第六章课后习题

逻辑斯蒂回归于最大熵模型

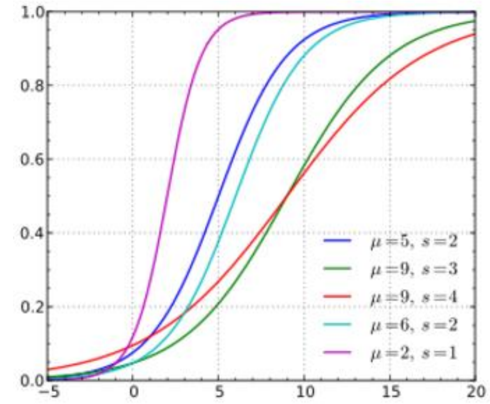
6.1确认逻辑斯蒂分布属于指数分布族

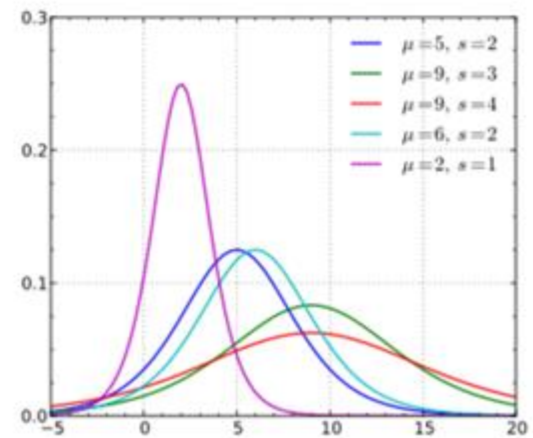
1. 逻辑斯蒂分布

设X是连续随机变量，X服从逻辑斯蒂回归分布是指X具有下列分布函数和密度函数：式中，μ为位置参数，γ>0为形状参数。是针对生物种群繁殖的速度变化规律而得到的一个分布函数

$$F(x) = P(X \leq x) = \frac{1}{1 + e^{-(x-\mu)/\gamma}}$$

$$f(x) = F'(x) = \frac{e^{-(x-\mu)/\gamma}}{\gamma(1 + e^{-(x-\mu)/\gamma})^2}$$





2. 指数分布族

指数分布族在上世纪30年代中期被提出，在概率和统计学上，它是一些有着特殊形式的概率分布的集合，包括许多常用的分布，如正态分布、伯努利分布、指数分布、泊松分布等。

下面我们用一个重要分布的例子来说明下指数分布族。假设有一个正态分布，均值为0，服从 $X - N(0,\sigma^2)$，则其概率密度函数PDF为：

$$\boxed{f(x|\sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{x^2}{2\sigma^2}}}$$

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

当μ为1时的高斯分布

这个概率密度函数由一个参数σ来定义。我们可以把该式子作如下变形：

$$f(x|\sigma) = \frac{1}{\sqrt{2\pi}} e^{-log\sigma} e^{-\frac{x^2}{2\sigma^2}} = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2\sigma^2} - log\sigma} = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2\sigma^2}x^2 - log\sigma}$$

令：$h(x) = \frac{1}{\sqrt{2\pi}}$，$\eta(\sigma) = -\frac{1}{2\sigma^2}$，$T(x) = x^2$，$A(\sigma) = log\sigma$；则上式可以用如下的形式表达：

$$f(x|\sigma) = h(x)exp(\eta(\sigma)T(x) - A(\sigma))$$

我们把参数一般化为θ，则上式为：

$$f(x|\theta) = h(x)exp(\eta(\theta)T(x) - A(\theta))$$

$$f(x|\theta) = h(x)exp(\eta(\theta)T(x) - A(\theta))$$

这就是指数分布族的概率密度函数PDF或概率质量函数PMF的通用表达式框架。

分布函数框架中的$h(x)$,$\eta(\theta)$,$T(x)$和$A(\theta)$并不是任意定义的，每一部分都有其特殊的意义。
θ是**自然参数(natural parameter)**，通常是一个实数；
$h(x)$是**底层观测值（underlying measure）**；
$T(x)$是**充分统计量（sufficient statistic）**；
$A(\theta)$被称为**对数规则化（log normalizer）**。

