**《机器学习》上机实验报告**

**学号：10930130448 姓名：吕思勤**

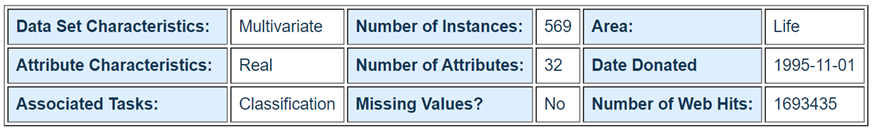
**第二次实验：乳腺癌聚类**

**一、 数据探索**

1.数据来源

https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Diagnostic%29

2.数据描述



假设在这组数据中，只有病人的病理特征，而没有病人的乳腺癌是良性还是恶性的标签信息，请用聚类算法对数据进行聚类，用聚类结果对数据进行标签，将这一标签结果与数据集中的标签进行比较

**二、数据预处理**

1. 数据读取

一張含有 文字 的圖片

自動產生的描述

1. 建立矩阵

一張含有 文字 的圖片

自動產生的描述

聚类时时，由于ID与判断是否得病两列会影响聚类，因此把两列特征值删去

1. 建立标签

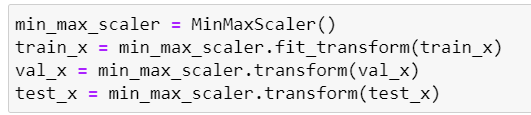


1. 查看标签

一張含有 文字 的圖片

自動產生的描述

1. 归一化



MinMaxScaler的基本上都应该理解数据归一化，本质上是将数据点映射到了[0,1]区间（默认）

计算公式：

一張含有 文字 的圖片

自動產生的描述

**三、模型训练（K-means聚类）**

1.K-means模型参数

n\_clusters：int型，生成的聚类数，默认为8

max\_iter：int型，执行一次k-means算法所进行的最大迭代数。 默认值为300

n\_init：int型，用不同的聚类中心初始化值运行算法的次数，最终解是在inertia意义下选出的最优结果。 默认值为10

init：有三个可选值：‘k-means++’、‘random’、或者传递一个ndarray向量。

１）‘k-means++’ 用一种特殊的方法选定初始质心从而能加速迭代过程的收敛

２）‘random’ 随机从训练数据中选取初始质心。

３）如果传递的是一个ndarray，则应该形如 (n\_clusters, n\_features) 并给出初始质心。

默认值为‘k-means++’。

tol：float型，默认值= 1e-4　与inertia结合来确定收敛条件。

n\_jobs：int型。指定计算所用的进程数。内部原理是同时进行n\_init指定次数的计算。

（１）若值为 -1，则用所有的CPU进行运算。若值为1，则不进行并行运算，这样的话方便调试。

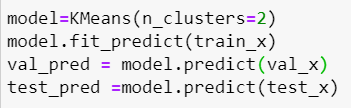
（２）若值小于-1，则用到的CPU数为(n\_cpus + 1 + n\_jobs)。因此如果 n\_jobs值为-2，则用到的CPU数为总CPU数减1。

random\_state：整形或 numpy.RandomState 类型，可选

用于初始化质心的生成器（generator）。如果值为一个整数，则确定一个seed。此参数默认值为numpy的随机数生成器。

2. KMeans模型

(k=2)

****

(k=3)

**一張含有 文字 的圖片

自動產生的描述**

**四、分析与总结**

1. K-means的优缺点

优点：

1.算法快速、简单;

2.对大数据集有较高的效率并且是可伸缩性的;

3.时间复杂度近于线性，而且适合挖掘大规模数据集。K-Means聚类算法的时间复杂度是O(n×k×t) ,其中n代表数据集中对象的数量，t代表着算法迭代的次数，k代表着簇的数目

缺点：

1、在k-measn算法中K是事先给定的，但是K值的选定是非常难以估计的。

2、在 K-means 算法中，首先需要根据初始聚类中心来确定一个初始划分，然后对初始划分进行优化。这个初始聚类中心的选择对聚类结果有较大的影响，一旦初始值选择的不好，可能无法得到有效的聚类结果，这也成为 K-means算法的一个主要问题。

3、当数据量很大时，算法的开销是非常大的。

1. 评价指标NMI

一張含有 文字 的圖片

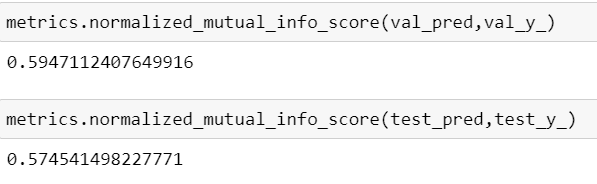
自動產生的描述

1. 结果分析

k=2时:

验证集NMI=0.594

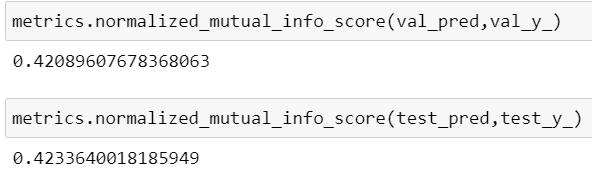
测试集NMI=0.574



**k=3时:**

验证集NMI=0.420

测试集NMI=0.423



显然，由实验结果可知，当K值=3时，聚类效果较佳。