文章编号：2095-2163(2021)09-0071-04 中图分类号：TP399 文献标志码：A

基于对称百分误差的线性回归与印染工业应用

宋丛威，张晓明
（北京雁栖湖应用数学研究院，北京 101408）

摘 要：印染工业现存的问题，是如何用染料在样品试验中的浓度预测真实面料中的浓度。理想情况下，染出相同颜色的染料浓度，无论在样品中还是真实面料中应该是等同的，或者至少呈现线性关系。虽然普通线性回归可以基本解决该问题，但其为最小化绝对误差的平方和，而工业上要求用百分误差，此时不能用代数方法求解线性模型。本文选用对称百分误差，并为对应的损失函数构造了梯度下降算法，其解优于用普通线性回归算法，从而提高预测性能。

关键词：对称百分误差；线性回归；梯度下降法；印染工业

Linear regression based on symmetric percentage error and its application in printing and dyeing industry

SONG Congwei, ZHANG Xiaoming
(1 Yanqi Lake Beijing Institute of Mathematical Sciences And Applications, Beijing 101408, China)

【Abstract】 There is a existing problem in the printing and dyeing industry: predicting the concentration of dyes in the real fabric via that in test samples. In the ideal case, the concentration of dyes that produces the same color should be equivalent in both the sample and the real fabric, or at least have linear relation. Ordinary linear regression can basically solve this problem. But ordinary linear regression minimizes the sum of squares of the absolute error, whereas the percentage error is demanded in industry. The problem cannot be solved with the linear model algebraically. In this paper we select symmetric percentage error, and construct a gradient descent algorithm for the corresponding loss function. Its solution is better than the ordinary linear regression algorithm, and sequentially improve the prediction performance.

【Key words】 symmetric percentage error; linear regression; gradient descent method; printing and tyeing industry

0 引 言

目前印染工业面临的问题，是根据已知的颜色，给出染料浓度配比，使得布料染出所需颜色[1]。文献[1]中已经建立了一个高效的线性模型用于预测染料浓度，本文则讨论另一种预测浓度的方案。工人通常会在正式染色前，对样品进行试验性的染色，得到一组浓度配比。这本身也是目前印染业标准生产流程的环节。理想情况下，样品试验得到的是正式染色的浓度配比。但在实际效果上却总是存在一定的误差。为了利用样品试验数据来预测正式染色时的染料浓度配比，提出了线性模型[2-3]：

$$y \sim \hat{y}(\boldsymbol{x}) = a \cdot \boldsymbol{x} + b$$

其中，$\boldsymbol{x}$、y 分别是样品浓度配比和正式染色的浓度配比。

注意：$\boldsymbol{x}$ 是向量，代表样品试验浓度配比，而 y 是数量，代表正式染色的浓度配比的一个分量。即使 y 是向量，也假定各分量是独立的，则分别研究每个分量即可。

为了便于理论分析，把所有参数整合成一个向量 $\boldsymbol{\theta} = (a, b)$，并把线性模型表示为[4]：

$$\hat{y}(\boldsymbol{x}) = \boldsymbol{\theta} \cdot \varphi(\boldsymbol{x}) \tag{1}$$

其中，$C\varphi(\boldsymbol{x})$ 表示由 $\boldsymbol{x}$ 决定的一组基。在本文中由 $\boldsymbol{x}$ 的分量和 1 构成。考虑到下述对数线性模型：

$$\hat{y}(\boldsymbol{x}) = e^{\boldsymbol{\theta} \cdot \varphi(\boldsymbol{x})} \tag{2}$$

相当于对 y 做对数化预处理。由于 $\boldsymbol{x}$、y 有相同的物理意义，$\boldsymbol{x}$ 也会做对数化预处理。

设 $\boldsymbol{\theta}$ 为 p 维向量，对于 N 个样本，则有：

$$\hat{y}(\boldsymbol{x}) = \boldsymbol{\Phi}(\boldsymbol{x})\boldsymbol{\theta} \tag{3}$$

