THU-70250043, Pattern Recognition (Spring 2020) **Project: 2**

大作业：降水量预测

Lecturer: Changshui Zhang zcs@mail.tsinghua.edu.cn

Student:

1 作业主题

天气预报中的降水量预测（Rainfall Prediction or Rainfall Forecasting）与我们的生活密切相关，它属于时空序列预测类问题。在机器学习和数据挖掘领域中，这一类问题多年来得到很多关注，不乏近年来一些顶刊、顶会上的工作。

文献 [1, 2, 3] 介绍了部分气象预报的方法，请大家在阅读这些文献的基础上，广泛阅读相关文献，深入思考，更好地完成本次训练。

在本次课程项目中，我们聚焦于用机器学习方法预测降水量。作业要求完整地完成数据处理、建模、实验、分析报告整个过程。

本项目对计算机算力要求较低，可在无 GPU 的个人计算机上完成实验。

2 作业内容

要求以小论文的格式提交实验报告，你的报告需要包括但不限于以下内容（基础分 80 分），在基本要求以外的实验、分析、讨论等均可以加分（满分 100 分，加满为止）。

注意由于任务定义和评价指标选择是开放的，评分时将更看重报告的完整度，在评价算法结果时将更看重改进的模型相对于基础模型的提高，而非最终结果的绝对值。同时，为了使得相对提高明显而故意把基础模型性能做差的行为是严格禁止的，这一点在评分时会有体现。

一、题目，摘要，关键词

二、简介（5’）

- 降水量预测任务及主流的解决方法。
- 机器学习方法在降水量预测中的应用（在本文档提供的文献之外，至少调研 2 篇相关文献）。

- 本文的主要贡献。

三、任务定义（5’）

- 可以将任务定义为预测降水量数值的回归问题，也可以将降水量划分为小雨、大雨等几个等级，将任务定义为分类问题。
- 考虑如何处理时空序列的输入和输出。可以预测下一小时的降水量，也可以预测下一天或下一周每天的降水量；可以运用气象站间的地理位置，同时预测多个气象站的降水量，并也可以对每个气象站进行单独建模处理。
- 用符号和公式定义降水量预测任务的输入和输出。

四、数据整理（5’）

- 数据来源、内容。
- 数据清洗、缺失值处理。
- 可视化地观察数据分布。

五、特征提取（5’）

- 数据归一化处理。
- 可尝试特征间的组合。
- 列出最终使用的特征维度和每一维的含义。

六、模型设计（25’）

- 使用至少 4 种不同的模型（包含至少一种基于树的方法，至少一种神经网络）。可选择但不限于方法：
    1. ARIMA: Auto-Regressive Integrated Moving Average model.
    2. SVR: Support Vector Regression.
    3. FNN: Feed forward neural network.
    4. LSTM: Long Short-Term Memory recurrent neural network.
    5. GBRT: Gradient Boosting Regression Tree.
    6. XGBoost.
- 使用至少 2 种 ensemble 方法组合不同模型。

- （加分项）尝试更多更复杂的模型，如图卷积神经网络（GCN）、注意力机制等；尝试对现有方法做出自己的改进。

七、实验设计及结果（15’）

- 数据集如何划分成训练、验证、测试集。
- 评价指标的定义。对回归/分类问题，应选择对应的至少 3 个评价指标。
- 每种模型单独的最好结果对比（列出图表并进行讨论）。
- Ensemble 后的最好结果对比（列出图表并进行讨论）。
- 每种模型在不同超参数下的表现（列出图表并进行讨论）。
- （加分项）其他实验设计及结果。

八、实验结果分析（15’）

实验结果分析的方式可以包括但不限于：

- 讨论数据和模型中每一部分的贡献（如果删去/更换部分数据输入/实验设定/模型结构，结果会发生什么变化）。
- 特征的重要性分析。
- 错误分析（模型在哪些数据下预测准确率高，哪些数据下预测准确率低）。
- 案例分析（在具体的案例上，不同模型表现的区别在哪里）。
- 模型和结果可视化分析。
- （加分项）其他方式的实验结果分析。

九、代码接口（5’）

具体要求见第4节。

十、小组成员贡献

十一、结论

3 数据采集

气象数据的种类多种多样，主要包括地面观测资料、天气雷达资料、气象卫星资料和数值预报产品。本次作业提供地面观测数据，（数据集已上传至清华云盘链接），数据集中各列的含义参见 abbreviated.txt。该数据集包

