机器学习
关于参数模型
与非参数模型研究

文 | 索信达 YVONNE YANG

引言

在大数据时代，我们常常面临成万上亿的数据，伴随着的是高维度的变量，当今很多学术和技术领域都致力于解决针对大数据的模型构造，例如神经网络、深度学习。但对于金融和商业领域，变量是如何影响响应变量的（可解释性）、模型是否可靠等因素尤为重要，特别是当我们的数据集只包含了屈指可数的几个变量，那么特征变量的选取应更严谨，变量与响应变量间关系的量化也是重要的指标，此时可以运用更为精细的模型探索方法。广义线性模型和广义加性模型分别由线性模型和加性模型推广而来，能更广泛地应用于不同分布的数据，并辅以似然比检验逐步优化模型。本文将从参数模型和非参数模型的角度，以广义线性模型和广义加性模型为例，配以相应的案例，对模型探索以及优化方法进行简要介绍。

1 参数回归模型

1.1 传统线性模型

统计学中，参数模型是一类可以通过结构化表达式和参数集表示的模型。为确定两种或两种以上变量之间的定量关系，我们希望通过手中已有的数据去“拟合”出一个“线性方程”，众所周知的经典线性回归模型（LINEAR REGRESSION MODEL）就属于参数模型，

$$\mathbf{Y} = \mathbf{X}\beta + \epsilon, \quad \epsilon \sim \mathbf{N}(\mathbf{0}, \boldsymbol{\Sigma})$$

它要求响应变量Y是实值的且连续的，用于我们通常所说的“回归问题”。

1.2 广义线性模型

1.2.1 模型介绍

日常生活中的许多问题的数据形式并不符合“连续”这个要求，并且面临很多“分类问题“，即该把某对象预测为属于哪一类，这时传统的线性回归模型便显得约束过于强而导致应用范围的狭窄。此外，对于一些较为特殊的分布，如偏态分布和常为重尾分布的金融数据，该如何选择模型呢?广义线性模型（GENERALIZED LINEAR MODEL）应运而生，它将线性回归的思想推广到探索多种形式的响应变量和回归变量之间的关系。其向量形式为：

$$g(E(\mathbf{Y}|\mathbf{X})) = \mathbf{X}\beta, \quad \mathbf{Y}|\mathbf{X} \sim f(\mathbf{y}|\mathbf{x})$$

其中 $g(.)$ 被称为连接函数（LINK FUNCTION），满足平滑（简单来说，图像光滑）且可逆（反函数存在的函数是可逆的）的条件，$f(.)$ 为给定样本下Y的分布。

可以看到，当Y为连续变量并且我们选择 $N(\mu, \sigma^2)$ 作为Y|X 的分布，连接函数取恒等函数时，$g(\mu) = \mu$充的线性回归模型；逻辑回归也是其特例，连接函数为 $log(\frac{p}{1-p})$；而当Y为离散值且选择分布为 $Poisson(\lambda)$，连接函数为 $g(\mu) = ln(\mu)$ 时，模型变为：

$$\mu(x) = E(y|x), \quad ln(\mu(\mathbf{X}))) = \mathbf{X}\beta, \quad \mathbf{Y}|\mathbf{X} \sim poisson(\mu(\mathbf{x}))$$

恰好是泊松回归模型。可以见得，以上常见模型都是广义线性模型的一种特殊形式，广义线性模型通过连接函数将模型变得更灵活而具有普适性。

1.2.2 分布选择

分布的选择依赖于给定样本中Y值的分布，Y的分布观察可利用直方图和核密度估计（KERNEL DENSITY ESTIMATOR）。特别的，对于均值与方差相差甚远的离散响应变量Y，选择POISSON分布不再是一个明智的选择（因为服从POISSON分布的随机变量均值与方差应相等），因此可尝试用负二项分布。

