古代玻璃制品的成分分析与鉴别

摘要

在我国史料中很早就有玻璃的相关记载，称作“璆琳琅玕”、“流离”等。由于最早的制作的玻璃是是玉的仿制品且易受埋藏环境的影响而风化，而在丝绸之路引入了大量西方玻璃制品，导致我国古代玻璃制品鉴别难度较大。要科学地考察中国古代玻璃的起源和发展，必须首先要经过科学鉴定，区分人工制造的玻璃、釉砂和玻砂，以及天然的玉石和宝石。本文根据出土玻璃的外观与表面化学成分特征，建立了科学的回归分类模型，并为相关研究者提供参考。

对于问题一，本文对运用附件 1 中玻璃样品的类型、纹饰和颜色等特征数据对表面风化状态进行回归分析，建立了 **Logistic 线性回归**模型，并结合神经网络**反向传输**的特性，利用**梯度下降**算法计算出最优拟合参数。首先对附件 1 中数据进行清洗并得到 54 条有效数据，再针对特征的特点进行标准化，得到样本特征矩阵。在此基础上建立 Logistic 线性回归模型

关键字： TOPSIS 法　熵权法　LSTM 模型　线性规划

一、 问题重述

玻璃的主要原料是石英砂，主要化学成分是二氧化硅（SiO_2）。由于纯石英砂的熔点较高，为了降低熔化温度，在炼制时需要添加助熔剂。古代常用的助熔剂有草木灰、天然泡碱、硝石和铅矿石等，并添加石灰石作为稳定剂，石灰石煅烧以后转化为氧化钙（CaO）。添加的助熔剂不同，其主要化学成分也不同。

古代玻璃极易受埋藏环境的影响而风化。在风化过程中，内部元素与环境元素进行大量交换，导致其成分比例发生变化，从而影响对其类别的正确判断。

现有一批我国古代玻璃制品的相关数据，考古工作者依据这些文物样品的化学成分和其他检测手段已将其分为高钾玻璃和铅钡玻璃两种类型。

请你们团队依据附件中的相关数据进行分析建模，解决以下问题：

- **问题 1:** 根据附件 1，分析这些玻璃文物的表面风化与其玻璃类型、纹饰和颜色的关系。再结合玻璃的类型，查阅附件 2，分析有无风化时化学成分含量的统计规律，并根据风化点检测数据，预测其风化前的化学成分的变化。
- **问题 2:** 首先分析高钾玻璃、铅钡玻璃的分类规律。再根据合适的化学成分对每类玻璃进行亚类划分，分析结果的合理性和敏感性。
- **问题 3:** 分析附件 3 中未知类别玻璃文物的化学成分，并鉴别其所属类型，分析结果的敏感性。
- **问题 4:** 针对划分的不同类别玻璃文物样品，分析同类别中样品化学成分之间的关联关系，再比较不同类别之间关联关系的差异性。

二、 问题分析

2.1 问题一

三、 符号说明

四、 数据预处理

4.1 附件 1

4.1.1 处理缺失值

首先，附件 1 中提供的某些文物数据在颜色项存在缺失，这可能是由于文物出土条件较差，导致颜色已经无法辨别。由于此缺失项无益于回归模型的建立，故去除所有包

含缺失值的条目（文物编号为 19、40、48、58）。

4.1.2 变量二值化

为了便于数学描述，本文构建了二值变量 $y^{(i)}$,$x_1^{(i)}$,$x_2^{(i)}$,$x_3^{(i)}$，以表示样品 i 的表面风化与否，类型和纹饰状态。

1. 风化状态

$$y^{(i)} = \begin{cases} 1, & \text{文物}i\text{ 风化;} \\ 0, & \text{文物}i\text{ 无风化.} \end{cases} \tag{1}$$

2. 类型

$$x_1^{(i)} = \begin{cases} 1, & \text{文物}i\text{ 样品类型为高钾玻璃;} \\ 0, & \text{文物}i\text{ 样品类型为铅钡玻璃.} \end{cases} \tag{2}$$

3. 纹饰

$$x_2^{(i)} = \begin{cases} 1, & \text{文物}i\text{ 纹饰为 A ;} \\ 0, & \text{文物}i\text{ 纹饰不为 A .} \end{cases} \tag{3}$$

$$x_3^{(i)} = \begin{cases} 1, & \text{文物}i\text{ 纹饰为 B;} \\ 0, & \text{文物}i\text{ 纹饰不为 B.} \end{cases} \tag{4}$$

