

数理统计大作业

|  |  |
| --- | --- |
| 学生姓名 | 周奕龙 |
| 学号 | ZB2423365 |
| 班级 | ZB24233 |
| 任课教师 | 李文玲 |

目录

[1. 背景与问题定义 3](#_Toc184462266)

[2. 数据收集与整理 3](#_Toc184462267)

[2.1 数据基本情况 3](#_Toc184462268)

[2.2 数据预处理 4](#_Toc184462269)

[3. 模型构建 9](#_Toc184462270)

[3.1 模型要素 9](#_Toc184462271)

[3.2 变分推断 10](#_Toc184462272)

[4. 实现过程 10](#_Toc184462273)

[4.1 模型代码 10](#_Toc184462274)

[4.2 评价指标 12](#_Toc184462275)

[5. 结果分析 13](#_Toc184462276)

[5.1 结果展示 13](#_Toc184462277)

[5.2 分析与结论 16](#_Toc184462278)

[6. 参考文献 17](#_Toc184462279)

## 背景与问题定义

传统的收入预测模型通常依赖于一系列简单的规则和假设，如个人的工作类型、教育水平、性别和年龄等。这些传统方法往往认为，较高的教育水平、稳定的工作职位和较多的工作经验是带来高收入的重要因素，但往往忽视了其他因素带来的潜在影响。随着数理统计方法与机器学习的发展，越来越多的复杂模型被应用于收入预测领域，这些模型能够更好地处理不确定性并捕捉变量之间的复杂关系。

在过去的研究中，贝叶斯方法已被广泛应用于分类问题。特别是当我们处理具有不确定性和多重因素的复杂数据时，贝叶斯模型通过其先验分布、似然函数以及后验分布的推导，能够有效地处理数据中的不确定性，从而做出更为准确的预测。

本次报告所研究的问题如下：通过贝叶斯模型来分析并预测1994年美国居民的当年收入是否超过了50000美元。通过本研究，我们希望能够构建一个有效的贝叶斯模型，并在实际应用中提供较为精准的分类结果，以展示贝叶斯模型强大的推理能力，并为未来在社会经济数据分析中应用贝叶斯模型提供有益的经验和指导。

## 数据收集与整理

### 数据基本情况

本次报告使用的是来自Kaggle平台的“Adult Census Income”数据集[1]，相关链接见页脚。该数据集是从1994年美国人口普查局数据库中所提取，并且使用了以下几个约束条件：((AAGE>16) && (AGI>100) && (AFNLWGT>1) && (HRSWK>0))，其相关解释如下：

1. AAGE > 16：即样本的年龄超过16周岁，有一定的收入能力；
2. HRSWK > 0：每周工作时长大于0的公民，即拥有具体的工作，能够取得相应劳动报酬；

但对于AGI和AFNLWGT两项指标在Kaggle平台以及源数据库中并未找到相关的含义说明。

数据集中共有32561条数据，15项变量，其中包含因变量1个（income），自变量14个，涵盖年龄、教育背景、工作类型、婚姻状况、种族、性别等可能直接或间接影响个人收入水平的因素。全部变量以及相关说明如下：

1. age：每个样本的工作年龄，数值型；
2. workclass：样本的工作性质，这里有私人的、当地政府等，文本型；
3. fnlwgt：直译为最终权重，但其实是指在一个州内，dataset的一个观测代表的人数，表示最终结果所占的权重，数值型；
4. education：每个样本的教育程度，包含从幼儿园到博士共16种受教育程度，文本型；
5. education.num：每个样本的教育程度的编码，数值型；
6. marital.status：每个样本的婚姻状况，包括未婚、已婚、分居等7种婚姻状态，文本型；
7. occupation：表示每个样本从事的职业，共15种，文本型；
8. relationship：每个样本所处的家庭状态，共6种，文本型；
9. race：样本的种族，有黑人和白人，共5种，文本型；
10. sex：样本的性别，男/女，文本型；
11. capital.gain：样本的资本收益，数值型；
12. capital.loss：样本的资本损失，数值型；
13. hours.per.week：样本每周工作时长，数值型；
14. native.country：表示样本的国籍，文本型；
15. income：样本的收入，仅区分大于50K和小于等于50K，数值型（0/1）；

### 数据预处理

#### 数据检查

在前期的数据检查过程中，有如下发现：

1. 见下图1和图2，在"occupation"、"workclass"、"native.country"这三种变量取值中包含空值（即缺省的"?"），且在对应列中占比分别为5.7%，5.6%和1.8%，仅占样本总数的小部分；



