# 问题定义：

根据已有的数据集，选择合适的分类算法，建立数学模型，对乳腺癌患者是否会病情复发进行预测。

# 数据获取：

该数据集来自南斯拉夫卢布尔雅那大学医疗中心肿瘤研究所，由M. Zwitter 与M. Soklic 二者提供。该数据集可在UCI数据库中找到。网址：<https://archive.ics.uci.edu/ml/datasets/Breast+Cancer>

该数据集有2大类，9个属性，共286个实例。class：类别，分别是：

乳腺癌复发（recurrence-events）和未复发（no-recurrence-events）

age：年龄，有 20-29, 30-39, 40-49, 50-59, 60-69, 70-79，六个区

menopause：绝经期，分为prememo（未绝经），ge40（40岁之后绝经），lt40（40岁之前绝经）

tumor-size：肿瘤大小

inv-nodes：淋巴结个数

node-caps：结节冒有无

deg-malig：肿瘤恶性程度，分为1、2、3三种，3恶性程度最高

breast： 分为left和right

breast-quad：所在象限irradiat：是否有放射性治疗经历

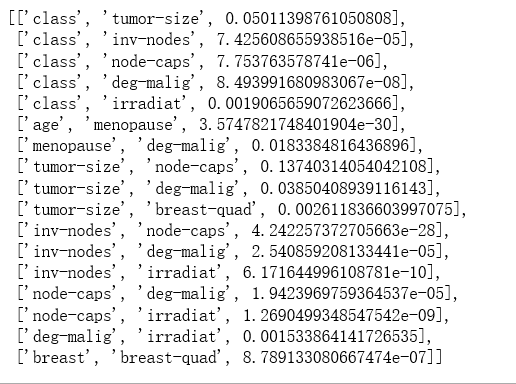
# 数据清洗：

## (1)数据清理

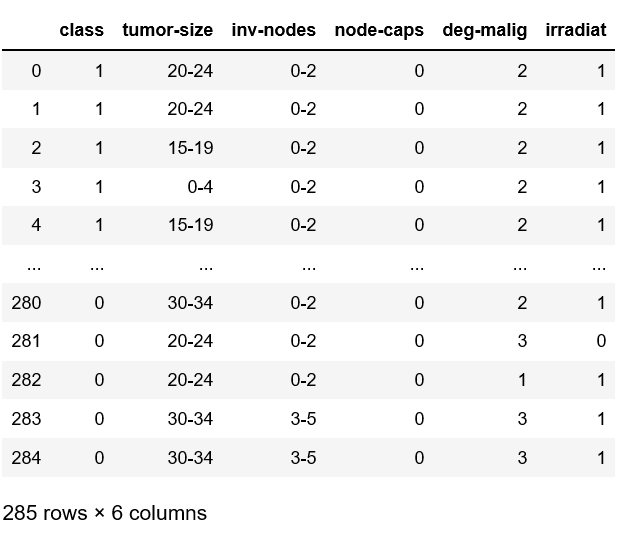
        对于缺失值的处理：使用最有可能的值进行填充，认为众数是最有可能的值

## (2)数据集成

冗余和相关分析：对于标签属性，我们使用卡方检验对属性之间的独立性进行检验，在90%的置信区间内，属性间两两相关的结果如下（列表的第三个值为两属性相互独立的概率）：

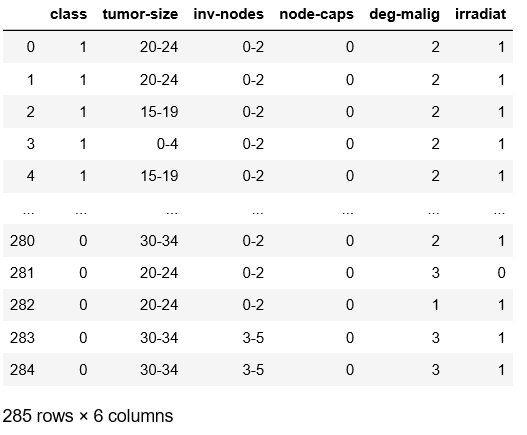


观察到有些属性之间关联性很强，比如age和menopause，认为这些属性存在冗余，只保留其中一个。与class相关性很弱的属性也不应该在研究范围内，应去除，去除后数据集：

