《交通信息融合与挖掘》课程第四次作业

（布置时间：2019 年 11 月 6 日，提交时间：2019 年 11 月 20 日上课前）

第一部分：数据相似性/相异性分析

1. 对于下面的向量 X 和 Y，计算指定的相似性或距离变量。（15 分）
   1) X=(1, 1, 1, 1), Y=(2, 2, 2, 2)：余弦、相关、欧几里得；
   2) X=(0, 1, 0, 1), Y=(1, 0, 1, 0)：余弦、相关、欧几里得、Jaccard;
   3) X=(0, -1, 0, 1), Y=(1, 0, -1, 0)：余弦、相关、欧几里得；
   4) X=(1, 1, 0, 1, 0, 1), Y=(1, 1, 1, 0, 0, 1)：余弦、相关、Jaccard;
   5) X=(2, -1, 0, 2, 0, -3), Y=(-1, 1, -1, 0, 0, -1)：余弦、相关。

(作业要求：给出计算公式、计算过程及答案)

2. 行程速度是表征道路交通状态的重要参数。图 1 为上海市延安高架路上某路段的行程速度数据示例，其中 series_A、series_B 与 series_C 分别表示不同的三天（2 个工作日、1 个非工作日），时间段为 5:30~24:00（间隔为 5 分钟），源数据文件为数据集 5。（25 分）

Time	series_A	series_B	series_C
5:30	78.49	75.75	78.83
5:35	79.57	74.15	77.25
5:40	76.55	81.37	83.6
5:45	80.52	84.73	86.82
5:50	81.69	79.63	82.49
5:55	78.69	80.01	81.24
6:00	84.29	75.81	83.71

图 1 三个时间序列数据示例

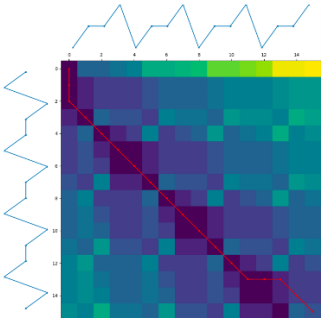

图 2 DTW 计算结果示意图

   1) 将 series_A、series_B 与 series_C 的数据可视化（时间序列图）；
   2) 运用动态时间规整方法（DTW），计算三个时间序列之间的两两相异性；
   3) 对比计算结果，简要说明你认为的造成两两之间差异性的可能原因。

(作业要求：提交 DTW 程序源代码和计算结果，并画出图 2 所示的计算结果图)

第二部分：数据预处理

3. 根据数据集 6（快速路线圈检测数据），运用 R 或 Python 完成如下工作：（30 分）
   1) 分别运用（1）最大值-最小值规范化法、（2）Z-Score 规范化法（标准偏差）、（3）小数定标规范化法，对检测数据（速度）进行标准化处理。
   2) 分别运用阈值法（五分位数法、3 倍标准差法（3σ 原则）），剔除异常检测数据（速度）。
   3) ~~分别运用（1）时间序列法（具体方法可自选）、（2）基于历史数据的修补方法、（3）基于空间位置的修补方法，修复周五的缺失检测数据（速度）（包括运用阈值法剔除异常数据后产生的空缺数据）~~。（这道题这次不需要做）

(作业要求：提交程序源代码和处理好的新数据集，以及简单的数据预处理过程说明文档)

4. 根据数据集 1，运用 R 或 Python 完成如下工作（30 分）：

以 Q15（交通安全政策支持程度）为例，运用主成分分析法，提取主成分，并简要讨论各因素的影响程度。（PCA 具体过程可参考教材-R 语言实战-第 14 章）

(作业要求：提交程序源代码以及简单的分析过程和结果说明文档)