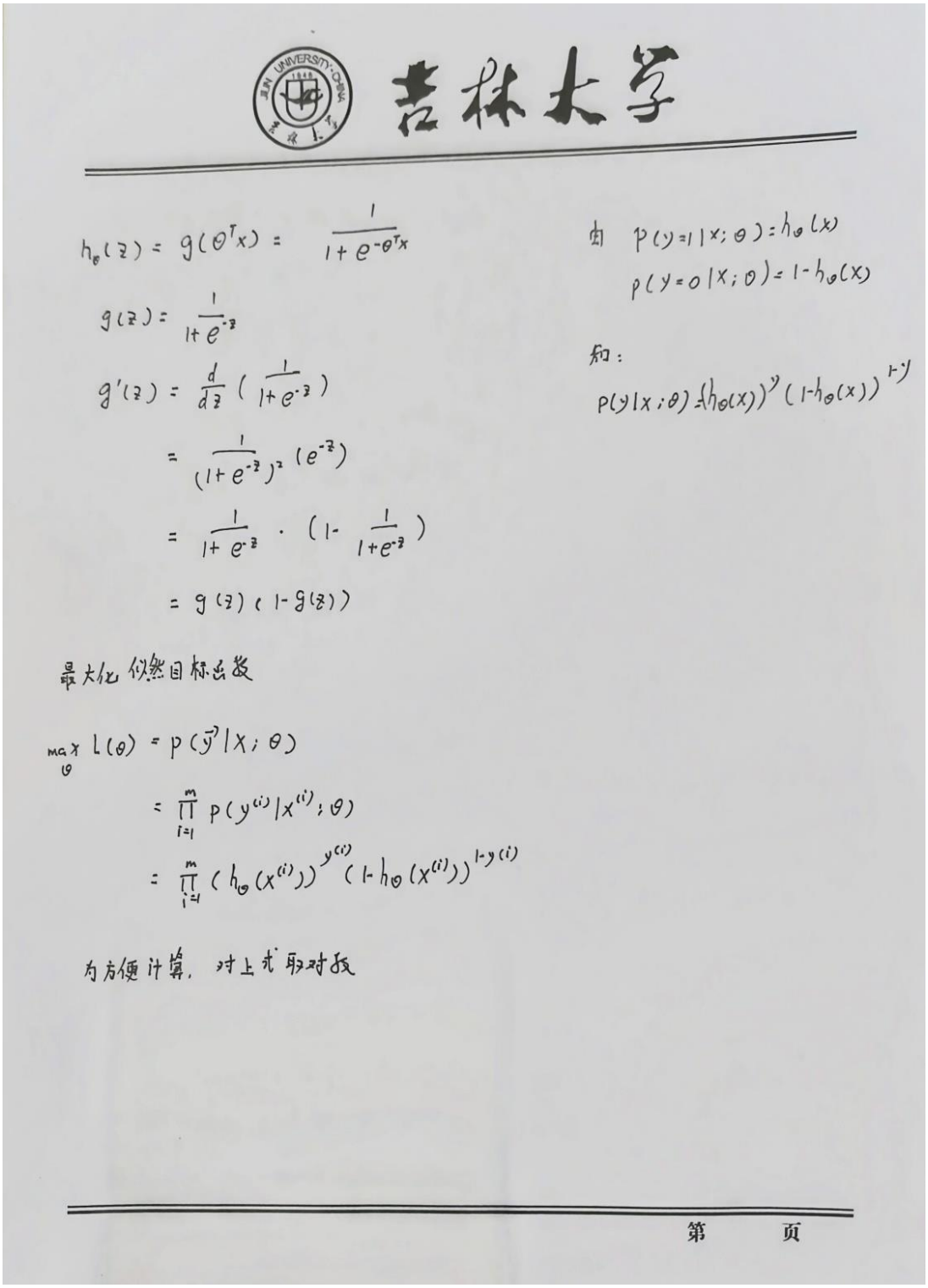
使用逻辑回归的方法实现垃圾邮件分类任务

1 理论分析

似然函数及参数的更新公式推导:



