**人工智能与机器学习报告**

1. **双向障碍检测**

3220101111 洪晨辉

1. **实验介绍**
   1. 实验背景

双相障碍属于心境障碍的一种疾病，表示既有躁狂发作又有抑郁发作的一类疾病。当前研究发现，在双相障碍发生过程中遗传因素、环境或应激因素之间的交互作用、以及交互作用的出现时间点等都产生重要的影响。

双相障碍检测，即通过医学检测数据预测病人是否双相障碍，或双相障碍治疗是否有效。医学数据包括医学影像数据与肠道数据。由于缺少医学样本且特征过多，因此选取合适的特征对双模态特征进行整合并训练合适的分类器进行模型预测具有较强的现实需求与医学意义。本实验力图完成少样本、多特征下的监督学习。

* 1. 实验要求

a) 实现双模态特征选择与提取整合。

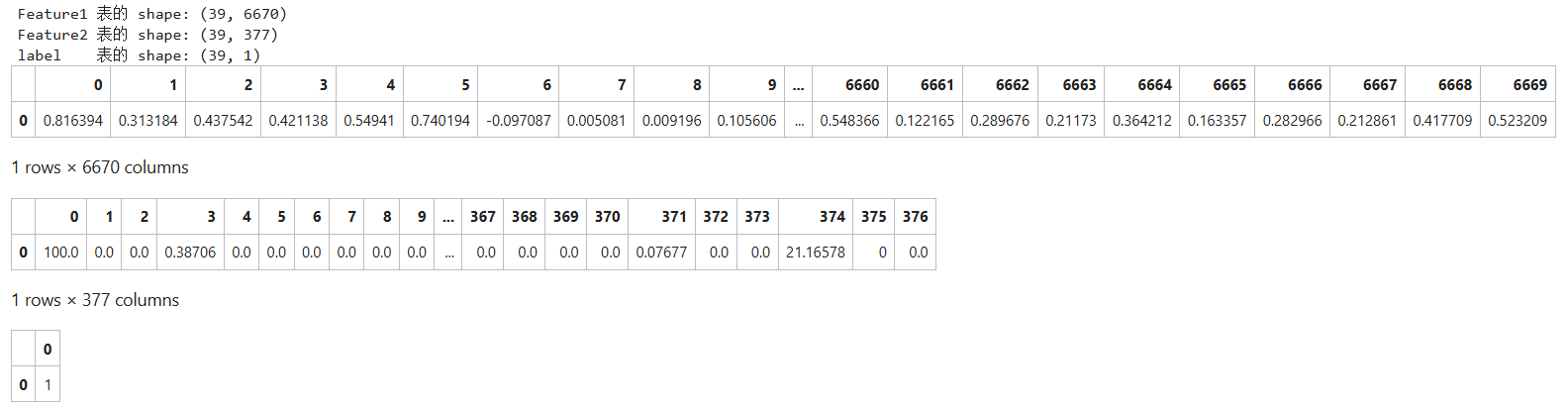
b) 选择并训练机器学习模型进行准确分类。

c) 分析不同超参数以及特征选择方法对模型的结果影响。

* 1. 实验环境

Python3. Numpy 库进行相关数值运算，sklearn 库进行特征选择和训练机器学习模型等。

1. **实验操作**
   1. 导入数据

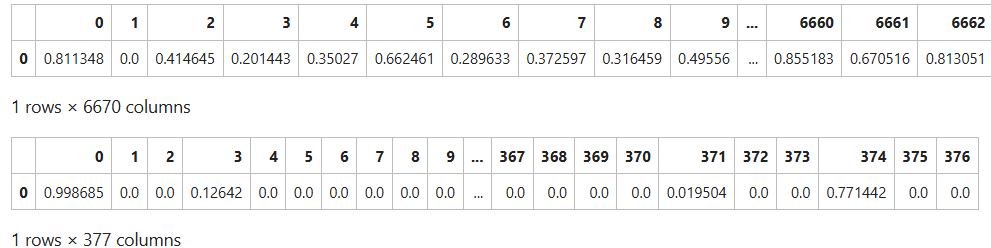


可以看到，医疗数据中的样本和特征数量存在着极大的不平衡。其中医疗影像数据共 6670 维，肠道数据共 377 维，而样本仅有 39 个，其中正样本标签为 1 ，负样本标签为 -1 。因此，特征的筛选和组合以及机器学习模型的选择优化对提高模型的性能极其重要。

