**实验 三**

**实验要求**

Task Overview

本次实验的目的是练习实现或调用数据挖掘算法或工具对传感器采集的真实数据进行分析，完成一项分类任务。完整过程请参考以下的流程图。

**警告：请独立完成作业，抄袭视为零分**

|  |  |  |
| --- | --- | --- |
|  |  |  |
| 1. 阅读数据描述文档，下载数据 |  |
| 2. 实现算法和实验输出 |  |
|  |  |
| 3. 撰写实验报告 |  |
| 4. 提交实验报告源代码及实验结果 |  |
|  | 思考如何评价你的实验结果 |

# 1. 数据描述与下载

**任务起源:** 我们每天都携带大量的传感器, 比如手机中的加速度传感器. 通过使用这些传感器采集的数据, 智能手机或智能程序可以准确推断用户行为为我们的生活提供便利。在本次实验中，我们将提供给你一批传感器采集的真实数据和它们的标签，而你的任务是预测那些没有(行为)标签的传感器数据所对应的标签(行为)，在数据挖掘或机器学习领域中这是一个典型的分类问题。

**下载数据:**

文件夹中有 5 个从不同用户身上采集到的训练数据集, 每一个 ZIP 文件包含 一个.feature file(特征文件) 和一个.label file(标签文件).

Traindata-A

Traindata-B

Traindata-C

Traindata-D

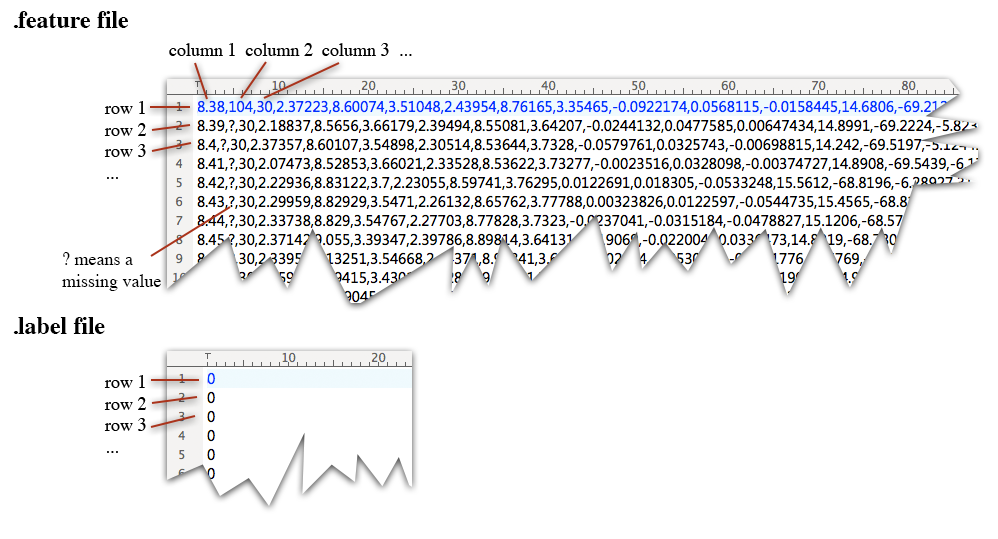
Traindata-E

此外还有 3 组测试文件 X, Y, Z, 也都是从不同人身上采集而来. ZIP文件中包含有3个 .feature file(特征文件)，但是没有.label files(标签文件)

    Testdata-XYZ

**文件描述:**

以上所有文件均为文本文件，均可通过文本编辑器打开，如下图所示。



* 在 .feature 文件中有41 列 (or called features). 其中第一列是数据的时间戳. 第二列是心率. 第3-15列, 第16-28列, 和第 29-41列分别是三组集成传感器采集的数据. 每组集成传感器采集的数据分别有一列体温数据, 3 列I 3D型加速度数据, 3列II 3D型加速度数据, 3 列 3D-gyroscope 型数据,和3 列3D-magnetometer型数据组成(共计13列)。
* 在.label文件中, 每一行代表与.feature文件中那行记录相对应的分类标签. 这也就是说.label file中的行数与.feature 文件的行数相等. 标签id 为0 to 24的整数代表不同的行为,0意为无行动。

**Your task** 为：从 A, B, C, D, E 数据中建立起分类模型用来预测训练数据X, Y, Z的标签。预测结果需要被保存在三个命名为 "X.label", "Y.label", and "Z.label" 文件中，同实验报告一起上传到课程主页。

**2. 算法实现与输出**

**算法实现:**

实现或者选择调用任一种或者多种你所知道的分类算法完成实验。

*(You are also encouraged to analyze the difficulties in this prediction task, find appropriate learning algorithm or invent your own algorithms. Problem analysis and innovative thoughts are helpful to get you high score.)*

**输出:**

输出结果存放在三个独立的文本文件中，分别命名为"X.label", "Y.label", and "Z.label", 每个文件存放的是对X, Y, Z 数据的预测结果，即每行都是对相应的feature file那行数据的预测结果.

**输出结果评价:**

对每一个预测出来的活动结果(标号从1-24)与真实结果比较，有以下下三种评价标准:

* Precision: 所有预测出的活动结果中，预测对的结果所占的比例。
* Recall: 所有真实的活动样例中, 正确预测出来的结果所占的比例;
* F1-measure :F1(a) = 2\*Precision(a)\*Recall(a)/( Precision(a)+Recall(a) ),
* *Precision: in the instances you predicted as the activity, how many of them are the true activity;*
* *Recall: in all true activity instances, how many of them are correctly predicted as the activity;*
* *The F1-measure of the activity is then calculated as F1(a) = 2\*Precision(a)\*Recall(a)/( Precision(a)+Recall(a) ), which is the harmonic mean of precision and recall.*

# 3. 实验报告要求

a) 对问题的理解分析;

b) 所采用算法的背景介绍(包括算法描述细节及伪代码);

c) 算法评估;

d) 用老师提供的word报告模版撰写，提交word转换的pdf 文件

**提交内容报括:**

       a) "X.label", "Y.label", and "Z.label" 文件;

       b) 实验报告;

       c) 算法源代码.

**注意**: 文件 "X.label", "Y.label", and "Z.label"严格按要求命名

以'姓名\_学号.ZIP'的命名方式zip所有文件上传至课程主页，实验三作业提交处