文物 i 共有 3 三种纹饰状态，即 A、B、C，若设置三个二值变量，则会导致完全多重共线性，从而落入虚拟变量陷阱[1]。故本文使用上述两个二值变量以表示以上三种状态，例如，当文物 i 的纹饰为 C 时，此时 $x_2^{(i)}$、$x_2^{(i)}$ 即表示为 0。从附件 1 中 44 件文物样品的纹饰分布来看，纹饰为 C 的样品共有 28 件，较其他两种纹饰更多，因此可以将 $x_2^{(i)}$、$x_2^{(i)}$ 的初始值均设为 0，当纹饰为 A、C 时，则令其对应的二值变量取 1。

4.1.3 颜色 RGB 标准化

样品的颜色受到多种因素影响，呈现出的色彩各异。同样地，若两个样品的化学组分和结构相似，表现出的色彩差异也较小（如深绿与浅绿）。因此，直接将颜色当作独立的分类变量并不妥当，本文将色彩量化为 RGB 值以体现颜色之间的差异程度，并据此建立后文模型。

由于 RGB 数据量级与其他数据差距较大，因此对其进行正态标准化

$$x_j^{(i)} = \frac{X_j^{(i)} - \mu_j}{\sigma_j}. \tag{5}$$

其中，$X_j^{(i)}(j = 4, 5, 6)$ 分别对应文物 i 颜色 RGB 值的三个分量，μ_j 及 σ_j 为对应分量的期望和标准差。

表 **1**　**RGB** 值对应表

颜色	R	G	B
蓝绿	0	131	143
深绿	0	100	0
深蓝	18	18	110
浅蓝	135	206	250
浅绿	204	224	153
紫	143	7	131
绿	0	255	0
黑	0	0	0

$$\begin{gathered}\mu_j = \frac{1}{54}\sum_{i=1}^{54} X_j^{(i)};\\ \sigma_j = \sqrt{\frac{1}{53}\sum_{i=1}^{54}\left[X_j^{(i)} - \mu_j\right]^2};\\ \begin{pmatrix} x_4^{(i)} & x_5^{(i)} & x_6^{(i)} \end{pmatrix} = \begin{pmatrix} \frac{R^{(i)}-\mu_4}{\sigma_4} & \frac{G^{(i)}-\mu_5}{\sigma_5} & \frac{B^{(i)}-\mu_6}{\sigma_6} \end{pmatrix}\end{gathered} \tag{6}$$

五、（问题一）

5.1 模型的建立

题目要求根据附件 1，分析这些玻璃文物的表面风化与其玻璃类型、纹饰和颜色的关系。

为了解决玻璃文物样品表面风化与否的二分类问题，本文建立了基于深度学习的 Logistic 回归模型。

5.1.1 数据标准化

整理附件 1 中数据，将在4.1中得到的风化状态、类型、纹饰和颜色的 R、G、B 值，记入样本特征矩阵 $\boldsymbol{x}^{(i)}$

$$\boldsymbol{x}^{(i)} = \begin{pmatrix} x_1^{(i)} & x_2^{(i)} & x_3^{(i)} & x_4^{(i)} & x_5^{(i)} & x_6^{(i)} \end{pmatrix}. \tag{7}$$

5.1.2 建立多元线性回归模型

本文首先从多元线性回归出发，建立玻璃文物的表面风化与否，与其玻璃类型、纹饰和颜色关系的线性回归模型。一般地，由于多元线性回归得到的实数值不能直接进行二分类，则需引入阶跃函数 $f(x)$

$$f(x)=\begin{cases}0, & \beta_0+\sum_{j=1}^{6}\beta_j x_j^{(i)}\leqslant 0;\\ 1, & \beta_0+\sum_{j=1}^{6}\beta_j x_j^{(i)}>0.\end{cases} \tag{8}$$

其基本思想是在空间中构造一个合理的超平面，把空间区域划分为两个子空间，每一种类别数据都在平面的某一侧，通过查看结果的划分区域，得到二分类的结果。但此阶跃函数在 o 点不连续且不可导，为此我们引入了 Logistic 回归模型。

5.1.3 建立 Logistic 回归模型

Logistic 回归模型作为一种广义的线性回归模型，常被用于数据分析、预测等领域。为了建立 Logistic 回归模型，我们首先以5.1.1中的样本特征矩阵 $\boldsymbol{x}^{(i)}$，确立线性自变量 $z^{(i)}=\beta_0+\boldsymbol{\beta}x^{(i)}=\beta_0+\sum_{j=1}^{6}\beta_j x_j^{(i)}$。