其中，$\hat{y}(\boldsymbol{x}) = (\hat{y}_1(\boldsymbol{x}), \hat{y}_2(\boldsymbol{x}), \cdots, \hat{y}_N(\boldsymbol{x}))^{\mathrm{T}}$ 为 N 维列向量，$\boldsymbol{\Phi}(\boldsymbol{x}) = (\boldsymbol{x}_{ij})$ 为 $N \times p$ 的设计矩阵，习惯

基金项目：浙江省自然科学基金(LQ19F050004)。

作者简介：宋丛威(1986-)，男，博士，讲师，主要研究方向：小波分析、调和分析、机器学习；张晓明(1965-)，男，博士，研究员，主要研究方向：大数据、数字孪生。

收稿日期：2021-07-20

上最后一列为 1。

传统的线性回归采用均方误差,本质上是最小化绝对误差 $|y-\hat{y}|$。但在现实生产活动中认为,对于较小的数值要求更高的精度,也就是样本的 $|y|$ 值越小权重越大。相对误差 $|(y-\hat{y})/y|$ 更为合适,也称为百分误差[5-8]。

本文对相对误差做一些改良。此时,最小二乘法的代数方法不再适用,需用梯度下降法(GD)进行优化。这就是本文要解决的技术问题。最后的数值实验说明,在该误差下,本文方法优于普通的线性回归。

本文做如下符号约定。$\hat{y}$ 表示 y 的估计。没有歧义,不写出求和符号 $\sum_i$ 的求和范围。本文用下指标给样本编号,如 y_i。梯度 ∇ 是对每个函数的每个变量求偏导数得到的列向量。$y \sim f(y)$ 表示 y 服从概率密度函数为 f 的分布,通常不是常用分布。C 总是代表某个不必写明的次要常数。

1 百分误差与损失函数

百分误差 $|(y-\hat{y})/y|$ 已经被广泛讨论和应用。但本文选择其近似的对称百分误差[9],并讨论如何改造误差函数,最终导出回归模型的损失函数。

1.1 矩阵分析

考虑到 y、$\hat{y}$ 都不一定能单独决定样本权重,方案选用对称百分误差:

$$\epsilon(y,\hat{y})=\begin{cases}2\dfrac{\hat{y}-y}{\hat{y}+y}=2\dfrac{1-y/\hat{y}}{1+y/\hat{y}}, & y\in\neq\hat{y}\\ 0, & y=\hat{y}\end{cases} \tag{4}$$

其中,系数 2 起到归一化的作用,在某些场合下可以省略,并不影响分析。限定 $\hat{y}\geqslant 0$,模型输出 $\hat{y}<0$ 时,会被强制取 0。因此,实际的线性模型为:

$$\hat{y}(x)=\max(\theta\cdot\varphi(x),0) \tag{5}$$

如果延拓定义域到 $\hat{y}<0$,则可直接取 $\varepsilon=1$。对于不同 y,误差曲线变化规律是不同的。图 1 反映了该误差曲线的“不对称性”。

如此定义相对误差的好处,就是对异常值不敏感。不对误差取绝对值,是为了更准确地反映误差的分布。

数值计算中,小数除法会导致溢出,必要时改造为:

$$\frac{\hat{y}-y}{\max(\hat{y}+y)/2,C}\text{ 或}\frac{\hat{y}-y}{(\hat{y}+y)/2+C}$$

其中,$C>0$ 是一个合理的常数,在程序实现时考虑这一点。

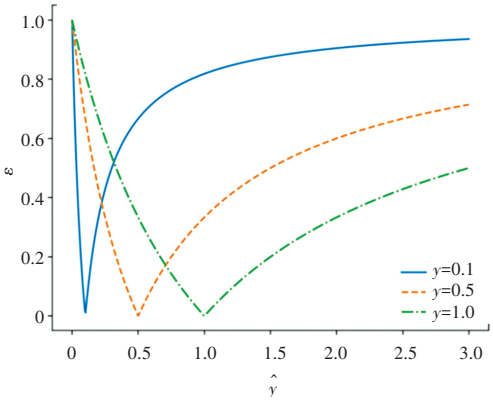

图 1 不同 y 值产生的误差绝对值

Fig. 1 The absolute value of errors produced by different y

一般情况下 $y-\hat{y}$ 都很小,因此可以根据近似关系构造误差函数:$\varepsilon(y,\hat{y})\sim\ln\hat{y}-\ln y$,即数据对数化处理后的绝对误差和原数据的相对误差近似[10]。

