关联规则挖掘

——数据挖掘第二次作业

姓名：于文楠

学号：2120151057

学院：计算机学院

班级：硕士2班

邮箱：[ywnbit@163.com](mailto:ywnbit@163.com)

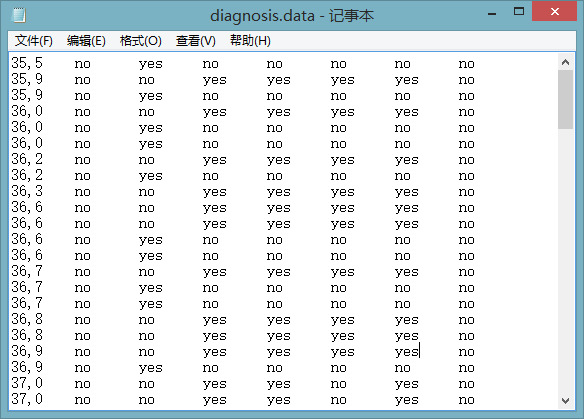
## 数据集处理：

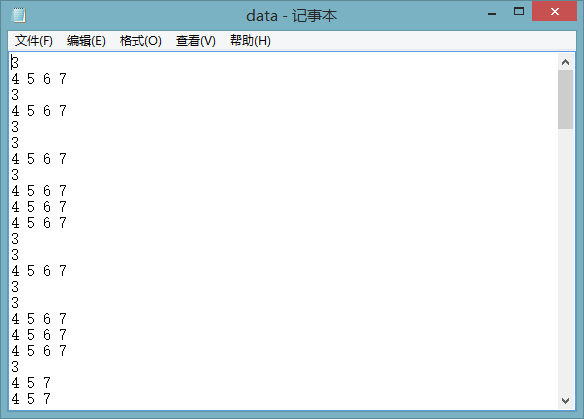
采用“Acute Inflammations Data Set”数据集，数据集一共有8个属性：Temperature of patient（病人的体温），Occurrence of nausea（恶心），Lumbar pain（腰痛），Urine pushing (continuous need for urination) （尿频），Micturition pains（尿痛），Burning of urethra, itch, swelling of urethra outlet（尿道灼烧），Inflammation of urinary bladder（膀胱炎），Nephritis of renal pelvis origin（肾炎）。

1、处理数据集：

1. 判断体温，若大于等于38则为yes，小于38则为no。
2. 提取出每一行中所有属性为yes的下标。

2、如图所示：





## 频繁项集

1、其中包含两个指标：

1）支持度： 包含频繁项集F的集合的数目

2）可信度：频繁项F与某项j的并集 （即F U {j}）的支持度 与 频繁项集F的支持度的比值

2、频繁项集两个定理：

1）子集定理：频繁项集的子集都是频繁项集，而非频繁项的超集都是非频繁项集。

2）合并/连接定理：由k-1项集，向k项集进行合并。当两个k-1项集，拥有k-2个相同元素时，才能合并成k项集。

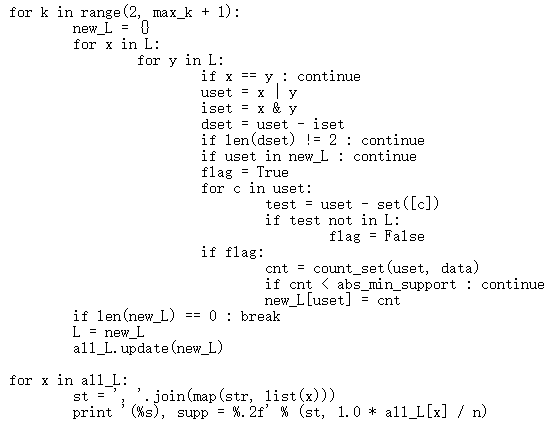
注：定义最小support值为0.4，confidence值为0.7，筛去数据中不符合support值的数据。

3、对求出的频繁项集进行合并，采取逐步合并的方式，流程类似如下：

原始数据 --> 一级构建 --> 一级过滤 --> 二级构建 --> 二级筛选 --> 三级......