区别于5.1.2中使用阶跃函数 $f(x)$，本模型借助阶跃函数的平滑版本，即 Sigmoid 函数，

$$a^{(i)}=f_{\boldsymbol{\beta}}\left(\boldsymbol{x}^{(i)}\right)=\frac{1}{1+e^{-z^{(i)}}}=\frac{1}{1+e^{-\left(\beta_0+\sum_{j=1}^{6}\beta_j x_j^{(i)}\right)}}. \tag{9}$$

其中，$\boldsymbol{\beta}=(\beta_1\beta_2\ldots\beta_6)^T$ 为连接权，即广义线性模型中自变量的参数。

通过 Sigmoid 函数将输入数据数据压缩至 0 至 1 之间，以确定输入数据属于铅钡玻璃或高钾玻璃（即 0 或 1）可能性。令 $y^{(i)}$ 服从 $\pi_i=f_{\boldsymbol{\beta}}\left(\boldsymbol{x}^{(i)}\right)$ 的 0–1 型分布，概率分布函数为

$$P\left\{y\mid \boldsymbol{x}^{(i)};\boldsymbol{\beta}\right\}=f_{\boldsymbol{\beta}}\left(\boldsymbol{x}^{(i)}\right)^{y^{(i)}}\left(1-f_{\boldsymbol{\beta}}\left(\boldsymbol{x}^{(i)}\right)\right)^{1-y^{(i)}},\quad y=0\text{ 或}1. \tag{10}$$

当 $y=0$ 时，$P\left\{y=0\mid \boldsymbol{x}^{(i)};\boldsymbol{\beta}\right\}=1-f_{\boldsymbol{\beta}}$，记为 $1-p$；当 $y=1$ 时，$P\left\{y=1\mid \boldsymbol{x}^{(i)};\boldsymbol{\beta}\right\}=f_{\boldsymbol{\beta}}$，记为 p。

因此，由上式可以得到

$$\ln\frac{p}{1-p}=\ln\frac{P\left\{y=1\mid \boldsymbol{x}^{(i)}\right\}}{P\left\{y=0\mid \boldsymbol{x}^{(i)}\right\}}=\beta_0+\sum_{j=1}^{6}\beta_j x_j^{(i)}. \tag{11}$$

上式即为 Logistic 回归模型，式左是与概率相关的对数值，故无法使用通常的最小二乘法拟合未知参数向量 $\boldsymbol{\beta}$，而应采用极大似然估计法。

5.1.4 构造似然函数

数据集中每个样本点都是相互独立的，则联合概率就是各样本点事件发生的概率乘积，故似然函数可以表示为

$$L(\boldsymbol{\beta})=\prod_{i=1}^{58}P\left\{y\mid \boldsymbol{x}^{(i)};\boldsymbol{\beta}\right\}=\prod_{i=1}^{58}\left[f_{\boldsymbol{\beta}}\left(\boldsymbol{x}^{(i)}\right)^{y^{(i)}}\left(1-f_{\boldsymbol{\beta}}\left(\boldsymbol{x}^{(i)}\right)\right)^{1-y^{(i)}}\right]. \tag{12}$$

对于此似然函数，无法求出当 $L(\boldsymbol{\beta})$ 最大时，拟合参数 $\boldsymbol{\beta}$ 的解析解。故转而尝试求其数值解，本文在这里采用基于最速下降法的神经网络算法对拟合参数进行求解。

5.2 模型的求解

神经网络的本质即为运筹，考虑到梯度下降算法常作为神经网络反向传播时使用的最优化算法，本文希望通过建立神经网络优化数值方法对于参数向量 $\boldsymbol{\beta}$ 的拟合效果.

