计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.200400017

基于逻辑回归的金融风投评分卡模型实现

边玉宁　陆利坤　李业丽　曾庆涛　孙彦雄
北京印刷学院　北京 102600

摘　要　文以当前银行信贷业务中客户违约问题为出发点,将客户违约率和信贷评分卡分值的关系合理映射。运用逻辑回归建立评分卡预测模型,使用梯度下降算法来实现银行风险投资中客户评分卡的构建。首先加载数据并对数据进行分析,接着划分数据集,并使用跨时间验证集作为模型最后的验证。最后使用 KS 值和 AOC 曲线双向评价模型的稳定性。实验证明,采用所提方法构建的评分卡模型具有较好的稳定性。

关键词:金融风投;逻辑回归;机器学习;评分卡
中图法分类号　TP391

Implementation of Financial Venture Capital Score Card Model Based on Logistic Regression

BIAN Yu-ning,LU Li-kun,LI Ye-li,ZENG Qing-tao and SUN Yan-xiong
Beijing Institute of Graphic Communication,Beijing 102600,China

Abstract　This paper takes the problem of customer default in the current bank credit business as the starting point and maps the relationship between customer default rate and credit score card value reasonably. The logistic regression is used to build the prediction model of the score card and the gradient descent algorithm is used to construct the customer score card in the bank venture capital. The data is first loaded and analyzed,then the data set is partitioned and the cross-time validation set is used as the final validation of the model. Finally,KS value and AOC curve are used to evaluate the stability of the model. Experimental results show that the score card model constructed by the proposed method has good stability.

Keywords　Financial venture capital,Logistic regression,Machine learning,Score card

1　引言

现在是人工智能和信用经济时代,各大传统行业纷纷向人工智能转型。其中,以银行的信贷业务为例,加强对信贷业务的监管有利于银行业更快地融入人工智能的经济时代。银行互联网信贷是指利用现代网络技术,给不同的借贷者提供货币借贷的经济行为。在“银行+互联网”模式下,银行利用互联网电商提供的数据,结合自身的信用评价标准来对借贷者的信用等级进行评分。信贷投资是一个风险与挑战并存的业务,合理有效地规避信贷投资风险是我们需要考虑的问题。因此,稳定而准确的银行信贷评价体系是银行规避借贷风险的重要保障[1]。银行的信贷评分系统可帮助银行对借贷者进行信用等级划分,从而有助于银行尽早地预测出借贷者的偿还能力,以规避风险,减少损失。

2　逻辑回归算法与评分模型的实现

2.1　逻辑回归算法

逻辑回归算法是用以解决分类问题的一种常用算法。对于输入的数据,当这个数据大于我们的阈值时,输出 1,小于阈值时则输出 0。该模型的输出变量的范围始终在 0～1 之间。逻辑回归模型的假设函数为:

$$h_\theta(x)=g(\theta^{\mathrm{T}}\boldsymbol{X}) \tag{1}$$

其中,$\boldsymbol{X}$ 代表特征向量,$g(z)$代表逻辑函数。$g(z)$是一个常用的逻辑函数为 S 形的函数,表达式为 $g(z)=\frac{1}{1+e^{-z}}$。将这两个公式合起来,可得到逻辑回归模型的假设[2]。$h_\theta(x)$的作用是:对于给定的输入变量 x,根据选择的参数计算输出变量=1 的可能性,即 $h_\theta(x)=P(y=1|x;\theta)$。例如,对于输入的 x,通过给定的参数计算得出 $h_\theta(x)=0.65$,说明 y 是正向类的概率为 65%,相应的 y 是负向类的概率是 35%。

逻辑回归的代价函数为:$J(\theta)=\frac{1}{m}\sum_{i=1}^{m}Cost(h_\theta(x^{(i)}),y^{(i)})$,$y=1$ 时,$Cost(h_\theta(x),y)$的取值为 $-\log(h_\theta(x))$;$y=0$ 时,其取值为 $-\log(1-h_\theta(x))$。$h_\theta(x)$与 $Cost(h_\theta(x),y)$之间的关系如图 1 所示。