图1 数据一览

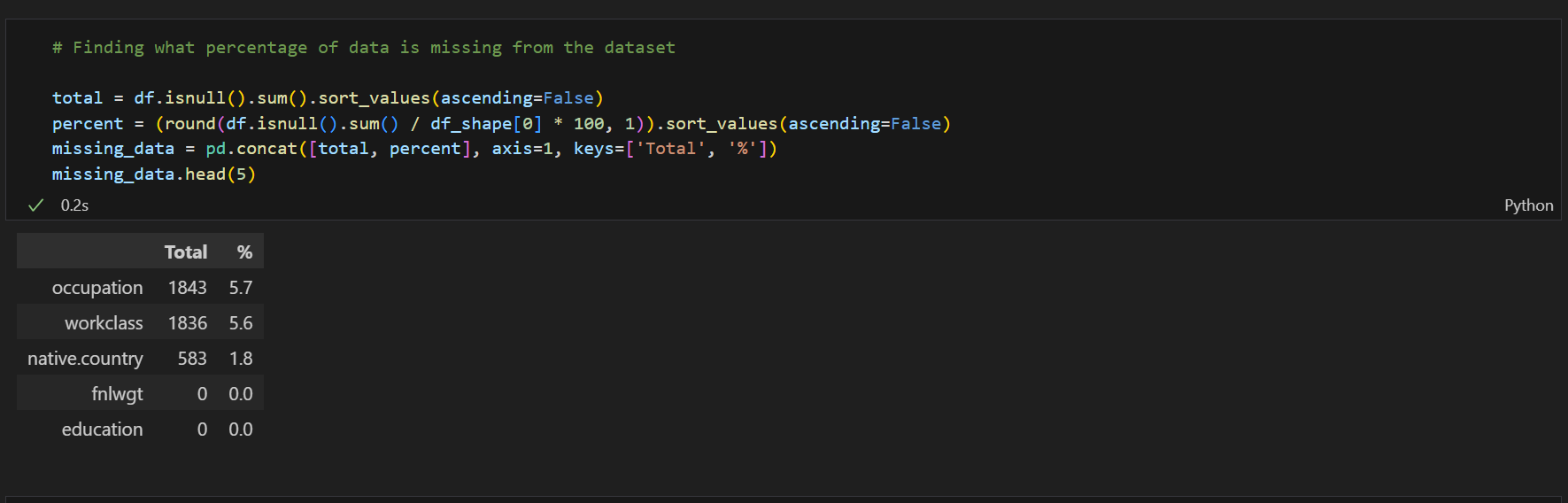


图2 数据空值占比

1. 在变量中涉及资本收益（capital.gain）和资本损失（capital.loss），都是描述样本当年资本获益情况，属于意义相近的变量。同理受教育程度education和education.num意义相同；
2. 多数变量取值类型为文本型，且观察发现不同的类别之间没有明显的顺序信息，若要使用贝叶斯模型进行推导，需要重新编码为离散的数值型变量；
3. 通过Python的value\_counts函数发现，部分文本变量的取值种类繁多，如受教育程度"education"有16种，职业"occupation"有15种。大量的变量取值种类可能会导致贝叶斯模型内部的状态繁多，使模型收敛速度变慢；但是相对的，若采用区间归纳方法，可能会丢失变量原本的信息。因此特地设计了相关的实验部分，检验对多取值变量区间归纳前后的效果。

#### 数据清洗

该步骤主要对在数据检查时发现的问题进行清洗，主要包括：

1. **空值填充**：缺省的空值仅占总数的小部分，使用众数填充；
2. **相近变量归纳**：为减小模型迭代收敛的难度，应相应减少变量个数，将意义相近的变量进行合并。对于描述样本资本获益情况的变量合并为资本差值capital.diff = capital.gain - capital.loss，描述样本受教育程度的education和education.num表示含义一样，舍弃education.num列方便后续进行变量区间归纳；
3. **标签编码**：将每个文本型变量的取值映射到整数值，从0开始递增，适配贝叶斯模型，且后续不需要进行归一化操作；
4. **变量取舍**：使用Python计算变量间相关性的结果如图3所示，观察到权重"fnlwgt"和国籍"native.country"两个变量对因变量"income"相关性为0，说明对最终结果没有影响，进行舍弃。同理工作性质"workclass"和种族"race"对结果的相关性不高，同样舍弃；