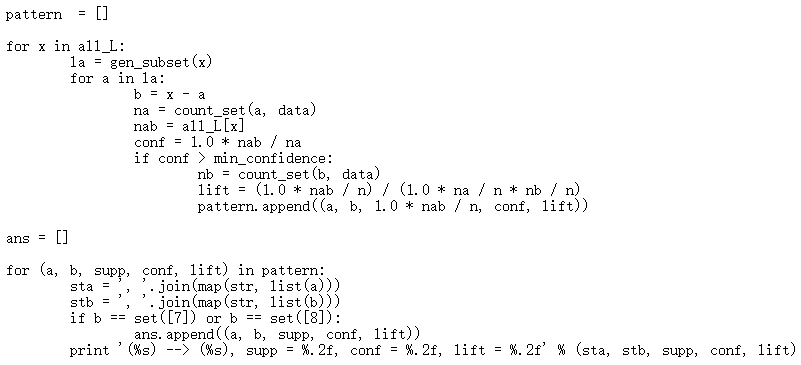
第k级会生成 k项集并过滤掉非频繁项集，保留下k项频繁项集。每一级的构建都会引起数据量的增加，而每一级的过滤后，会大大减少数据量，便于下一级的运算。逐步过滤和合并，直至生成极少数量的频繁项集，或者极大频繁项集。

4、采用Apriori算法，实现如下：



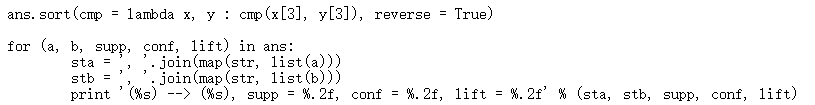
## 求出关联规则

将频繁项集拆开，进行枚举，对所有关联求其confidence，筛去不符合定义的最小confidence值的关联规则。



## 去除冗余规则

筛选出导出为第7属性和第八属性的关联规则：



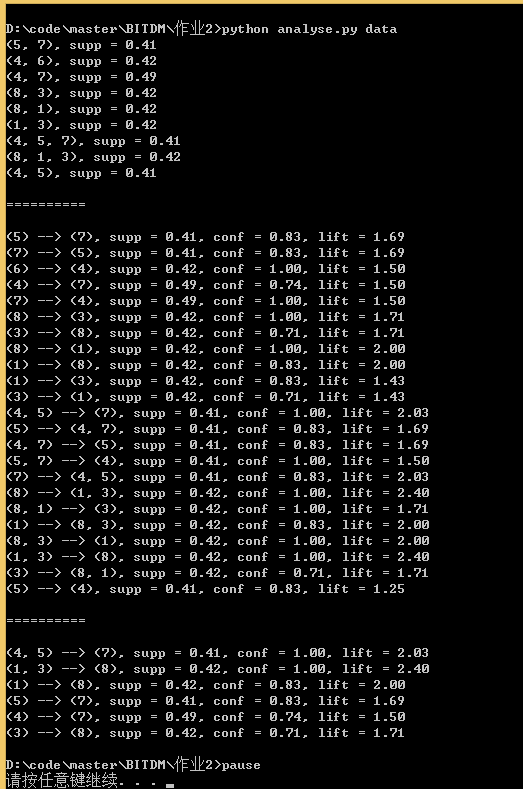
## 评价规则

使用lift对规则进行评价：

lift(X -> Y) = lift(Y -> X) = conf(X -> Y)/supp(Y) = conf(Y -> X)/supp(X) = P(X and Y)/(P(X)P(Y))

称为A条件对于B事件的提升度,如果该值=1，说明两个条件没有任何关联,如果<1，说明A条件(或者说A事件的发生)与B事件是相斥的，一般在数据挖掘中当提升度大于3时，我们才承认挖掘出的关联规则是有价值的。

## 挖掘结果及分析



从结果中可得出，第一、三属性可推出第八属性，即发烧和腰痛症状有较大概率为肾炎，第四、五属性可推出第七属，即尿频、尿痛症状有江大概率为膀胱炎。