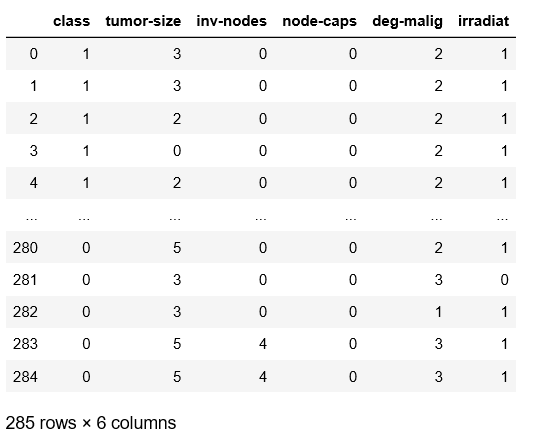


## (3)数据变换和数据离散化

        数据离散化：对于tumor-size、inv-nodes属性分别进行等距分箱处理，箱子的间距分别为5、3。



数据变换：对非数字型和数字型标签值进行标准化，处理后的数据集：



# 特征选择

### 在数据庞大的情况下，必须进行研究属性的选择，数据和特征决定了学习的上限，而模型和算法只是逼近这个上限而已，因此进行合理的特征选择是至关重要的。首先进行数据探索，寻找与患者复发率有关的属性。

## 数据探索：

### 

### 图1

### 由图1可以看出，乳腺癌复发的概率和年龄关系不大，但在30-39这个年龄段，可能因为绝经的原因,而导致复发率相比较高，但影响不大。

### 

### 图2

### 图2表明，经过放疗会较大程度地增加复发的概率。

### 

### 图3

### 图3表明，肿块大小，淋巴个数多少，结节冒有无，都影响着复发率。

### 综合之前卡方检验结果，此处准备选择tumor-size、inv-nodes、node-caps、deg-malig、irradiat这5个属性。

### 

# ****数据集划分：****

### 数据集分割（训练集与测试集比例7：3）

### 

### 训练集

### 

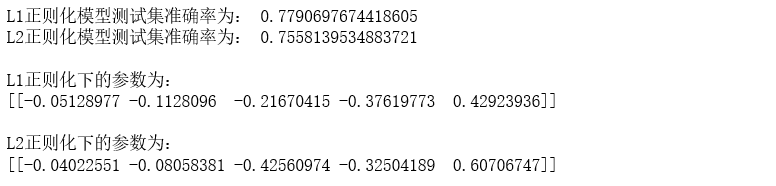
### 测试集

# ****模型建立：****

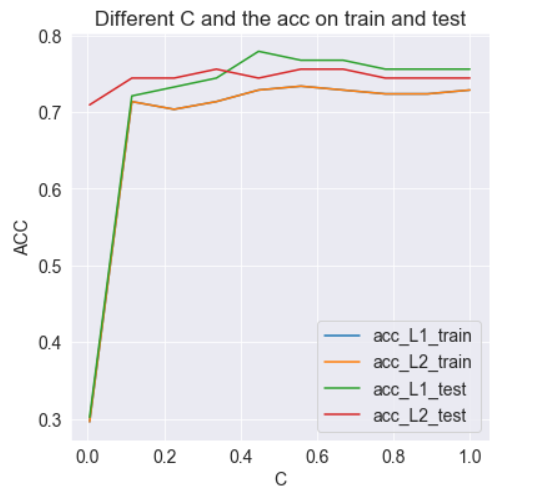
## 逻辑回归二分类：

### 参数测试：

正则化是用来防止模型过拟合的过程，常用的有L1正则化和L2正则化两种选项，分别通过在损失函数后加上参数ω向量的L1范式和L2范式的倍数来实现。这个增加的范式，被称为“正则项”，也被称为"惩罚项"。损失函数改变，基于损失函数的最优化来求解的参数取值必然改变，以此来调节模型拟合的程度。分别使用L1和L2两种不同的正则化方式进行训练，经过训练发现，经过L1正则化进行训练的模型准确率高于L2正则化训练模型的准确率，训练准确率和训练参数如下：



