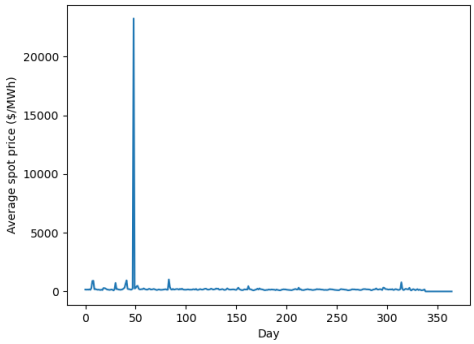
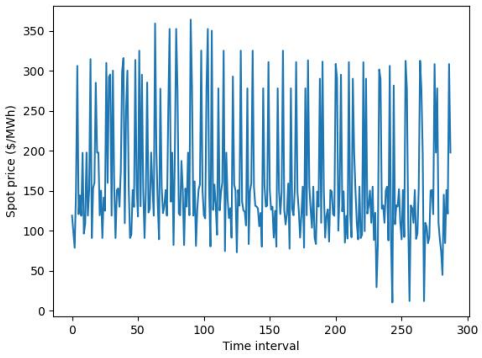
2024 秋 大数据技术与应用 课程大作业说明

作业背景

2024 年 10 月 15 日，我国省间电力现货市场转入正式运行，标志着全国统一电力市场建设取得重大进展。现货市场是电力市场的重要组成部分，现货市场的价格往往波动很大，甚至会在某些时段出现负电价的现象。对于电力现货市场的参与主体，各种市场决策都依赖于对未来现货市场价格的预测。下图是某地区现货市场的实际电价数据。下图 1 表示日均现货价格在一年内的波动；下图 2 表示实时的现货价格在一天之内的波动（每 5min 一个电价数据，所以一天共 288 个数据点）。可以看到无论是在年的时间尺度，还是日的时间尺度，电价的波动都较大。对电价的错误估计会导致市场主体的盈利能力下降。因此本学期的课程大作业为对某地区现货市场电价的预测。



一年内的日均现货价格



一天内的现货价格

作业任务

由于数据量较大，本次大作业使用数据库格式文件提供原始数据。提供 data_2021.db 和 data_2022.db 两个数据库文件。大作业所需的各类文件可以在下方清华云盘链接中下载（https://cloud.tsinghua.edu.cn/d/68319a796a4141bebf37/）。

提供了某地区 2021-2022 年的电力市场数据，包括各运行区域的负荷、各新能源机组的预测出力、各机组的投标曲线、各运行区域的电价。同学们可以根据历史数据训练模型，实现对电价的预测。

对于某一市场的市场主体，在构建训练模型的时候可以利用历史的投标数据做数据挖掘，但是在实时预测阶段，市场主体一般无从立刻得知各发电商的投标曲线，市场主体只能获得**非新能源机组容量**，**新能源机组预测出力**，**各地区的预测负荷**对价格进行预测（样例如'test_input_demo.db'所示）。

同学们需要完成：

1. 从数据库中读取数据，对数据进行预处理和分析
2. 根据历史数据，合理划分训练-验证-测试集，选择算法、构建模型、进行训练，对各个运行区域（共 5 个）的电价进行预测
3. 调整算法和其他超参数，尽力提升算法精度

数据说明

数据库文件中共有三张表，分别是机组表 Gen，负荷表 Demand，价格表 Price，这三张

表所包含的数据如下所示

机组表 Gen

列名	数据说明
DUID	机组编号
type	机组类型 1：新能源机组 0：其他机组
region	机组所在区域
time	运行时间
capacity	对于新能源机组，为新能源预测出力；对于其他机组，为最大出力
price_band_{i}	分段投标的价格，i=1-10 ($/MWh)
volume_band_{i}	分段投标的电量，i=1-10 (MW)

负荷表 Demand

列名	数据说明
region	区域
time	时间
demand	区域对应的负荷 (MW)

价格表 Price

列名	数据说明
region	区域
time	时间
price	能量价格 ($/MWh)