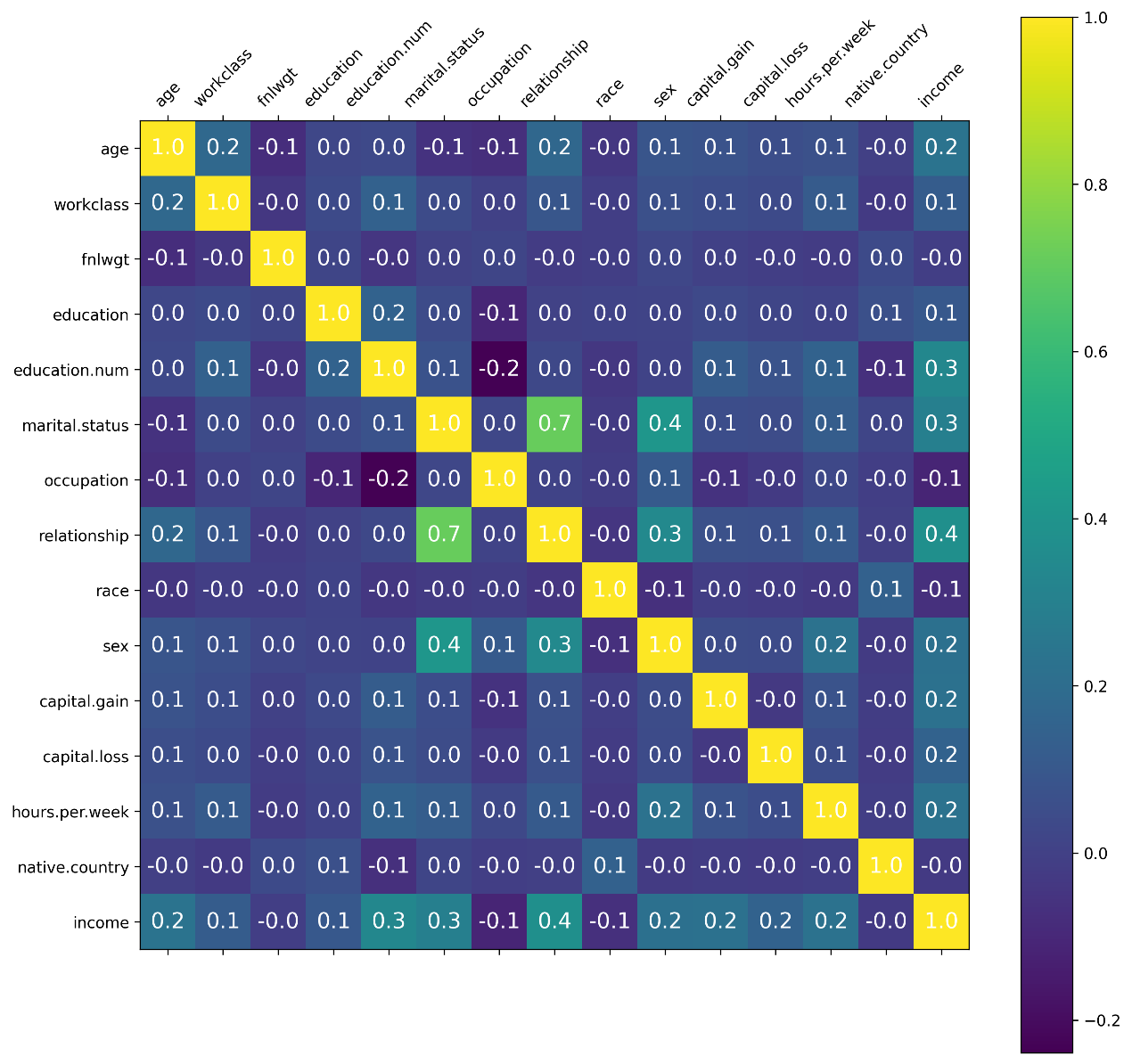


图3 变量相关性分析图

1. **区间归纳**：在数据检查的第四点中我们提到，部分变量的取值种类繁多，可能不利于贝叶斯模型的状态推导，因此设计实验检验区间归纳的必要性。对照组设置为不进行区间归纳，维持原本取值种类，直接进行标签编码。实验组即进行区间归纳，归纳后进行标签编码。归纳类别如下表所示，其中前三个变量为离散型变量，后三个变量为连续型变量：

表1 区间归纳取值

|  |  |  |
| --- | --- | --- |
| 列名 | 取值类别 | 归纳类别 |
| occupation | Craft-repair工艺与修理 | Tech |
| Machine-op-inspct机器操作与检查 |
| Tech-support技术支持 |
| Other-service其他服务 | Service |
| Protective-serv安全服务 |
| Priv-house-serv私人家庭服务 |
| Sales销售 |
| Exec-managerial高级管理人员 | Office |
| Adm-clerical行政文员 |
| Prof-specialty专业人员 |
| Farming-fishing农渔业 | Labor |
| Transport-moving运输与搬运 |
| Handlers-cleaners清洁工与搬运工 |
| Armed-Forces武装力量 |
| workclass | Local-gov地方政府 | Gov |
| State-gov州政府 |
| Federal-gov联邦政府 |
| Without-pay无薪水 | No\_Pay |
| Never-worked从未工作 |
| Self-emp-not-inc非股份公司员工 | Self-emp |
| Self-emp-inc股份公司员工 |
| race | White白人 | White |
| Black黑人 | Black |
| Amer-Indian-Eskimo | Other |
| Asian-Pac-Islander |
| Other其他 |
| age | 16 ~ 25 | Young |
| 26 ~ 65 | Working\_Age |
| 66 ~ 100 | Old |
| capital diff | -5000 ~ 5000 | Minor |
| 5000 ~ 10000 | Major |
| hours per week | 0 ~ 30 | Lesser Hours |
| 30 ~ 40 | Normal Hours |
| 40 ~ 100 | Extra Hours |

1. **数据归一化**：在实验组中我们将所有变量区间归纳，不需考虑数据归一化。但在对照组中对于连续型变量，需要进行数据归一化处理，避免状态间的权重不均。在本次项目中我们假设数据服从正态分布，使用样本的均值和方差进行Z-score标准化，即

