

第 4 章 第四次作业

本次作业只需要写代码，无需写文档。请提交代码（如果有多个文件，请放到一个压缩包中）至人大云盘 (<https://pan.ruc.edu.cn:443/link/2EE479549DDF23447D4DCF075E2F860F>，有效期限：2024-11-15 23:59，访问密码：IAB9)

文件命名为“多模态机器学习 24 秋学期-第四次作业-[学号]-[姓名]”，截止时间为 2024 年 11 月 15 日 23:59 (UTC+8)。

4.1 动态时间规整算法 (Dynamic Time Warping)

本次作业要求使用代码实现课上所讲的 DTW 算法，并能对测试集中的序列进行分类，压缩包中附有已经划分好的训练集和测试集。一些说明如下：

- 数据集的格式为 $\langle \text{label}, d_1, d_2, d_3, \dots, d_{2000} \rangle$ ， $d_1 - d_{2000}$ 表示一个维度为 2000 的序列，label 是这个序列的标签 (类别)。
- 本次作业提供两个规模的数据集。data_large 文件夹下的数据集训练集有 104 条序列，测试集有 208 条序列，共有 52 个类别。data 文件夹下的数据集是从 data_large 数据集中采样两个类别得到的，训练集有 33 条序列，测试集有 130 条序列。data 文件夹的数据量较少，方便同学们更好的检查和验证自己的实现。
- 分类的依据是序列的相似性大小。具体来说，基于自己实现的 DTW 算法，对于测试集中的每一条序列 X ，计算该序列和训练集中每一个类别 y 的所有序列距离的平均值，作为序列 X 与类别 y 的距离，距离越大，相似度越低。取相似度最高的类别作为序列 X 的预测类别，最终的指标是整个测试集中预测正确的比例。
- 因此，本次作业和前面不同，无需模型的训练，只需要实现该算法，能够计算出两个序列的距离，就可以很容易完成后面的预测了。参考链接：[Kaggle_DTW](#)

这里简单举一个例子，希望能帮助大家理解：取测试集中的一条序列 $X = [d_1, d_2, \dots, d_{2000}]$ ，其真实的类别是 Y 。训练集中包含两个类别 y_1 和 y_2 ，属于 y_1 的序列有两条 $\{X_a, X_b\}$ ；属于 y_2 的序列有三条 $\{X_c, X_d, X_e\}$ 。那么序列 X 与类别 y_1 和 y_2 的距离分别为：

$$D_1 = \frac{\text{DTW}(X, X_a) + \text{DTW}(X, X_b)}{2} \quad (4.1)$$

$$D_2 = \frac{\text{DTW}(X, X_c) + \text{DTW}(X, X_d) + \text{DTW}(X, X_e)}{3} \quad (4.2)$$

根据 D_1 和 D_2 的大小可以决定序列 X 属于哪个类别，如果和真实的标签 Y 一致，那么就算预测正确，否则预测错误。如果计算出来的相似度最高的类别有多个，且包含正确的类别，那么也算作预测正确。