第一大题为概念简单题，选 5 个
第二到第十题为问答和计算题，按之前的卷子应该是六道题目比较合适
第十一题为开放式问答题
所有题目的考察点已经在题目前标出

一、简答题 （概念题选五个）

1. 两个仅包含二元属性的对象之间的相似性度量称为相似系数，简述三种(含)以上相似系数的计算方法与应用场景

又称为 Jaccard 相似系数（Jaccard similarity coefficient）用于比较有限样本集之间的相似性与差异性。Jaccard 系数值越大，样本相似度越高。比较文本相似度，用于文本查重与去重；计算对象间距离，用于数据聚类等。

$$J(A,B)=\frac{|\mathrm{A}\cap\mathrm{B}|}{|\mathrm{A}\cup\mathrm{B}|}=\frac{|\mathrm{A}\cap\mathrm{B}|}{|\mathrm{A}|+|\mathrm{B}|-|\mathrm{A}\cup\mathrm{B}|}$$

又称为余弦相似性，是通过计算两个向量的夹角余弦值来评估他们的相似度。余弦相似度将向量根据坐标值，绘制到向量空间中，如最常见的二维空间。最常见的应用就是计算文本相似度。将两个文本根据他们词，建立两个向量，计算这两个向量的余弦值，就可以知道两个文本在统计学方法中他们的相似度情况。实践证明，这是一个非常有效的方法。

$$\cos\theta=\frac{\sum_1^n(\mathrm{A}_i*\mathrm{B}_i)}{\sqrt{\sum_1^n A_i^2}\times\sqrt{\sum_1^n B_i^2}}$$

也称皮尔森积矩相关系数(Pearson product-moment correlation coefficient) ，是一种线性相关系数。皮尔森相关系数是用来反映两个变量线性相关程度的统计量。相关系数用 r 表示，其中 n 为样本量，分别为两个变量的观测值和均值。r 描述的是两个变量间线性相关强弱的程度。r 的绝对值越大表明相关性越强。利用样本相关系数推断总体中两个变量是否相关，可以用 t 统计量对总体相关系数为 0 的原假设进行检验。若 t 检验显著，则拒绝原假设，即两个变量是线性相关的；若 t 检验不显著，则不能拒绝原假设，即两个变量不是线性相关的。

$$r=\frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{X_i-\bar{X}}{s_X}\right)\left(\frac{Y_i-\bar{Y}}{s_Y}\right)$$

2. 简述关联规则中支持度和置信度的概念，并解释为什么采用这两种度量来表示关联规则的强度

假设 $I=\{I_1,I_2,\ldots,I_m\}$ 是项的集合。给定一个交易数据库 $D=\{t_1,t_2,\ldots,t_n\}$，其中每个事务（Transaction）t 是 I 的非空子集，即 $t\subseteq I$，每一个交易都与一个唯一的标识符 TID（Transaction ID）对应。关联规则是形如 $X\Rightarrow Y$ 的蕴涵式，其中 $X,Y\subseteq I$ 且 $X\cap Y=\varnothing$，X 和 Y 分别称为关联规则的先导(antecedent 或 left-hand-side, LHS)和后继(consequent 或 right-hand-side, RHS) 。关联规则 $X\Rightarrow Y$ 在 D 中的支持度（support）是

D 中事务包含的百分比 $X \cap Y$，即概率 $P(X \cap Y)$；置信度（confidence）是包含 X 的事务中同时包含 Y 的百分比，即条件概率 $P(X \mid Y)$。如果同时满足最小支持度阈值和最小置信度阈值，则认为关联规则是有趣的。

3.简述 Apriori 算法的计算复杂度受哪些因素影响，并加以解释　P213(DM)

4.在分类算法的评价指标中，recall 和 precision 分别是什么含义　P30 (ML)

