Chap 4

4 数据分析过程

数据准备

首先我们应当注意到, 我们有 country.csv 以及 daily_info.csv 两个数据文件. 而新冠疫情, 人们所最为关心的一点, 总是后来一天的新增病例. 为了实现这个目标, 我们应当尽可能地利用上已有的信息. country.csv 给我们提供了各个国家的特定信息, 我们首先就应当提取其中重要的部分.

注意到主成分分析能够帮助我们实现上述的要求. 我们将读取数据, 处理缺失值, PCA, 并使用 k-means 进行聚类集成在 Divide_types.m 中. 此处可以调整的参数为 PCA 后截断的参数以及分类的总类数. 这里我们选取 PCA 后的前三项性质将所有 daily_info.csv 中出现过的国家分为六类. 具体的分类可以参见 Project.mlx.

接下来我们应当考虑我们将使用什么数据进行预测. 这里首先我们会用到上面截断出的参数. 再考虑到总体的病例以及预测日期之前几天的新增病例数对预测有重要作用, 我们再加上 total_cases 以及预测日之前四天的新增病例数作为 features. 这里我们把数据整合起来的功能集成于 Data_Prepatraion.m 中. 这里接受输入参数为希望读取的国家, 是否随机化以及是否光滑化数据.

网络准备

lstm

预测检验

我们应当认为被分为同一类中的国家容易被我们所训练的网络预测.