中医诊断系统

数据处理部分

1. **数据库结构**
2. **数据库结构**
3. **表结构**

user表

medicine表

symptom表

test表

result表

1. **数据预处理**

从表中读出所有记录，分别将每条的记录由一对多（一种药材对应多种症状）分解成一对一（一种药材对应一种症状）的多条记录并且将这些记录存入数组中，最后所有的基础数据全部处理结束之后，将这个数组做成一个矩阵，方便接下来的操作。

向量化、矩阵化操作是机器学习的追求。从数学表达式上看，向量化、矩阵化表示更加简洁；在实际操作中，矩阵化更高效。

首先我们需要一张词典，该词典囊括了训练数据集中的所有必要药材。其次，根据词典矩阵化每个处理后的记录。具体的，每条记录都定义为词典大小，分别遍历记录中的每个药材并统计出现次数；最后得到一个和词典一样大小的矩阵，这些矩阵由一个个整数组成，每个整数代表了词典上一个对应位置的药材在训练数据中的出现频率。最后，统计记录总数，某一数据药材频数除以相应类别的记录总数，即得到相应的条件概率

1. **朴素贝叶斯**

贝叶斯定理的许多应用之一就是贝叶斯推断，一种特殊的统计推断方法，随着信息的增加，贝叶斯定理可以应用于更新假设的概率。在决策理论中，贝叶斯推断与主观概率密切相关，通常称为贝叶斯概率

贝叶斯推断根据先验概率和统计模型导出的似然函数的结果，再有贝叶斯定理计算后验概率：

P(A|B)=

这里A表示药材，B表示症状。

P(A|B)：后验概率，即病人有B症状时，药方里含有A药材的概率

P(B|A)：药方中有A药材的情况下，观测到病人有B症状的概率

P(B)：边际似然，所有的假设都相同，因此不参与决定不同假设的相对概率，即可以忽略。

我们需要求的是当P(A|B)取得最大值时A的取值

A\*=P(B|A)P(A)

P(A)在数据预处理的时候已经得到所以只需要求得P(B|A)即可。

对连续数据建模的合理猜测是多元高斯或多元正太分布：

P(B|A)=

由于唯独很高时概率非常小，将使用对数似然：

D表示维度

对A\*=P(B|A)P(A)两边取对数得：

A\*=

1. **计算并生成结果**

计算前首先检测test表里面是否有数据，如果有数据则读出表中的记录并且针对每一条记录中的症状，算出相应的所有的药材的概率，并按概率的大小进行排序，取概率最大的15种药材写入result表中，最后删除test表中的已处理的记录。

1. **总结**

不同于其他分类器，朴素贝叶斯是一种基于概率理论的分类算法；

特征之间的条件独立性假设，显然这种假设显得“粗鲁”而不符合实际，这也是名称中“朴素”的由来。然而事实证明，朴素贝叶斯在有些领域很有用。

在具体的算法实施中，要考虑很多实际问题。比如因为“下溢”问题，需要对概率乘积取对数等

总体来说，朴素贝叶斯原理和实现都比较简单，学习和预测的效率都很高，是一种经典而常用的分类算法。