吉林大学

$h_\theta(z) = g(\theta^T x) = \frac{1}{1+e^{-\theta^T x}}$

$g(z) = \frac{1}{1+e^{-z}}$

$g'(z) = \frac{d}{dz}\left(\frac{1}{1+e^{-z}}\right)$

$= \frac{1}{(1+e^{-z})^2}(e^{-z})$

$= \frac{1}{1+e^{-z}} \cdot \left(1 - \frac{1}{1+e^{-z}}\right)$

$= g(z)(1-g(z))$

由 $p(y=1|x;\theta) = h_\theta(x)$

$p(y=0|x;\theta) = 1-h_\theta(x)$

和:

$p(y|x;\theta) = (h_\theta(x))^y (1-h_\theta(x))^{1-y}$

最大化似然目标出发

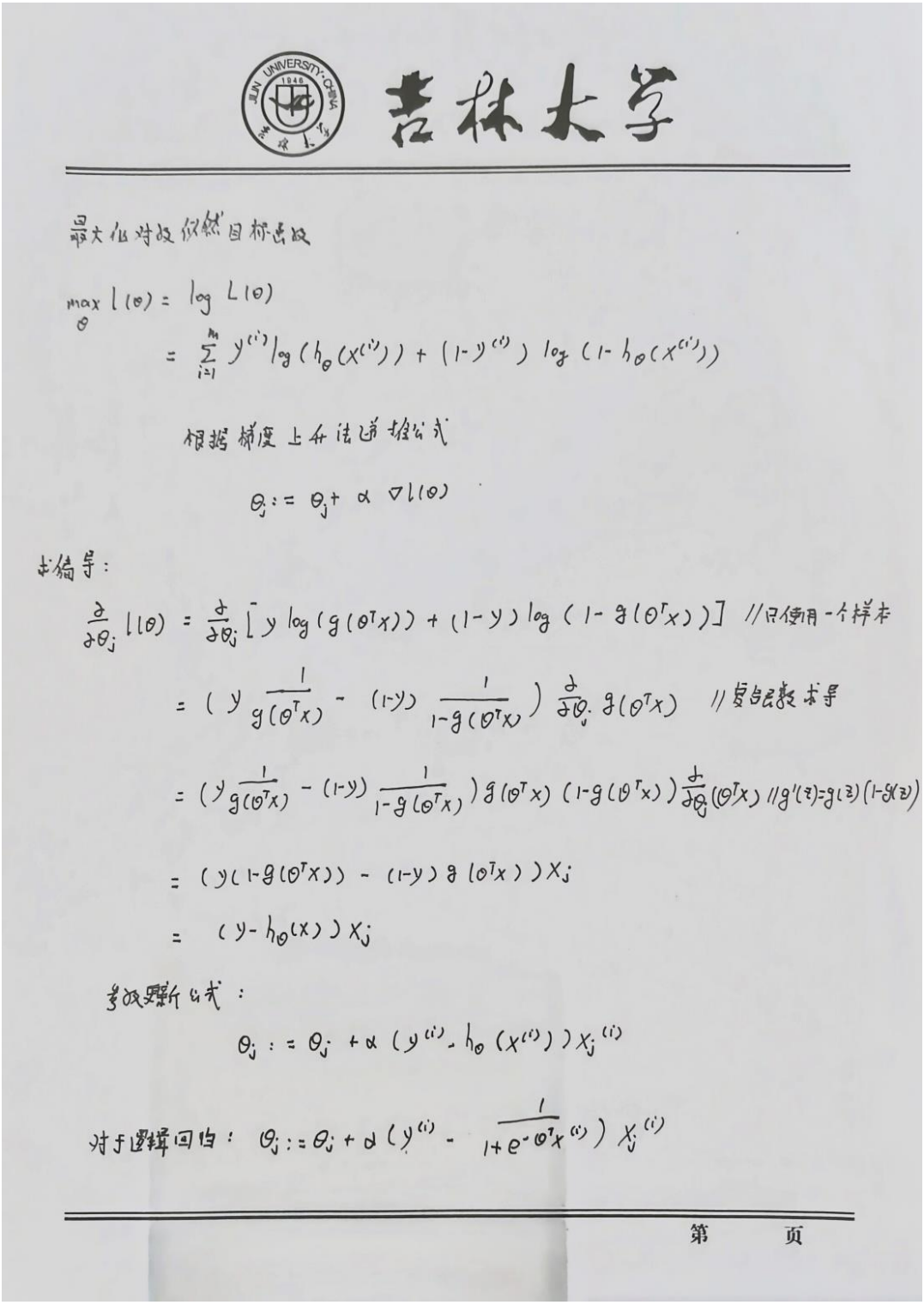
$\max_\theta L(\theta) = p(\vec{y}|X;\theta)$

