广义线性模型 Generalized linear models

概述

广义线性模型 (GLM) 是简单线性回归的扩展，在广义线性模式中，假设每个样本的观测值 Y 来自某个指数族分布，而$E(y)=\mu=g(x^T\theta)$则作为给定x的的情况下对 y 的估计。广义线性模型的主要部分为：

1. 指数族的分布函数f
2. 线性自然参数$\eta=x^T\theta$
3. 链接函数 g 使得 $E(y)=\mu=g(x^T\theta)$

指数族

指数族的概率密度函数(p.d.f.) 都可以写为：

$$f(y;\theta,\tau)=exp(\frac{a(y)b(\theta)}{h(\tau)}+d(y,\tau))$$

τ 为尺度参数，a,b,c,d,h 为已知函数，$\eta=b(\theta)$为自然参数，$a(y)$为充分统计量，且分布的支撑不依赖于 θ.

由于通常情况下不用考虑 τ，所以引入CS229 中 Andrew Ng 对指数族的简化表示：

$$f(y;\eta)=b(y)exp[\eta^TT(y)-a(\eta)]$$

η 称为自然参数, $T(y)$ 为充分统计量

线性回归和逻辑回归的模型都可以写为上面的形式：

线性回归

$$p(y;x,\mu)=\frac{1}{\sqrt{2\pi}}\exp[-(y-\mu)^2/2]$$

$$p(y;x,\mu)=\frac{1}{\sqrt{2\pi}}\exp(-\frac{1}{2}y^2)\exp[\mu y-\frac{1}{2}\mu^2]$$

$$\eta=\mu,T(y)=y,a(\eta)=\frac{1}{2}\eta^2$$

$$b(y)=\frac{1}{\sqrt{2\pi}}\exp(-\frac{1}{2}y^2)$$

逻辑回归

$$p(y;\phi)=\phi^y(1-\phi)^{1-y}$$

$$p(y;\phi)=\exp[y\ln\phi+(1-y)\ln(1-\phi)]$$

$$p(y;\phi)=\exp[y\ln\tfrac{\phi}{1-\phi}+\ln(1-\phi)]$$

$$T(y)=y,b(y)=1,\eta=\ln\tfrac{\phi}{1-\phi}$$

$$a(\eta)=-\ln(1-\phi)=\ln(1+e^\eta)$$

其中

$$\eta=\ln\frac{\phi}{1-\phi}$$

又称为lnit函数，由此能引出lnistic函数

$$\phi=\frac{1}{1+e^\eta}$$

构造广义线性回归模型

考虑一个回归或分类问题，亦即用一个x的函数预测一个随机变量$y, y=f(x)$.
构造广义线性模型基于以下三条假设：

1. $y|x;\theta\sim ExponentialFamily(\eta)$
2. 给定x，预测目标是充分统计量$T(y)$，在多数情况下，$T(y)=y$，也就是用hypothesis $h(x)$ 预测 $E[y|x]$ ，线性回归里，$h_\theta(x)=x^T\theta$；在逻辑回归里则是$h_\theta(x)=p(y=1|x;\theta)$
3. 自然参数$\eta=x^T\theta$，如果η为向量，则$\eta_i=x^T\theta_i$

第3条就是广义线性模型的线性所在，之所以人为限制自然参数的形式，主要是这种形式有比较好的性质，当然，如果改掉第3条，同样能给出一些模型，但这些模型就不算是广义线性模型了。

常见广义线性模型

Y的分布	名称	链接函数	均值函数
高斯	恒等	$x^T\theta=\mu$	$\mu=x^T\theta$
Gamma	倒数	$x^T\theta=\mu^{-1}$	$\mu=(x^T\theta)^{-1}$
逆高斯	二次倒数	$x^T\theta=\mu^{-2}$	$\mu=(x^T\theta)^{-1/2}$
Possion	自然对数	$x^T\theta=\ln\mu$	$\mu=\exp(x^T\theta)$
二项分布	logit	$x^T\theta=\ln\frac{\mu}{1-\mu}$	$\mu=\frac{\exp(x^T\theta)}{1+\exp(x^T\theta)}$

模型正则化

在训练数据不够多时，或者overtraining时，常常会导致overfitting（过拟合）。其直观的表现如下图所示，随着训练过程的进行，模型复杂度增加，在训练集上的误差渐渐减小，但是在验证集上的误差却反而渐渐增大，这是因为训练出来的网络过拟合了训练集。正则化是一种常用的避免过拟合的方法，典型的就是L1,L2正则项。

从Cost function的角度看广义线性模型，则模型求解等同于即最小化Cost function，假设原Cost funtion的形式为：

$$C(\theta) = C_0(y, h_\theta(x))$$

则求解的结果为：

$$\hat{\theta} = \operatorname{argmin}_\theta C(\theta)$$

加入正则化项后，Cost function为：

$$C(\theta) = C_0(y, h_\theta(x)) + \lambda R(\theta)$$

此时最小化C不仅要考虑C_0，还要考虑λR，R取L2范数时会限制θ的大小，使得θ没有那个分量显著太大，R取L1范数时则倾向于令θ的某些分量趋于零，因而往往具有变量选择的效果。

岭回归 **Ridge Regression**

在线性回归模型中添加L2正则项得到的模型称为岭回归
sklearn.linear_model.Ridge

套索 **Lasso**

在线性回归模型中添加L1正则项得到的模型称为Lasso
sklearn.linear_model.Lasso

弹性网 **Elastic Net**

在线性回归模型中同时添加L1，L2正则项得到的模型称为弹性网
sklearn.linear_model.ElasticNet