* 1. 准备数据

数据预处理是一种数据挖掘技术，它是指把原始数据转换成可以理解的格式。在这个过程中一般有数据清洗、数据变换、数据组织、数据降维和格式化等操作。对于本数据集，没有无效或丢失的条目；然而需要我们进行特征的筛选和整合。这些预处理可以极大地帮助我们提升机器学习算法模型的性能和预测能力。本例中主要是归一化：

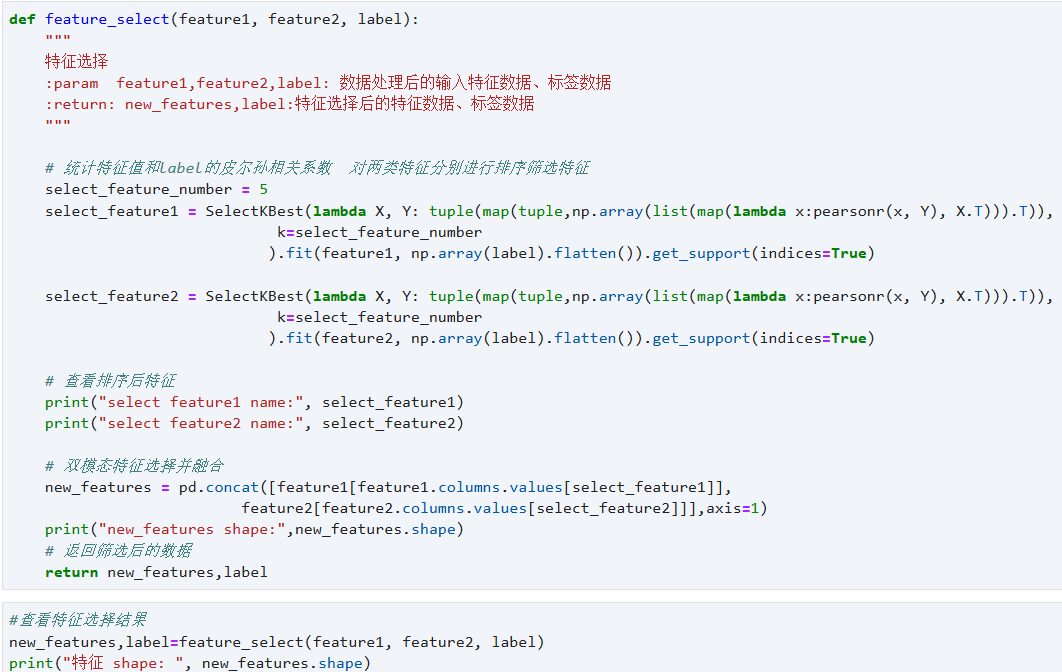




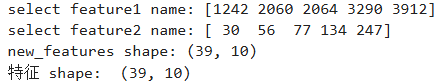
* 1. 特征选择

使用监督学习算法的一个重要的任务是决定哪些数据特征能够提供最强的预测能力。专注于少量的有效特征和标签之间的关系，我们能够更加简单具体地理解标签与特征之间的关系，这在很多情况下都是十分有用的。

可以看到：医疗数据中的样本和特征数量存在着极大的不平衡，其中医疗影像数据共 6670 维，肠道数据共 377 维，而样本仅有 39 个。因此，为了训练预测模型，特征的筛选和组合以及机器学习模型的选择优化极其重要。本报告采用简单的皮尔逊相关系数法进行双模态选择融合，两者共计选择十个特征。



