|  |
| --- |
| 哈尔滨工业大学（深圳） |
| **大数据导论大作业报告** |
| 题目：特定疾病的回归和分类 |
| |  |  | | --- | --- | | **姓 名** | **许峻玮** | | **学 号** | **190110105** | | **报告日期** | **2021年12月31日** | |

1. 实验目的

本实验需要我们充分利用理论课里所学的知识，将抽象模型具体化，使用代码通过对患病病人的临床数据和体检指标，来预测人群指示病情程度的指标。 在完成基本预测的前提下，需要进一步地设计高效、解释性强的算法使得模型的预测能力更上一层楼。

1. 实验内容分析

本次实验分为两个部分——回归和预测。

回归部分，给出数据集，针对原始数据进行合理且充分的清洗，在此基础上构建模型，评估模型的准确率和召回率，最后对测试数据的标签列进行预测。

类似地，预测部分在构建模型后需要评估模型的均方误差，最后对测试数据的标签列进行预测。

图示

描述已自动生成通读要求不难发现实验可按以下步骤进行。

接下来的两部分实验均遵照上述流程图。

1. 实验过程及结果

包括算法实现的主要步骤，算法实现的关键代码，算法运行结果截图，算法性能曲线图及结果分析等。

3.1 预测部分

3.1.1 数据预处理

3.1.1.1 缺省值

首先，观察数据发现某些区域明显相对其他区域“稀疏”，考虑对缺省数据进行处理，纵向的缺省数据清洗可以使特征降维，在提高模型准确性的同时避免参数过多造成过拟合或训练困难的问题；横向的缺省数据清洗则可以滤除效用不大的个体数据，这一部分数据由于缺省字段过多，很可能不仅对模型的构建帮助甚微，甚至会有帮倒忙的现象。因此我们需要对数据中的缺省进行处理。

文本

描述已自动生成缺省数据统计由下面的函数完成。

表格

描述已自动生成表格

中度可信度描述已自动生成结果如下，篇幅有限，仅列出缺省数据最多的10个行列。

经分析，删除特征“RBP4”、“分娩时”、“SNP22”、“SNP21”、“SNP23”、“SNP54”、“SNP55”、 “ACEID”.

针对行缺失，设定阈值，即删去超过20个缺省字段的行数据。

文本

描述已自动生成 删除代码如下图所示。

尽管删除了缺省过多的行与列，但数据中仍存在缺省值，这对后面的模型构建是不可接受的，因此使用均值对缺省值进行填充。特别地，在训练集中将数据根据标签列分为两部分，两部分的均值分别计算。实践证明单独对每一类计算均值相比不分类较优。

文本

描述已自动生成 填充缺省值的代码如下。

3.1.1.2 相关性分析

文本

描述已自动生成 易知，大数量的特征中极易出现与标签特征不相关或相关极大的特征，这些特征的存在或多或少对模型存在负面影响，因此下面进行相关性分析。

表格

描述已自动生成 基于数据不服从正态分布且多为非连续变量的事实，实验采用Kendall相关系数，结果如下，篇幅有限，仅列出正相关前10的数据。

删去相关系数小于0.01的特征：“SNP49”、“SNP30”、“SNP9”、“SNP7”、“SNP44”、“SNP37”、“SNP5”、“LDLC”、“SNP8”和“SNP19”.

另外我们观察到，“身高”、“孕前体重”、“BMI分类”和“孕前BMI”表述内容相近，仅保留“BMI分类”这一特征。

3.1.1.3 标准化

图形用户界面, 应用程序

描述已自动生成 不同特征的数据存在量级上的差异，对此采用标准化的方式缩放数据。

这一步也是为下一步的方差分析做准备。

3.1.1.4 方差滤波

图形用户界面, 应用程序

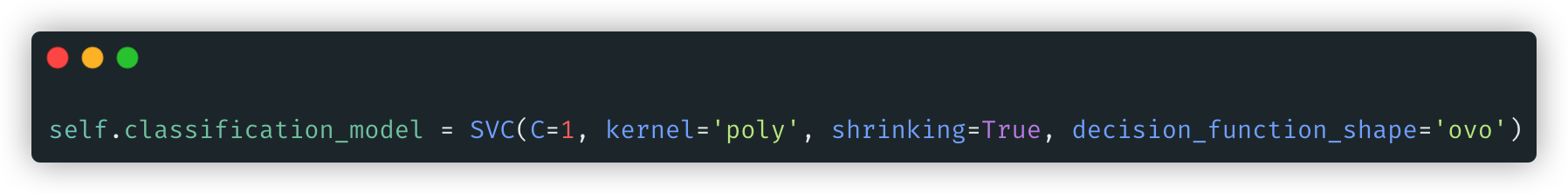
描述已自动生成 方差较小的特征意味着这一特征在所有个体中的变化并不明显，也就是说它可能不是一个较好的分类指标，因此需要将这样的特征删去。

表格

描述已自动生成 方差最大的10个特征如下图所示。

过滤方差小于0.01的特征：“糖筛孕周”、“ALT”、“CHO”和“ApoB”.

3.1.2 模型的选择与实现

 选择SVM作为本次实验的二分类器。

惩罚系数设定为 1，核函数选择多项式。

3.1.4 模型的预测

表格

低可信度描述已自动生成 模型对测试集的预测如下，篇幅有限，仅列出前10项。

3.2 回归部分

3.2.1 数据预处理

3.2.2 模型的选择与实现

3.2.3 模型的修正与改进

3.2.4 模型的预测

1. 实验心得

实验完成后的感悟与总结。