**工业机理模型评估的朴素贝叶斯分类模型使用说明**

1. 模型用途

根据工业机理模型的应用访问记录，用户评价等数据来对工业机理模型等级进行分类。

1. 数据说明

输入数据用变量x表示，共有8个特征：

第一维元素x(0)表示：模型上架使用时间,

第二维元素x(1)表示：模型辐射度 ,

第三维元素x(2)表示：用户评价 ,

第四维元素x(3)表示：模型应用访问记录,

第五维元素x(4)表示：模型参数参数,

第六维元素x(5)表示：模型属性(可用1、2、3、4表示),

第七维元素x(6)表示：模型属性个数

第八维元素x(7)表示：模型辐射度 (一个0-1之间的数)。

输出数据用y表示，只有1个特征：

工业机理模型等级。

1. 算法原理

朴素贝叶斯分类是一种十分简单的分类算法，叫它朴素贝叶斯分类是因为这种方法的思想真的很朴素，朴素贝叶斯的思想基础是这样的：对于给出的待分类项，求解在此项出现的条件下各个类别出现的概率，哪个最大，就认为此待分类项属于哪个类别。通俗来说，就好比这么个道理，你在街上看到一个黑人，我问你你猜这哥们哪里来的，你十有八九猜非洲。为什么呢？因为黑人中非洲人的比率最高，当然人家也可能是美洲人或亚洲人，但在没有其它可用信息下，我们会选择条件概率最大的类别，这就是朴素贝叶斯的思想基础。

（一）数据预处理

工业机理模型评估的朴素贝叶斯分类模型的输入数据共有八个维度，然而各个特征之间量级有很大的差异。本模型使用的算法是基于距离的，因此其对于数据的数量级十分敏感，因为数量级的差异将导致量级较大的特征占据主导地位。所以我们需要对与这些数据特征进行标准化，将它们都缩放到相似的范围。

业机理模型评估的朴素贝叶斯分类模型可以使用两种方法对各个特征进行数据的标准化：min-max标准化和z-score标准化。其中，min-max标准化的公式为:, min和max分别为第i个特征的最小值和最大值。这可以将原始值映射到[0,1]区间内。z-score标准化的公式为,和分别为第i个特征对应样本集的均值和标准差。这使得原始值满足正态分布。

（二）数学建模

在工业机理模型评估的朴素贝叶斯分类模型中，当给定训练集时，我们无非就是先计算出所有的先验概率和条件概率，然后把它们存起来（当成一个查找表）。当来一个测试样本时，我们就计算它所有可能的后验概率，最大的那个对应的就是测试样本的类别，而后验概率的计算无非就是在查找表里查找需要的值。以此来根据焊接历史数据来对焊接工艺质量等级进行分类。

1. 使用方法

运行train.py文件依据历史数据得到模型参数（仅用于训练模型，正常使用时不需运行），具体输入命令：

**python train.py --train\_dir 训练数据存放地址**

例如：

**python train.py --train\_dir ./trainDataCTTZ.txt**

之后根据模型参数，运行model.py文件得到预测工时，具体输入命令：

**python model.py --train\_dir 训练数据存放地址 --predict\_data 预测数据存放地址**

例如：

**python model.py --train\_dir ./trainDataCTTZ.txt --predict\_data ./predictData.txt**

预测结果会存放在**./result/predictResult.txt**文件中