#### 数据集划分

使用了Python的sklearn库中的train\_test\_split函数，设置测试集占20%的比例，随机种子取为42，保证实验结果的可重复性；

在数据清洗完成后，得到的结果如下所示，分为是否进行区间归纳两种情况进行展示：

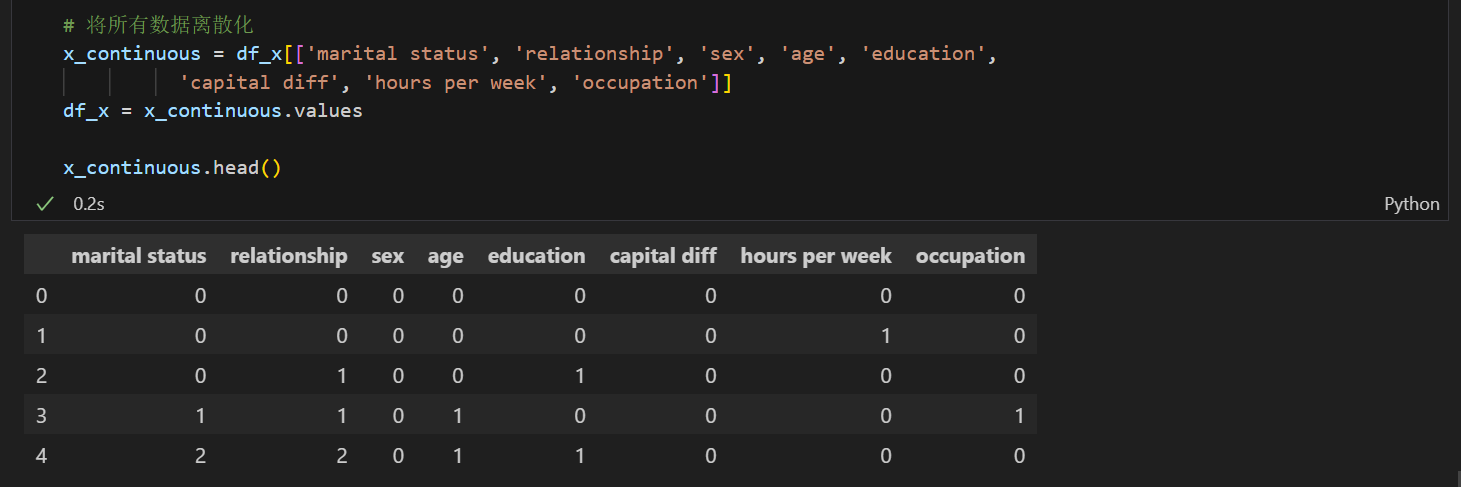


图4 区间归纳后数据清洗结果

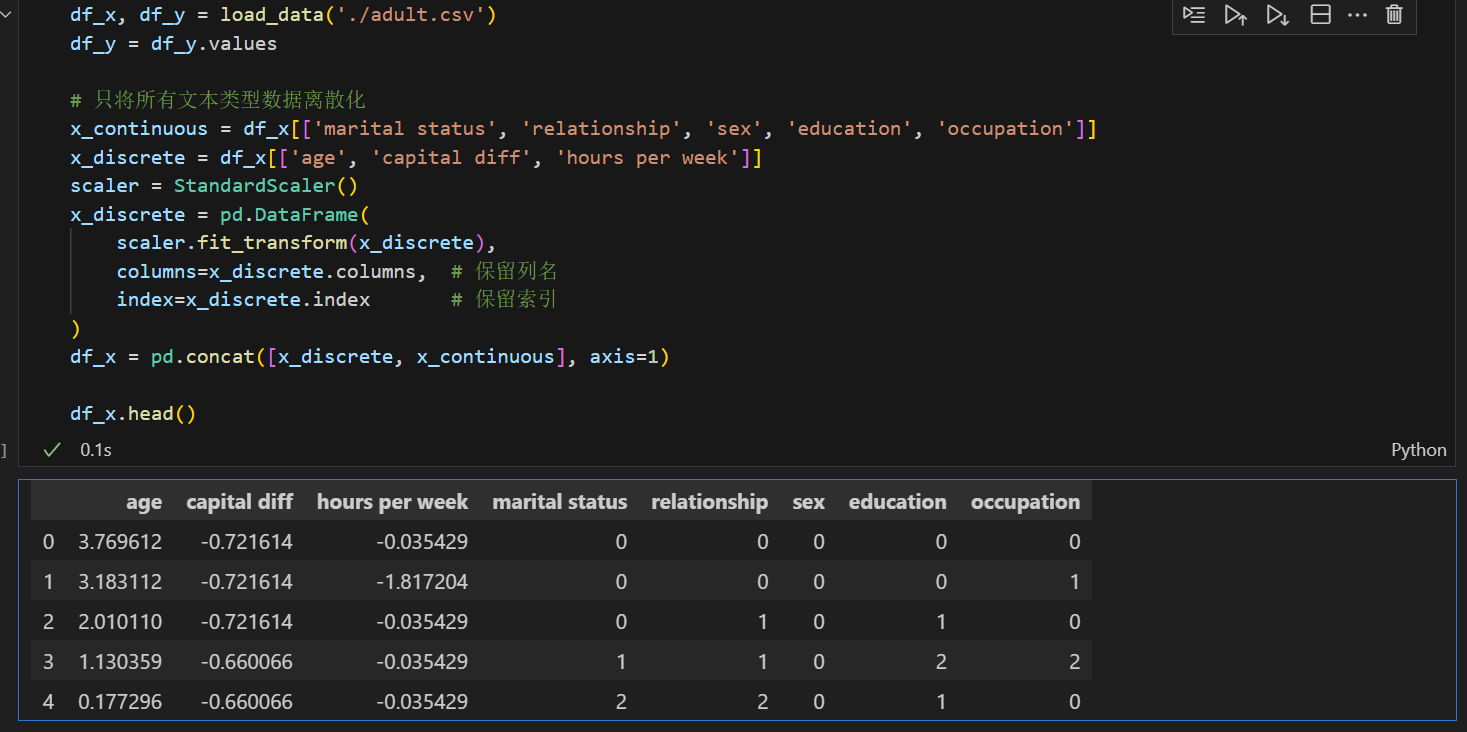


图5 非区间归纳数据清洗结果

## 模型构建

### 模型要素

根据前面数据清洗的结果，最终得到的输入数据共有8维，数据总量32561条，训练集26048条占比80%，测试集6513条占比20%。该问题的线性回归模型应该为：

其中为回归系数，为截距。

现构建贝叶斯模型，主要包括先验分布、条件概率和后验分布三部分。

先验分布：表示我们对模型参数（即回归系数和截距项 intercept）在观察到数据之前的知识或假设。具体来讲，我们分别用每个特征的均值和标准差作为回归系数 的均值和标准差，设定如下：。截距项intercept的先验分布假设为服从一个均值为0，标准差为10的正态分布，即。

条件概率：条件概率是指在给定模型参数的情况下，观测数据的概率。在逻辑回归中，给定回归系数和截距项 intercept后，使用Sigmoid函数作为激活函数，则每个样本的预测值为：

在本问题中，预测年收入是否超过50000美元属于二分类问题，我们认为观测值符合伯努利分布。

后验分布：先验分布反映在抽样前对的认识，而后验分布是总体和样本对先验分布作调整的结果，反映抽样后对的认识，其在本任务数据离散形式时的公式如下：