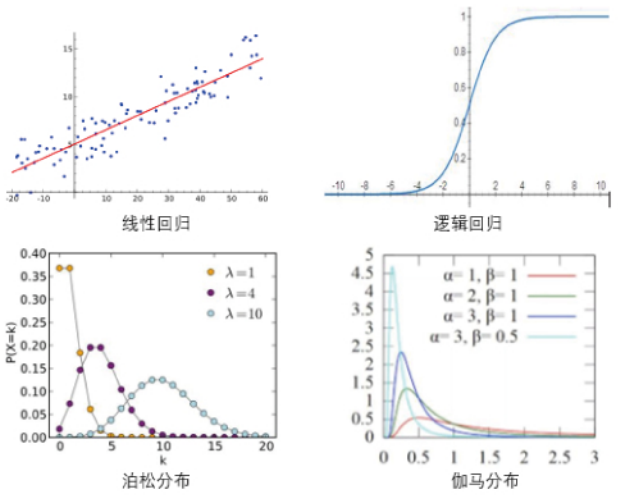


图1. 几种常见分布

1.2.3 连接函数

对于连接函数的选择，此处引入指数族分布的概念：概率密度函数（P.D.F.）或者概率质量函数（P.M.F.）能化成如下形式的分布，被称为指数族分布，

$$f(y;\theta,\varphi)=exp\{\frac{\theta y-b(\theta)}{\varphi}\}+c(y,\varphi)$$

其中 θ 是自然参数，φ 是尺度参数，且满足条件：

$$\int exp\{\frac{\theta y}{\varphi}+c(y,\varphi)\}dy<\infty$$

以下给出一些常见的指数族分布及其标准连接函数（CANONICAL LINK FUNCTION）：

分布	标准连接函数
高斯分布 Gaussian $N(\mu,\sigma^2)$	μ
二项分布 Binomial $b(n,p)$	$log(\frac{p}{1-p})$
泊松分布 Poisson $P(\lambda)$	$log(\lambda)$
伽马分布 Gamma $\Gamma(r,\lambda)$	$\frac{1}{\mu}$

为简化得到系数 β 的数学计算，通常我们选择标准连接函数，特别地，对于要求Y值大于0的响应变量，可以选择连接函数 $g(\mu)=log(\mu)$ 以保证预测值仍然满足大于0。

1.2.4 模型优化之似然比检验

对于参数不显著的项(例如X1)，说明其对于目标变量的影响不显著，可是去掉之后模型是否显著变好，有时用肉眼无法甄别，此时可借助似然比检验。我们希望检验一个更“小”的模型是否可行，此处假设设定为：

原假设： $H_0:\beta_1=0$

备择假设： $H_1:\beta_1\neq 0$

若 $2ln\Lambda(H_0,H_1)>\chi^2(\alpha)$，P值小于显著性水平时，拒绝原假设，认为X1去掉模型效果将显著变差。

1.3 示例

例如在一个案例中（数据来自：HTTPS://WWW.KAGGLE.COM/MIRICHOI0218/INSURANCE），我们要探索保险公司给付的保费（CHARGES）与年龄、性别、BMI、地区、抽烟习惯、小孩数量之间的函数关系，目标变量CHARGES为连续变量，下图为核密度估计曲线，发现其近似GAMMA分布，因此可选择GAMMA分布与其标准连接函数 $g(\mu)=\frac{1}{\mu}$.

