2024 年研究生机器学习课程项目

提交日期：2025 年 1 月 20 日中午 12 点之前（实际完成需至少 1 个月，视调试经验或有长短）。以邮件提交时间为准，逾期恕不接收。邮箱地址：20244027012@stu.suda.edu.cn

合作方式：最多可 2 人组队，请在报告上详细列明各位作者的贡献，并注明各自所属的研究领域。团队内部讨论为主，但这并不妨碍团队间的交流。如果从其他团队有借鉴之处，必须在参考文献中明确标注，否则将被视为剽窃。

问题来源：共享单车系统革新了传统的自行车租赁服务，实现了从用户注册、租赁自行车到归还的全流程高度自动化。借助这一系统，用户可以方便地在指定站点租借自行车，并在任意站点归还。截至目前，全球已有超过 500 个共享单车项目，共计投放了 50 多万辆自行车。由于共享单车在缓解交通拥堵、改善环境以及促进公众健康等方面发挥了重要作用，越来越受到社会各界的关注和欢迎。

对共享单车租赁数量进行预测可以帮助运营商合理规划和优化车辆调度，提升运营效率和服务质量。而共享单车租赁需求与环境和季节高度相关。例如，天气条件、降水、星期几、季节、一天中的时间段等都会影响租赁行为。核心数据集与美国华盛顿特区 Capital Bikeshare 系统的两年历史日志（对应于 2011 年和 2012 年）有关，可在 http://capitalbikeshare.com/system-data 上公开获取。我们以每小时为基础汇总数据，然后提取并添加相应的天气和季节信息。天气信息摘自 http://www.freemeteo.com。该数据集包含从 2011 年 1 月 1 日到 2012 年 12 月 31 日每天每小时的骑行人数。骑行人数分为临时骑行者和注册骑行者，汇总在 cnt 列中。

具体数据形式如下：

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
1	instant	dteday	season	yr	mnth	hr	holiday	weekday	workingda	weathersit	temp	atemp	hum	windspeed	casual	registered	cnt
2	1	2011/1/1	1	0	1	0	0	6	0	1	0.24	0.2879	0.81	0	3	13	16
3	2	2011/1/1	1	0	1	1	0	6	0	1	0.22	0.2727	0.8	0	8	32	40
4	3	2011/1/1	1	0	1	2	0	6	0	1	0.22	0.2727	0.8	0	5	27	32
5	4	2011/1/1	1	0	1	3	0	6	0	1	0.24	0.2879	0.75	0	3	10	13
6	5	2011/1/1	1	0	1	4	0	6	0	1	0.24	0.2879	0.75	0	0	1	1
7	6	2011/1/1	1	0	1	5	0	6	0	2	0.24	0.2576	0.75	0.0896	0	1	1
8	7	2011/1/1	1	0	1	6	0	6	0	1	0.22	0.2727	0.8	0	2	0	2
9	8	2011/1/1	1	0	1	7	0	6	0	1	0.2	0.2576	0.86	0	1	2	3
10	9	2011/1/1	1	0	1	8	0	6	0	1	0.24	0.2879	0.75	0	1	7	8
11	10	2011/1/1	1	0	1	9	0	6	0	1	0.32	0.3485	0.76	0	8	6	14
12	11	2011/1/1	1	0	1	10	0	6	0	1	0.38	0.3939	0.76	0.2537	12	24	36
13	12	2011/1/1	1	0	1	11	0	6	0	1	0.36	0.3333	0.81	0.2836	26	30	56
14	13	2011/1/1	1	0	1	12	0	6	0	1	0.42	0.4242	0.77	0.2836	29	55	84
15	14	2011/1/1	1	0	1	13	0	6	0	2	0.46	0.4545	0.72	0.2985	47	47	94
16	15	2011/1/1	1	0	1	14	0	6	0	2	0.46	0.4545	0.72	0.2836	35	71	106
17	16	2011/1/1	1	0	1	15	0	6	0	2	0.44	0.4394	0.77	0.2985	40	70	110
18	17	2011/1/1	1	0	1	16	0	6	0	2	0.42	0.4242	0.82	0.2985	41	52	93
19	18	2011/1/1	1	0	1	17	0	6	0	2	0.44	0.4394	0.82	0.2836	15	52	67
20	19	2011/1/1	1	0	1	18	0	6	0	3	0.42	0.4242	0.88	0.2537	9	26	35
21	20	2011/1/1	1	0	1	19	0	6	0	3	0.42	0.4242	0.88	0.2537	6	31	37
22	21	2011/1/1	1	0	1	20	0	6	0	2	0.4	0.4091	0.87	0.2537	11	25	36
23	22	2011/1/1	1	0	1	21	0	6	0	2	0.4	0.4091	0.87	0.194	3	31	34
24	23	2011/1/1	1	0	1	22	0	6	0	2	0.4	0.4091	0.94	0.2239	11	17	28
25	24	2011/1/1	1	0	1	23	0	6	0	2	0.46	0.4545	0.88	0.2985	15	24	39
26	25	2011/1/2	1	0	1	0	0	0	0	2	0.46	0.4545	0.88	0.2985	4	13	17
27	26	2011/1/2	1	0	1	1	0	0	0	2	0.44	0.4394	0.94	0.2537	1	16	17

