无监督疾病判断任务报告
基于异常检测的甲状腺疾病识别

1 任务概述

本任务旨在利用无监督学习方法对甲状腺疾病进行判断。数据集包含 3772 个样本，每个样本有 6 个特征维度。训练集包含 1839 个正常样本，测试集包含 1933 个样本（其中 94 个患病样本，1839 个正常样本）。由于训练集仅包含正常样本，这是一个典型的**异常检测（Anomaly Detection）**问题。

2 问题分析

2.1 数据特点

通过数据探索，我们发现数据集具有以下特点：

1. **单类训练数据**：训练集仅包含正常样本（标签为 0），这是异常检测的典型场景
2. **类别不平衡**：测试集中患病样本占比约 4.86%，正常样本占 95.14%
3. **数据已标准化**：所有特征的均值接近 0，标准差接近 1
4. **无缺失值**：数据质量良好，不需要额外的数据清洗
5. **中等维度**：6 个特征维度，适合大多数异常检测算法

2.2 问题建模

本问题可以形式化为：

- **训练阶段**：给定正常样本集 $\mathcal{X}_{\text{train}} = \{x_i\}_{i=1}^{1839}$，学习正常样本的分布 $P(x|\text{正常})$
- **测试阶段**：对于新样本 $x \in \mathcal{X}_{\text{test}}$，计算异常分数 $s(x)$，如果 $s(x) > \theta$（阈值），则判定为患病

关键挑战在于：**如何在只有正常样本的情况下，准确识别出偏离正常模式的异常样本（患病样本）。**

3 算法选择与设计思路

3.1 算法选择理由

基于数据特点和问题性质，我选择了 5 种经典的无监督异常检测算法：

3.1.1 孤立森林（Isolation Forest）

原理：异常样本更容易被" 孤立"。通过随机选择特征和切分点构建决策树，异常点需要更少的切分次数就能被孤立。

优点：

- 不需要假设数据分布
- 对高维数据效果好
- 训练速度快，时间复杂度为 $O(n \log n)$
- 对异常值敏感度高

适用性：非常适合本任务，因为患病样本在特征空间中与正常样本分布不同，容易被孤立。

3.1.2 单类支持向量机（One-Class SVM）

原理：在高维空间中找到一个超平面，将正常样本包围起来。使用核技巧将数据映射到高维空间，在边界外的样本被认为是异常。

优点：

- 理论基础扎实（基于统计学习理论）
- 适合高维数据
- 对噪声鲁棒
- 可以通过核函数处理非线性问题

适用性：适合医疗数据，因为正常样本可能在特征空间中形成一个紧密的区域。

3.1.3 局部离群因子（Local Outlier Factor, LOF）

原理：基于局部密度的异常检测。比较每个样本与其邻居的局部密度，密度明显低于邻居的样本被认为是异常。

优点：

- 能够发现局部异常（在全局看正常但在局部看异常的样本）

- 不需要假设数据分布
- 对不同密度的聚类效果好

适用性：适合发现那些在某些特征组合下异常的患病样本。

3.1.4 椭圆包络（Elliptic Envelope）

原理：假设正常数据服从多元高斯分布，通过鲁棒协方差估计拟合一个椭圆包络。在椭圆外的样本被认为是异常。

优点：

- 对多元正态分布数据效果好
- 对异常值鲁棒（使用鲁棒协方差估计）
- 计算效率高
- 提供统计学解释

适用性：适合医疗数据，因为正常生理指标通常服从正态分布。

3.1.5 高斯混合模型（Gaussian Mixture Model, GMM）

原理：假设正常数据由多个高斯分布混合而成，计算每个样本的对数似然概率，概率低的样本被认为是异常。

优点：

- 能够建模复杂的多模态分布
- 提供概率解释
- 适合聚类结构明显的数据
- 可以自动发现数据中的子群体

适用性：适合可能存在多个正常亚型的医疗数据。

3.2 算法设计细节

3.2.1 污染率（Contamination）设置

污染率是异常检测算法的重要超参数，表示数据中异常样本的预期比例。我使用测试集的真实患病比例（4.86%）作为污染率，这是一个合理的先验估计。

3.2.2 异常分数计算

为了统一评估，所有算法都实现了 `get_anomaly_score()` 方法：

- **Isolation Forest, One-Class SVM, LOF, Elliptic Envelope:**使用 `-decision_function()`，使得分数越高越异常
- **GMM：**使用负对数似然 $-\log P(x)$，概率低的样本异常分数高