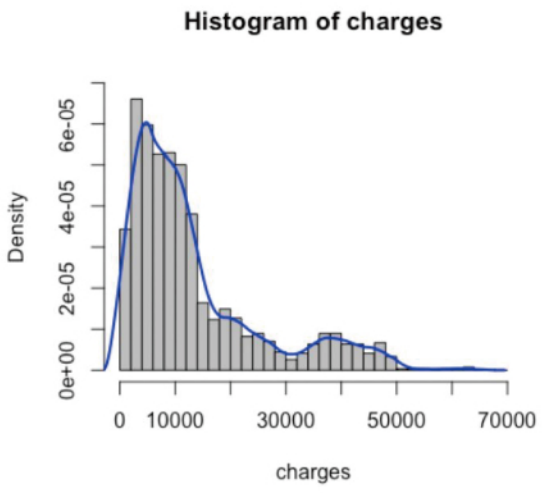


图2: 目标变量CHARGES分布

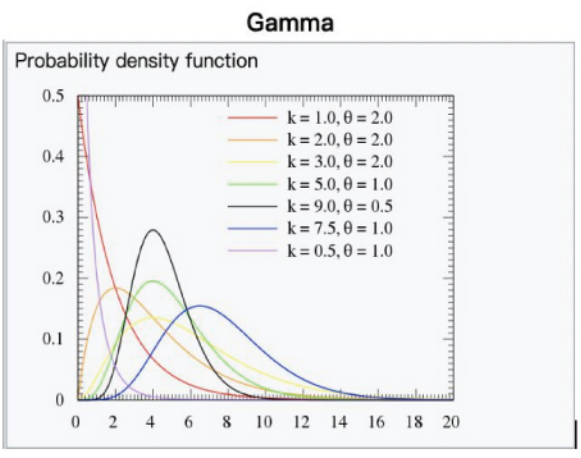


图3: GAMMA分布图（来自：WIKIPEDIA）

对于上图中的数据集，当模型优化至此步，发现不显著项（P值大于0.05）大部分与REGION（地域）有关（如左图所示）。去掉REGION这一项后，重新拟合模型，在参数显著性上有所提升，然而AIC略微上升（如右图所示）：

```
Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)         3.052e-04  1.640e-05  18.618  < 2e-16 ***
age                -3.939e-06  2.110e-07 -18.669  < 2e-16 ***
bmi                 9.214e-08  4.334e-07   0.213  0.83170
children           -1.066e-05  1.785e-06  -5.974 3.15e-09 ***
smokeryes          -2.251e-04  1.809e-05 -12.444  < 2e-16 ***
regionnorthwest     1.429e-06  3.592e-06   0.398  0.69080
regionsoutheast     4.006e-06  3.353e-06   1.195  0.23241
regionsouthwest     3.054e-06  3.606e-06   0.847  0.39717
age:smokeryes       3.670e-06  2.330e-07  15.755  < 2e-16 ***
children:smokeryes  1.083e-05  2.156e-06   5.024 5.94e-07 ***
bmi:smokeryes      -1.337e-06  4.684e-07  -2.855  0.00438 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for Gamma family taken to be 0.4642672)

    Null deviance: 861.69  on 1069  degrees of freedom
Residual deviance: 270.53  on 1059  degrees of freedom
AIC: 21106
```

```
Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)         3.053e-04  1.632e-05  18.700  < 2e-16 ***
age                -3.941e-06  2.111e-07 -18.675  < 2e-16 ***
bmi                 1.616e-07  4.300e-07   0.376   0.7072
children           -1.058e-05  1.781e-06  -5.941 3.83e-09 ***
smokeryes          -2.254e-04  1.809e-05 -12.460  < 2e-16 ***
age:smokeryes       3.673e-06  2.330e-07  15.765  < 2e-16 ***
children:smokeryes  1.061e-05  2.143e-06   4.951 8.59e-07 ***
bmi:smokeryes      -1.323e-06  4.680e-07  -2.826   0.0048 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for Gamma family taken to be 0.4648259)

    Null deviance: 861.69  on 1069  degrees of freedom
Residual deviance: 271.25  on 1062  degrees of freedom
AIC: 21103
```

图4: 模型结果输出-优化前（左）优化后（右）

看起来小小牺牲AIC能换来参数显著性的提高，那么把地域特征去除真的能达到优化模型的效果吗？

我们希望检验一个更“小”的模型是否可行，当显著性水平设为 $\alpha=0.05$ 时，根据如下输出结果，P值为0.3989，不拒绝原假设，认为“地区（REGION）”可从模型中移除，至此模型优化结束。似然比检验能更客观地告诉我们某一特征能否从模型中移除。

```
> lrtest(Gafit5,Gafit4) #reduce model is ok
Likelihood ratio test

Model 1: charges ~ age + bmi + children + smoker + age:smoker + children:smoker +
    bmi:smoker
Model 2: charges ~ age + bmi + children + smoker + region + age:smoker +
    children:smoker + bmi:smoker
  #Df LogLik Df Chisq Pr(>Chisq)
1   9 -10542
2  12 -10541  3 2.953     0.3989
```