伯努利分布又叫做两点分布或者0-1分布，是一个离散型概率分布，若成功则随机变量取1，概率取φ；如果失败，则随机变量取0，概率取1-φ；指数分布族为

$$
\begin{aligned}
\mathrm{P}(\mathrm{y};\varphi) &= \varphi^{y}(1-\varphi)^{1-y} = \exp(\log\varphi^{y}(1-\varphi)^{1-y}) \\
&= \exp(\mathrm{y}\log\varphi + (1-\mathrm{y})\log(1-\varphi)) \\
&= \exp(\mathrm{y}\log\frac{\varphi}{1-\varphi} + \log(1-\varphi))
\end{aligned}
$$

对比指数分布族，有

$$P(y;\eta) = b(y)exp(\eta^{T}T(y) - a(\eta))$$

$$\mathrm{b}(\mathrm{y}) = 1$$

$$\mathrm{T}(\mathrm{y}) = \mathrm{y}$$

$$\eta = \log\frac{\varphi}{1-\varphi} \Rightarrow \boxed{\varphi = \frac{1}{1+e^{-\eta}}}$$

正则响应函数

$$\mathrm{a}(\eta) = -\log(1-\varphi) = 1 + e^{-\eta}$$

$$\varphi = \frac{1}{1+e^{-\eta}}$$

当η为(x-μ)/γ

$$F(x) = P(X \le x) = \frac{1}{1+e^{-(x-\mu)/\gamma}}$$

$$f(x) = F'(x) = \frac{e^{-(x-\mu)/\gamma}}{\gamma(1+e^{-(x-\mu)/\gamma})^2}$$

为什么要选用伯努利模型呢？
这是因为 logistic模型对问题的前置概率估计其实就是伯努利分布

6.2 写出逻辑斯蒂回归模型学习的梯度下降算法

输入：目标函数$f(x)$,梯度函数$g(x) = \nabla f(x)$,计算精度ε

输出：$f(x)$的极小值点x^*.

（1） 取初始值$x^{(0)} \in R^n$,置 k = 0

（2） 计算$f\left(x^{(k)}\right)$

（3） 计算梯度$g_k = g\left(x^{(k)}\right)$,当$||g_k|| < \varepsilon$时，停止迭代，令$x^* = x^{(k)}$;否则，令$p_k = -g(x^{(k)})$,

求λ_k，使$f\left(x^{(k)} + \lambda_k p_k\right) = \min\limits_{\lambda \geq 0} f(x^{(k)} + \lambda p_k)$

（4） 置 $x^{(k+1)} \leftarrow x^{(k)} + \lambda_k p_k$ ， 计 算 $f\left(x^{(k+1)}\right)$ 当 $\left|\left|f\left(x^{(k+1)}\right) - f\left(x^{(k)}\right)\right|\right| < \varepsilon$ 或

$\left|\left|x^{(k+1)} - x^{(k)}\right|\right| < \varepsilon$时，停止迭代，令$x^* = x^{(k+1)}$

（5） 否则，置 k = k+1，转（3）

当目标函数是凸函数时，梯度下降法是全局的最优解，一般情况下，其解不保证是全局最优解，梯度下降法的收敛速度也未必是很快的。

6.2 写出最大熵模型学习的DFP算法

（1）给定初始点 x_0，初始矩阵 H_0（通常取单位阵），计算 g_0，令 k=0，给定控制误差 ε。

（2）令 $p_k = -H_k g_k$。

（3）由精确一维搜索确定步长 α_k， $f(x_k + \alpha_k p_k) = \min_{\alpha \geq 0} f(x_k + \alpha p_k)$

（4）令 $x_{k+1} = x_k + \alpha_k p_k$。

（5）若 $\|g_k\| \leq \varepsilon$，则 $x^* = x_{k+1}$ 停；

否则令 $s_k = x_{k+1} - x_k$ ， $y_k = g_{k+1} - g_k$ 。

（6）由 DFP 修正公式得 H_{k+1}。令 k=k+1，转步骤（2）