在给定训练数据后，我们使用PyMC库中的自动变分推断（ADVI, Automatic Differentiation Variational Inference）不断近似拟合。

### 变分推断

当模型变得复杂或数据量增大时，像贝叶斯推断这种精确计算后验概率分布往往变得不可行，因为这需要对高维空间进行积分，这在计算上是非常昂贵甚至不可行的。进而产生了近似推断方法，如基于Markov链的Monte-Carlo（MCMC）方法和变分推断方法。相较于MCMC法，变分推断通过将复杂的后验概率分布推断问题转化为优化问题，使用简化的分布来近似真实的后验分布，更加灵活与高效。

为了衡量近似分布与真实后验分布之间的相似度，变分推断使用了（KL, Kullback-Leibler）散度。变分推断先定义一个简单的分布族，其中 是可调节的变分参数，目标是使更接近后验分布。KL散度描述了两个概率分布之间的差异，是变分推断期望最小化的目标函数，定义为：

由于是一个常数项，最小化KL散度等价于最大化变分下界（ELBO, Evidence Lower Bound），其公式为：

而本次报告所使用的ADVI 是一种通用的变分推断方法，结合了自动微分技术，其主要特点是将所有的变量变换为标准化的空间，并使用梯度优化变分参数，更加快速，效率更高。

## 实现过程

### 模型代码

本次代码使用Python3.10编写，主要使用了PyMC和sklearn两个库，模型搭建相关代码如下图6所示。在这里定义了模型的基本要素，包括先验分布的参数和、回归函数、概率值pred、观测变量observed以及变分推断trace。代码集成程度较高，主要通过调用pymc库的内置函数实现。