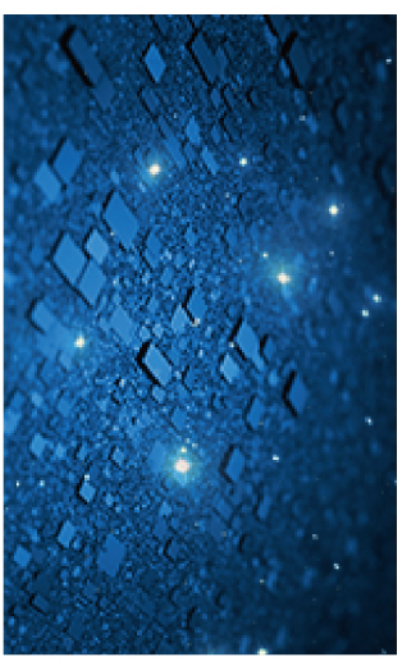
图5: 似然比检验结果

2 非参数回归模型

2.1 非参数模型

参数模型与非参数模型的区别在于：参数模型预先设定了模型的形式，后通过最小化SCORE FUNCTION求得参数，而非参数模型不对随机变量预先假设任何模型形式，预测器的构建都依赖于数据，可以自由地从数据中学习出模型，具有极大的灵活性。在构造目标函数时，非参的方法寻找合适的训练数据，同时保留一些对数据的泛化能力，因此，这些非参方法能够拟合大多数的函数形式。例如K近邻算法就是典型的非参模型，对于一个新的数据实例X，该算法寻找离X最邻近的K个样本，以“多数取胜”策略来确定X的类别。

受限于先验模型的形式，参数模型有时无法全面地捕捉到数据样本的特征，且有时模型形式难以预先确定。对于回归问题而言，常用的非参数回归方法有局部多项式回归（POLYNOMIAL REGRESSION）和样条回归（SPLINE REGRESSION）。对于P个变量的回归问题，非参数方法能得到最终的回归方程 $m(x_1, x_2, \ldots, x_p)$，或向量表示为 $m(X)$。



2.2 从加性模型到广义加性模型

非参数回归的模型形式自由，完全由数据驱动，适应力强，但也有显著的缺点，例如“高维诅咒”：当维数较高时，前述的两种非参数回归方法开始变得不稳定且收敛减慢，同时最终的回归方程解释性和可视化能力弱。解决维度问题的一种方法是利用加性模型（ADDITIVE MODEL）：

$$m(X) = c + \sum_{i=1}^{p} g(x_i)$$

其中 $g(x_i)$ 为单变量非参数方程。该方法相当于将1个P维变量的方程转换成了P个单变量方程。

相似地，通过连接函数 $G(.)$，可推广到广义加性模型（GENERALIZED ADDITIVE MODEL）：

$$G(E(Y|\mathbf{X})) = c + \sum_{i=1}^{p} g(x_i), \quad \mathbf{Y}|\mathbf{X} \sim f(\mathbf{y}|\mathbf{x})$$

2.3 示例

采用广义加性模型，能克服非参数模型常有的解释性弱和可视化能力差的问题，可以研究单个变量的非参数项。例如探究房价与10个变量的关系，采用样条函数平滑每一个变量，变量显著性如下图所示：

```
Approximate significance of smooth terms:
          edf Ref.df      F  p-value
s(x1)   3.595  4.569  9.028 6.01e-07 ***
s(x2)   4.810  5.844  3.612   0.0020 **
s(x3)   8.878  8.989 13.458  < 2e-16 ***
s(x4)   8.281  8.855 23.392  < 2e-16 ***
s(x5)   1.770  2.243  0.652   0.4782
s(x6)   7.927  8.702 12.026  < 2e-16 ***
s(x7)   7.504  8.351  8.690 3.10e-11 ***
s(x8)   1.000  1.000 55.987 3.37e-13 ***
s(x9)   1.000  1.000  4.300   0.0387 *
s(x10) 8.093  8.776 28.511  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’
```

