《机器学习》大作业

北京交通大学《机器学习》课程组

大作业1：多元时间序列预测

任务：电力部门的二氧化碳排放量回归预测

- **要求：**

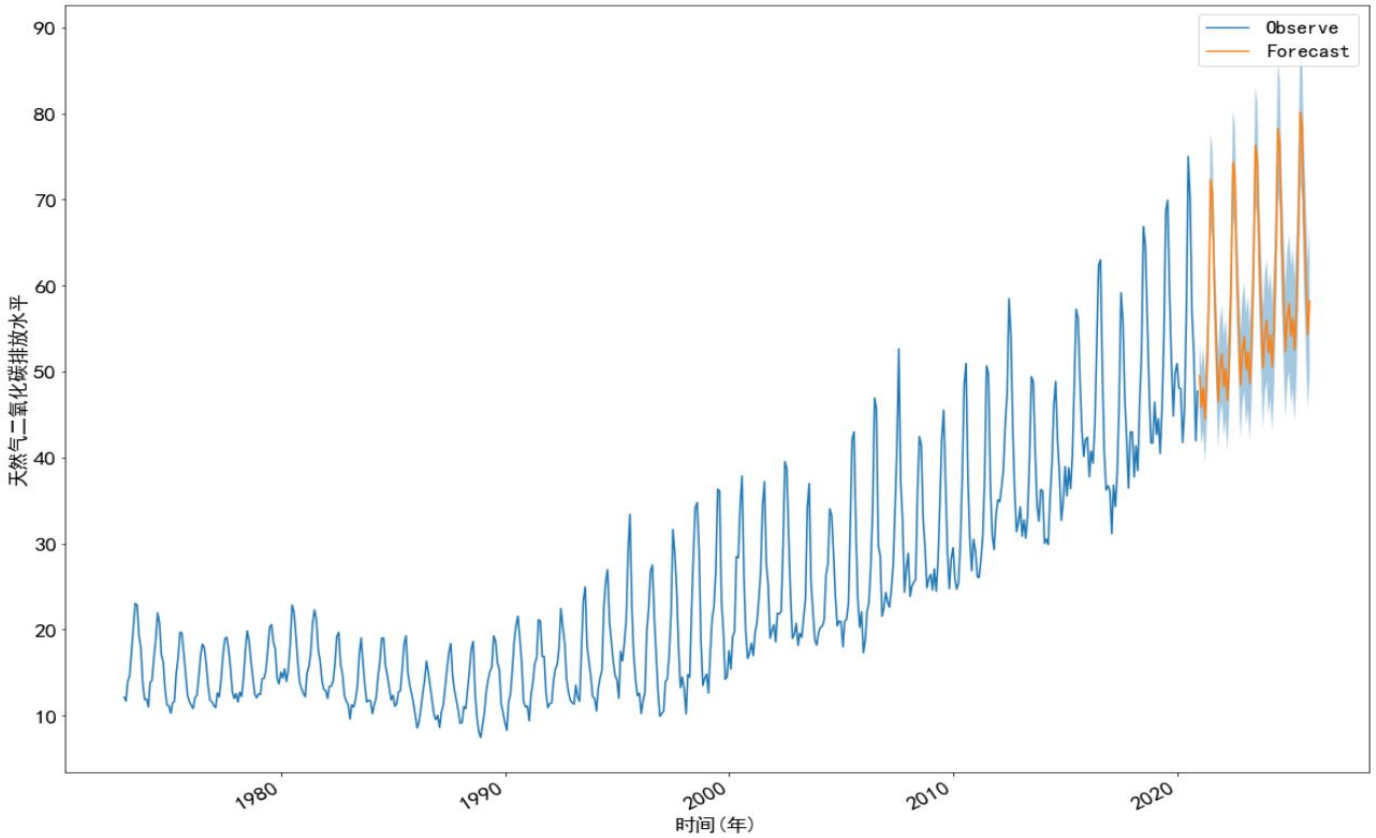
1. 数据时间跨度从1973年1月到2021年12月，按月份记录。
2. 数据集包括“煤电”，“天然气”，“馏分燃料”等共9个指标的数据（其中早期的部分指标not available）
3. 要求预测从2022年1月开始的半年时间的以下各个部分的排放量

① Coal Electric Power Sector CO2 Emissions
② Natural Gas Electric Power Sector CO2 Emissions
③ Distillate Fuel, Including Kerosene-Type Jet Fuel, Oil Electric Power Sector CO2 Emissions
④ Petroleum Coke Electric Power Sector CO2 Emissions
⑤ Residual Fuel Oil Electric Power Sector CO2 Emissions
⑥ Petroleum Electric Power Sector CO2 Emissions
⑦ Geothermal Energy Electric Power Sector CO2 Emissions
⑧ Non-Biomass Waste Electric Power Sector CO2 Emissions
⑨ Total Energy Electric Power Sector CO2 Emissions

主题：多元时间序列预测

关键点：

1. 这9个指标相互之间可能存在相关性，以“天然气”为例，除了它的历史值可用于自身建模预测，其它指标的历史值也可能用于其自身预测，请同学们分析。
2. 考虑数据本身季节性和趋势性
3. 自行划分2022年之前的数据（可以不全部使用），训练集进行模型训练，以及验证集进行模型选择，最后用自己认为最优的模型预测2022上半年六个月的排放量
4. 模型算法可使用线性模型，决策树，SVM，神经网络，集成学习等，最后只需介绍用于提交预测结果所对应的模型算法



