2016年数据仓库与数据挖掘期末试卷（回忆版）

一、填空题（50分）

1. 五数概括（注意：卡方检验，*Pearson* 系数，协方差）
2. 最小-最大规范化、小数规范化（注意：*z*-分数规范化）
3. 闵可夫斯基距离的表现形式（$p=1$，$p=2$，$p=+\infty$）
4. 关联规则的计算（给定支持度计数、关联规则，填空支持度和置信度）
5. 贝叶斯网络的约简表示，一个五维向量（每个向量有3种值），4条边，若不采用贝叶斯网络则需要多少行(243)？问最少(39)、最多要存储多少行(255)?
6. 有 n 个变量，每个变量两个值，$y=f(x_1, x_2, x_3, ..., x_n)$，$y\in\{0, 1, 2\}$，问有 f 有多少种不同的函数？假设有 m 个样本($m<2^n$)，则有多少不同的函数，若采用最小描述长度表示，则有多少个参数？（不懂.....）
7. 提升度 *lift* 计算公式，保留（大于）1的时候的频繁项集。（注意：混淆矩阵、准确率、精度、召回率、*F* 分数，上述会结合 *KNN* 先做分类，然后画出混淆矩阵）
8. f 近似函数 h，评估 h 优劣的准则有哪三个（准确性、复杂性、完备性）？由于计算复杂性不能再所有数据上进行计算，则可以在什么上面计算来评估 h（测试数据集）？

二、贝叶斯网络（10分）

1.考察贝叶斯网络中联合概率分布的形式 $P(U, W, X, Y, Z) = P(U)\cdot P(W \mid X)\cdot P(X \mid Y, Z)$.....类似这种形式。

2.如何用SQL语句得到某个条件概率 $P(U=u, W=w \mid X=x, Y=y)$的值，给出计算过程。

（注意：边缘概率、条件概率、联合概率分布表的存储行数计算）

三、频繁项集挖掘算法（11分）

1.对于给定的事务表格和支持度，利用 *Apriori* 算法进行频繁项集挖掘，写出主要步骤过程（不用伪代码表示过程）

2.给出一个闭频繁模式、一个极大频繁模式和一个关联规则

（注意：如果只给定一棵 *FP* 树，要求你给出频繁项集和关联规则，指定支持度和置信度）

四、分类问题（14分）

对于给定数据集 $D=(x_i, y_i)$, $i=1, 2, ..., |D|$，$y_i\in\{+1, -1\}$：

1.假设用线性函数 h 来做分类，考虑正则化，请给出该优化问题的形式化描述

2.假设采用朴素贝叶斯分类器，画出网络结构，并给出相应的训练/学习算法

（注意：如何利用决策树分类或贝叶斯网填补表格缺失值？）

五、聚类问题（15分）

给定8个数据点的二维坐标：(0, 1), (2, 4), (2, 1), (1, 3), (4, 3), (4, 4), (6, 3), (6, 5)。

1.假设 $k=2$，中心点为(0, 1)和(2, 4)，使用 k 中心点算法聚类，给出前三次迭代的过程。

2.聚类完成后，数据点(4, 3)的轮廓系数

3.若 $Minpts=2$，$\varepsilon=2$，这些数据点中有哪些是核心对象（3个）？