人工智能基础

编程作业 2

完成截止时间：2020/6/27
提交方式：bb 系统中提交
助教： 褚晓萌 cxmeng@mail.ustc.edu.cn
姚舜一 ustcysy@mail.ustc.edu.cn
于博文 yubowen@mail.ustc.edu.cn
段逸凡 dyf0202@mail.ustc.edu.cn

P1：监督学习问题——学生表现预测

实验目的：

本部分实验目的为加强同学们对于 SVM，KNN 以及其他经典机器学习算法的掌握，感受数据科学的魅力。

数据集介绍：

本次实验采用数据集 Student Performance Data Set (https://archive.ics.uci.edu/ml/datasets/Student+Performance)，数据属性包括学生成绩、家庭背景，生活习惯等知识，目的为预测学生最终的成绩（G3）。两个数据集提供了两个不同的科目的数据：数学(mat)和葡萄牙语(por)。其中目标属性 G3 与属性 G2、G1 有很强的相关性，这是因为 G3 是最终成绩(在第三期发布)，而 G1 和 G2 为第一阶段和第二阶段的成绩。在本次实验的实验目的为预测最终成绩 G3，为了简化要求，我们将成绩做二等级制处理，大于等于 10 分为合格，小于 10 分为不合格。

实验要求：

1. 提交一个 main.py，在其中实现数据的读取，测试集训练集的划分(7:3)，算法的调用，结果的评价等。可以对数据进行适当的预处理，对数据属性进行任意加工处理，比如删减、降维、组合等。
2. 提交一个 KNN.py 文件，在其中实现 K 近邻算法模块来解决二分类的问题：
自己实现 knn 算法并在 main.py 文件中调用解决预测学生的 G3 成绩是否合格的问题。允许使用 sklearn.preprocessing 中的 LabelEncoder()函数将数据集中的字符型属性转换成整型，其他不允许调库，请自己实现。
3. 提交一个 SVM.py 文件，在其中实现 SVM 模块解决二分类问题：
    1. 要求实现支持软间隔与除线性核外至少一种核函数的 SVM。根据数据的特点，选择你认为合适的核函数进行实现。
    2. 函数的参数应至少包含 trainset，trainlabel，testset，C（软间隔的参数），kernel（使用的核函数），以及其他在你的算法中对结果起重要影响的参数，方便在实现算法后进行调参优化。返回值为 predictlabel，在 main.py 中进行评测。
    3. 在实验报告中，关于本部分内容应至少包括
        (1) 采用核函数与否对实验结果的影响，和你使用该核函数的原因。（如果没有原因，可以多实现几种核函数进行比较测试）
        (2) 对你实现的算法进行描述。并在代码中进行注释，至少让助教可以看懂每一块代码的功能。
    4. 在实现算法的过程中，不允许调用 SVM 算法库与计算优化库。
4. 提交一个 other.py，实现其他的机器学习算法。
    (1) 选择一个你感兴趣的机器学习算法（课内或课外），进行相关资料的查询，学习相关库的使用。在本数据集中，选择一个你感兴趣的标签进行预测。
    (2) 在实验报告中，对你设计的任务，所用的方法与实验结果进行描述。
5. 对于 KNN 与 SVM 算法，应评测使用属性 G1、G2 和不使用 G1、G2 时的性能。评价指标如下：

$$\text{F1 score} = \frac{2 \times P \times R}{P + R}$$

准确率 P = TP / (TP + FP)，召回率 R = TP / (TP + FN)

真正例（True Positive，TP）：真实类别为正例，预测类别为正例。
假正例（False Positive，FP）：真实类别为负例，预测类别为正例。
假负例（False Negative，FN）：真实类别为正例，预测类别为负例。
真负例（True Negative，TN）：真实类别为负例，预测类别为负例。

P2：无监督学习问题(30%)