5.请写出构建决策树时不纯度度量的三种指标　信息增益、增益率、基尼系数　p75-79（ML）

6.SVM 中核函数的作用是什么？

7.介绍 k-means 算法对初始点敏感的缺点（可以图示辅助分析）

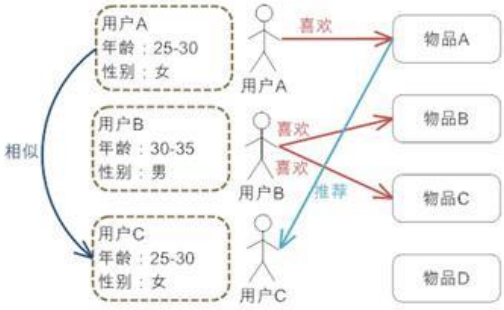
代次数,n 是所有样本的个数.然而,传统的 k-means 算法对于初始聚类中心的选择是随机的,或者是由用户主观指定的.k-means 算法对初始聚类中心的依赖性很强,初始聚类中心选取的不同往往导致聚类结果相当不稳定,同时也会使迭代的次数大幅度增加;

8.传统的推荐系统算法主要是哪两种？

基于人口统计学的推荐：

这是最为简单的一种推荐算法，它只是简单的根据系统用户的基本信息发现用户的相关程度，然后将相似用户喜爱的其他物品推荐给当前用户。

系统首先会根据用户的属性建模，比如用户的年龄，性别，兴趣等信息。根据这些特征计算用户间的相似度。比如系统通过计算发现用户 A 和 C 比较相似。就会把 A 喜欢的物品推荐给 C。



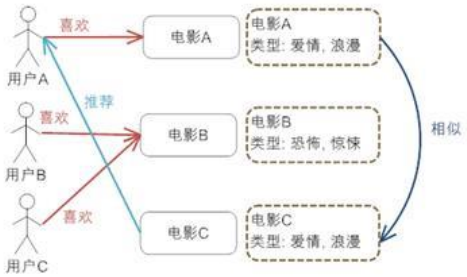
优缺点：不需要历史数据，没有冷启动问题；不依赖于物品的属性，因此其他领域的问题都可无缝接入；算法比较粗糙，效果很难令人满意，只适合简单的推荐

基于内容的推荐：

与上面的方法相类似，只不过这次的中心转到了物品本身。使用物品本身的相似度而不是用户的相似度。

系统首先对物品（图中举电影的例子）的属性进行建模，图中用类型作为属性。在实际应用中，只根据类型显然过于粗糙，　还需要考虑演员，导演等更多信息。通过相似度计算，发现电影 A 和 C 相似度较高，因为他们都属于爱情

类。系统还会发现用户 A 喜欢电影 A，由此得出结论，用户 A 很可能对电影 C 也感兴趣。于是将电影 C 推荐给 A。



优缺点：对用户兴趣可以很好的建模，并通过对物品属性维度的增加，获得更好的推荐精度；物品的属性有限，很难有效的得到更多数据；物品相似度的衡量标准只考虑到了物品本身，有一定的片面性；需要用户的物品的历史数据，有冷启动的问题

协同过滤：

协同过滤是推荐算法中最经典最常用的，分为基于用户的协同过滤和基于物品的协同过滤。那么他们和基于人口学统计的推荐和基于内容的推荐有什么区别和联系呢？

基于用户的协同过滤——基于人口统计学的推荐：基于用户的协同过滤推荐机制和基于人口统计学的推荐机制都是计算用户的相似度，并基于“邻居”用户群计算推荐，但它 们所不同的是如何计算用户的相似度，基 于人口统计学的机制只考虑用户本身的特征，而基于用户的协同过滤机制可是在用户的历史偏好的数据上计算用户的相似度，它的基本假设是，喜欢类似物品的用户 可能有相同或者相似的口味和偏好。

基于物品的协同过滤——基于内容的推荐：基于项目的协同过滤推荐和基于内容的推荐其实都是基于物品相似度预测推荐，只是相似度计算的方法不一样，前者是从用户历史的偏好推断，而后者是基于物品本身的属性特征信息。

协同过滤的优势：它不需要对物品或者用户进行严格的建模，而且不要求物品的描述是机器可理解的，所以这种方法也是领域无关的；这种方法计算出来的推荐是开放的，可以共用他人的经验，很好的支持用户发现潜在的兴趣偏好