对于多维情形 $y=(y^{(1)},\cdots,y^{(n)})$,对每个分量计算误差 $\epsilon(y,\hat{y})=(\epsilon(y^{(1)},\hat{y}^{(1)}),\cdots,\epsilon(y^{(n)},\hat{y}^{(n)}))$,或直接用其 2 范数平方根取代:

$$\epsilon(y,\hat{y})=\sqrt{\frac{1}{n}\sum_j\epsilon(y^{(j)},\hat{y}^{(j)})^2} \tag{6}$$

注:本文用下指标给样本编号,用上指标表示分量。

1.2 损失函数

在定义误差之后,问题转化为最小化误差期望 $L(\theta)=E\,\varepsilon(y,\hat{y}(\theta))^2$。其由经验误差近似计算。对于一组独立同分布的样本 (x_i,y_i),$i=1,\cdots,N$ 来说,最终的损失函数为:

$$L(\theta)=\frac{1}{N}\sum_i\epsilon(y_i,\hat{y}_i(\theta))^2 \tag{7}$$

这个损失函数被称为(对称)均方百分误差。

用代数方法可以求解绝对误差下的损失函数最小化问题,但不能求解损失函数(7)的最小化问题。相反,本文采用一种常用的梯度下降法(GD)——Adam 算法[11-12],预测可以得到比最小二乘法更好的解。

利用 GD 的关键,是计算损失函数的梯度。$\varepsilon(y,\hat{y})^2$ 关于参数 θ 的梯度为:

$$(\hat{y}-y)\nabla\hat{y}(\theta)\begin{cases}\dfrac{y}{m^3}, & M>C\\ \dfrac{2}{C}, & m\leqslant C\end{cases}$$

其中：$m=\dfrac{y+\hat{y}}{2}$ 对线性模型 $y\sim\theta\cdot x$ 来说，$\nabla\hat{y}(\theta)=x$；而对对数线性模型来说，$\nabla\hat{y}(\theta)=\hat{y}(\theta)x$（不要忘记对 x 进行预处理）。故有：

$$\nabla L(\theta)=\frac{1}{N}\sum_i(\hat{y}_i-y_i)\nabla\hat{y}_i(\theta)\begin{cases}\dfrac{y_i}{m_i^3}, & m_i>C\\ \dfrac{2}{C^2}, & m_i\leqslant C\end{cases}\tag{8}$$

其中 $m_i=\dfrac{y_i+\hat{y}_i}{2}$。

注：导数中可能出现小数的三次方除法，故在程序设计时需考虑溢出和梯度值异常。

上述损失函数等价于，在 y 服从以 $\hat{y}>0$ 为中心，取值于[0，∞）偏态分布假设下的极大似然估计：

$$y\sim e^{-C\left(\frac{\hat{y}-y}{\hat{y}+y}\right)^2}$$

式中省略了归一化参数，C 大致为方差的倒数，这个值基本决定了置信度。该分布形状类似于逆 Gamma 分布，如图2所示（因为归一化因素的存在，y 轴不必显示刻度）。令相对误差 ε 服从“紧支撑的正态分布”$\varepsilon\sim\exp\{-C\varepsilon^2\}$，$-1\leqslant\varepsilon\leqslant1$，而绝对误差变成了偏态的。