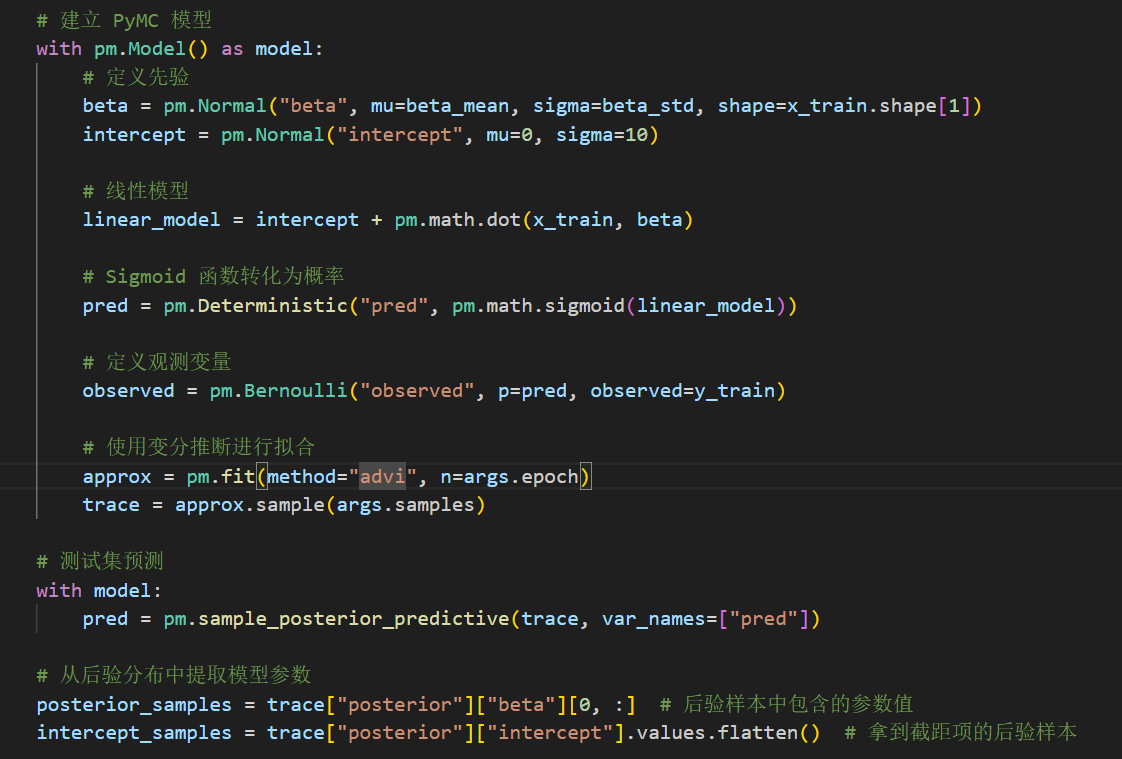


图6 模型代码实现

模型预测阶段的代码如下图7所示。首先从拟合完的后验分布中提取模型中和参数，随后对每一个后验样本计算测试集的概率，接着使用sigmoid激活函数将线性模型的输出转换为概率作为逻辑回归的输出，维数为，其中为后验采样的个数，为测试集的数量，在本次实验中为6513。最后对于每个测试样本，计算所有后验样本的预测概率均值，得到测试集中每个样本的最终预测概率，并将模型预测概率值大于0.5的样本认为是阳性，最终维度为。



图7 模型预测阶段代码实现

### 评价指标

本次实验所使用的评价指标有三个，分别是：正确率Accuracy、召回率Recall和 “受试者工作特征”曲线下面积AUC-ROC。首先引入“混淆矩阵”的概念。

表2 混淆矩阵

|  |  |  |
| --- | --- | --- |
| 真值 | 预测值 | |
| 真 | 假 |
| 真 | 真正例(True Positive, TP) | 假负例(False Negative, FN) |
| 假 | 假正例(False Positive, FP) | 真负例(True Negative, TN) |

“混淆矩阵”根据分类结果和样本真值将实验结果分为四类，分别为真正例(TP)、假负例(FN、假正例(FP)和真负例(TN)。

正确率定义为模型分类与真值一致的数量，占总样本的比例，即

召回率定义为所有真实为正的样本中，被正确预测为正的比例，即

ROC曲线纵坐标为真正率(True positive rate, TPR)，横坐标为假正率 (False positive rate, FPR)，其中TPR和FPR的定义如下：

FPR与TPR两种指标互斥，模型几乎无法同时在两个指标上达到最优。因此研究者通常使用ROC曲线下的面积(AUC-ROC)对TPR、FPR两个指标进行综合度量。该值取值在0和1之间，越接近1，表示TPR相对越高，FPR相对越低，模型的分类性能便相对越好。