$= \prod_{i=1}^{m} p(y^{(i)}|x^{(i)};\theta)$

$= \prod_{i=1}^{m} (h_\theta(x^{(i)}))^{y^{(i)}} (1-h_\theta(x^{(i)}))^{1-y^{(i)}}$

为方便计算，对上式取对数

第 页



2 数据准备

（1）数据集格式：总的数据集一共有 4458 条数据，将按照 8:2 进行划分训练集和验证集。数据集的格式有三列分别是 ID，Label(pam、spam)，Email。

（2）数据清洗：去掉停用词、去掉 URL、去掉 HTML 标签、去掉特殊符号、去掉表情符号、去掉长重复字、将缩写补全、去掉单字、提取词干等等。

（3）词袋模型：使用 CountVectorizer 将文本数据转换为词袋模型。词袋模型是一种简单而强大的文本表示方法，它将文本转换为单词出现次数的向量。
（4）TF-IDF 转换：使用 TfidfTransformer 将词袋模型转换为 TF-IDF 矩阵。TF-IDF 是一种统计方法，用以评估一个词语对于一个文件集或一个语料库中的其中一份文件的重要程度。
（5）标签编码：使用 LabelEncoder 对标签进行编码，将文本标签转换为数值。这是为了让模型能够处理分类问题。

3 模型实现

使用 sklearn 中的 LogisticRegression 模型进行训练，计算以下指标用于评估模型：
（1）准确率（LR score）：使用模型在验证集上计算准确率，这是模型预测正确的样本数与总样本数的比例。
（2）ROC 曲线：绘制接收者操作特征曲线（ROC Curve），它是一个用于选择最佳模型的图形，显示了在不同阈值设置下的假正例率（FPR）和真正例率（TPR）。
（3）AUC 值：计算 ROC 曲线下的面积（AUC），AUC 值越接近 1，模型的分类性能越好。
（4）混淆矩阵：计算混淆矩阵，它是一个实际类别与预测类别的矩阵，可以进一步分析模型的性能，如真正例、假正例、真负例和假负例。

