\log p(\tilde{y}_i|w^T x_i) = -x_i\left(\frac{\phi(\mu_i)^2}{\Phi(\tilde{y}_i\mu_i)^2} + \frac{\tilde{y}_i\mu_i\phi(\mu_i)}{\Phi(\tilde{y}_i\mu_i)}\right)x_i^T
$$

9.4 Probit regression

隐变量解读 (Latent variable interpretation)

我们可以用以下的方式解读 probit 和 logistic 模型，首先让每个样本 x_i 都关联两个潜在效用 (latent utilities)，u_{0i} 和 u_{1i}，分别对应着 $y_i = 0$ 和 $y_i = 1$ 的可能选择。然后我们假定输出选择取决于那个输出具有更大的效用 (utility)，具体如下：

$$u_{0i} \triangleq w_0^T x_i + \delta_{0i}$$

$$u_{1i} \triangleq w_1^T x_i + \delta_{1i}$$

$$y_i = \mathbb{I}(u_{1i} > u_{0i})$$

其中 δ 是误差项，这个被称为随机效能模型 (random utility model, RUM)

定义 $z_i = w_1^T x_i - w_0^T x_i + \epsilon_i$，其中 $\epsilon_i = \delta_{1i} - \delta_{0i}$

如果 δ 服从高斯分布，则 ϵ 也服从，所以有：

$$z_i \triangleq w^T x_i + \epsilon_i$$

$$\epsilon_i \sim \mathcal{N}(0, 1)$$

$$y_i = 1 = \mathbb{I}(z_i \geq 0)$$

这被称为变种RUM (difference RUM, dRUM)

9.4 Probit regression

隐变量解读 (Latent variable interpretation)

我们可以边缘化 z_i，从而得到 probit 模型：

$$
\begin{aligned}
p(y_i = 1|x_i, w) &= p(\mathbb{I}(z_i \geq 0)|x_i, w) \\
&= \int \mathbb{I}(z_i \geq 0)\mathcal{N}(z_i|w^T x_i, 1)dz_i \\
&= p(z_i \geq 0) \\
&= p(w^T x_i + \epsilon \geq 0) \\
&= p(\epsilon \geq -w^T x_i) \\
&= 1 - \Phi(-w^T x_i) \\
&= \Phi(w^T x_i)
\end{aligned}
$$

隐变量解读的意义：

如果我们对 δ 使用一个耿贝尔分布 (Gumbel distribution)，我们将会推出 ϵ 服从 logistic 分布，模型将会变成 logistic regression

9.4 Probit regression

Ordinal probit regression

Probit regressiion 的隐变量解读方式的一个好处就是可以方便地扩展到响应变量是有序的情况（即响应变量可以取 C 个离散值，这些离散值是可以用某种方式进行排序的，例如：低，中，高）。这被称为有序回归 (ordinal regression)。

- 思考：我们为什么不直接用多分类模型解决?
- 构建有序 Probit 回归模型

基本思路如下，我们引入 $C+1$ 个阈值 γ_j，然后令：

$$y_i = j \quad \text{if} \quad \gamma_{j-1} < z_i \leq \gamma_j$$

其中，$\gamma_0 \leq \cdots \leq \gamma_C$，基于易于辨识的原因，我们令

$\gamma_0 = -\infty, \gamma_1 = 0, \gamma_C = \infty$

9.5 Multi-task learning

定义

- 有时候，我们需要拟合一些相关的分类或回归问题。通常可以合理地假设这些不同的模型有一个相似的输入输出映射，所以我们可以通过同时拟合所有的参数来达到更好地效果。
- 在机器学习里，这一般称为多任务学习 (multi-task learning)，迁移学习 (transfer learning)，Learning to learn。在统计学中，这通常使用层次贝叶斯 (Hierarchical Bayesian) 进行求解

9.5 Multi-task learning

定义

具体地，假定 $\mathbb{E}[y_{ij}|x_{ij}] = g(x_{ij}^T \beta_j)$，其中 g 是 GLM 的联结函数，然后，假定 $\beta_j \sim \mathcal{N}(\beta_*, \sigma_j^2 \mathrm{I})$ 和 $\beta_* \sim \mathcal{N}(\mu, \sigma_*^2 \mathrm{I})$，在这个模型中，数据量小的群组可以借用数据量大的统计强度，因为所有 β_j 通过隐藏共同上级 β_* 而关联在一起。其中 σ_j^2 控制群组 j 依赖于共同上级的程度，σ_*^2 控制总体先验的强度。

为了简化，假定 $\mu = 0$ 且 σ_j^2 和 σ_*^2 是已知的，则整个对数概率如下：

$$\log p(\mathcal{D}|\beta) + \log p(\beta) = \sum_j \left[\log p(\mathcal{D}_j|\beta_j) - \frac{||\beta_j - \beta_*||^2}{2\sigma_j^2} \right] - \frac{||\beta_*||^2}{2\sigma_*^2}$$

9.5 Multi-task learning

运用

- 个人垃圾邮件过滤

  多任务学习的一个有趣运用是个人垃圾邮件。假定我们想要对每个用户拟合一个分类器，β_j。因为大多数用户都没有标注自己那些邮件是垃圾邮件，那些不是，所以很难独立地拟合这些模型。所以，我们让 β_j 有一个共同的先验，代表通用用户的参数。

  使用一个简单的trick来模拟上面的模型，将特征拷贝两份，一个跟用户关联，另一个不关联

  $$\mathbb{E}[y_i|x_i,\mu] = (\beta_*, w_1, \cdots, w_J)^T [x_i, \mathbb{I}(u=1)x_i, \cdots, \mathbb{I}(u=J)x_i]$$

  其中 u 是用户id.