协同过滤的缺点：方法的核心是基于历史数据，所以对新物品和新用户都有“冷启动”的问题；推荐的效果依赖于用户历史偏好数据的多少和准确性；在大部分的实现中，用户历史偏好是用稀疏矩阵进行存储的，而稀疏矩阵上的计算有些明显的问题，包括可能少部分人的错误偏好会对推荐的准确度有很大的影响等等；对于一些特殊品味的用户不能给予很好的推荐；由于以历史数据为基础，抓取和建模用户的偏好后，很难修改或者根据用户的使用演变，从而导致这个方法不够灵活

9.请写出两个 social network 方向的研究内容，如影响力分析

社会网络分析是研究一组行动者的关系的研究方法。一组行动者可以是人、社区、群体、组织、国家等，他们的关系模式反映出的现象或数据是网络分析的焦 点。从社会网络的角度出发，人在社会环境中的相互作用可以表达为基于关系的一种模式或规则，而基于这种关系的有规律模式反映了社会结构，这种结构的量化分析是社会网络分析的出发点。

现在来看，社会网络分析可以解决或可以尝试解决下列问题：

1-人际传播问题，发现舆论领袖，创新扩散过程；

2-小世界理论，六度空间分割理论；

3-Web 分析，数据挖掘中的关联分析，形成交叉销售，增量销售，也就是啤酒和尿布的故事；

4-社会资本，产业链与价值链；

5-文本的意义输出，通过追问调查研究文本的关联和意义；

6-竞争情报分析；

7-语言的关联，符号意义；

8-相关矩阵或差异矩阵的统计分析，类似得到因子分析和 MDS 分析；

9-恐怖分子网络；

10-知识管理与知识的传递，弱关系的力量；

11-引文和共引分析；

二、（关联规则）Apriori 算法使用产生—计数的策略找出频繁项集。通过合并一对大小为 k 的频繁项集得到一个大小为 k+1 的候选项集（称作候选产生步骤）。在候选项集剪枝步骤中，如果一个候选项集的任何一个子集是不频繁的，则该候选项集将被丢弃。假定将 Apriori 算法用于表中所示数据集，最小支持度为 30%，即任何一个项集在少于 3 个事务中出现就被认为是非频繁的。

事务 ID	购买项
1	{a,b,d,e}
2	{b,c,d}
3	{a,b,d,e}
4	{a,c,d,e}
5	{b,c,d,e}
6	{b,d,e}
7	{c,d}
8	{a,b,c}
9	{a,d,e}
10	{b,d}

(a) 画出表示表中所示数据的项集格，用下面的字母标记格中的每个结点。

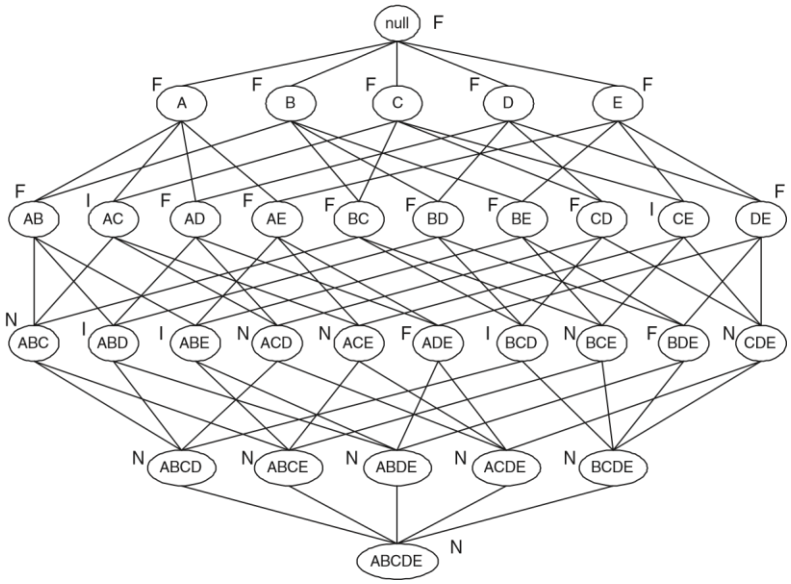
- **N**：如果该项集被 Apriori 算法认为不是候选项集。一个项集不是候选项集有两种可能的原因：它没有在候选项集产生步骤产生，或它在候选项集产生步骤产生，但是由于它的一个子集是非频繁的而在候选项集简直步骤被丢掉
- **F**：如果该候选项集被 Apriori 算法认为是非频繁的
- **I**：如果经过支持度计数后，该候选项集被发现是非频繁的