含了对 122 个气象站从 2000 年到 2016 年逐小时观测数据 (并不是所有气象站都是从 2000 年开始观测的)，包含 17 个气象参数。

请将数据集自行划分为训练集、验证集和测试集（建议以时间顺序划分），划分比例自定，如 7:2:1，测试集**不允许**参与模型的训练。另外，如果算力不够，允许取出完整数据集中的一部分数据作为 mini 数据集，即可以只保留一部分气象站全部年份的数据，或者只保留全部气象站的其中几年的数据，只要合理即可，然后在 mini 数据集中完成此次任务。

要求在报告中说明对数据集的处理方式。

4 作业提交

大作业报告提交截止时间为5 月 29 日 23:59。

4.1 作业说明

- 编程语言不限。
- 可以个人或 2 人小组的形式完成作业，每小组提交一份即可并用独立段落说明成员贡献。注意，对小组完成的作业要求更高。
- 完成作业过程中可以参考已有的代码，但要在报告中用独立段落给出详尽说明，具体到自己提交的代码中哪一个文件哪些行，并提供参考来源的链接。若参考的代码过多，将影响评分；若在未标明参考他人的部分发现了雷同现象，本次作业将判定零分。

4.2 提交代码说明

- 自己完成的全部训练、测试代码均需要进行提交。库函数的代码无需提交，除非自己修改了某个已有的库。
- 需要手工从测试集中挑选五个气象站的测试数据 `station1` ~ `station5`，作为示例放入最终提交的压缩包中，五个气象站数据要分开存放。除了这五个气象站以外的测试数据无需提交。
- 训练数据如有特殊处理（如划分出 mini 数据集），需在报告中说明，无需上传，中间文件和生成的结果文件不需要上传。
- 需要从所做的模型中选择一个 ensemble 前单个性能最优的，作为可直接运行的测试模型提交。只需提交最优模型训练好的参数，其他模型提交代码即可。
- 提交的代码必须是自包含的（self-contained）。即，从下载原始数据开始，不依赖于任何中间结果，必须要能够完整地复现训练、测试过程。

4.3 提交文件说明

- 最终作业以 zip 压缩包形式提交。除压缩包命名外，目录名和文件名采用英文，防止因为操作系统不同造成乱码。提交的压缩包展开后目录结构应当如下：

  - 张三-李四-降水量预测.zip
    - codes/
      - README.md
      - testset/
        - station1 - station5
      - ...
    - report.pdf

  其中，report.pdf 为报告，... 部分为代码，可以包含多个文件或目录。

- 需要写一个Markdown格式的 README.md，内容至少要包括：
  - 按模块描述压缩包中每部分文件的作用。
  - 运行代码所需的软件环境和软件版本，并提供从裸的操作系统开始配置所需环境的命令。
  - 下载原始训练测试数据和整理数据的命令。
  - 训练所提交的最优模型的命令。按此命令训练出的模型，不应当与用提交的模型参数加载出的最优模型性能差异过大。
  - 用上一步训练出来的模型在五个气象站测试集数据上测试的命令。
  - 用提交的最优模型在五个气象站测试集数据上测试的命令，以及预期的结果。
- README.md 必须包含关于环境、数据、训练、测试的所有命令。README.md 中的命令依次复制粘贴到命令行里执行，应当能够复现整个流程。提交前建议反复检查并确认这一点，复现失败将极大影响评分。
- 最终提交的压缩包体积不得超过 50MB，如模型或者测试数据太大导致超出这一限制，需要将较大的文件移出压缩包直至满足这一限制为止，并将这些文件放在互联网上（例如清华云盘）且在 README.md 中提供下载命令。

参考文献

[1] Ramana R V, Krishna B, Kumar S R, et al. Monthly rainfall prediction using wavelet neural network analysis[J]. Water resources management, 2013, 27(10): 3697-3711.

[2] Yu Zheng, Xiuwen Yi, Ming Li, Ruiyuan Li, Zhangqing Shan, Eric Chang, and Tianrui Li. Forecasting fine-grained air quality based on big data. In Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pages 2267–2276. ACM, 2015.

[3] Yuxuan Liang, Songyu Ke, Junbo Zhang, Xiuwen Yi, Yu Zheng, GeoMAN: Multi-level Attention Networks for Geo-sensory Time Series Prediction. In International Joint Conference on Artificial Intelligence (IJCAI), 2018.