数据中各列的含义:

Field	instant	dteday	season	yr	mnth	hr	holiday	weekday
Description	Index Number	date	Season Code	years	month	Hour	Holiday logo	Day of the week

Field	workingday	weathersit	temp	atemp	hum	windspeed	casual	registered	cnt
Description	Working day logo	Weather Condition Code	Actual temperature	Feeling temperature	humidity	Wind speed	Non-registered user - Vehicle	Registered User - Vehicle	Total rental quantity

注：

season：季节编码，映射规则为：1：春季 2：夏季 3：秋季 4：冬季

yr：年份，用数字表示相对年份，例如：0：第一个年份（2011）1：第二个年份（2012）

holiday：节假日标识，为二值变量：0：非节假日 1：节假日

workingday：工作日标识，为二值变量：0：非工作日（周末或节假日）1：工作日（非周末且非节假日）

weathersit：天气情况编码，为以下类别：1：晴朗或少云 2：多云或有雾 3：雨天或雪天

casual：非注册用户租借的自行车数量。

registered：注册用户租借的自行车数量。

cnt：总租借数量，等于 casual 和 registered 的总和。

temp、atemp、hum 以及 windspeed 都为归一化值，范围为 0 到 1。

预测任务：根据所提供的数据对未来单车租赁数量（cnt）进行预测。基于过去 I=96 小时的数据曲线来预测未来（i）O=96 小时（短期预测）和（ii）O=240 小时（长期预测）两种长度的变化曲线（需要分别训练，即长期预测的模型参数不能用于短期预测）。按照方法分为三部分。

前两部分为基础题，第三部分为开放题，各占总分的三分之一：

1. 使用 LSTM 模型进行预测；
2. 使用 Transformer 模型进行预测；
3. 使用自己提出的改进模型进行预测，结构不限。此部分以原理的新颖程度为首要评价标准，性能为次要评价标准。

训练与测试：数据集主要分为 train 和 test 两部分（具体见文件“train_data.csv”和“test_data.csv”）。请使用两种评价标准进行测试，即均方误差（MSE）与平均绝对误差（MAE）。至少进行五轮实验，并对结果取平均值，同时提供标准差（std）以评估结果的稳定性。

提交方式：实验报告应由以下四部分构成：1.问题介绍、2.模型（可以包含少量伪代码）、3.结果与分析、4.讨论。同时，需要提交代码（可以给出 Github 链接）。结果需以截图形式贴在报告中，并绘制出单车租赁数量（cnt）预测与真实值（Ground Truth）曲线的对比图。请注意对三种方法进行比较。如果自行提出的方法虽然新颖但性能不佳，但原因分析有力，同样可以获得较高的分数。务必注明参考文献，否则将视为抄袭，每抄袭一处将扣除 33 分。允许使用 ChatGPT 一类工具撰写报告，但仅限于撰写部分，并请注明，必要的参考文献仍然不可或缺。

注意：所提供的数据中可能会存在一天中只有 20 个小时数据的情况，但这并不对我们的预测任务产生影响，同学们只需将每一行当作一个小时的数据进行处理即可。此外，一个 Sample 的大小为（input+output），如果对时间序列数据处理方面有所疑问的同学，可以和上一届参加过机器学习课程的同学进行交流，如果对窗口大小和步长等概念不太清楚的同学可以参考以下博客[1][2][3]。

[1]https://blog.csdn.net/qq_47885795/article/details/143462299

[2]https://datac.blog.csdn.net/article/details/105498685?fromshare=blogdetail&sharetype=blogdetail&sharerId=105498685&sharerefer=PC&sharesource=weixin_44709585&sharefrom=from_link

[3]https://datac.blog.csdn.net/article/details/105928752?fromshare=blogdetail&sharetype=blogdetail&sharerId=105928752&sharerefer=PC&sharesource=weixin_44709585&sharefrom=from_link