浙江工业大学

数据挖掘

计算机科学与技术学院

基于 python 的数据分析

一、目的

熟悉基于 python 的数据分析基本操作：包括 NumPy 的 ndarray 对象的向量和矩阵操作，pandas 的输入导入和基于 DateFrame 对象，sklearn 中数据和算法调用，以及基于 matplotlib 的画图。

二、内容

1、基于 numpy 中 ndarray 对象基本操作

1.1 创建向量 v= [48, 6, 51, 32, 4, 85]，并转成 ndarray 对象类型；

1.2 查看向量 v 的形状、数据类型；

1.3 将向量 v 转成浮点数向量，并查看向量 v 的形状、数据类型；

1.4 对向量 v 前三项求和、后四项的平方求均值；

1.5 向量 v 的奇数项（第一项下标为 0）变成原来的算术平方根；

1.6 求向量 v 最小、最大值及其所在位置（打印下标）；

1.7 对向量 v 排序，并输出排序后的向量以及排序前的索引。

1.8 创建矩阵 M，并转成 ndarray 对象类型:

```
[
    [45,62,31,753],
    [78,43,12,546],
    [146,785,2475,7]
]
```

1.9 查看矩阵 M 的形状、数据类型；

1.10 将矩阵 M 转成浮点数矩阵，并查看矩阵 m 的形状、数据类型；

1.11 矩阵 M 按列、行求和；

1.12 对矩阵 M 的第 0 行和第 1 行求欧式距离（基于 numpy 向量化运算）；

1.13 调用 np.linalg.norm 实现 6)中计算，对比结果。

注意尝试用向量化运算实现 2.7、2.8。

1.14 把矩阵 M 的每一行归一化为单位向量,即向量长度（模）为 1，并打印确认。

1.15 计算 M 中每两行之间的相似度矩阵 S，其中S_{ij}表示 M 中第 i 行与第 j 行的余弦相似度。

提示：两个列向量 x, y 之间的余弦相似度由以下公式计算：

$$s(x,y) = \frac{x^T y}{\|x\|\|y\|}$$

两个向量在已经单位化（向量长度为 1）的情况下，以上公式的分母为 1，所以余弦相似度退化为点积，即以上公式中的分子。

2、基于 pandas 对象 DataFrame 的基本操作

2.1 用 pandas 读取 flights.csv 文件保存为 df，并打印前 5 行。

2.2 用.info()方法查看数据基本统计信息，观察每个特征的是否存在缺失值。

2.3 输出存在缺失值的前 3 个记录（一行为一个记录）。

2.4 计算每个属性的缺失率，即该属性值缺失的记录数与总记录数的百分比，并输出缺失率最大的 3 个属性及其缺失率。

2.5 用 0 填充 df 所有的缺失值，并验证 df 是否还存在缺失值。

2.6 求 df 每个数值列的均值并输出。

2.7 输出 hour 列的中位数

2.8 统计 flights 的特征 distance 在区间[0, 100)、[100, 200)、[200, 500)、[500, 1000)、[1000, 2000)、[2000, 5000)的样本数量，绘制直方图 figure 1。

提示：用 matplotlib.pyplot.hist 来画图。调用 hist 时可以指定参数 bins=6 来指定区间数目。

3、sklearn 中的 iris 数据集

3.1 用 sklearn.datasets 导入 iris 数据集，并输出训练数据（前 5 个）、标签/target、特征名/feature_names、类名/target_names。

3.2 将 iris 数据集的训练数据、标签分别转成 pandas.DataFrame 类型，记为 dfx、dfy，其中 dfy 的列名记为‘target’，并分别输出。

3.3 将 dfx、dfy 拼起来，dfy 在最右边，记为 df，并输出（提示：可以用.concat()方法来拼接）。

说明：在下面题目中，第 i 行/列中的 i 指下标（从 0 开始）

3.4 输出 df 第 1、3、5 行。

3.5 输出 df 最后三列。

3.6 输出 df 第 1 到第 5 行的第 1、3、target 列。

3.7 输出 df 第 3 行第 3 列的元素。

3.8 将 df 保存成 csv 文件，文件名为‘iris_new.csv’。

提示：df.to_csv（‘d:/iris_new.csv’）

3.9 绘制散点图 figure 2a，其中第 0、1 列分别为 x、y 坐标，target 等于 0、1、2 的分别为红、绿、蓝色。