C是正则化强度的倒数，必须是一个大于0的浮点数，不填写默认1.0，即默认正则项与损失函数的比值是1：1。C越小，损失函数会越小，模型对损失函数的惩罚越重，正则化的效力越强，参数会逐渐被压缩得越来越小。对于L1和L2两种不同的正则化方式分别对训练集和测试集用不同的正则化项系数的倒数进行测试，经过验证，发现C值小于0.2的时候，准确率很低，推测出现了过拟合现象，测试结果如下：



### 结果分析：

L1正则化和L2正则化虽然都可以控制过拟合，但效果并不相同。当正则化强度逐渐增大（即C逐渐变小），参数的取值会逐渐变小，但L1正则化会将参数压缩为0，L2正则化只会让参数尽量小，不会取到0。在L1正则化在逐渐加强的过程中，携带信息量小的、对模型贡献不大的特征的参数，会比携带大量信息的、对模型有巨大贡献的特征的参数更快地变成0，所以L1正则化本质是一个特征选择的过程，掌管了参数的“稀疏性”。L1正则化越强，参数向量中就越多的参数为0，参数就越稀疏，选出来的特征就越少，以此来防止过拟合。

## 决策树模型：

### 参数测试：

为了对决策树进行剪枝，需要设置三个参数max\_depth、min\_samples\_split、min\_samples\_leaf。

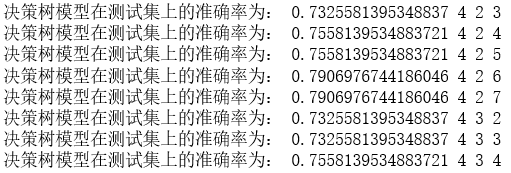
max\_depth：限制树的最大深度，超过设定深度的树枝全部剪掉，这是用得最广泛的剪枝参数，在高维度低样本量时非常有效。决策树多生长一层，对样本量的需求会增加一倍，所以限制树深度能够有效地限制过拟合。

min\_samples\_split：min\_samples\_split限定，一个节点必须要包含至少min\_samples\_split个训练样本，这个节点才允许被分枝，否则分枝就不会发生。

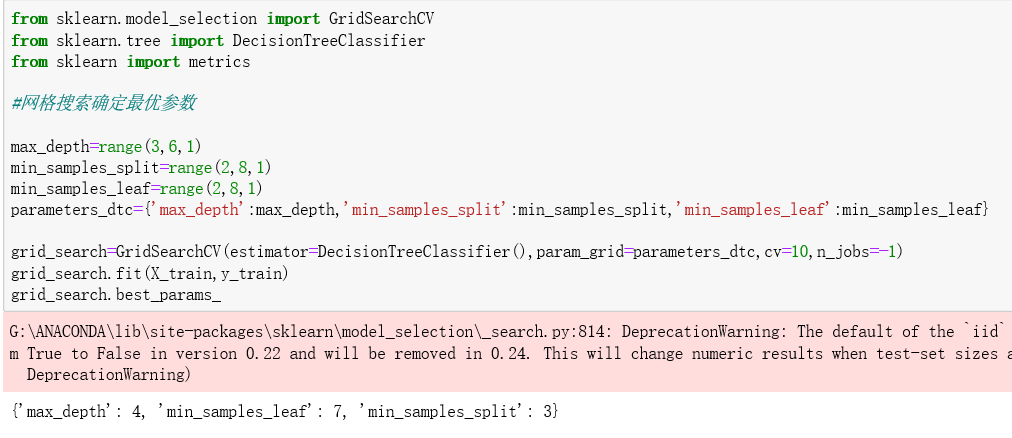
min\_samples\_leaf：min\_samples\_leaf 限定，一个节点在分枝后的每个子节点都必须包含至少min\_samples\_leaf个训练样本，否则分枝就不会发生，或者，分枝会朝着满足每个子节点都包含min\_samples\_leaf个样本的方向去发生

