6.1 确认逻辑斯蒂分布属于指数分布族

> 定义6.1（逻辑斯谛分布） 设X是连续随机变量，X服从逻辑斯谛分布是指X具有下列分布函数和密度函数：
>
> $$F(x)=P(X\leqslant x)=\frac{1}{1+\mathrm{e}^{-(x-\mu)/\gamma}}$$
> $$f(x)=F'(x)=\frac{\mathrm{e}^{-(x-\mu)/\gamma}}{\gamma(1+\mathrm{e}^{-(x-\mu)/\gamma})^2}$$
>
> 式中，μ为位置参数，$\gamma>0$为形状参数。

指数分布族的定义

> - Reference：https://en.wikipedia.org/wiki/Exponential_family
> - 对于随机变量x，在给定参数θ下，其概率分别满足如下形式：
>
> $p(x|\theta)=h(x)\exp(\theta^T T(x)-A(\theta))$ 称之为指数分布族
>
>   其中：$g(\theta)$表示归一化系数，$h(x)>0$

证明

二项逻辑斯蒂回归模型的分布函数为：
$P(Y=1|X)=\frac{\mathrm{e}^{\beta_0+\beta^TX}}{1+\mathrm{e}^{\beta_0+\beta^TX}}$
$P(Y=0|X)=\frac{1}{1+\mathrm{e}^{\beta_0+\beta^TX}}$

似然函数为：

$$\begin{aligned}L(\beta)&=\prod_{i=1}^{N}P(Y=y_i|X=x_i)\\&=\prod_{i=1}^{N}(\frac{\mathrm{e}^{\beta_0+\beta^TX}}{1+\mathrm{e}^{\beta_0+\beta^TX}})^{y_i}(\frac{1}{1+\mathrm{e}^{\beta_0+\beta^TX}})^{1-y_i}\\&=exp(\sum_{i=1}^{N}(y_i(\beta_0+\beta^TX)-\log(1+\mathrm{e}^{\beta_0+\beta^TX})))\end{aligned}$$

满足指数分布族的形式，其中

- $h(x)=1$
- $\theta=(\beta_0,\beta)$
- $T(x)=y_i(1,x)$
- $A(\theta)=\log(1+\mathrm{e}^{\beta_0+\beta^TX})$
  多项逻辑斯蒂回归模型：
  $P(Y=k|X)=\frac{\mathrm{e}^{\beta_{k0}+\beta_k^TX}}{1+\sum_{l=1}^{K-1}\mathrm{e}^{\beta_{l0}+\beta_l^TX}}$

证明过程与二项逻辑斯蒂回归模型类似，满足指数分布族的形式

6.2 写出逻辑斯谛回归模型学习的梯度下降算法

定义6.2（逻辑斯谛回归模型） 二项逻辑斯谛回归模型是如下的条件概率分别:

$$P(Y=1|x)=\frac{\exp(w\cdot x+b)}{1+\exp(w\cdot x+b)}$$

$$P(Y=0|x)=\frac{1}{1+\exp(w\cdot x+b)}$$

> 这里，$x\in R^n$是输入，$Y\in\{0,1\}$是输出，$w\in R^n$和$b\in R^n$是参数，w称为权值向量，b称为偏置，$w\cdot x$为w和x的内积

逻辑斯谛回归模型学习的梯度下降法

逻辑斯谛回归模型的对数似然函数为:

$L(w)=\sum_{i=1}^{N}\left[y_i(w\cdot x_i)-\log(1+\exp(w\cdot x_i))\right]$

将对数似然函数求偏导，可得

$$\frac{\partial L(w)}{\partial w^{(k)}}=\sum_{i=1}^{N}\left[x_i\cdot y_i-\frac{\exp(w^{(k)}\cdot x_i)\cdot x_i}{1+\exp(w^{(k)}\cdot x_i)}\right] \tag{1}$$

梯度函数为

$$\nabla L(w)=\left[\frac{\partial L(w)}{\partial w^{(0)}},\cdots,\frac{\partial L(w)}{\partial w^{(n)}}\right] \tag{2}$$

根据书中附录**A** 梯度下降法得逻辑斯谛回归模型学习的梯度下降算法:

- 输入：目标函数$f(w)$，梯度函数$g(w)=\nabla f(w)$，精度ε
- 输出：$f(w)$的极大值w^*
  1. 初值$w^{(0)}\in R^n$，$k=0$
  2. 计算$f(w^{(k)})=\sum_{i=1}^{N}\left[y_i(w^{(k)}\cdot x_i)-\log(1+\exp(w^{(k)}\cdot x_i))\right]$
  3. 计算梯度$g_k=g(w^{(k)})=\sum_{i=1}^{N}\left[x_i\cdot y_i-\frac{\exp(w^{(k)}\cdot x_i)\cdot x_i}{1+\exp(w^{(k)}\cdot x_i)}\right]$：

     3.1 当$\|g_k\|<\varepsilon$时，停止迭代，令$w^*=w^{(k)}$

     3.2 否则，令$p_k=-g(w^{(k)})$，求λ_k，使 $f(w^{(k)}+\lambda_k p_k)=\max_{\lambda\geqslant 0} f(w^{(k)}+\lambda p_k)$
  4. 置$w^{(k+1)}=w^{(k)}+\lambda_k p_k$，计算$f(w^{(k+1)})$:

     当$\|f(w^{(k+1)})-f(w^{(k)})\|<\varepsilon$或 $\|w^{(k+1)}-w^{(k)}\|<\varepsilon$时，停止迭代，令$w^*=w^{(k+1)}$
  5. $k=k+1$，转 $step\ 3$