高传染性传染病的传播趋势预测

Prediction of the Spread Trend of Highly Contagious Diseases

背景

Background

传染病（Contagious Diseases）的有效防治是全人类面临的共同挑战，如何通过大数据，特别是数据的时空关联特性，来精准预测传染病的传播趋势和速度，将极大有助于人类社会控制传染病，保障社会公共卫生安全。希望借助此次竞赛，充分发挥全球选手的聪明才智，运用大数据技术助力传染病的传播预测和控制，增强人类社会合作抗风险的意识和能力。

It’s a common challenge for all human beings to effectively prevent and control contagious diseases. Accurately predicting their spreading speed and trends with big data, especially the data with spatio-temporal correlations, will greatly help human society to control the contagious diseases and ensure the security of public health. We hope that the contestants from all over the world will fully utilize your intelligence and wisdom in helping to predict and control the spreading of contagious diseases, and to enhance the awareness and ability of human society in cooperating against risks using big data technology.

任务描述

Task Description

针对赛题所构造的若干虚拟城市，构造传染病群体传播预测模型，根据该地区传染病的历史每日新增感染人数、城市间迁徙指数、网格人流量指数、网格联系强度和天气等数据，预测群体未来一段时间每日新增感染人数。

In the regions of several virtual cities, the contestants are required to construct a contagious diseases spreading model to predict the number of newly infected persons per day in the future according to the historical number of newly infected persons per day, intercity migration scale index, grid population flow index, grid association intensity, and weather data.

赛题共涉及 11 个虚拟城市 90 天的感染情况，每个城市有若干重点区域。初赛要求针对所提供的 5 个城市，利用每个城市各区域前 45 天的样本数据进行训练，预测每个城市各区域后 30 天每天的新增感染人数。复赛要求针对包含初赛城市在内的 11 个城市，利用每个城市各区域前 60 天的样本数据进行训练，预测每个城市各区域后 30 天每天的新增感染人数。

This competition involves the infection of 11 virtual cities in 90 days and each city has several key regions. In the preliminary competition, only 5 cities are involved, and the first 45 days of labeled data of each region in each city are available for training to predict the number of newly infected persons of each region in each city over the last 30 days. In the playoff, the participants are required to take labeled data of each region in each city (11 cities in total including the preliminary cities) over the first 60 days as the training set to predict the number of newly infected persons per day of each region in each city over the last 30 days.

数据集

Data Set

初赛阶段

Preliminary Stage

训练数据（train_data）：

Training Data（train_data）：

训练集共包括 5 个城市，每个城市目录下的数据集总体说明：

The training set involves 5 cities in total. The general description of the data set under the directory of each city is as follows:

1. 各区域每天新增感染人数。文件名：**infection.csv**。提供前 45 天每天数据，文件格式为，城市 ID, 区域 ID, 日期, 新增感染人数；“, ”分割。

The number of newly infected persons per day of each region. File name: **infection.csv**. The daily data for 45 days are provided, and the file format is "city_id, region_id ,date,, number of newly infected persons", separated by the comma.

2.城市间迁徙指数。文件名：**migration.csv**。提供 45 天每天数据。文件格式为迁徙日期, 迁徙出发城市, 迁徙到达城市, 迁徙指数；“, ”分割。

Intercity migration scale index. File name: **migration.csv**. The daily data for 45 days are provided, and the file format is "migration date, migration departure city, migration arrival city, migration scale index", separated by the comma.

3.网格人流量指数。文件名：**density.csv**。提供 45 天内每周两天抽样数据，文件格式为日期，小时，网格中心点经度, 网格中心点纬度, 人流量指数；“, ”分割。

Grid population flow index. File name: **density.csv**. Data sampled twice a week within 45 days are provided, and the file format is "date, hour, longitude of grid center point, latitude of grid center point, population flow index", separated by the comma.

4.网格关联强度。文件名：**transfer.csv**。城市内网格间关联强度数据，文件格式为小时，出发网格中心点经度，出发网格中心点纬度，到达网格中心点经度，到达网格中心点纬度，迁移强度；“, ”分割。

Grid association intensity. File name: **transfer.csv**. The data on the intensity of the association between grids within the city are provided, and the file format is "hour, longitude of departure grid center point, latitude of departure grid center point, longitude of arrival grid center point, latitude of arrival grid center point, transfer intensity", separated by the comma.