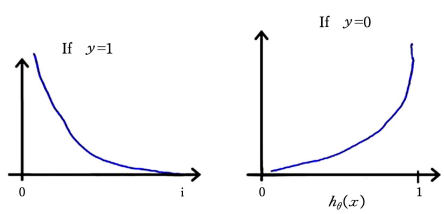

图 1　$h_\theta(x)$与 $Cost(h_\theta(x),y)$的关系

Fig. 1　Relationship between $h_\theta(x)$ and $Cost(h_\theta(x),y)$

基金项目:北京科技创新服务能力建设项目(PXM2016_014223_000025);广东省科技重大专项项目(190826175545233)
This work was supported by the Beijing Science and Technology Innovation Service Capacity Building Project(PXM2016_014223_000025) and Major Special Projects of Science and Technology in Guangdong Province(190826175545233).
通信作者:边玉宁(774200483@qq.com)

这样构建的代价函数的特点是，当实际上 $y=1$ 且 $h_\theta(x)$ 也为 1 时，误差为 0，当 $y=1$ 但 $h_\theta(x)$ 不为 1 时，误差随着 $h_\theta(x)$ 的变小而变大；当实际上 $y=0$ 且 $h_\theta(x)$ 也为 0 时，代价为 0，当 $y=0$ 但 $h_\theta(x)$ 不为 0 时，误差随着 $h_\theta(x)$ 的变大而变大。将构建的 $Cost(h_\theta(x),y)$ 化简如下：

$$Cost(h_\theta(x),y)=-y*\log(h_\theta(x))-(1-y)*\log(1-h_\theta(x)) \tag{2}$$

将其带入逻辑回归的代价函数中得到：

$$J(\theta)=-\frac{1}{m}\sum_{i=1}^{m}[y^{(i)}\log(h_\theta(x^{(i)}))+(1-y^{(i)})\log(1-h_\theta(x^{(i)}))] \tag{3}$$

在得到这样一个代价函数之后，我们使用梯度下降算法来求得能使代价函数最小的参数。

梯度下降法就是每次沿着下降速度最快的方向前进，多次迭代后到达曲线的最低点。为了求出式(3)中最小的参数 θ，每次计算时先随机地初始化一个值，接着迭代更新 θ 的值。更新过程为：$\theta_j=\theta_j-\alpha\frac{\phi}{\phi\theta_j}J(\theta)$（$\alpha$ 为学习率）[3]。

2.2 逻辑回归评分模型的映射

常见的信用评分卡分为 A、B、C 卡，这 3 种卡的主要区别在于各自发生的时间不同，因此模型中对应 y 的含义也不同。A 卡发生在贷前，用客户历史逾期天数的最大值表示；B 卡发生在贷中，用多期借款中逾期最长的一次来表示；C 卡发生在贷后。不同的用途下 y 的含义也不同[4]。

信贷业务评估的是客户的违约率（Percent of Default，PD），是[0，1]的概率，如 2%表示 100 个客户中有 2 个违约，简称 p。

评分卡中不直接用客户违约率 p，而是用违约概率与正常概率的比值，称为 Odds，即 $Odds=\frac{p}{1-p}$。评分卡的背后逻辑是 Odds 的变动与评分变动的映射（把 Odds 映射为评分），分值是根据 Odds 的前提条件算出来的，而不是人工取的。文献[5]中以单个客户在整张评分卡的得分的变动（如评分从 50 分上升到 70 分）来反映 Odds 的变动（如 Odds 从 5%下降至 1.25%），以及相对应的客户 PD 的变动（如 PD 从 4.8%下降到 1.2%）。从业务上来说，文献[6]中的 PD 可解释不强，而评分的可解释性就很强。评分卡与违约率的映射如表 1 所列。

表 1 评分卡与违约率的映射

Tabel 1 Mapping of scorecard and default rate

Score	Odds
80	0.625∶100
70	1.25∶100
60	2.5∶100
50	5∶100
40	10∶100
30	20∶100
20	40∶100
10	80∶100

