KNN 算法 - 细胞测序

Requirement

对给出的 “单中期分析数据标记-细胞测序数据.csv” 和 “中期分析数据-单细胞测序数据Labels.csv” 进行分类

Solution

Usage: `$ set PYTHONIOENCODING=utf8 && python3 -u ./knn-classification.py`

见 `knn-classification.py`

更名

`单中期分析数据标记-细胞测序数据.csv -> rawdata.csv`
`中期分析数据-单细胞测序数据Labels.csv -> label.csv`

数据导入&处理

使用 `pandas` 库导入 `rawdata.csv`，并删除数据的第一列（属性名称）

导入 `label.csv`，并设置 `header = None`

将数据分类：测试集25%，训练集75%

参数设置 `numOfNearest = 3`

定义下列属性：

- distance: 某一点到训练集内点的距离
- mins: 下表，距离最短n点
- mins_label: 类别

```
distance = np.zeros((numOfTrain,))
mins = np.zeros((numOfNearest,))
mins_label = np.zeros((numOfNearest,), dtype='int32')
```

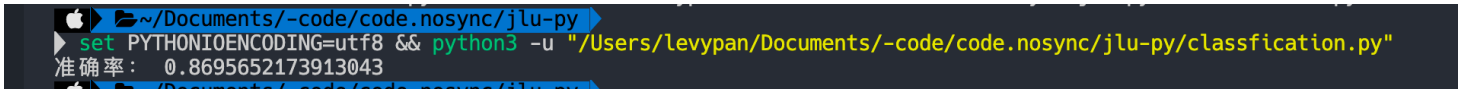
分类测试

将数据应用到模型中进行分类测试，求出最小距离n个点的下标准，并判断它属于哪一类细胞

```
# Find the min
for i in range(0, numOfTrain):
    distance[i] = (sum((X_train[i] - X_test[j]) ** 2)) ** (1/2)
    mins = distance.argsort()[:numOfNearest]


# Classify
for i in range(0, numOfNearest):
    mins_label[i] = y_train[mins[i]]
    result = np.argmax(np.bincount(mins_label))
```

最后给出准确率结果



Reference

- K-近邻算法

Dependency

- sklearn
- pandas
- numpy