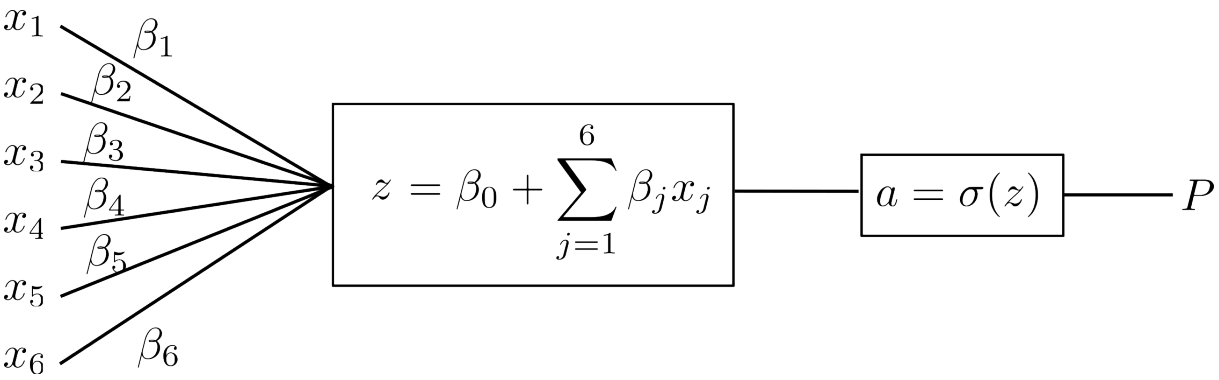

图 **1**　神经网络结构示意图

5.2.1 神经网络模型结构

搭建如图 1 所示的神经网络，需要注意的是，图中只显示了每个输入及其权重如何连接到输出，而隐去了偏置的值. 在图 1 所示的神经网络中，输入为 $\boldsymbol{x}^{(i)}$，因此输入层中的输入数（或称为特征维度）为 6. 网络的输出为 P，预测文物 i 表面风化的可能性，因此输出层的输出数是 1. 特别地，计算神经元只有一个，不仅为单层神经网络，且为全连接层.

5.2.2 连接权

$$\boldsymbol{\beta}=\begin{pmatrix}\beta_0 & \beta_1 & \cdots & \beta_6\end{pmatrix}^{\mathrm{T}}. \tag{13}$$

其不仅在广义线性回归模型中作为线性自变量（预测部分）的系数，也为神经网络中输入的权重，是需要拟合的参数向量. 在神经网络中，也作为最优化算法的决策变量，使得目标函数交叉熵取得最小值.

5.2.3 网络输入

$$z^{(i)} = \beta_0 + \sum_{j=1}^{6} \beta_j x_j^{(i)}. \tag{14}$$

在广义线性回归模型中作为线性自变量，也称为线性预测部分. 在深度学习中，本质是一个多元函数，将 $\boldsymbol{x^{(i)}}$ 和 $\boldsymbol{\beta}$ 经过线性组合降维成实数 $z^{(i)}$ 传入激活函数.

5.2.4 激活函数

激活函数也称激励函数、活化函数，用来执行对神经元所获得的网络输入的变换，S 形函数是常见的一种. 其为一元函数，可以将输入的自变量为 $z^{(i)}$ 数据压缩至 0 到 1 的范围内. 公式为

$$a^{(i)} = \sigma(z^{(i)}) = \frac{1}{1+e^{-z}} = \frac{1}{1+e^{-\left(\beta_0+\sum_{j=1}^{6}\beta_j x_j^{(i)}\right)}} = f_{\boldsymbol{\beta}}(\boldsymbol{x^{(i)}}). \tag{15}$$

可知 Logistic 回归中的 Sigmoid 函数即为 σ 和 z 的复合函数.

5.2.5 损失函数

回忆化学中关于酸碱度 pH 值的概念，算子 p 表示取对数值的相反数. 令损失函数

$$\mathcal{L}(\boldsymbol{\beta}) = \mathrm{p}L(\boldsymbol{\beta}) = -\sum_{i=1}^{58}[y^{(i)} \ln f_{\boldsymbol{\beta}}(\boldsymbol{x^{(i)}}) + (1-y^{(i)}) \ln(1 - f_{\boldsymbol{\beta}}(\boldsymbol{x^{(i)}}))]. \tag{16}$$

恰好是二分类问题常用的交叉熵损失函数. 由此可见，似然函数和损失函数具有同一性，最值同时取得.

5.2.6 最速下降法

最速下降法以负梯度方向作为极小化算法的下降方向，因此需要计算损失函数关于决策变量的偏导数.