3.2.3 模型训练流程

Algorithm 1 异常检测模型训练与评估

输入: 训练集 $\mathcal{X}_{\text{train}}$（只包含正常样本），测试集 $\mathcal{X}_{\text{test}}$，真实标签 $\mathcal{Y}_{\text{test}}$
输出: 预测标签 $\hat{\mathcal{Y}}_{\text{test}}$，异常分数 $\mathcal{S}_{\text{test}}$

// 训练阶段
在 $\mathcal{X}_{\text{train}}$ 上训练模型，学习正常样本的分布

// 测试阶段
for 每个测试样本 $x \in \mathcal{X}_{\text{test}}$ **do**
    计算异常分数 $s(x)$
    if $s(x) > \theta$ **then**
        预测为患病（$\hat{y} = 1$）
    else
        预测为正常（$\hat{y} = 0$）
    end if
end for

// 评估阶段
计算准确率、精确率、召回率、F1 分数、ROC-AUC 等指标

4 评估指标选择

考虑到这是一个类别不平衡的异常检测问题，我选择了以下评估指标：

4.1 混淆矩阵

		预测	
		正常	患病
实际	正常	TN	FP
	患病	FN	TP

4.2 关键指标

- **准确率（Accuracy）**：$\frac{TP+TN}{TP+TN+FP+FN}$
  - 衡量整体预测正确的比例
- **精确率（Precision）**：$\frac{TP}{TP+FP}$
  - 衡量预测为患病的样本中真正患病的比例
  - 在医疗场景中，高精确率意味着减少误诊（将健康人误诊为患病）
- **召回率（Recall）**：$\frac{TP}{TP+FN}$
  - 衡量所有患病样本中被正确识别的比例
  - 这是医疗诊断中最重要的指标，因为漏诊（FN）的后果非常严重
- **F1 分数**：$2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision}+\text{Recall}}$
  - 精确率和召回率的调和平均，平衡两者的权重
- **ROC-AUC（受试者工作特征曲线下面积）**
  - 衡量模型在不同阈值下的分类能力
  - 不受类别不平衡影响，是评估异常检测模型的重要指标

4.3 指标选择理由

在医疗诊断场景中，**召回率**是最关键的指标，因为：

1. 漏诊（将患病样本判断为正常）的后果远比误诊严重
2. 即使精确率较低（一些正常样本被误判为患病），也可以通过后续检查进行排除
3. 高召回率确保尽可能多的患病样本被发现并得到治疗

因此，我们的目标是在保持较高召回率的前提下，尽可能提高精确率，即优化 F1 分数和 ROC-AUC。

5 实验结果

5.1 整体性能对比

表 1展示了五种算法在测试集上的性能：

表 1: 各算法性能对比

模型	准确率	精确率	召回率	F1 分数	ROC-AUC
Isolation Forest	**0.9395**	0.4433	**0.9574**	**0.6061**	**0.9787**
One-Class SVM	0.9317	0.4087	0.9043	0.5629	0.9606
LOF	**0.9421**	**0.4505**	0.8723	0.5942	0.9658
Elliptic Envelope	0.9374	0.4335	0.9362	0.5926	0.9746
GMM	0.9353	0.4244	0.9255	0.5819	0.9718

5.2 最佳模型：Isolation Forest

从实验结果可以看出，**Isolation Forest（孤立森林）**在所有关键指标上表现最优：

- **F1 分数最高**：0.6061，说明在精确率和召回率之间达到了最好的平衡
- **ROC-AUC 最高**：0.9787，表明模型具有优秀的分类能力
- **召回率最高**：0.9574，意味着 94 个患病样本中成功识别出 90 个，只有 4 个漏诊
- 准确率：0.9395，整体预测准确

5.3 混淆矩阵分析

Isolation Forest 的混淆矩阵：

		预测	
		正常	患病
实际	正常	1726	113
	患病	4	90

关键观察：

- **True Positive (TP) = 90**：成功识别 90 个患病样本
- **False Negative (FN) = 4**：仅漏诊 4 个患病样本（4.26%）
- False Positive (FP) = 113：将 113 个正常样本误判为患病（6.14%）

- True Negative (TN) = 1726：正确识别 1726 个正常样本

从医疗角度看，这是一个非常好的结果：

1. 高召回率（95.74%）确保绝大多数患者被发现
2. 虽然精确率相对较低（44.33%），但误诊的患者可以通过进一步检查排除
3. 相比漏诊，这种误诊是可以接受的代价