，可以让模型变得更加平滑。这个参数的数量设置得太小会引起过拟合，设置得太大就会阻止模型学习数据。

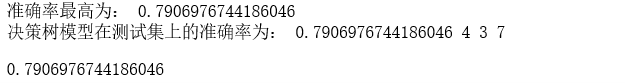
将max\_depth、min\_samples\_split、min\_samples\_leaf的取值区间定为2-6、2-8、2-8，进行穷举，计算出每种参数设置情况下的准确率，得到准确率最高为0.79，最低为0.73。



经过穷举的结果很不稳定，于是又使用网格搜索的办法确定最优参数，得到三个参数分别为4、3、7，计算结果如下：



代入计算得到准确率为0.79，与之前的计算的最高准确率相同，所以认为准确率0.79是决策树模型对该问题的最优解



### 结果分析：

### 在max\_depth为2的条件下，无论另外两个参数怎么变化，准确都为最低状态，说明可能出现了欠拟合现象，可以看到随着max\_depth升高，准确率有所提升

### 

### 在max\_depth为4、min\_samples\_split为2的情况下， 准确率随min\_samples\_leaf的值的增大而增大，已经知道当min\_samples\_leaf小的时候可能会出现过拟合现象，于是可以推测min\_samples\_leaf小的时候引起了过拟合，导致准确率下降。

### 

### 决策树算法构建出的树形结构如下：

### 

## 随机森林：

鉴于决策树容易过拟合的缺点，随机森林采用多个决策树的投票机制来改善决策树。训练可以高度并行化，对于大数据时代的大样本训练速度有优势；由于随机选择特征时只考虑特征子集，因此样本特征维度很高时仍然高效；训练后，可以给出各个特征的重要性；由于采用了随机采样，方差小，泛化能力强

### 参数测试：

### 通过网格搜索进行交叉验证，确定最优参数

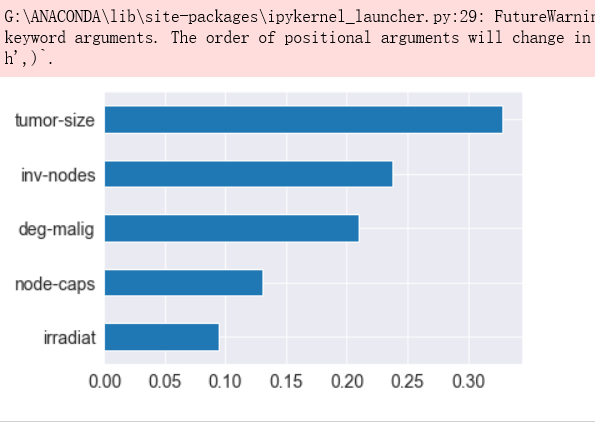
### 

### 结果分析：

### 经过训练之后，随机森林的准确率达到0.78，比决策树低一点。



计算出属性对于患者复发的重要性进行排序，得到如下结果：



# 总结

通过以上的分析得知，乳腺癌患者病情复发的影响因素主要有tumor-size、inv-nodes、node-caps、deg-malig、irradiat，影响程度从大到小分别为tumor-size、inv-nodes、deg-malig、node-caps、irradiat。由于本次分析的数据量较小，可能对分析结果的准确率有所影响，应尽可能的手机更多的数据进行分析，使分析结果更加准确。