5.网格归属区域。文件名：**grid_attr.csv**。城市内网格对应的归属区域 ID，文件格式为网格中心点经度，网格中心点纬度，归属区域 ID；“，”分割。

The region attribute of each grid. File name: **grid_attr.csv.** The grids with corresponding region id are provided, and the file format is "grid_x, grid_y, region_id", separated by the comma.

6.天气数据。文件名：**weather.csv**。提供 45 天每天数据，文件格式为日期, 小时, 气温, 湿度, 风向, 风速, 风力, 天气；“，”分割。

Weather data. File name: **weather.csv**. The data on the weather data for 45 days are provided, and the file format is "date, hour, temperature, humidity, wind direction, wind speed, wind force, weather", separated by the comma.

文件数据示例详细说明：

Detailed Description of the File Data Examples:

1. infection.csv

字段名称 Field name	含义 Description	示例 Example
city_id	城市 ID City ID	A
region_id	区域 ID Region ID	1
date	日期 Date	21200501
index	新增感染人数 Number of newly infected persons	20

2. migration.csv

字段名称 Field name	含义 Description	示例 Example
date	迁徙日期 Migration date	21200501
departure_city	迁徙出发城市 Migration Departure city	A
arrival_city	迁徙到达城市 Migration Arrival city	B
index	迁徙指数 Migration scale index	0.0125388

3. density.csv

字段名称 Field name	含义 Description	示例 Example
date	日期 Date	21200501
hour	小时 Hour	10
grid_x	网格中心点经度 Longitude of grid center point	166.306128
grid_y	网格中心点纬度 Latitude of grid center point	20.331142
index	人流量指数 Population flow index	3.1

4. transfer.csv

字段名称 Field name	含义 Description	示例 Example
hour	小时 Hour	10
start_grid_x	出发网格中心点经度 Longitude of departure	166.306128

	grid center point,	
start_grid_y	出发网格中心点纬度 Latitude of departure grid center point	20.331142
end_grid_x	到达网格中心点经度 Longitude of arrival grid center point,	171.678139
end_grid_y	到达网格中心点纬度 Latitude of arrival grid center point	17.812359
index	迁移强度 Transfer intensity	0.2

5. grid_attr.csv

字段名称 Field name	含义 Description	示例 Example
grid_x	网格中心点经度 Longitude of grid center point	166.306128
grid_y	网格中心点纬度 Latitude of grid center point	20.331142
region_id	区域 ID Region ID	1

6. weather.csv

字段名称 Field name	含义 Description	示例 Example
date	日期 Date	21200501
hour	小时 Hour	10
temperature	气温，摄氏度 Temperature, ℃	12
humidity	湿度 Humidity	77%
wind_direction	风向 Wind direction	Southeast
wind_speed	风速 Wind speed	16-24km/h
wind_force	风力 Wind force	<3
weather	天气 Weather	sunny

选手需要提交的数据 (result_data)

Data to be Submitted by the Contestants (result_data)

选手需要预测后 30 天每天每个城市对应区域的新增感染人数，提供的文件格式为：城市 ID,区域 ID,日期,每日新增感染人数；“, ”分割。选手提交结果文件命名为 submission.csv，内容示例如下：

Contestants are required to predict the number of newly infected persons of each region in each city per day over the next 30 days, and submit the result files in csv format with the name "submission.csv". File format: City ID, Region ID, date, number of newly infected persons per day, separated by the comma, with an example as bellow:

```
A,1,21200701,0
A,1,21200702,0
etc.
```

复赛阶段

Playoff Stage

新增 6 个城市，训练集的城市数量从 5 增加到 11；训练集的时间窗口从 45 天增加到 60 天；其他不变。

6 cities are added and the number of cities in the training set increases from 5 to 11. The time window of the training set increases from 45 days to 60 days. Others remain unchanged.

评估标准

Evaluation

均方根对数误差（root mean squared logarithmic error），公式如下：

RMSLE (root mean squared logarithmic error), calculated as

$$\sqrt{\frac{1}{n}\sum_{i=1}^{n}(\log(p_i+1)-\log(a_i+1))^2}$$

其中：

where:

n 是观测总数

n is the total number of observations

p_i是预测值

p_i is your prediction

a_i是实际值

a_i is the actual value

$\log(x)$是 x 的自然对数

$\log(x)$ is the natural logarithm of x