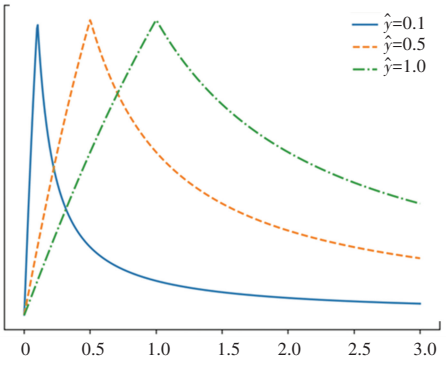

图 2　不同估值下 y 的分布

Fig. 2　Distribution of y with respect to different estimat

偏态分布比正态分布更符合基本要求，即当 $\hat{y}$ 较小时，应该比较大时更接近 y 的准确值。

最后，损失函数的多维形式为：

$$L(\theta)=\frac{1}{nN}\sum_{ij}\epsilon(y_i^{(i)},\hat{y}_i^{(j)}(\theta))^2$$

就本文的线性模型而言，多维输出不是本质的。因为每个分量均独立，所以只需单独对每个分量应用 GD，然后对其对应的损失函数值求平均即可。

1.3　损失函数其它形式

损失函数：

$$L(\theta)=E|\varepsilon(y,\hat{y}(\theta))|\sim\frac{1}{N}\sum_i|\epsilon(y_i,\hat{y}_i(\theta)|\tag{9}$$

称为（对称）平均绝对值百分误差[13]。此时，梯度为（已省略次要系数）：

$$\sim\sum_i\operatorname{sgn}(\hat{y}_i-y_i)\nabla\hat{y}_i(\theta)\begin{cases}\dfrac{y_i}{m_i^2}, & m_i>C\\ 1, & m_i\leqslant C\end{cases}$$

其中，符号函数：

$$\operatorname{sgn}(t)=\begin{cases}1, & t>0\\ -1, & t<0\\ 0, & t=0\end{cases}$$

等价于，假定 y 服从下述分布的极大似然估计。

$$y\sim e^{-C\left|\frac{\hat{y}-y}{\hat{y}+y}\right|}$$

通常可以考虑：

$$L(\theta)=E|\varepsilon(y,\hat{y}(\theta))|^p,\ p>0\tag{10}$$

不过对 GD 而言，$p=2$ 依然是首选。

2　算法与实验

本文算法利用 Python3.8 实现，运行于 MacOS10.15 上，程序设计遵从 scikit-learn 的 API 设计规范[14-15]。已在 Gitee 上公开了所用数据、源代码和运行结果，网址为 https://gitee.com/williamzjc/relinear。

2.1　算法

在此，虽然最小二乘法不再适用，但可以用来初始化参数。对于可微性较好的相对误差，采用梯度下降法；对于可微性较差的相对误差，采用遗传算法等智能算法。

算法设计采用 Adam 算法优化误差函数。基本处理过程概括如下：

（1）输入 x、y。

（2）可用最小二乘法初始化 θ 或随机初始化。

（3）根据式（8）计算梯度，并用 Adam 算法优化 $L(\theta)$，得到最终的 θ。

（4）预测测试数据。本算法还具有增量学习功能。学习未来产生的数据时，可以从当前存储的 θ 开始迭代，无需初始化。

2.2 实验

实验数据来自绍兴的一家印染工厂。输入变量 X 是样品在小缸中进行实验的染料浓度,输出变量 Y 是正式染色时的染料浓度。两者维度为 3,因此包含 3 个线性模型,每个模型有 4 个参数,总共 134 条数据。

数据被随机分成训练数据(80%) 和测试数据(20%)。损失函数选用式(7)来计算。将本文算法和多种常用的线性模型相关算法进行比较,测试重复进行 50 次,产生 50 份实验数据。最后取这 50 份实验数据的中位数作为最终的结果,并保留 4 位有效数字,实验结果见表 1。

表 1 数值实验报告

Tab. 1 Report on numerical experiments

算法/模型	测试误差	训练误差	训练耗时/ms
本文算法	0.027 14	0.026 41	~500
线性回归	0.310 1	0.292 9	2.98
脊回归	0.257 6	0.224 2	3.047
Bayes 脊回归	0.302 3	0.294 3	7.796
Huber 回归	0.095 44	0.089 44	70.98
Theil-Sen 回归	0.231 7	0.222 6	>1 000