## 结果分析

### 结果展示

如前文所述，本次的实验分为两组，对照组为不进行区间归纳，保持数据的原始取值类别；实验组进行了区间归纳，为减小模型推理开销。下表3为实验结果：

表3 实验结果

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 条件设置 | | 原始类别 | | | 区间归纳 | | |
| 采样数 | 迭代数 | Acc | Recall | AUCROC | Acc | Recall | AUCROC |
| 100 | 100 | 0.238 | 1.0 | 0.595 | 0.237 | 1.0 | 0.743 |
| 1000 | 0.237 | 1.0 | 0.634 | 0.237 | 0.999 | 0.670 |
| 5000 | 0.245 | 1.0 | 0.614 | 0.440 | 0.875 | 0.663 |
| 10000 | 0.615 | 0.415 | 0.592 | 0.768 | 0.288 | 0.731 |
| 50000 | 0.803 | 0.407 | 0.844 | 0.829 | 0.449 | 0.871 |
| 100000 |  |  |  |  |  |  |
| 1000 | 100 | 0.238 | 1.0 | 0.631 | 0.237 | 1.0 | 0.756 |
| 1000 | 0.238 | 1.0 | 0.608 | 0.237 | 0.999 | 0.716 |
| 5000 | 0.245 | 1.0 | 0.650 | 0.421 | 0.881 | 0.671 |
| 10000 | 0.594 | 0.458 | 0.599 | 0.760 | 0.248 | 0.730 |
| 50000 | 0.804 | 0.403 | 0.844 | 0.828 | 0.449 | 0.871 |
| 100000 |  |  |  |  |  |  |
| 5000 | 100 | 0.237 | 1.0 | 0.608 | 0.237 | 1.0 | 0.717 |
| 1000 | 0.237 | 1.0 | 0.620 | 0.237 | 0.999 | 0.686 |
| 5000 | 0.245 | 1.0 | 0.685 | 0.417 | 0.881 | 0.664 |
| 10000 | 0.587 | 0.466 | 0.598 | 0.764 | 0.302 | 0.731 |
| 50000 | 0.804 | 0.404 | 0.844 | 0.828 | 0.449 | 0.871 |
| 100000 |  |  |  |  |  |  |
| 10000 | 100 | 0.238 | 1.0 | 0.594 | 0.237 | 1.0 | 0.708 |
| 1000 | 0.237 | 1.0 | 0.621 | 0.237 | 0.999 | 0.695 |
| 5000 | 0.244 | 1.0 | 0.660 | 0.417 | 0.881 | 0.667 |
| 10000 | 0.589 | 0.462 | 0.598 | 0.733 | 0.350 | 0.734 |
| 50000 | 0.804 | 0.414 | 0.844 | 0.828 | 0.449 | 0.871 |
| 100000 |  |  |  |  |  |  |
| 20000 | 100 | 0.237 | 1.0 | 0.624 | 0.237 | 1.0 | 0.710 |
| 1000 | 0.237 | 1.0 | 0.643 | 0.237 | 0.999 | 0.695 |
| 5000 | 0.245 | 1.0 | 0.651 | 0.419 | 0.881 | 0.666 |
| 10000 | 0.591 | 0.458 | 0.597 | 0.764 | 0.301 | 0.732 |
| 50000 | 0.804 | 0.412 | 0.844 | 0.828 | 0.450 | 0.871 |
| 100000 |  |  |  |  |  |  |

