

**软件工程**

**个人项目说明文档**

项目名称：\_\_\_机器学习方法预测糖尿病\_\_

学 号： 2018229049\_\_\_\_\_

姓 名： 吴振宇 \_\_ \_\_\_

国际工程师学院

2019年11月24日

个人项目的github网址是：<https://github.com/luoyehuanlaiyiji/PersonalProject-of-SoftwareEngineering>

**1 引言**

该个人项目的实现使用了机器学习的方法来预测房屋的价格，方法包括以下几个步骤：读取数据，检查数据，数据可视化，数据预处理，特征选取，评估，选择模型，保存模型等多个步骤，最终对机器学习的流程更加明确。

接下来的内容主要是对代码的解释。

**2代码解释**

**2.1读取数据**

对应代码文件是code/pandas\_read.py。

首先是导入项目所需要的类库，然后就是导入pima数据集，这里使用Pandas类库导入CSV文件的数据，使用pandas.read\_csv()函数返回关于数据的DataFrame。

**2.2检查数据**

对应代码文件是code/check\_data.py。

通过对数据的简单审视和观察，可以发现数据之间的内在关系，有助于对数据进行整理。代码中可以查看前十条数据，数据的维度，数据属性和类型，数据的描述性统计结果，数据分组分布，数据属性之间的相关性，数据的离散偏离程度。

**2.3数据可视化**

对应代码文件是code/data\_visual.py。

为了生成最优化的算法模型，必须对数据进行理解，最快，最有效的方式是通过数据的可视化，本代码使用Matplotlib库来可视化数据。直方图：用横轴表示数据类型，纵轴表示分布情况，可以非常直观地展示每个属性的分布状况，可以看出数据是高斯分布、指数分布还是偏态分布；密度图：一种表现与数据值对应的边界或域对象的图形表示方法，一般用于呈现连续变量，用平滑的线来描述数据的分布；箱线图：用于显示数据离散情况的统计图；相关矩阵图：用来展示两个不同属性相互影响的程度；散点矩阵图：表示因变量随着自变量变化的大致趋势。

**2.4数据预处理**

对应代码文件是code/data\_preprocess.py。

数据预处理大致分为三个步骤：数据的准备、数据的转换、数据的输出。本代码使用scikit-learn类库来处理数据，并将处理后的数据应用到后边的算法中去。

调整数据尺度：将不同计量单位的数据统一成相同的尺度，利于对事物的分类和分组，也就是对数据进行缩放的过程。

正态化数据：有效的处理符合高斯分布的数据的手段。

标准化数据：归一化处理，适合处理稀疏数据。

二值化数据：使用阈值将数据转化为二值，大于阈值设置为1，小于阈值设置为0。

**2.5特征选取**

对应代码文件是code/select\_feature.py。

特征选取能够选择有助于提高预测结果准确度的特征数据，如果数据中包含无关的特征属性，会降低算法的准确度。

卡方检验-单变量特征选定：统计样本的实际观测值与理论观测值之间的偏离程度。

递归特征消除：使用一个基模型来进行多轮训练，每轮训练后消除若干权值系数的特征，通过每个基模型的精度，找到最终的预测结果影响最大的数据特征。

主要成分分析：也成数据降维，本质是将原始的样本映射到维度更低的样本空间中，使用线性代数来转换压缩数据。

袋装决策树：有专门的函数来计算数据特征的重要性。

**2.6评估**

对应代码文件是code/evaluation.py。

分离训练数据集和评估数据集：简单地将原始数据分为两部分，第一部分用来训练算法生成模型，第二部分通过模型预测结果，并于已知的结果进行比较。

K折交叉验证分离：将原始数据分为K组（一般是均分），将每个数据子集分别做一次验证集，其余的K-1组最为训练集,这样会得到K个模型，再用K个模型分类准确率的平均数作为性能指标。

弃一交叉验证分离：其实就是N-1个交叉验证。

重复随机分离评估数据集与训练数据集。

**2.7选取模型**

对应代码文件是code/select\_model.py。

给出了一种对于多种算法准确性进行分析比较的方法，通过这个方法可以找到一种或者两种算法对给定数据集能够生成准确度最高的模型，从而选择合适的算法。

**2.8保存模型**

对应代码文件是code/save\_model.py。

提供了两种对生成的模型进行序列化的方法，一个是通过pickle序列化和反序列化（加载模型）学习模型，另一种是通过joblib序列化和反序列化（加载模型）学习模型。