(b) 频繁项集的百分比是多少？（考虑格中所有的项集）

(c) 对于该数据集，Apriori 算法的剪枝率是多少？（剪枝率定义为由于如下原因不认为是候选的项集所占的百分比：在候选项集产生时未被产生，或在候选剪枝步骤被丢掉）

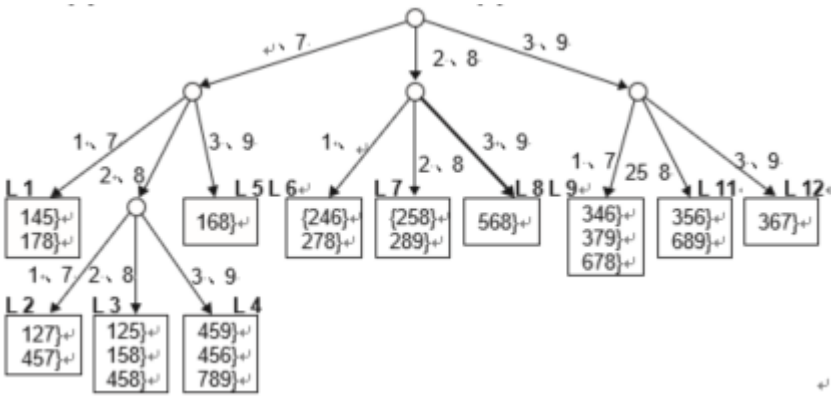
(d) 假警告率是多少？（假警告率是指经过支持度计算后被发现是非频繁的候选项集所占的百分比） DM Ch6 Ex.8

(a)Lattice 的结构如下图所示。



(b) 比例的频繁 itemsets=16/32=50.0%(包括空集)。

(c) 修剪比是 *N* 的总数。itemsets 自数 *N* =11,所以修剪的比率是 11/32=34.4%



(d) 虚假报警率的比率是 我 总数 itemsets。 自数 I =5,因此虚假报警率是 5/32=15.6%

三、（朴素贝叶斯）试由下表的训练数据学习一个朴素贝叶斯分类器并确定 x = （2，S）T 的类判别结果 y。表中 $X^{(1)}$，$X^{(2)}$ 为特征，Y 为类标记。

	1	2	3	4	5	6	7	8
$X^{(1)}$	1	1	1	2	2	2	3	3
$X^{(2)}$	S	M	M	S	M	M	L	M
Y	-1	-1	1	1	-1	1	1	1

P50 （统计学习方法 Ep4.1）

四、（SVM）已知正例点 $x_1 = (2.5, 2.5)^T$，$x_2 = (5, 2)^T$，负例点 $x_3 = (1.5, 1.5)^T$，试用 SVM 对其进行分类，求最大间隔分离超平面，并指出所有的支持向量。

五、（决策树）下表是一个由 15 个贷款申请训练数据，数据包括贷款申请人的四个特征属性：分别是年龄，是否有工作，是否有自己的房子以及信贷情况，表的最后一列为类别，是否同意贷款。

ID	年龄	有工作	有自己的房子	信贷情况	类别
1	青年	否	否	一般	否
2	青年	否	否	好	否
3	青年	是	否	好	是
4	青年	是	是	一般	是
5	青年	否	否	一般	否
6	中年	否	否	一般	否
7	中年	否	否	好	否
8	中年	是	是	好	是
9	中年	否	是	非常好	是
10	中年	否	是	非常好	是
11	老年	否	是	非常好	是
12	老年	否	是	好	是
13	老年	是	否	好	是
14	老年	是	否	非常好	是
15	老年	否	否	一般	否