还有几点其他的说明

1. 市场每 5min 进行一次出清，价格/投标/负荷数据每 5min 就有一组数据
2. 在这里，分段投标可以理解为，机组愿意以 price_band_{i} 的价格出售 volume_band_{i}的电量，在该电力市场中机组可以报分 10 段进行投标
3. 数据集中共包含 5 个运行区域
4. 市场中的机组数量可能随着时间的变化而变化
5. 某些时段的数据可能会存在缺失

评价指标

本次大作业在评价模型预测精度的时候，将提供 12*288 组数据，含义是提供 12 天完整的运行数据（**非新能源机组容量**，**新能源机组预测出力**，**各地区的预测负荷**），完整的含义是这一天的数据从 00:00 开始，到 23:55 结束，共 288 个时间点。为了保证同学们的模型在不同月份都具有良好的预测能力，这 12 天测试集相互独立，并不是连续产生，而是通过**随机抽样**的方式得到。同学们需要应用自己的模型预测这 12 天的电价，预测得到的电价将通过评价指标 S 进行排名。测试时，将提供 db 文件，该文件仅包含 Gen 表和 Demand 表，同时 Gen 表的 price_band 和 volume_band 字段将被隐去。

本次作业的评价指标 S 基于均方误差，具体的计算公式如下所示

$$S=\frac{1}{1+\sqrt{\sum_{r}^{5} w_r \frac{\sum_{t}^{T}(\lambda_{r,t}-\hat{\lambda}_{r,t})^2}{\sum_{t}^{T}(\lambda_{r,t})^2}}}$$

式中 $\lambda_{r,t}$ 表示的是区域 r 时段 t 的真实电价，$\hat{\lambda}_{r,t}$ 表示的是区域 r 时段 t 的预测电价，w_r 表示的区域 r 的预测权重。

由于 5 个运行区域的负荷并不相同，5 个区域内的电价预测精度具有不同的重要性，本次作业简单地认为负荷较高区域的预测精度具有更重要的权重，5 个运行区域的预测权重分别是

运行区域	**r**	w_r
NSW1	1	0.36
QLD1	2	0.29
SA1	3	0.06
TAS1	4	0.06
VIC1	5	0.23

不难看到，指标 S 的取值在 0-1 之间，预测结果越精确，预测结果越接近 1，如果预测结果完全准确指标 S 将会等于 1。

竞争性实验说明

本次大作业将以竞争性实验的方式开展，并依托 DataFountain 平台进行。

竞争性实验的时间线如下表所示

周次（校历）	**内容**
第 11 周	大作业布置
第 13 周	平台开启，竞争性实验第 1 周，发布测试集 test_input_1.db
第 14 周	竞争性实验第 2 周，发布测试集 test_input_2.db
第 15 周	竞争性实验第 3 周，发布测试集 test_input_3.db

在校历的第 13-15 周，每周一 DataFountain 平台将给出本周的测试集（格式如测试集样例'test_input_demo.db'所示），同学们每天可以提交 3 次预测结果，预测 5 个运行区域 12 个运行日的现货电价，提交预测结果的格式需要参考'test_output_demo.csv'，文件命名为'test_output.csv'。同学们提交的预测结果将实时得到评价，生成预测精度指标 S，并根据该指标实时排名。每周周日晚将封榜固定本周排名。次周周一，将公布上周测试集对应的真实电价数据，供同学们进一步对模型进行改进。

在校历第 16 周，全部 3 次竞争性实验都已经完成，最终排名将根据三周的综合预测指标进行排名，具体计算方式如下

$$S_{total} = \sum_{k=1}^{3} \alpha_k S_k$$

其中，S_k 表示的是第 k 周的预测评价指标，α_k 表示的是第 k 周预测结果权重，本次作业中，预测结果权重分别为

预测结果权重	**取值**
α_1	0.2
α_2	0.3
α_3	0.5

值得说明的是，尽管本次大作业采用动态排名的方式开展，但是最终大作业的给分**不参考该排名**，将基于同学们最终提交的作业报告，根据是否合理应用了大数据方法、思考深度等方面进行给分。