主题：多元时间序列预测

- **大作业提交内容：**

1. 提交算法模型的预测结果和源代码
2. 课上展示，简要说明所使用的算法（包括主要超参数）和数据处理方式
3. 大作业报告

- 从2022年1月至6月的二氧化碳排放量待预测

	A	B	C	D	E	F
571	2020 May	48.501	45.671	0.246	0.746	
572	2020 June	66.23	56.808	0.294	0.931	
573	2020 July	90.095	73.453	0.316	0.961	
574	2020 August	91.279	69.642	0.298	0.894	
575	2020 September	70.008	55.433	0.226	0.509	
576	2020 October	61.279	51.75	0.269	0.441	
577	2020 November	62.059	42.139	0.266	0.651	
578	2020 December	78.735	48.32	0.323	0.86	
579	2021 January	82.218	47.641	0.263	0.82	
580	2021 February	87.181	42.976	0.843	0.819	
581	2021 March	62.652	41.048	0.256	0.778	
582	2021 April	54.615	41.242	0.259	0.453	
583	2021 May	64.59	44.522	0.274	0.591	
584	2021 June	87.095	59.249	0.273	0.579	
585	2021 July	102.259	67.423	0.248	0.804	
586	2021 August	101.985	68.814	0.346	0.875	
587	2021 September	80.443	54.323	0.251	0.732	
588	2021 October	64.52	51.508	0.286	0.722	
589	2021 November	59.326	48.144	0.301	0.919	
590	2021 December	62.391	48.368	0.339	0.696	
591	2022 January					
592	2022 February					
593	2022 March					
594	2022 April					
595	2022 May					
596	2022 June					
597						
598						
599						

大作业2：外科手术术后并发症预测

- 2017 年《中国心血管病报告》数据显示，过去 10 年间我国心血管外科手术量从 8 万例增长到近 21 万例，
- 临床工作中时常遇到合并受损心血管、肺功能状态的患者。
- 术前需要根据患者的疾病检测结果，评估并预测患者围术期并发症风险以及决定是否适合进行外科手术。

数据集介绍

	label	手术类型	手术时间	性别	身高	体重	性别.1	年龄	血红蛋白	贫血	...	心脏特征4	心脏特征5	心脏特征6	心脏特征7	心脏特征8	血氧1	血氧2	血氧3	血氧4	呼吸增量
0	0	1	62	1	168	54.0	1	58	150.0	0	...	0	25	1	32	0	NaN	NaN	NaN	NaN	6.01
1	1	1	94	2	152	63.0	2	78	110.0	0	...	1	24	0	39	0	96.10	93.2	95.0	92.0	20.16
2	0	1	58	2	162	59.0	2	66	121.0	0	...	1	23	0	32	0	95.50	95.0	95.0	92.0	3.16
3	1	1	68	2	158	46.0	2	67	137.0	0	...	1	25	1	45	1	94.57	94.0	93.0	93.0	6.28
4	1	1	57	1	163	49.5	1	71	135.0	0	...	1	28	1	41	0	97.80	97.6	98.0	95.0	8.22
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
134	0	1	110	1	158	61.0	1	65	142.0	0	...	1	20	0	32	0	92.17	93.0	93.5	85.0	9.26
135	0	1	62	2	145	53.5	2	63	130.0	0	...	1	19	0	29	0	95.95	95.0	96.0	93.0	5.82
136	0	1	74	1	168	54.0	1	75	104.0	0	...	1	20	0	33	0	100.00	99.6	94.5	93.0	6.63
137	0	1	205	1	168	93.0	1	42	151.0	0	...	1	21	0	37	0	96.13	95.2	95.0	93.0	14.46
138	0	1	238	1	168	67.0	1	65	141.0	0	...	1	26	1	36	0	95.93	97.2	97.0	94.0	7.04

139 rows × 25 columns

分类任务——患者检测

- **任务一：利用《机器学习》这门课所学分类方法进行建模，至少选取三种，并计算模型的准确率，AUC值及F1值，recall。【采用五折交叉验证】**
- **任务二：特征筛选**
- 1）利用数据本身性质进行特征筛选，例如相关系数法或方差选择法等。具体方法望同学们自行调研和学习。（至少两种方法）
- 2）选取分类模型后，利用模型性能进行特征筛选，在分类性能尽量保持的前提下，筛选出最多6个重要特征。
- **任务三：对模型结果及特征筛选进行解释。**
- 1）得到一组性能较高且特征最少的组合，并对所用到的特征选择方法以及特征筛选依据和过程进行详细说明。
- 2）尝试对所选特征组合进行实际意义上的解释。

北京交通大学《机器学习》课程组成员

景丽萍：lpjing@bjtu.edu.cn，http://faculty.bjtu.edu.cn/8249/

桑基韬：jtsang@bjtu.edu.cn，http://faculty.bjtu.edu.cn/9129/

王　晶：wj@bjtu.edu.cn，http://faculty.bjtu.edu.cn/9167/

李晓龙：lixl@bjtu.edu.cn，http://faculty.bjtu.edu.cn/9089/

黄晓雯：xwhuang@bjtu.edu.cn，http://faculty.bjtu.edu.cn/9545/