**工业机理模型评估的K-最近邻分类模型使用说明**

1. 模型用途

根据模型应用访问记录，用户评价等信息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

焊缝质量等级的K-最近邻分类模型的核心思想是将焊缝质量的历史数据的属性用n个属性描述，每个元组代表n维空间中的一个点，所有的训练元组存放在n维的模式空间。当给定一个未知元组时，搜索该模式空间，找出最接近未知元组的k个训练元组。未知元组指派到它的k个最近邻中的多数类。“邻近性”用距离度量，如欧几里德距离。

（一）数据预处理

工业机理模型等级的K-最近邻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。

工业机理模型等级的K-最近邻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

工业机理模型等级的K-最近邻分类模型只是假设数据都分布在欧式的特征空间内（以特征值为坐标区分不同样本的空间），然后我们恰好又知道全部数据在这个空间中的位置以及其中一部分数据的类别。那么现在我们怎么判断剩余那些数据的类别呢？

为了让分类进行下去，这里其实我们假设：空间中距离越近的点属于一类的可能性越大。

有了这条“公理”，那事情就好办多了。我们只需要计算每个待分类数据到全部已知类别数据的距离就好了。经过多次迭代，使得损失误差减小到合理的程度，返回该回归模型的参数，以此对新的焊缝质量等级进行预测。

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