图6：GAM模型拟合结果

通过绘制部分预测图，可以探索单个变量的效应，其中每幅图横轴为单一变量取值，纵轴为对应的：

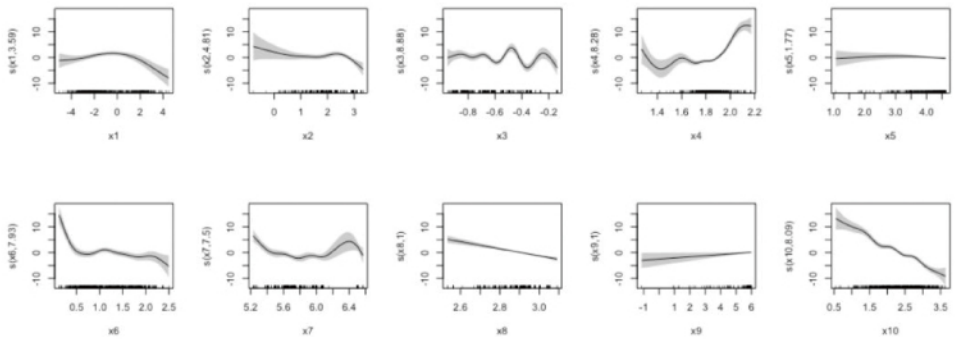


图7: 部分预测图

.

3 半参数模型

3.1 模型介绍

在非参数模型模型优化的过程中，有些变量呈现出强烈的线性性，对该x_i变量应用线性项替代非线性项βx_i放入模型中。此时得到的是“半参数模型”，它同时含有线性项和非线性项，作为非参数模型和参数模型之间的一类模型，半参数模型既继承了非参数模型的灵活性，又继承了参数模型的可解释性，可以进一步改善非参数模型的缺陷。半参数模型常具有以下的形式：

$$G(E(Y|\mathbf{X})) = c + \sum_i \beta_i x_i + \sum_j g(x_j)$$

其中为线性项，为非线性项。

3.2 示例

在2.3示例中，非参数模型拟合后发现X5和X9的非参数项不够显著，进一步观察部分预测图发现变量X5对模型响应变量没有贡献（因为其纵轴刻度始终都在0附近），变量X8和X9表现出线性性（因为其部分预测图近似直线），而其他变量表现出非线性性。据此移除变量X5，并将X8和X9的项替换为线性项，优化得到一个半参模型，所有项都是显著的，结果如下所示：

```
Parametric coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  60.3611     5.5091  10.957  < 2e-16 ***
x8          -13.8665     1.8350  -7.557  2.3e-13 ***
x9            0.4338     0.2111   2.055   0.0404 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Approximate significance of smooth terms:
          edf Ref.df      F  p-value
s(x1)   3.504  4.461  9.033 7.18e-07 ***
s(x2)   4.678  5.703  3.467  0.00312 **
s(x3)   8.854  8.987 13.335  < 2e-16 ***
s(x4)   8.317  8.868 23.642  < 2e-16 ***
s(x6)   7.883  8.681 11.885  < 2e-16 ***
s(x7)   7.728  8.505  8.562 4.02e-11 ***
s(x10) 8.070  8.766 32.887  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

图8: 优化后参数显著性

4 小结

传统的参数模型(如线性回归)只能处理一些简单的变量间呈现特定关系的数据，当面临的问题更复杂的时候，变量关系说不清道不明，参数模型不一定能达到目标效果。非参数模型可以规避上述问题，具有更好的灵活性，并可通过广义加性模型获得更好的性能。此外，半参数模型是介于参数模型和非参数模型之间的一类，常由非参模型优化得来，兼具灵活性和可解释性。

对于样本量足够大而变量数量不大的数据集，或者对一些需要追踪指标变化原因的场景，这些统计模型及其优化方法或许能派上用场。其通过分布选择与连接函数推广到更具有普适性的模型，并能利用统计方法去检测变量的选择是否具有合理性。无论是广义线性模型和广义加性模型，都能学习到一个既定的模型，通过变量参数或者部分预测图去发现变量如何影响响应变量，同时对于新的数据集可以产生相应的预测值。

注：本文使用的分析工具为R语言，有兴趣的读者可自行了解。