1）请根据上表的训练数据，以错误率作为划分标准来构建预测是否进行放贷的决策树。

2）按照所构建的决策树，对属性值为（中年，无工作，无自己的房子，信贷情况好）的申请者是否进行放贷

3） 在构建决策树的时候，可能会出现过拟合的问题，有哪些方法可以避免或者解决？p80-81(LM)

有几种途径用来避免决策树学习中的过度拟合。它们可被分为两类：

(1). 及早停止增长树法，在 ID3 算法完美分类训练数据之前停止增长树；

(2). 后修剪法（post-prune），即允许树过度拟合数据，然后对这个树后修剪。

尽管第一种方法可能看起来更直接，但是对过度拟合的树进行后修剪的第二种方法被证明在实践中更成功。这是因为在第一种方法中精确地估计何时停止增长树很困难。无论是通过及早停止还是后修剪来得到正确大小的树，一个关键的问题是使用什么样的准则来确定最终正确树的大小。解决这个问题的方法包括：

(1). 使用与训练样例截然不同的一套分离的样例，来评估通过后修剪方法从树上修剪结点的效用。

(2). 使用所有可用数据进行训练，但进行统计测试来估计扩展（或修剪）一个特定的结点是否有可能改善在训练集合外的实例上的性能。例如，Quinlan （1986）使用一种卡方（chi-square）测试来估计进一步扩展结点是否能改善在整个实例分布上的性能，还是仅仅改善了在当前的训练数据上的性能。

(3). 使用一个明确的标准来衡量训练样例和决策树编码的复杂度，当这个编码的长度最小时停止增长树。这个方法基于一种启发式规则，被称为最小描述长度（Minimum Description Length）的准则。Quinlan & Rivest（1989）和 Mehta et al.（1995）也讨论了这种方法。

上面的第一种方法是最普通的，它常被称为训练和验证集（training and validation set）法。下面我们讨论这种方法的两个主要变种。这种方法中，可用的数据被分成两个样例集合：一个训练集合用来形成学习到的假设，一个分离的验证集合用来评估这个假设在后续数据上的精度，确切地说是用来评估修剪这个假设的影响。这个方法的动机是：即使学习器可能会被训练集合中的随机错误和巧合规律性所误导，但验证集合不大可能表现出同样的随机波动。所以，验证集合可以用来对过度拟合训练集中的虚假特征提供一个防护检验。当然，很重要的一点，验证集合应该足够大，以便它本身可提供具有统计意义的实例样本。一种常见的做法是取出可用样例的三分之一用作验证集合，使用另外三分之二用作训练集合。

4） 对于含有连续型属性的样本数据，决策树有哪些处理方法？

离散化（等区间离散、等数据离散），高斯分布模拟

六、聚类分析（聚类）

(1) 在聚类分析中，传统的 K-means 算法都有哪些局限性？有哪些相应的改进方法？

(1)对于离群点和孤立点敏感；

(2)k 值选择;

(3)初始聚类中心的选择；

(4)只能发现球状簇。

（1）离群点检测的 LOF 算法，通过去除离群点后再聚类，可以减少离群点和孤立点对于聚类效果的影响。

（2）必须首先给出 k（要生成的簇的数目），k 值很难选择。事先并不知道给定的数据应该被分成什么类别才是最优的；初始聚类中心的选择是 K-means 的一个问题。算法思路是这样的：可以通过在一开始给定一个适合的数值给 k，通过一次 K-means 算法得到一次聚类中心。对于得到的聚类中心，根据得到的 k 个聚类的距离情况，合并距离最近的类，因此聚类中心数减小，当将其用于下次聚类时，相应的聚类数目也减小了，最终得到合适数目的聚类数。可以通过一个评判值 E 来确定聚类数得到一个合适的位置停下来，而不继续合并聚类中心。重复上述循环，直至评判函数收敛为止，最终得到较优聚类数的聚类结果。

（3）选择批次距离尽可能远的 K 个点。具体选择步骤如下：首先随机选择一个点作为第一个初始类簇中心点，然后选择距离该点最远的那个点作为第二个初始类簇中心点，然后再选择距离前两个点的最近距离最大的点作为第三个初始类簇的中心点，以此类推，直至选出 K 个初始类簇中心点。

