一、简介

有关逻辑回归的解释、数学描述以及 sklearn.linear_model.LogisticRegression 的使用参见我的博客：https://blog.csdn.net/qq_38384924/article/details/97499694
其它步骤有不理解的也可以查看我的 CSDN 博客的一些笔记：
https://blog.csdn.net/qq_38384924

二、主要步骤

第一步，数据预处理：对 KDDCUP99 原始数据集进行处理，使其成为适合进行学习的形式。进行特征选择，降低维度。
第二步，调参：网格搜索最优参数组合。
第三步，模型训练与测试：利用训练集进行模型训练，利用测试集进行测试，计算各种评价指标，绘制 ROC 曲线等。

三、具体执行

3.1、数据预处理

由 data_processing.py 和 select_features.py 两部分组成。
前者对数据集进行处理，得到适合学习的形式。后者进行特征选择，进行降维。

data_processing.py

- 首先读取原始数据集；
- 然后对其类标进行转换，正常和攻击的标称转化为 0 和 1；
- 对连续属性进行离散化处理，使用聚类离散化的方法；
- 对标称离散属性进行 one-hot 独热编码处理；
- 全体数据标准化、归一化；
- 将处理后的数据集导出。

select_features.py

- 首先读取经过 data_processing 后的数据集；
- 利用随机森林进行特征选择；
- 将选择的特征输出，以待后续使用。

3.2、调参

利用网格搜索进行调参，logistic_search.py

logistic_search.py

- 首先利用 load_data(file_train,file_fea)函数读取经过 data_processing 后的训练数据集以

及选择的特征，然后得到训练数据集 x_train 和类标集 y_train；

- 利用 optimize_params(x_train,y_train)函数，进行网格参数搜索，得到最优参数。

3.3、模型训练与测试

logisitic_train_test.py

- 首先利用 load_data(file_train,file_test,file_fea)函数读取经过 data_processing 后的训练数据集以及选择的特征，然后得到 x_train,y_train,x_test,y_test ;
- 利用 classify(x_train,y_train,x_test)对逻辑回归模型进行训练，得到模型 LR，并对测试集 x_test 进行预测得到 y_predict ；
- 利用 calculate_metrics(y_test, y_predict,y_score) 函数计算各种指标（报告）：accuracy_score 、 confusion_matrix 、 classification_report 、 hamming_loss 、 jaccard_similarity_score、matthews_corrcoef、zero_one_loss、log_loss 等；
- 利用 ROC_curve(X_test,y_test,LR)绘制 ROC 曲线计算 AUC。

;