以对照组为例，绘制了模型在采样数20000和迭代次数50000时的训练损失曲线、ROC曲线以及混淆矩阵，如下图8图9图10所示。

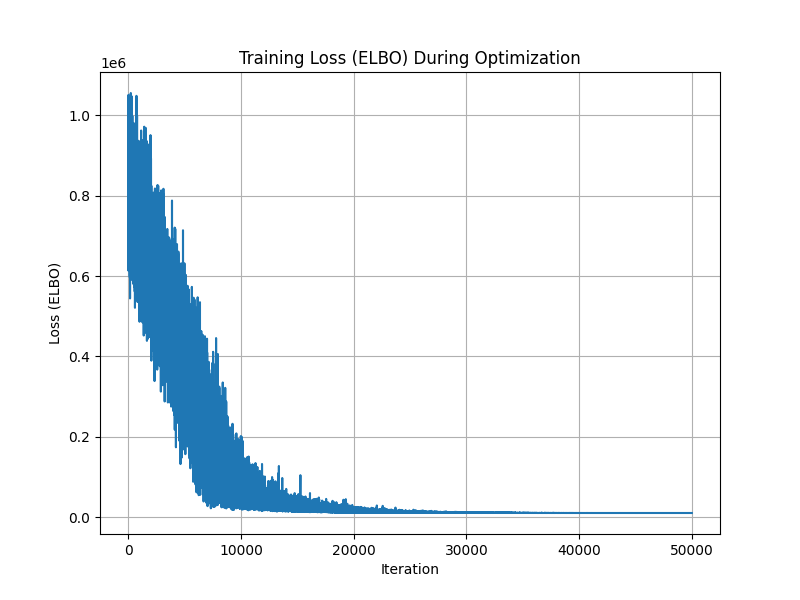


图8 训练损失曲线

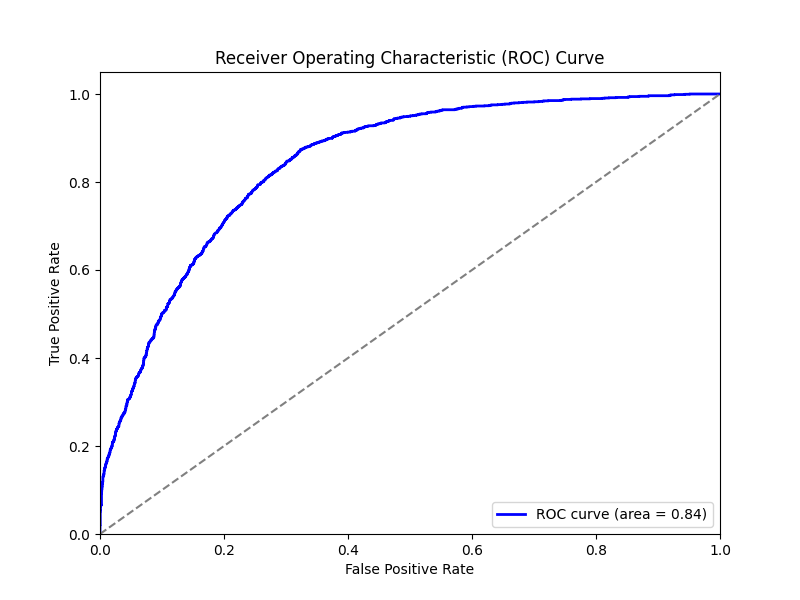


图9 ROC曲线

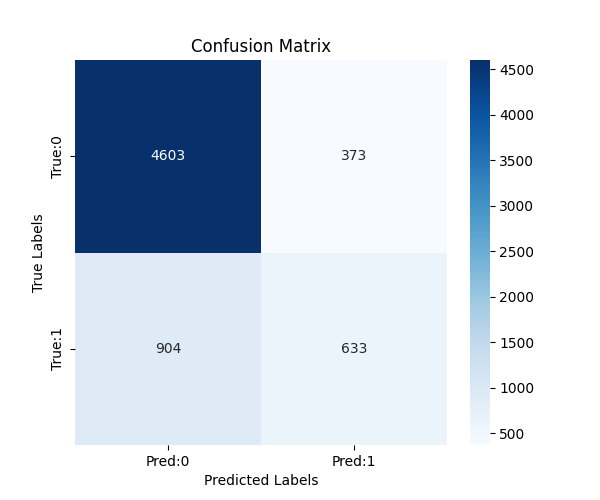


图10 最终混淆矩阵

### 分析与结论

#### 采样数与迭代数的关系

小采样数（100、1000）：在较低的采样数下，模型表现出的准确率通常较低。尤其是在较小采样数下，Recall 始终为 1，显示出模型在大部分情况下完美地识别了所有正类样本，但准确率和 AUCROC 值较低，可能是由于过拟合或样本不均衡问题。

中等采样数（5000、10000）：当采样数增加时，模型的准确率通常提升，尤其是在 10000 采样数下，准确率和 AUCROC 达到较高水平。然而，召回率出现了下降，可能是因为模型开始更加精确地区分正负类样本，但未能完全识别出所有正类样本。

大采样数（50000、100000）：随着采样数进一步增大，准确率稳定在较高水平（约为 0.80 左右），而 AUCROC 值也趋于较高的稳定值（约为 0.84和0.87）。然而，Recall 依然偏低，这可能表明模型在极大样本下的表现趋于优化，但对正类样本的召回依然不理想。

#### 数据处理方式的对比：

原始类别：在使用原始类别时，模型的准确率较高，特别是在采样数为 10000 或更多时，准确率大多能够维持在 0.8 左右。但 Recall 仍然较高且稳定，这表明在大部分情况下模型能够较好地识别正类样本。AUCROC 变化较大，表明模型对负类样本的区分能力较弱。

区间归纳：在进行区间归纳时，模型的表现略有不同。尤其在较小的采样数下，Accuracy 和 AUCROC 显著提高，且 Recall 稍有下降，且在多数情况下AUCROC值都优于原始类别，表明区间归纳处理后的数据可能帮助模型减少了过拟合，提高了对样本的区分能力。然而，随着采样数增加，Recall 的值下降较多，这表明区间归纳在处理极大样本时无法保持对正类的良好召回。

#### 模型的敏感性：

采样数和迭代数的影响：较大采样数对模型表现有显著提升，尤其是在 10000 和 50000 采样数时，模型的准确率和 AUCROC 提升较为明显。迭代数对模型性能的影响不如采样数那样显著，可能是因为贝叶斯模型已经能够通过较少的迭代次数收敛。

数据预处理的影响：区间归纳方法显著提高了准确率，但却牺牲了召回率。这表明该方法可能减少了模型的偏差，但也带来了召回能力的下降。原始类别方法在模型训练中维持了较高的召回率，但由于过拟合的问题，导致了准确率的下降。

#### 主要结论

采样数增加提升模型表现：随着采样数的增加，模型的准确率、AUCROC 逐渐稳定并提高，而在较小采样数下，模型表现不稳定，准确率较低。

数据预处理（区间归纳）提高准确性，但降低召回率：区间归纳方法帮助模型提高了对数据的区分能力，减少了过拟合，但也带来了对正类的召回能力下降。

模型仍需进一步调优：尽管准确率较高，但模型的 Recall 值始终未能达到理想水平，可能需要进一步的改进，特别是在召回正类样本方面，可以考虑使用其他方法（如调整阈值、重新采样等）。

AUCROC 是更可靠的指标：由于模型的准确率受样本不均衡的影响较大，AUCROC 提供了更全面的性能评估，尤其是在样本分布不均的情况下，AUCROC 更能反映模型的实际表现。

## 参考文献

1. Kaggle. Adult Income Census[EB/OL].(2016.8.27)[2024.12.7]. https://www.kaggle.com/datasets/uciml/adult-census-income
2. 孙海燕，周梦，李卫国，冯伟. 数理统计[M]. 北京：北京航空航天大学应用数学与系统科学学院，2015: P160-P171
3. Liangliang Lee. 确定近似推断：变分贝叶斯[EB/OL].(2021.2.16)[2024.12.7]. https://learn.lianglianglee.com/专栏/机器学习40讲/36 确定近似推断：变分贝叶斯.md
4. CSDN. 贝叶斯统计——先验分布与后验分布[EB/OL].(2020.7.27)[2024.12.7]. https://blog.csdn.net/weixin\_44293582/article/details/107619180
5. 知乎. 一文搞懂变分推断（Variational inference）[EB/OL].(2024.2.16)[2024.12.7]. https://zhuanlan.zhihu.com/p/682453554