## 《综合项目实践》指导书（任务书）

**课程题目：**基于机器学习的径流预报

**课程目标：**

**整体目标**：能够综合利用特征工程、机器学习和集成学习等技术实现径流预报。

**具体目标**如下：

* 数据分析：了解数据集，进行数据统计分析和探索性数据分析。
* 特征工程：分别实现2种特征归一化方法和2种特征选择方法。
* 机器学习：选择2种机器学习算法实现给定单个流域的径流预报，并确定预报精度最优的特征工程方法+机器学习算法组合。
* 流域聚类：要求利用1种聚类算法将水文单元中的流域划分为不同的簇。
* 空间多尺度建模：针对给定水文单元内的多个流域，分别进行分区建模和全区建模，并确定预报精度最优的特征工程方法+机器学习算法组合。
* 集成学习：使用集成学习当中的Stacking和简单平均法将2种不同机器学习模型的预报结果进行集成。

**算法技能目标：**能够掌握相关特征工程、机器学习算法和集成学习算法。

**编程技能目标：**能够熟练使用python编程语言进行数据的处理、分析、建模和评估。

**思政目标：**掌握机器学习技术本领，服务于祖国和人民。

**任务流程：**

一、**数据分析**

1、 通过给定数据的网址或文档，理解所用数据集（dataset02.zip）的内容。

https://daac.ornl.gov/DAYMET/guides/Daymet\_Daily\_V4R1.html

https://nas.er.usgs.gov/hucs.aspx

2、对数据进行统计分析和探索性数据分析，并进行可视化，深入了解数据分布、异常及相关性等情况。

数据集中csv文件较多，如果对每个csv做数据分析将花费大量的时间，因此建议将所有csv文件求平均（可以把每个csv视为一个矩阵，所有矩阵做加法运算，然后再除以矩阵个数），例如将所有csv文件中2000-01-01的Discharge值求平均得到2000-01-01的Discharge平均值，依次类推得到所有日期所有字段各自的平均值），最后做整体数据分析。

<https://mp.weixin.qq.com/s/V9W0nZzqPdvpp2IGjJRb5w>

https://mp.weixin.qq.com/s/9lxM8U5OqRKVodzE1Q5FCQ

二、**特征工程**

1、特征归一化：总结常用的特征归一化方法，并实现最小-最大归一化、Z-score标准化方法。

2、特征选择：总结常用的特征选择方法，并实现2种特征选择方法。

https://mp.weixin.qq.com/s/hNkvlqcTC9uS\_G2zl93yJA

https://mp.weixin.qq.com/s/sL7hZjko5GG96K\_99xiNAg

https://mp.weixin.qq.com/s/YybyrF4t6DzVl2fhRvhplg

https://mp.weixin.qq.com/s/HF7KMtH0e-EbNG8t9NFzsA

https://mp.weixin.qq.com/s/19tiow6ETVaoWIox8sBLig

**三、机器学习回归**

1、调研适用于径流预报的机器学习算法。

2、实现2种机器学习算法（BP神经网络、支持向量机）。

3、利用Grid Search CV（网格搜索）调超参数。

4、基于编号01333000流域的数据（01333000.csv），进行径流预报。

5、针对编号01333000流域，确定预报精度最优的特征工程方法+机器学习算法组合。

**四、流域聚类**

1、使用聚类算法对编号02水文单元内的多个流域（对应dataset02.zip中的所有csv）进行聚类分析

2、确定最优的聚类数目，根据最优聚类数目将流域划归为不同的簇。

**五、空间多尺度建模**

1、分区建模：为编号02水文单元的每个簇分别构建基于不同特征工程方法和不同机器学习的径流预报模型；利用Grid SearchCV调超参数；为每个簇确定预报精度最优的特征工程方法+机器学习算法组合，该组合称之为最优分区建模。

2、全区建模：为编号02水文单元构建基于机器学习的径流预报模型。利用Grid SearchCV调超参数；为编号02水文单元确定预报精度最优的特征工程方法+机器学习算法组合，该组合称之为最优全区建模。

3、分区建模与全区建模的精度对比：在同一套测试数据集上，对比最优分区建模和最优全区建模的径流预报精度。

**六、集成学习**

1、实现基于Stacking的集成学习算法。

2、实现基于简单平均法的集成学习算法。

3、基于编号01333000流域的数据（01333000.csv文件），分别进行基于Stacking和简单平均法的径流预报。

4、对比集成学习算法与单一机器学习算法的径流预报精度。

杭州电子科技大学

《综合项目实践》

实践报告

|  |  |
| --- | --- |
| 题 目 | 基于机器学习的径流预报 |
| 专 业 |  |
| 学 号 |  |
| 姓 名 |  |
| 指导教师 | 江净超、黄经州 |
| 成 绩 |  |

二〇二四 年 月 日

一、任务要求

（简述本小组所选择的题目；对题目作出解析，叙述本小组对题目的理解（如属于监督/无监督学习、分类/回归任务等），及解题思路）

二、详细分工

（列出本小组成员各自要完成的具体任务，商定工作量比例）

三、过程设计

（分解题目，列出解题步骤，并给出各个步骤所用到的工具/算法，及处理目标）

四、编码实现及结果

（叙述编写代码所用的操作系统/开发环境/算法库；给出与第二部分（详细设计）对应的关键步骤的代码和注释；请利用可视化工具展示过程和结果）

五、程序调试

（叙述解题过程中所遇到的问题及解决方法）

六、总结

（1. 以列表形式列出本次实践中所用到的技术/模型/算法；2. 指出当前所用模型和取得的结果中可能存在的问题，设计后续的改进方案；3. 说明在本次实践过程中，对机器学习课程新的认识，以及自己的心得体会）

注：

1. 报告中请务必写明每位同学的分工及工作量比例。
2. 在实践课程答辩结束后，本实践报告纸质版（每组1份）请各班长收齐后交至科技馆609江净超、黄经州老师，同时请小组长将实践报告电子版、汇报PPT及代码文件打包提交至网络教学平台；【压缩包以“组长姓名+成员1姓名+成员2姓名+成员3姓名”方式命名】。

不交实践报告者，本小组成员成绩将全部判为不及格。