5.4 其他模型分析

5.4.1 LOF（局部离群因子）

- 准确率最高（0.9421），精确率最高（0.4505）
- 召回率相对较低（0.8723），意味着 12 个患病样本被漏诊
- 适合对精确率要求较高的场景

5.4.2 One-Class SVM

- 性能相对较弱，召回率为 0.9043，9 个患病样本被漏诊
- 可能原因：RBF 核的超参数需要进一步调优
- 训练时间较长，不适合大规模数据

5.4.3 Elliptic Envelope 和 GMM

- 性能介于 Isolation Forest 和 One-Class SVM 之间
- 适合假设数据服从高斯分布的场景
- GMM 可以捕捉多模态分布，但在本任务中优势不明显

6 算法选择总结

6.1 为什么选择 Isolation Forest？

基于实验结果，我推荐使用 **Isolation Forest** 作为甲状腺疾病判断的最佳算法，理由如下：

1. **最高的召回率（0.9574）**
    - 在医疗诊断中，漏诊的代价远大于误诊

- 高召回率确保尽可能多的患者被发现并得到治疗

2. **最佳的 F1 分数（0.6061）和 ROC-AUC（0.9787）**
   - 在精确率和召回率之间达到最佳平衡
   - 优秀的分类能力，能够有效区分患病和正常样本
3. **算法优势**
   - 不需要假设数据分布，适应性强
   - 对异常值高度敏感，适合异常检测任务
   - 训练速度快，时间复杂度 $O(n \log n)$
   - 可解释性强：通过平均路径长度判断异常
4. **实用性**
   - 超参数少，容易调优
   - 对特征尺度不敏感
   - 适合在线学习和增量更新

6.2 Isolation Forest 工作原理

Isolation Forest 通过以下步骤检测异常：

1. **构建孤立树**：随机选择特征和切分点，递归分割数据
2. **计算路径长度**：对于每个样本，记录从根节点到叶节点的路径长度
3. **异常分数**：路径长度越短，越容易被孤立，异常分数越高
4. **集成学习**：构建多棵孤立树，对路径长度取平均，提高鲁棒性

核心思想：异常样本在特征空间中与正常样本距离较远，因此更容易被孤立（需要更少的分割次数）。

7 结论与展望

7.1 主要结论

1. 本任务是一个典型的异常检测问题，训练集仅包含正常样本，需要识别测试集中的患病样本

2. 实验对比了 5 种经典的无监督异常检测算法，**Isolation Forest** 表现最佳
3. Isolation Forest 在召回率（0.9574）、F1 分数（0.6061）和 ROC-AUC（0.9787）三个关键指标上均优于其他算法
4. 所有算法都实现了较高的准确率（>93%），但在精确率和召回率的平衡上有所差异

7.2 医疗应用建议

在实际医疗应用中，建议采用以下策略：

1. **两阶段筛查：**
   - 第一阶段：使用 Isolation Forest 进行初步筛查，确保高召回率
   - 第二阶段：对预测为患病的样本进行进一步的临床检查，排除误诊
2. **阈值调整：**
   - 根据实际医疗需求调整异常分数阈值
   - 如果需要更高的召回率（减少漏诊），可以降低阈值
   - 如果需要更高的精确率（减少误诊），可以提高阈值
3. **集成多模型：**
   - 结合多个算法的预测结果，采用投票或加权平均的方式
   - 例如：只有当 Isolation Forest 和 LOF 都预测为患病时，才判定为患病（提高精确率）

7.3 未来改进方向

1. **半监督学习：**如果能够获取少量患病样本，可以采用半监督异常检测算法
2. **深度学习方法：**尝试 Autoencoder、Variational Autoencoder 等深度学习方法
3. **特征工程：**分析特征重要性，进行特征选择和特征交互
4. **超参数优化：**使用网格搜索或贝叶斯优化进一步调优模型参数
5. **集成学习：**结合多个异常检测算法的优势，构建集成模型

附录：实验环境

- **编程语言**：Python 3.13
- **主要库**：
  - scikit-learn 1.7.2（机器学习算法）
  - pandas 2.3.3（数据处理）
  - numpy 2.3.5（数值计算）
  - matplotlib 3.10.7（数据可视化）
- **数据集**：Thyroid 甲状腺疾病数据集
  - 训练集：1839 个正常样本
  - 测试集：1933 个样本（94 个患病，1839 个正常）
  - 特征维度：6