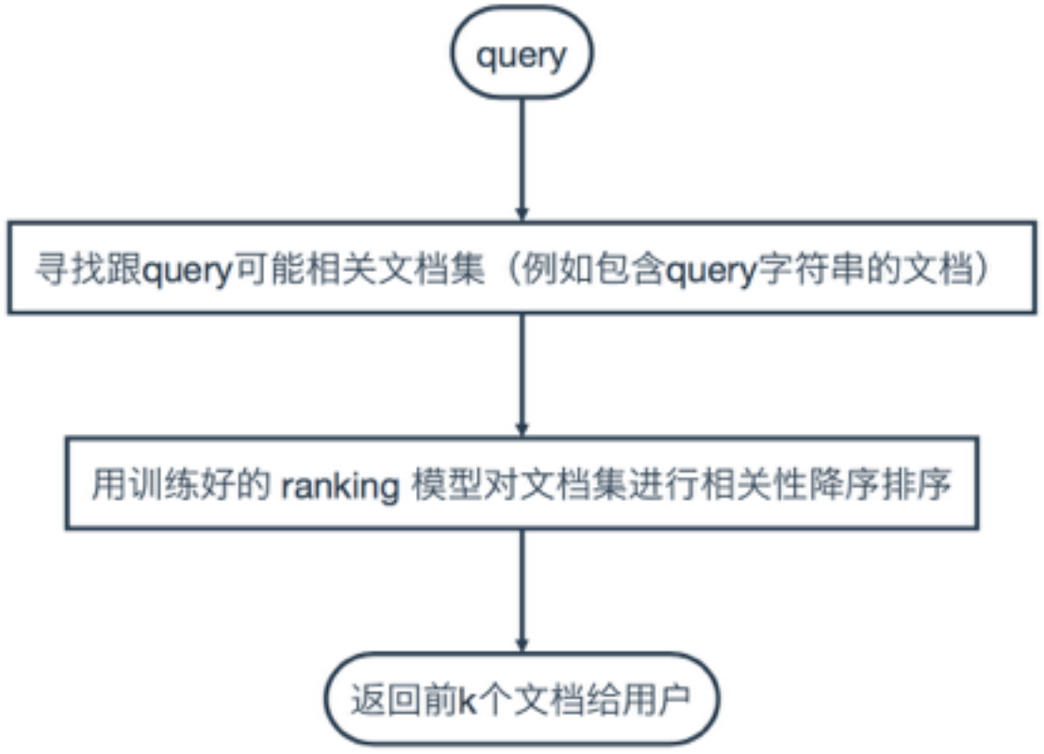
- 领域自适应

  领域自适应：基于不同领域的数据训练一些不同的分类器，是多任务学习的一个特例，但它的不同任务只是领域有区别而已。

9.7 Learning to rank *

定义

- Learning to rank 就是要学习一个可以对一些事物进行排序的函数 (ranking 模型)。最常见的运用就是信息检索 (Information retrieval), 信息检索的一般流程如下：



9.7 Learning to rank *

相关性测量方法

在这些计算相关性的方法中，一个标准的方法来测量一个文档 d 相对于查询 q 的相关性方法是使用一个基于词袋模型的概率语言模型 (probabilistic language model)。

具体地，定义 $sim(q,d) \triangleq p(q|d) = \prod_{i=1}^{n} p(q_i|d)$，其中 q_i 是查询的第 i 个词语。而对于 $p(q_i|d)$ 我们可以用下面表达式进行估计：

$$p(t|d) = (1-\lambda)\frac{\mathrm{TF}(t,d)}{\mathrm{LEN}(d)} + \lambda p(t|\mathrm{background})$$

- $\mathrm{TF}(t,d)$ 是词语 t 在文档 d 出现的频率
- $\mathrm{LEN}(d)$ 是文档的不同词汇数量
- $p(t|\mathrm{background})$ 是平滑项，关于词汇 t 的先验知识。
- $0 < \lambda < 1$ 是平滑参数

9.7 Learning to rank *

模型构建

- 逐点方法 (The pointwise approach)
  - 符号定义
    - 对于每个查询 q，假定检索出 m 个可能的相关文档 d_j, $j = 1 : m$。
    - 对于每个 (查询, 文档) 对，我们定义一个特征向量 $x(q, d)$，例如这可能包含了查询和文档的相似度分数和PageRank分数等。
    - 假定我们有 label 集合 y_j，代表文档 d_j 相对于查询 q 的相关性度数。这些标签可以是二元（相关/不相关），也可以是多元的，代表相关性的程度（非常相关，相关，不相关）。