问题描述：
本实验需要同学使用 PCA 算法对实验数据进行降维，并且使用 kmeans 算法对降维后数据进行聚类及可视化。请结合课上学习的内容以及自行查阅的资料完成实验。
数据集介绍：数据集是自意大利同一地区但来自不同品种的葡萄酒的化学分析，是一经典的分类数据集，数据集共 13 个维度，第一个维度为葡萄酒的实际品种，其他维度均为葡萄酒化学分析特征，数据集的其他相关信息可见
http://archive.ics.uci.edu/ml/datasets/Wine
实验要求
1 数据预处理：
该数据集未经过预处理，请使用数据除第一维以外部分，对数据进行缩放到合理范围，标准化。
2 实现 PCA 算法：
使用 pca 算法处理预处理后的数据，要求提交一个 python 函数 PCA（data, threshold）其中 threshold 表示特征值的累计贡献率。即选择前 m 个特征向量，使得

$$\frac{Sum(first\ m-1\ eigenvalues)}{Sum(all\ eigenvalues)} < threhold \leq \frac{Sum(first\ m\ \ eigenvalues)}{Sum(all\ eigenvalues)}$$

返回值为降维后的矩阵。

3 实现 kmeans 算法：
基于降维后的数据，使用 kmeans 算法将数据进行聚类，计算分为不同数量类别的轮廓系数(Silhouette Coefficient)和兰德系数，根据轮廓系数选择最优的聚类数量，输出聚类结果。
使用未经过 pca 算法处理的数据，使用 kmeans 算法将数据进行聚类，计算分为不同数量类别的轮廓系数(Silhouette Coefficient)，根据轮廓系数选择最优的聚类数量，输出聚类结果，比较该聚类结果与前者的兰德系数，分析结果。
要求提交一个 python 函数 KMeans(k，data），data 为需要聚类的数据，k 为聚类后的数量。要求以元组的形式返回聚类后的数据和聚类的轮廓系数。并将前者保存至 csv 文件中。然后根据数据真实的分类计算兰德系数。

备注：

1)兰德系数：$RI = \frac{a+d}{a+b+c+d}$

假设用 C 表示真实的分组情况，K 表示聚类结果，那么：
a 为在 C 中为同一类且在 K 中也为同一类别的数据点对数
b 为在 C 中为同一类但在 K 中却隶属于不同类别的数据点对数
c 为在 C 中不在同一类但在 K 中为同一类别的数据点对数 d 为在 C 中不在同一类且在 K 中也不属于同一类别的数据点对

2) 轮廓系数：$S(i) = \frac{b(i)-a(i)}{\max(a(i),b(i))}$

其中a(i) = average (i 向量到所有它属于的簇中其它点的距离)

其中b(i) = average (i 向量到与它相邻最近的一簇内的所有点的平均距离)

将所有点的轮廓系数求平均，就是该聚类结果总的轮廓系数

3）kmeans 距离度量使用欧式距离

4 实验报告要求：

1 分析不同$threhold$的降维结果

2 图表分析不同数量类别的轮廓系数。分析降维前后的 kmeans 聚类结果。

作业要求：

1. 使用 python 实现算法，**不可以**调用 sklearn 等机器学习库
2. 实验报告使用 PDF 格式提交，实验报告包含以下几点：
   1）算法思想
   2）实验结果说明与分析。

实验提交：

1. 提交方式：**bb 系统中提交**
2. **请组织好文件结构**，提交的目录结构树应如下例所示：

```
PBXXXXXXXX__张三_exp2\
    |---supervise\
        |---src\
            |---main.py
            |---KNN.py
            |---SVM.py
            |---other.py(可以自己取名字)
        |---data\
            |---数据（不用上传）
        |---report1.pdf
    |---unsupervise\
        |---src\
            |---{your_code}
        |---input\
            |---{your_input_file}
        |---output\
            |---
        |---{readme.txt}
        |---report2.pdf
```

   将文件夹 **PBXXXXXXXX__张三_exp2** 压缩为 **PBXXXXXXXX__张三_exp2.zip**，将压缩包提交
3. **请务必按时完成实验，不接受逾期提交的实验。**
4. 实验中有任何问题请联系助教。