机器学习

实训一：特征提取

【实训说明】

用于做机器学习实训的数据存储在 machine_data.csv 文件，该数据包含了几个城市 2014 年 1 月 1 日至 2022 年 4 月 30 日的天气数据，但不是所有特征用于训练机器学习模型都能取得好的效果，现在需要从该表中按要求提取相应的特征，以供后续建立机器学习算法模型使用。

【实训要求】

在“TianTq”项目“app1”应用下的“machine_learning”包中存在 train_model.py 文件，编写该文件的 feature_extraction()函数实现提取数据特征功能，用于机器学习建模的数据文件 machine_data.csv 存放在【app1/data/】目录。

（1）需要根据传递的城市参数，提取该城市用于机器学习建模的特征。

（2）进行数据特征提取时：

- 输入为前面七天的最高温、最低温、天气、风向、所属月份。
- 输出为该日的最高温、最低温、天气。
- 除此之外，还需要一列城市(city)作为数据的标志；

举例说明：假设当前日期为 2 月 10 日，则输入为其前七天（2 月 9 日-2 月 3 日）对应的最高温、最低温、天气、风向、所属月份，特征命名按时间倒序排序，对应输入特征命名参考见表 1。

表 1：列名命名规范

数据类型	日期	特征（列名）
特征值	2 月 9 日	day1_high_tem 、 day1_low_tem 、 day1_weather 、 day1_wind、day1_month
	2 月 8 日	day2_high_tem 、 day2_low_tem 、 day2_weather 、 day2_wind、day2_month
	2 月 7 日	day3_high_tem 、 day3_low_tem 、 day3_weather 、 day3_wind、day3_month
	2 月 6 日	day4_high_tem 、 day4_low_tem 、 day4_weather 、

		day4_wind、day4_month
	2月5日	day5_high_tem、day5_low_tem、day5_weather、day5_wind、day5_month
	2月4日	day6_high_tem、day6_low_tem、day6_weather、day6_wind、day6_month
	2月3日	day7_high_tem、day7_low_tem、day7_weather、day7_wind、day7_month
标签值	2月10日	cur_high_tem、cur_low_tem、cur_weather
城市	/	city

所以提取特征之后的数据是一个：
[‘city’,’day1_high_tem’,’day1_low_tem,……,‘cur_high_tem’ ，‘cur_low_tem’，‘cur_weather’]共39列的DataFrame，该数据将用于后续的机器学习；

注意：提取的数据顺序必须先提取最新日期的，比如长沙市的最新日期是2022年4月30日，则提取特征之后DataFrame中的第一条数据必须对应2022年4月30日这天的数据（包括当天的温度天气以及它过去7天的温度、天气等信息）。

（3）在提取特征之后，将“city”列为“长沙”的DataFrame的info()、describe()打印到控制台。

（4)并将DataFrame的内容保存到CSV文件中,命名为“changsha_feature.csv”，存储到【app1/data/feature/】目录，若目录不存在需自行创建。

【操作说明】

使用PyCharm打开桌面上的“TianTq”项目进行编码。

注意：实现功能后将结果截图粘贴到答题报告相应区域。

实训二：模型训练

【实训说明】

在提取特征之后，需要构建机器学习算法模型对最高温度进行预测。由于每个城市的温度差异很大，没有足够的规律性可言，因此不能通过一个统一的模型对多个城市进行预测，每个城市只能单独去训练模型；

又由于每个需要预测的目标值影响它们的因素有差异,所以不能使用一个模型同时预测这两个目标值，而应该分别创建模型进行预测。在创建模型的时候，每种模型所需要的特征也是不一样的。现需要基于实训一的结果来训练机器学习算法模型。

【实训要求】

使用实训一提取到的特征数据(共 39 个特征)构建模型进行训练。在“app1”应用下的“machine_learning”包中存在 train_model.py 文件，编写该文件中的 train_all_model()函数用于训练一个城市的最高温的机器学习模型。

（1）训练高温预测模型

- 训练高温预测模型需要的特征数据为：前七天的最高温、天气、风力、所属月份，总共 4*7=28 列特征值，目标值为当天的最高温。高温预测模型特征字段见表 2。

表 2:高温预测模型特征与目标值

数据类型	特征（列名）
特征值	day1_high_tem、day1_weather、day1_wind、day1_month、 day2_high_tem、day2_weather、day2_wind、day2_month、 day3_high_tem、day3_weather、day3_wind、day3_month、 day4_high_tem、day4_weather、day4_wind、day4_month、 day5_high_tem、day5_weather、day5_wind、day5_month、 day6_high_tem、day6_weather、day6_wind、day6_month、 day7_high_tem、day7_weather、day7_wind、day7_month
标签值	cur_high_tem

- 由于数据中既有类别型数据，又有数字型数据，因此需要在输入机器学习模型之前对数据进行编码操作，每一列应具有一个独立的编码器。编码规则如下：数据为整数或者整数型字符串直接编码为整数；字符串则通过标签编码器进行编码；
- 数据集划分方式：编码完成之后，需要划分训练集和测试集，测试集比例为 0.2，随机种子选择 7。
- 模型训练规则：自己选择模型（模型种类不限），使用训练集对模型进行训练，使用测试集 r2_score 分数对模型进行验证，保存最优模型。
- 模型训练完成后，将最终模型进行保存，模型保存到【app1/machine_learning/model_ckpt/】目录下(若目录不存在需自行创建目录)，命名为“城市名_high_model.pkl”，例如长沙的高温模型名为“长沙_high_model.pkl”。

（2）训练低温预测模型

- 训练低温预测模型需要的特征数据为:前七天的最低温、天气、风力、所属月份，总共 28 列，目标值为当天的最低温。低温预测模型特征字段见表 3。

表 3：低温预测模型特征与目标值

数据类型	特征（列名）

特征值	day1_low_tem、day1_weather、day1_wind、day1_month、 day2_low_tem、day2_weather、day2_wind、day2_month、 day3_low_tem、day3_weather、day3_wind、day3_month、 day4_low_tem、day4_weather、day4_wind、day4_month、 day5_low_tem、day5_weather、day5_wind、day5_month、 day6_low_tem、day6_weather、day6_wind、day6_month、 day7_low_tem、day7_weather、day7_wind、day7_month
标签值	cur_low_tem

- 编码规则、训练集划分方式以及模型训练规则与高温模型的方式一致。
- 训练模型后需要保存最终模型，模型保存到【app1/machine_learning/model_ckpt/】目录下，命名为“城市名_low_model.pkl”，例如长沙的低温模型名为“长沙_low_model.pkl”。

【操作说明】

使用 PyCharm 打开桌面上的“TianTq”项目进行编码。
注意：实现功能后将结果截图粘贴到答题报告相应区域。