（4）只能获取球状簇的根本原因在于，距离度量的方式。在李荟娆的硕士论文 K_means 聚类方法的改进及其应用中提到了基于 2 种测度的改进，改进后，可以去发现非负、类椭圆形的数据。但是对于这一改进，个人认为，并没有很好的解

决 K-means 在这一缺点的问题，如果数据集中有不规则的数据，往往通过基于密度的聚类算法更加适合，比如 DESCAN 算法。

(2) 请简要描述聚类与关联分析的主要相似点和不同点。

(3) 请举出一个采用聚类作为主要的数据挖掘方法的实际应用例子。

七、（决策树）证明：在决策树分类方法中，将结点划分为更小的后继结点后，结点熵不会增加　DM　Ch.4　Ex4

让 $Y=\{y_1,y_2,\cdots,yc\}$表示 c 类和 $X=\{x_1,x_2,\cdots,xk\}$表示 k 的属性值的属性的 X。 在节点上的拆分 X,熵是:

$$E(Y)=-\sum_{j=1}^{c}P(y_j)\log_2 P(y_j)=\sum_{j=1}^{c}\sum_{i=1}^{k}P(x_i,y_j)\log_2 P(y_j)\text{”} \qquad (4.1)$$

在那里我们用的是, $P(y_j)=\sum_{i=1}^{k}P(x_i,y_j$从法的总的概率。

拆分后的 X、熵的每个子节点 $X=xi$:

$$E(Y|x_i)=-\sum_{j=1}^{c}P(y_j|x_i)\log_2 P(y_j|x_i \qquad (4.2)$$

其中 $P(yj|xi)$是小部分的示例与 $X=xi$,属于类 yj。 熵分割之后在 X 的加权平均信息量的子节点:

$$\begin{aligned}E(Y|X) &= \sum_{i=1}^{k}P(x_i)E(Y|x_i)\\ &= -\sum_{i=1}^{k}\sum_{j=1}^{c}P(x_i)P(y_j|x_i)\log_2 P(y_j|x_i)\\ &= -\sum_{i=1}^{k}\sum_{j=1}^{c}P(x_i,y_j)\log_2 P(y_j|x_i),\end{aligned} \qquad (4.3)$$

我们在那里用已知的事实是从概率论, $P(xi,yj)=P(y_j|x_i)\times P(x_i$ of Y given $X.$)。 请注意, $E(Y|X)$也称为有条件的熵。

要回答这个问题,我们需要证明 $E(Y|X)\le E(Y)$。 让我们的计算之间的区别后 entropies 分割和分割之前,即 $E(Y|X)-E(Y)$、使用方程 4.1 和 4.3:

$$
\begin{aligned}
& E(Y|X) - E(Y) \\
= & -\sum_{i=1}^{k}\sum_{j=1}^{c} P(x_i, y_j) \log_2 P(y_j|x_i) + \sum_{i=1}^{k}\sum_{j=1}^{c} P(x_i, y_j) \log_2 P(y_j) \\
= & \sum_{i=1}^{k}\sum_{j=1}^{c} P(x_i, y_j) \log_2 \frac{P(y_j)}{P(y_j|x_i)} \\
= & \sum_{i=1}^{k}\sum_{j=1}^{c} P(x_i, y_j) \log_2 \frac{P(x_i)P(y_j)}{P(x_i, y_j)}
\end{aligned}
\tag{4.4}
$$

为证明这一等式 4.4 非积极的,我们可以使用以下属性一种对数函数:

$$
\sum_{k=1}^{d} a_k \log(z_k) \leq \log\left(\sum_{k=1}^{d} a_k z_k\right)\text{”} \tag{4.5}
$$