最终选出如下十个特征：



* 1. 混洗、切分数据

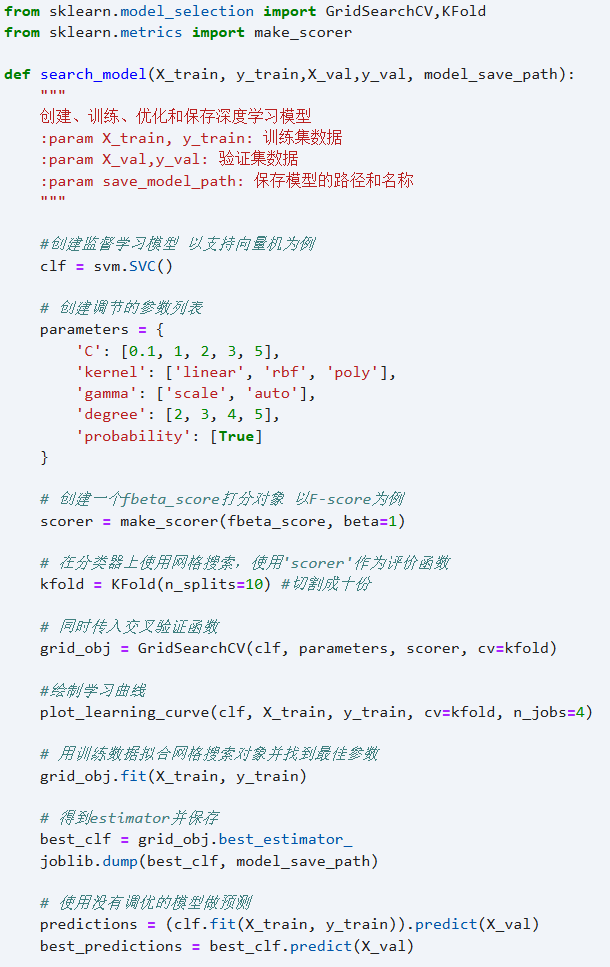
现在特征选择已经完成并得到了新的特征数据。那么下面将数据（包括特征和它们的标签）整合并切分成训练集和测试集。其中 80% 的数据将用于训练和 20% 的数据用于测试。然后再进一步把训练数据分为训练集和验证集，用来选择和优化模型。

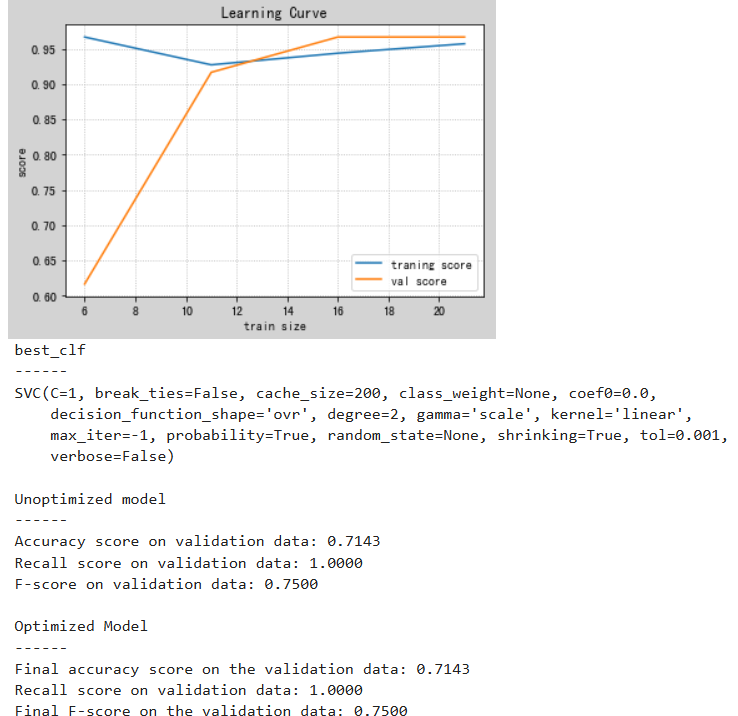


共39个样本，选择24/7/8分的方式，合理。

* 1. 监督模型选择

监督学习，也称为监督机式器学习，是机器学习和人工智能的一个子类别。它的定义是使用标记的数据集来训练算法，以准确分类数据或预测结果，符合本任务要求。本报告使用支持向量机，导入scikit-learn库完成模型训练。





得到模型。值得注意的是，优化后的模型并没有表现更好。同时，注意到模型因为训练数据集太少，出现了较为严重的过拟合现象（训练分数近于1的情况下，验证集仅有0.75左右）。

1. **结果分析**

提交测试，得到：



结果尚可。这与该模型在训练过程中的良好表现不无关系。

1. **总结与讨论**

本次报告探索并实践了一系列数据处理函数，对双向障碍患者的数据库进行了全面的分析和预测，仅使用了10个特征，取得显著成效。本实验力图教授数据降维的有效技巧和主成分提取的实用方法，将极大地帮助未来的研究和工作中更高效地处理和分析数据。