对数线性模型测试结果见表 2。

表 2 对数线性模型的数值实验报告

Tab. 2 Report on numerical experiments for log-linear models

算法/模型	测试误差	训练误差	训练耗时/ms
本文算法	0.009 494	0.007 8	>1 000
线性回归	0.009 539	0.007 8	3.448
脊回归	0.009 779	0.007 96	3.489
Bayes 脊回归	0.009 543	0.007 8	8.904
Huber 回归	0.009 668	0.008 407	58.94
Theil-Sen 回归	0.042 18	0.034 71	>1 000

实验验证了本文算法的有效性。在线性模型中,其表现显著超越所有算法。但在对数化模型中,所有算法整体上表现相当,其中本文算法的训练误差稍优于其它算法。就本问题而言,对数化处理似乎非常有效,以至于一定程度上掩盖了本文算法的作用。为了降低误差,设置较高的迭代次数,同时也增加了训练时间。

3 结束语

在工业生产中,损失函数通常有实际意义,比如 y 值越小样本权重越大。本文最终选择式 (4) 和式 (7) 作为误差函数和损失函数。对称百分误差导出的偏态分布一定程度上接近真实数据的分布情况。

本文算法的核心是通过 GD 优化误差函数,通过实验充分证实了本文算法的有效性,可应用于工业领域。在精度方面显著高于其它算法,但是效率较低,有待提高。

未来工作主要寻找并研究其它可行的损失函数。损失函数与误差的分布是联系在一起的。因此,构造合理的误差分布也将是未来的任务之一。

参考文献

[1] 宋丛威, 张晓明. 基于 PCA 的解大型超定线性方程组快速算法及应用[J]. 智能计算机与应用, 2019, 9(4):91-95.

[2] HASTIE T, TIBSHIRANI R, FRIEDMAN J. The Elements of Statistical Learning[M]. Springer, 2009.

[3] HU G. ParametricEstimation and Prediction Theory in Linear Model[M]. Beijing Science Press, 2018.

[4] RAO C R, TOUTENBURG H, SHALABH, HEUMANN C. Linear Models and Generalizations: Least Squares and Alternatives [M], Springer-Verlag, 2008.

[5] CLEGER-TAMAYO S, FERNÁNDEZ-LUNA J M, HUETE J F. On the use of weighted mean absolute error in recommender systems[C]// Mathematical Statistics: Basic Ideas and Selected Topics Volume II, CRC Press, 2016: 24-26.

[6] HYNDMAN R J, KOEHLER A B. Another look at measures of forecast accuracy[J]. International journal of forecasting, 2006, 22(4): 679-688.

[7] MYTTENAERE, ARNAUD D et al. Mean absolute percentage error for regression models[J]. Neurocomputing, 2016(192): 38-48.

[8] PARK H, STEFANSKI L A. Relative error prediction[J]. Statist. and Probab. Letters, 1998, 40: 227-236.

[9] Symmetric mean absolute percentage error. https://en. wikipedia. org/wiki/Symmetric _ mean _ absolute _ percentage _ error [OL], 2020.

[10] Tofallis C. A better measure of relative prediction accuracy for model selection and model estimation [J], Journal of the Operational Research Society, 2015, 66(8): 1352-1362.

[11] GOODFELLOW I, BENGIO Y, COURVILLE A. 深度学习[M], 北京: 人民邮电出版社, 2017.

[12] KINGMA D, BA J. Adam: A Method for Stochastic Optimization [J]. Computer Science, 2014.

[13] Symmetric mean absolute percentage error. https://en. wikipedia. org/wiki/Symmetric _ mean _ absolute _ percentage _ error [OL], 2020.

[14] BUITINCK L, LOUPPE G, BLONDEL M. API design for machine learning software: experiences from the scikit - learn project [C]//ECML PKDD Workshop: Languages for Data Mining and Machine Learning, 2013:108-122.

[15] Scikit-learn. https://scikit-learn.org/stable[OL]. 2021.