但条件是 $\sum_{k=1}^{d} a_k$=1。 此属性是一种特殊的情况下具有更普遍的定理涉及外接功能(其中包括对数函数)称为詹森的不平等。

通过应用 Jensen 的不平等、等式 4.4 可以在限定范围内,如下所示:

$$
\begin{aligned}
E(Y|X) - E(Y) & \leq \log_2\left[\sum_{i=1}^{k}\sum_{j=1}^{c} P(x_i, y_j)\frac{P(x_i)P(y_j)}{P(x_i, y_j)}\right] \\
& = \log_2\left[\sum_{i=1}^{k} P(x_i)\sum_{j=1}^{c} P(y_j)\right] \\
& = \log_2(1) \\
& = 0
\end{aligned}
$$

因为 $E(Y|X)- E(Y)\leq 0$,因此熵从来没有增加分裂后的属性。

八、（效果评价 ROC）请评价两个分类器 M1 和 M2 的性能。所选择的测试集包含 26 个二值属性，记作 A 到 Z。

表中是模型应用到测试集时得到的后验概率（图中只显示正类的后验概率）。因为这是二类问题，所以 P(-)=1-P(+),P(-|A,…,Z)=1-P(+|A,…,Z)。假设需要从正类中检测实例

(a) 画出 M1 和 M2 的 ROC 曲线（画在一幅图中）。哪个模型更多？给出理由

(b) 对模型 M1，假设截止阈值 t=0.5.换句话说，任何后验概率大于 t 的测试实例都被看作正例。计算模型在此阈值下的 precision，recall 和 F-score

(c) 对模型 M2 使用相同的截止阈值重复（b）的分析。比较两个模型的 F-score，哪个模型更好？所得结果与从 ROC 曲线中得到的结论一致吗》

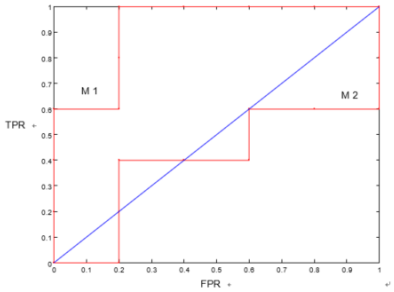
(d) 使用阈值 t=0.1 对模型 M2 重复（b）的分析。t=0.5 和 t=0.1 哪一个阈值更好?该结果和你从 ROC 曲线中得到的一致吗？

实例	真实类	P(+|A,…,Z,M1)	P(-|A,…,Z,M2)
1	+	0.73	0.61
2	+	0.69	0.03
3	-	0.44	0.68
4	-	0.55	0.31

5	+	0.67	0.45
6	+	0.47	0.09
7	-	0.08	0.38
8	-	0.15	0.05
9	+	0.45	0.01
10	-	0.35	0.04

DM Ch.5 Ex.17

(a) ROC 曲线为 *M*1 和 *M*2 中所示的图 5.5. *M*1 是更好,因为它的面积根据 ROC 曲线是较大的面积根据 ROC 曲线为 *M*2



(b) 当 t =0.5 的混淆矩阵为 M1 是如下图所示。

		+	-
实际	+	3	2
	-	1	4

Precision=3/4=75%。记得=3/5=60%。F=(2 × .75 × 6)/(.75+ 6)=0.667。

(c) 当 *t* =0.5 的混淆矩阵为 *M*2 是如下图所示。

		+	-
实际	+	1	4
	-	1	4

Precision=1/2=50%。记得=1/5=20%。F=(2 × 5 × 2)/(5+ 2)=0.2857。

(d) 当 t =0.1 的混淆矩阵为 M1 是如下图所示

		+	-
实际	+	5	0
	-	4	1

Precision=5/9=55.6%。记得=5/5=100%。F=(2 × .556 × 1)/(.556+1)=0.715。

根据 F 的测量, t =0.1 的比 t =0.5。

当 t =0.1、 FPR =0.8 及 TPR =1。 牾碱环矫,当 t =0.5、 FPR =0.2 和 TRP =0.6。 因为(0.2、0.6)较近的点(0,1),我们赞成 t =0.5。 这一结果是不符合的结果使用的措施。 我们也可以显示此计算下面的区域。

ROC 曲线:

为 t =0.5,area=0.6 ×(1-0.2)=0.6 × 0.8=0.48。

为 t =0.1、面积=1 ×(1-0)8=1 × 0.2=0.2。

由于该区的 t =0.5 是较大的面积, t =0.1、我们喜欢 t =0.5。

九（频繁项）考虑下面的候选 3-项集的集合：{1，2，3}，{1，2，5}，{1，2，6}，{1，3，4}，{2，3，4}，{2，4，5}，{3，4，6}，{4，5，6}

(a) 构造以上候选 3-项集的 Hash 树，假定 Hash 树使用这样一个 Hash 函数：所有奇数项都被散列到节点的左子女，所有的偶数项都被散列到右子女。一个候选 k-项集按照如下方法被插入到 Hash 树中：散列候选项集中的每个相继项，然后再按照散列值到相应的分支。一旦到达叶节点，候选项集将按照下面的条件插入：

- 条件 1：如果该叶节点的深度等于 k（假设根节点的深度为 0），则不管该节点已经存储了多少项集，将该候选插入该节点
- 条件 2：如果该叶节点的深度小于 k，则只要该节点存储的项集数不超过 maxsize，就把它插入到该叶节点。这里，假定 maxsize 为 2
- 条件 3：如果该叶节点的深度小于 k 且该节点已存储的项集数量超过 maxsize，则这个叶节点转变为内部节点，并创建新的叶节点作为老的叶节点的子女。先前老叶节点中存放的候选项集按照散列值分布到其子女中。新的候选项集也按照散列值存储到相应的叶节点

（b）候选 Hash 树中共多少个叶节点，多少个内部节点？

（c）考虑一个包含项集{1，2，3，4，5，6}的事务，使用（a）所创建的 Hash 树，则该事务要检查哪些叶节点？该事务包含哪些候选 3-项集 DM Ch.6 Ex7

(a) {{11、、22、、34、、45}}、{{11″、22″、34″、56}}、{{11″、22″、35″、66}}。

{1、3、4、5}、{1、3、4、6}、{2、3、4、5}。 {2、3、4、6}、{2、3、5、6}。

(b) {1、2、3、4}、{1、2、3、5}、{1、2、4、5}、{2、3、4、5}、{2、3、4、6}。

(c) {1、2、3、4}

十、(ensemble 组合方法)请简述构建组合（集成）分类器的几种方法，并说明集成分类器能够改善分类器性能的原因。

P181(LM)

十一、（开放课题）现有一个城市的数据集，包括交通卡、交通事故、出租车轨迹、公交车运行、地铁运行、空气质量、气象检测、新浪微博等（具体特征如下表）。

请利用你所学过的机器学习和数据挖掘的方法解决预测该城市空气质量的问题：

（1） 哪些数据或者特征可能用到，并简要说明原因

（2） 可以使用所学过的哪些机器学习方法解决该问题？

（3） 请简要给出一个解决方案（最大限度地利用现有数据）。

序号	数据集名称	具体数据项
1	城市道路交通指数	状态、区域、当前指数、参考指数、指数差值
2	地铁运行数据	线路、车站、换乘站数据、首末班车各站时刻表数据、站间运行时间数据、限流车站、封站数据、路网票价矩阵、列车实时到发站台时刻、线路拥挤及阻塞数据、出入口、厕所、残疾电梯数据
3	一卡通乘客刷卡数据	卡号、交易日期、交易时间、线路/地铁站点名称、行业名称（公交、地铁、出租、轮渡、P+R 停车场）、交易金额、交易性质（非优惠、优惠、无）
4	浦东公交车实时数据	设备号码,线路编码,站点编码,协议编号,进出站状态,方向,车载上报时间、编码对应表
5	强生出租汽车行车数据	车辆 ID、GPS 时间、经纬度、速度、卫星颗数、营运状态高架状态、制动状态
6	空气质量状况	序号，日期，PM2.5，PM10，O3，SO2，NO2，CO，AQI，质量评价，首要污染物
7	气象数据	日期、时间、监测点、天气类型、温度、风速、风向、降水量
8	道路事故数据	事故 ID、事故类型、事故地点、事故时间
9	高架匝道关闭数据	匝道 ID、位置信息、关闭时间、开放时间
10	新浪微博交通数据	涵盖路况、交通工具、天气等与交通相关的关键词的微博信息