课堂练习：计算信息增益

- **问题描述**：假设一组连续值及其所属类别如下表所示，利用信息增益求第一次划分的分隔数据点。

数据	0.243	0.245	0.437	0.481	0.608	0.666
类别	C0	C0	C1	C1	C1	C0

$$GAIN_{\text{split}} = Entropy(p) - \left(\sum_{i=1}^{k} \frac{n_i}{n} Entropy(i) \right)$$

$\log_2(1/3) = -1.585$ ，$\log_2(2/3) = -0.585$
$\log_2(3/5) = -0.737$ ，$\log_2(2/5) = -1.322$
$\log_2(3/4) = -0.415$ ，$\log_2(1/4) = -2.0$

小测：MLE与MAP

- 最大后验概率估计(MAP) — 理解先验p(θ)
  - 扔硬币的例子：10次实验，其中**正面朝上(参数：θ)**的次数为7次，反面朝上的次数为3次，结果记为(1,0,1,1,0,1,0,1,1,1)

- 最大后验概率估计(MAP)—理解先验p(θ)
  - 扔硬币的例子：我们期望待估计的参数θ的先验分布在0.5处取得最大值，可以选用**Beta分布**（θ服从Beta分布）即：

$$p(\theta|\alpha,\beta) \triangleq Beta(\alpha,\beta) = \frac{1}{B(\alpha,\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}$$

  - 取 $\boldsymbol{\alpha} = \boldsymbol{\beta} = 5$， 使得先验分布Beta分布在0.5处取得最大值
  - **使用MAP方法求解参数**

小测：计算频繁项集

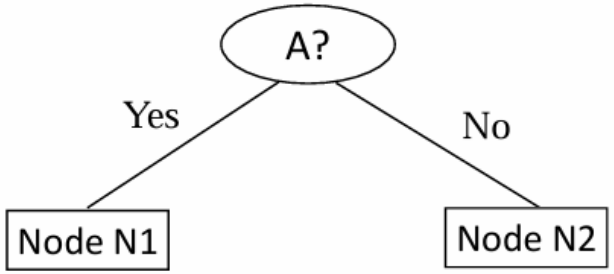
- 设min_sup = 50%求出右图事务列表中所有的频繁项集
- （包括1-频繁，2-频繁，3-频繁等，给出求解过程）

TID	Item
100	1 2 3 4
200	1 2 5
300	1 2 3 5
400	2 4 5
500	1 2 3

小测：决策树特征选择

- 决策树——分类错误率 vs Gini值

分别使用Gini值、分类错误率计算以下分裂能够取得的增益是多少?



A		Parent
	C1	7
	C2	3
	= ?	

	N1	N2
C1	3	4
C2	0	3
= ?		

$$GINI(t) = 1 - \sum_j [p(j \mid t)]^2$$

$$GINI_{split} = \sum_{i=1}^{k} \frac{n_i}{n} GINI(i)$$

$$Error(t) = 1 - \max_i P(i \mid t)$$

小测：贝叶斯分类

给定以下7个用户的数据，使用**朴素贝叶斯方法**预测用户8={有工作=否，婚姻状况=已婚}的拖欠贷款属性最有可能是Yes还是No，并给出求解过程。

用户ID	有工作	婚姻状况	拖欠贷款
1	否	已婚	No
2	是	单身	Yes
3	否	单身	No
4	是	已婚	Yes
5	否	单身	No
6	是	单身	Yes
7	否	已婚	No
8	**否**	**已婚**	**？**