3 逻辑回归模型的构建

本文使用的数据集来自 2018 年某互联网公司的金贷数据，如表 2 所列。拿到数据之后先对数据进行必要的脱敏，然后再来分析数据的特征。下文是对数据的初步探索。

表 2 数据集属性及部分数据

Tabel 2 Data set attributes and partial data

Obs_mth	Bad_ind	Uid	Td_score
2018-10-31	0.0	A1000005	0.675349
2018-07-31	0.0	A1000002	0.825269
2018-09-3	0.0	A1000011	0.315406
2018-07-31	0.0	A1000481	0.002386
2018-07-31	0.0	A1000069	0.406310

通过对数据的初步分析，我们得出以下结论：uid 是每个用户的唯一标识；bad_ind 用于判断一个用户是“好人”还是“坏人”，即式(3)中的 y；obs_mth 是每一行数据发生的时间。剩下的 10 个属性，我们可以通过模型中的特征来预测客户是坏人的风险。

待分析数据中共有 5 个月的数据，分别是 2018 年 6 月、2018 年 7 月、2018 年 9 月、2018 年 10 月和 2018 年 11 月的数据。我们将 2018 年 11 月的数据作为跨时间测试集，将剩下的前 4 个月的数据作为训练集来完成本次模型的构建[7]。

4 风投评分模型的评价

4.1 查准率(Precision)与查全率(Recall)

我们将算法预测的结果分成 4 种情况：

(1)正确肯定(True Positive，TP)：预测为真，实际为真；

(2)正确否定(True Negative，TN)：预测为假，实际为假；

(3)错误肯定(False Positive，FP)：预测为真，实际为假；

(4)错误否定(False Negative，FN)：预测为假，实际为真。

则查准率 $=TP/(TP+FP)$。例如，在我们预测会出现信贷问题的所有客户中，实际上出现信贷问题的人占的百分比越高越好。

查全率 $=TP/(TP+FN)$。例如，在所有实际上出现信贷问题的人中，成功预测出现信贷问题的人占的百分比越高越好。

4.2 KS 曲线和 ROC 曲线

KS(Kolmogorov-Smirnov)曲线：KS 用于对模型风险区分能力进行评估，即总体样本中好样本和坏样本之间的区分差异程度。好坏样本累计差异越大，KS 指标越高，则模型的风险区分能力越强。

KS 曲线的作图步骤：根据学习器的预测结果（注意，是正例的概率值，非 0/1 变量）对样本按从大到小的顺序进行排序——这就是截断点依次选取的顺序。按顺序选取截断点，并计算 TPR 和 FPR ——也可以只选取 n 个截断点，分别在 $1/n$，$2/n$，$3/n$ 等位置。横轴为样本的占比百分比（最大 100%），纵轴分别为 TPR 和 FPR，可以得到 KS 曲线[8-10]。

TPR 和 FPR 曲线分隔最开的位置就是最好的“截断点”，最大间隔距离就是 KS 值，通常 KS 大于 0.3 即可认为模型有较好的预测准确性。KS 值及其对应的含义如表 3 所列。

表 3 KS 值及其对应的含义

Table 3 KS value and its corresponding meaning

KS	Meaning
大于 0.3	prediction effect of the model is good
0.2～0.3	Models available
0～0.2	prediction effect of the model is not good
小于 0	Model error

受试者工作特征曲线（Receiver Operating Characteristic curve，ROC）又称为感受性曲线（sensitivity curve），由于曲线上各点反映着相同的感受性而得此名。它们都是对同一信号刺激的反应，只不过是在几种不同的判定标准下所得出的结果。接受者操作特性曲线是以假阳性概率（负例分错的概率）为横轴、查全率为纵轴组成的坐标图，和受试在特定刺激条件下因采用不同的判断标准而得出的不同结果所画出的曲线。横轴的计算公式为$\frac{FP}{FP+TN}$；纵轴的计算公式为$\frac{TP}{TP+FN}$。