4 性能评估

准确率

LR score: 0.9596412556053812

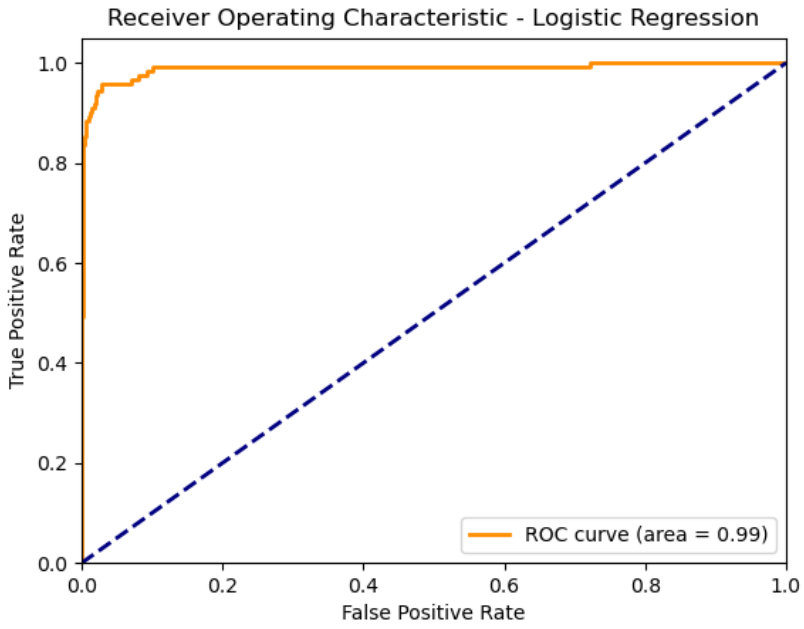
LR score 表示模型在验证集上的准确率为 95.96%。这意味着模型正确分类了大约 95.96%的邮件。

ROC 曲线

ROC 曲线是一种用于展示二分类系统性能的工具，它显示了在不同阈值设置下模型的真正例率（TPR）和假正例率（FPR）之间的关系。
（1）真正例率（TPR）：是模型正确预测为正类的样本数与实际为正类的样本总数的比例。计算公式：TPR=TP/（TP+FN），其中 TP 是真正例的数量，FN 是假负例的数量。
（2）假正例率（FPR）：是模型错误预测为正类的样本数与实际为负类的样本总数的比例。计算公式：FPR=FP/（FP+TN），其中 FP 是假正例的数量，TN 是真负例的数量。
（3）绘制 ROC 曲线：

- ROC 曲线的横轴是 FPR，纵轴是 TPR。
- 不同的阈值对应不同的点，将这些点连接起来就形成了 ROC 曲线。
- 曲线下面积（AUC）是衡量模型好坏的一个重要指标，AUC 值越高，模型性能越好。

绘制的 ROC 曲线如下图：



ROC 曲线非常接近左上角，这意味着在不同的阈值下，模型能够很好地平衡真正例率（TPR）和假正例率（FPR）。即使在保持高真正例率的同时，假正例率也很低。

AUC 值

LR AUC: 0.9890887800723865

LR AUC 表示模型的 AUC 值为 0.989，这是一个非常高的值。AUC 值范围从 0 到 1，接近 1 表示模型具有很好的区分能力，能够很好地区分垃圾邮件和非垃圾邮件。

混淆矩阵：
LR confusion: [[769 1] [35 87]]
表示混淆矩阵的结果如下：
真正例（TP）：87，即模型正确识别为垃圾邮件的垃圾邮件数量。
假正例（FP）：1，即模型错误地识别为垃圾邮件的正常邮件数量。
真负例（TN）：769，即模型正确识别为正常邮件的正常邮件数量。
假负例（FN）：35，即模型错误地识别为正常邮件的垃圾邮件数量。

5 结论总结

上述这些结果表明：

（1）模型具有很高的准确率，能够正确分类绝大多数邮件。
（2）ROC 曲线非常接近左上角，在不同的阈值下，模型能够很好地平衡真正例率和假正例率。
（3）AUC 值非常高，说明模型在不同阈值下的表现都很好，具有很好的区分能力。

（4）混淆矩阵显示模型在预测垃圾邮件时的假负例相对较少，而真负例非常多，这表明模型在识别正常邮件方面非常有效。

总的来说，逻辑回归模型在这个垃圾邮件分类任务上表现非常好，具有很高的准确率和区分能力。

附：代码地址 https://github.com/JLU-KDZ/MachineLearning4.git

参考：https://github.com/ljx02/Spam_Email_Classificaton