后文使用复合函数的链式求导法则，先给出

$$\frac{\mathrm{d}\mathcal{L}}{\mathrm{d}a^{(i)}} = -\frac{y^{(i)}}{a^{(i)}} + \frac{1-y^{(i)}}{1-a^{(i)}}. \tag{17}$$

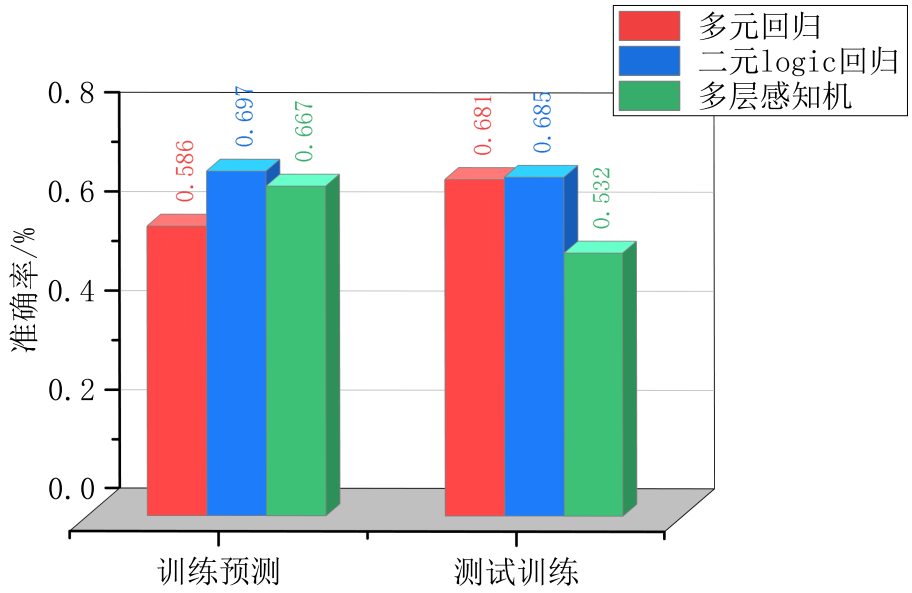

图 **2**　准确率对比

进而在每一次神经网络训练时，计算

$$\frac{\partial \mathcal{L}}{\partial \beta_j} = x_j^{(i)} \frac{\mathrm{d}\mathcal{L}}{\mathrm{d}z^{(i)}} = x_j^{(i)} \frac{\mathrm{d}\mathcal{L}}{\mathrm{d}a^{(i)}} \frac{\mathrm{d}a^{(i)}}{\mathrm{d}z^{(i)}} = x_j^{(i)}(a^{(i)} - y^{(i)}). \tag{18}$$

并更新决策变量

$$\beta_j := \beta_j - \alpha \frac{\partial \mathcal{L}}{\partial \beta_j}, \quad \text{其中}\alpha \text{ 为学习率.} \tag{19}$$

反向传播各个样本每次训练时进行一次，达到循环次数后结束，并进行神经网络的测试，通过后投入预测使用.

5.2.7 模型评价

表 **2**　方程中的变量

	玻璃类型	纹饰 A	纹饰 B	R	G	B	常量
β	10.955	-0.621	13.097	1.470	-0.286	-1.372	-1.834

六、 统计规律

题目 1 中还要求结合玻璃的类型，分析有无风化时化学成分含量的统计规律，并预测其风化前的化学成分的变化。

为了描述附件 2 中，在经受风化和未经受风化下，其表面化学成分含量呈现的规律，首先需要考察风化对表面化学成分的影响，即哪些表面化学成分在风化前后产生了显著差异。因此，本文对附件 2 中风化后取样点的 14 种化学成分与风化前取样点进行多因素方差分析。

6.1 多因素方差分析

方差分析常用于检验多种样品均值的差异是否具备统计学意义，在研究和生产中具备广泛的应用

Algorithm 1 Long short-term memory for out model.

Input: 从原始数据中获取原始的订购与供应数据; 从原始数据中挑取前一步筛选的 44 家供应商信息;

Output: 第 i 周第 j 家供应商的预测值 ψ_{ij};

1: 对数据的 A、B、C 分类，做均一化预处理;
2: 将数据按供应商拆分为一维时间序列数据;
3: 将一维时间序列数据先按时间窗口切分为二维数据，以同样的方法，再将每个时间窗口中的数据切分为特征向量;
4: 对升维后的数据做一次正态标准化缩放，保存缩放参数;
5: 将处理后的数据送入 LSTM 网络，训练完毕后，返回神经网络;
6: 每次预测 1 周，将预测的得到的 1 周数据重新送入神经网络，迭代指定此时以得到相应的预测步数;
7: 使用保存的缩放数据，将预测数据复原。
8: **return** ψ_{ij}.

七、（问题二）

参考文献

[1] 伍德里奇. 计量经济学导论: 现代观点[M]. 北京: 清华大学出版社, 2014: 185.

附录 A　源程序

1.1　ELOL.py