传统的诊断试验评价方法有一个共同的特点，即必须将试验结果分为两类，再进行统计分析。ROC 曲线的评价方法与传统的评价方法不同，无须此限制，而是根据实际情况，允许有中间状态，可以把试验结果划分为多个有序分类，如正常、大致正常、可疑、大致异常和异常 5 个等级再进行统计分析。因此，ROC 曲线评价方法的适用范围更广。ROC 的作图步骤为：根据学习器的预测结果（即正例的概率值，非 0/1 变量）对样本进行排序（从大到小）——这就是截断点依次选取的顺序；接着按顺序选取截断点，并计算 TPR 和 FPR。也可以只选取 n 个截断点，分别在 $1/n$，$2/n$，$3/n$ 等位置；最后连接所有的点（TPR，FPR），即为 ROC 图[11-13]。

根据以上评价标准，带入数据集后计算得出 KS 的训练值为 0.41，KS 的测试值为 0.37。在预测 KS 和跨时间验证集相差 4 个百分点的情况下，ROC 曲线如图 2 所示。

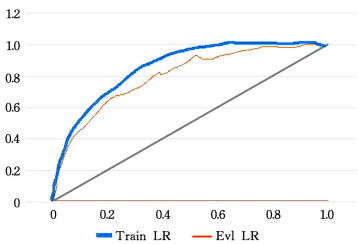

图 2 ROC 评价曲线

Fig. 2 ROC evaluation curve

结束语 本文以银行风险投资为背景，以某互联网公司的金贷真实数据集为基础做出了客户评分卡，用以预测借贷用户是否会出现信贷违约问题。使用梯度下降算法计算损失函数并使用 KS 值和 AOC 曲线双向评价来预测模型的稳定性[14]，为银行业预测投资风险提出了可靠的验证方法。

参 考 文 献

[1] YI S. Research on Internet credit risk early warning based on big data analysis [D]. Xianyang: Northwest A & F University, 2019.

[2] ZHOU H J. Research on risk management of credit card business of commercial banks [D]. Shanghai: Shanghai Jiaotong University, 2012.

[3] WANG Z C, XIAO Z J, YAN Z G. Research on Optimization of logistic regression classification identification [J]. Journal of Qilu University of technology, 2019(5): 47-51.

[4] LUO L G, PEI X J, HUANG R Q, et al. Landslide susceptibility by GIS based on certainty factor and logistic regression model in Jiuzhaigou scenic area[J/OL]. Journal of Engineering Geology, [2020-03-23]. https://doi.org/10.13544/j.cnki.jeg.2019-202.

[5] BAI J Y. Financial risk identification model based on classic score card and machine learning and its application [D]. Tianjin: Tianjin University of Commerce, 2019.

[6] TANG H. Campus Personalized Learning Resource Recommendation System Based on logistic regression model [J]. Electronic Technology and Software Engineering, 2019(22): 164-165.

[7] DU S M. Prediction of e-commerce users' repurchase behavior based on classification model [D]. Hangzhou: Hangzhou Normal University, 2019.

[8] PAPAKOSTA M A. Geographical variation in morphometry, craniometry, and diet of amammalian species (Stone marten, Martes foina) using data mining[J]. Turkish Journal of Zoology, 2018, 42: 99-106.

[9] THOMAS L. A Survey of Credit and Behavioural Scoring: Forecasting financial risk of lending to consumers[J]. International Journal of Forecasting, 2000, 16(2): 149-172.

[10] FAWCETT T. An introduction to ROC analysis [J]. Pattern Recognition Letters, 2006, 27(8): 861-874.

[11] GU G, HE Y. Research on the application of data mining to customer relationship management in the mobile communication industry[C]// IEEE International Conference on Computer Science and Information Technology. IEEE, 2010: 597-599.

[12] SOUZA J. Data Mining and Machine Learning to Promote Smart Cities: A Systematic Review from 2000 to 2018[J]. Sustainability, 2019, 11(4): 1077.

[13] PAPAKOSTA M A. Geographical variati on in morphometry, craniometry, and diet of amammalian species (Stone marten, Martes foina) using data mining[J]. Turkish Journal Ournal of Zoology, 2018, 42: 99-106.

[14] LEE E, LEE B. Herding Behavior in Online P2P Lending: An Empirical Investigation[J]. Electronic Commerce Research and Application, 2012, 11(5): 495-503.

BIAN Yu-ning, born in 1994, master. Her main research interests include data mining and so on.