- 构建模型
  - 对于二元相关标签，我们可以用标准的二元分类机制来估计 $p(y = 1|x(q, d))$
  - 对于有序相关性标签，我们可以使用有序回归来预测排名，$p(y = r|x(q, d))$

9.7 Learning to rank *

模型构建

- 成对方法 (The pairwise approach)

用一个形式为 $p(y_{jk}|x(q,d_j),x(q,d_k)$ 二元分类器来对数据进行建模，其中 $y_{jk}=1$ 如果 $\text{rel}(d_j,q)>\text{rel}(d_k,q)$，否则 $y_{jk}=0$，一个具体的建模方式如下：

$$p(y_{jk}=1|x_j,x_k)=\text{sigm}(f(x_j)-f(x_k))$$

其中，$f(x)$ 是一个评分函数，经常设为线性函数，即 $f(x)=w^Tx$

这是一种特殊的神经网络，称为 RankNet

我们可以通过最大化对数似然进行参数 w 的 MLE 估计，即等价于最小化交叉熵，其表达式如下：

$$L=\sum_{i=1}^{N}\sum_{j=1}^{m_i}\sum_{k=j+1}^{m_i}L_{ijk}$$

$$\begin{aligned}-L_{ijk}=&\ \mathbb{I}(y_{ijk}=1)\log p(y_{ijk}=1|x_{ij},x_{ik},w)\\&+\mathbb{I}(y_{ijk}=0)\log p(y_{ijk}=0|x_{ij},x_{ik},w)\end{aligned}$$

9.7 Learning to rank *

模型构建

- 成列方法 (The listwise approach)
  - 构建模型

    定义 π 是文档列表的某一排序顺序，为了对 π 的不确定性建模，我们用 $Plackett-Luce$ 分布，其形式如下：

$$p(\pi|s) = \prod_{j=1}^{m} \frac{s_j}{\sum_{u=j}^{m} s_u}$$

    其中 $s_j = s(\pi^{-1}(j))$ 是排在第 位置文档的相关性分数

  - 模型参数估计

    为了训练模型，我们可以让 y_i 是对于查询 q 的真实相关性分数，然后用最小化交叉熵：

$$-\sum_{i}\sum_{\pi} p(\pi|y_i)\log(\pi|s_i)$$

不可计算，所以只考虑 Top k

$$p(\pi_{1:k}|s_{1:m}) = \prod_{j=1}^{k} \frac{s_j}{\sum_{u=1}^{m} s_u}$$

9.7 Learning to rank *

评价标准

- Mean reciprocal rank (MRR) 平均倒数排名

  对于一个查询 q，我们定义 $r(q)$ 是出现第一个相关文档的位置。所以平均倒数排名即如下，这是一个相当简单的评价标准。

$$\mathrm{MRR} = \frac{1}{Q}\sum_{i=1}^{Q}\frac{1}{r(q_i)}$$

- Mean average precision (MAP) 平均准确率均值

  对于二元相关标签，定义对于某排序，在 k 位置的准确率如下：

$$P@k(\pi) \triangleq \frac{\text{num.relevant documents in the top } k \text{ positions of } \pi}{k}$$

  然后，我们定义平均准确率如下：

$$AP(\pi) \triangleq \frac{\sum_k P@k(\pi) * I_k}{\text{num.relevant documents}}$$

  其中 $I_k = 1$ 当且仅当第 k 个文档是相关的

9.7 Learning to rank *

评价标准

- Normalized discounted cumulative gain (NDCG)

  假定相关性 label 具有多个等级，定义对于某个排序的前 k 个项 的 **discounted cumulative gain** 如下：

$$\mathrm{DCG@}k(r) = r_1 + \sum_{i=2}^{k} \frac{r_i}{\log_2 i}$$

其中，r_i 是第 i 项的相关度，$\log_2$ 是用来对文档列表后面项的相关度大折扣。

但由于文档列表长度不一样，DCG 的值也不一样，所以通常会用 ideal DGG (IDGG) 归一化，其表达式如下：

$$\mathrm{IDCG@}k(r) = \mathrm{argmax}_{\pi}\mathrm{DCG@}k(r)$$

$$\mathrm{NDCG} = \frac{\mathrm{DCG}}{\mathrm{IDCG}}$$

9.7 Learning to rank *

评价标准

- Rank correlation

我们可以测量文档排名列表 π 和 真实排名列表 π^* 的相关性。其中一个测量两个排名的相关性方法是 weighted Kendall's τ 统计，其表达式如下：

$$\tau(\pi, \pi^*) = \frac{\sum_{u<v} w_{uv}[1 + \text{sgn}(\pi_u - \pi_v)\text{sgn}(\pi_u^* - \pi_v^*)]}{2\sum_{u<v} w_{uv}}$$

其中，w_{uv} 是文档对的权重。