机器学习实验报告

1. **问题描述：**

使用svm、XGBOOST、Random forest三种机器学习方法对大于1000的数据集进行学习和分类。

1. **数据集描述：**

本次实验，选用了我在本科毕业设计时，使用Gabor滤波方法对肺部结节CT影像提取的维纹理特征做数据集。纹理特征数据集是360维向量，和1或-1的标记（1代表患有肺癌，-1代表没有患肺癌）。数据集共3843条数据，其中1496条未患肺癌记录，2347条患肺癌记录。

数据集样例：

1.233433194 -0.271845552 …… -1.087078439 1.334201965 -0.164718962 1

1. **实验环境：**

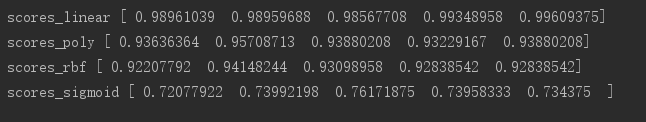
使用python，应用sklearn库。

1. **实验过程和结果：**

**（一）SVM方法**

对原始数据进行随机，使两类数据随机分布，使用sklearn中的svm库进行学习分类，使用cross\_val\_score()进行5组交叉验证。分别使用linear、poly、rbf、sigmoid四种kernel进行机器学习。得到结果如下表所示：

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | 第一组 | 第二组 | 第三组 | 第四组 | 第五组 | 平均值 |
| linear-kernel | 0.98961039 | 0.98959688 | 0.98567708 | 0.99348958 | 0.99609375 | 0.99089354 |
| poly-kernel | 0.93636364 | 0.95708713 | 0.93880208 | 0.93229167 | 0.93880208 | 0.94066932 |
| rbf-kernel | 0.92207792 | 0.94148244 | 0.93098958 | 0.92838542 | 0.92838542 | 0.93026416 |
| sigmoid-kernel | 0.72077922 | 0.73992198 | 0.76171875 | 0.73958333 | 0.73437500 | 0.73927566 |

****结果截图：

**（二）XGBOOST**

使用XGboost库，构建机器学习模型，进行了五组交叉实验，分别选取

param = {'max\_depth': 2, 'eta': 1, 'silent': 1, 'objective': 'binary:logistic'}，round=2.

param = {'max\_depth': 2, 'eta': 1, 'silent': 1, 'objective': 'binary:logistic'}，round=5.

param = {'max\_depth': 5, 'eta': 1, 'silent': 1, 'objective': 'binary:logistic'}，round=2.

param = {'max\_depth': 2, 'eta': 3, 'silent': 1, 'objective': 'binary:logistic'}，round=2.

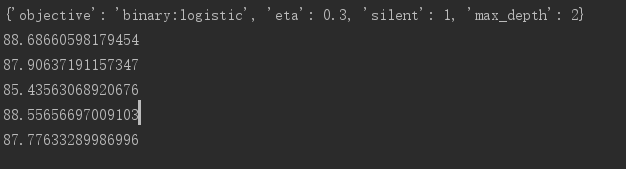
param = {'max\_depth': 2, 'eta': 0.3, 'silent': 1, 'objective': 'binary:logistic'}，round=2.

进行建模测试。

得到结果如下表所示：

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | 第一组 | 第二组 | 第三组 | 第四组 | 第五组 | 平均值 |
| 参数一 | 87.25617685 | 87.51625488 | 86.73602081 | 89.20676203 | 87.38621586 | 87.62028609 |
| 参数二 | 93.3680104 | 91.93758127 | 92.32769831 | 90.89726918 | 91.28738622 | 91.96358908 |
| 参数三 | 93.49804941 | 93.23797139 | 97.26918075 | 96.74902471 | 95.05851756 | 95.16254876 |
| 参数四 | 30.29908973 | 32.11963589 | 33.42002601 | 41.48244473 | 30.68920676 | 33.60208062 |
| 参数五 | 88.68660598 | 87.90637191 | 85.43563069 | 88.55656697 | 87.7763329 | 87.67230169 |

实验截图：

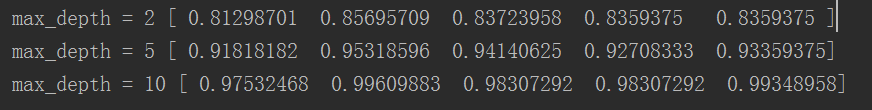


**（三）Random Forests方法**

使用sklearn中ensemble模块的RandomForestClassifier函数构建随机森林机器学习模型，使用cross\_val\_score()进行5组交叉验证。分别取max\_depth为2,5,10，使用gini指数作为评估标准，取10个树，进行机器学习。得到结果如下表所示：

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| max\_depth | 第一组 | 第二组 | 第三组 | 第四组 | 第五组 | 平均值 |
| 2 | 0.81298701 | 0.85695709 | 0.83723958 | 0.8359375 | 0.8359375 | 0.83581174 |
| 5 | 0.91818182 | 0.95318596 | 0.94140625 | 0.92708333 | 0.93359375 | 0.93469022 |
| 10 | 0.97532468 | 0.99609883 | 0.98307292 | 0.98307292 | 0.99348958 | 0.98621179 |

结果截图：



1. 实验总结

三种机器学习方法，在改变参数时的准确率变化规律符合算法和参数的特点，总体的准确率较为理想。通过本次实验，熟悉了sklearn和xgboost的使用，尤其在xgboost的安装编译过程中遇到各种问题也最终一一结果。

本次实验的源代码均已上传至github：<https://github.com/LumingSun/ML_homeword>