第二次竞赛文档

1. 实现思路
2. 收集数据：提供 csv 格式文件
3. 准备数据：编写函数 load\_csv(filename) ，读取训练集和测试集数据
4. 分析数据：用 excel 打开csv 文件中的数据，发现里面有的数据是以 ‘？’ 的形式存储的。在程序中需要对其实现转换
5. 训练算法：pca降维、 svm 支持向量机
6. 测试算法：编写函数使用文件test.csv中的数据作为测试样本，测试样本与非测试样本的区别在于测试样本是未知的类别，非测试样本是已知类别。
7. 使用算法
8. 实现过程：
9. 数据预处理
10. 使用pandas读取数据，此时数据为DataFrame 类型。

使用 Pandas 读数据的主要原因：可以使用 DataFrame 中的replace 方法来代替训练集或测试集中的 ‘？’。

1. 实现方法：由于要用每一列的均值来代替缺省值，而直接从 csv 中读到的数据中 ‘？’ 是 String 类型，无法计算均值。所以先用0 来代替 ‘？’ ，代替之后数据中所有的数据就都可以转换为数值类型。由于每一列是一个特征，所以用每一列的均值代替每一列中的缺省值。再将数据转换为 np.array() 类型。则此时数据则可以进行处理
2. PCA 算法
3. 导入PCA 库



1. 对训练集进行降维



pca 是对 PCA 模型的实例化，此时 根据 n\_components=9 可知此时数据集已经降维9 个特征值的数据。利用 x（读取出来的训练集）对pca 进行训练。最终 train\_X 是降维之后的训练集。

1. 对测试集进行降维



由于 pca 为已经训练好的模型，所以此时直接将测试集进行转换。

1. SVM 对数据进行训练及预测
2. 导入 SVC 库



1. 初始化 SVC 模型



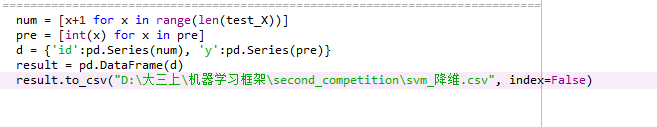
1. 使用 SVC对数据进行训练



1. 用 SVC 对测试集数据进行预测，并存在列表中



1. 将预测值存储在csv 文件中



1. 实验总结

因为此次的数据特征值有一点多，所以我认为应该对其进行适当地降维。鉴于之前的学习，SVC的正确率比KNN的正确率稍高些，所以选取SVM中SVC模型。此次实验让我了解了如何处理数据中的缺省值，以及对sklearn中的库理解更深刻。