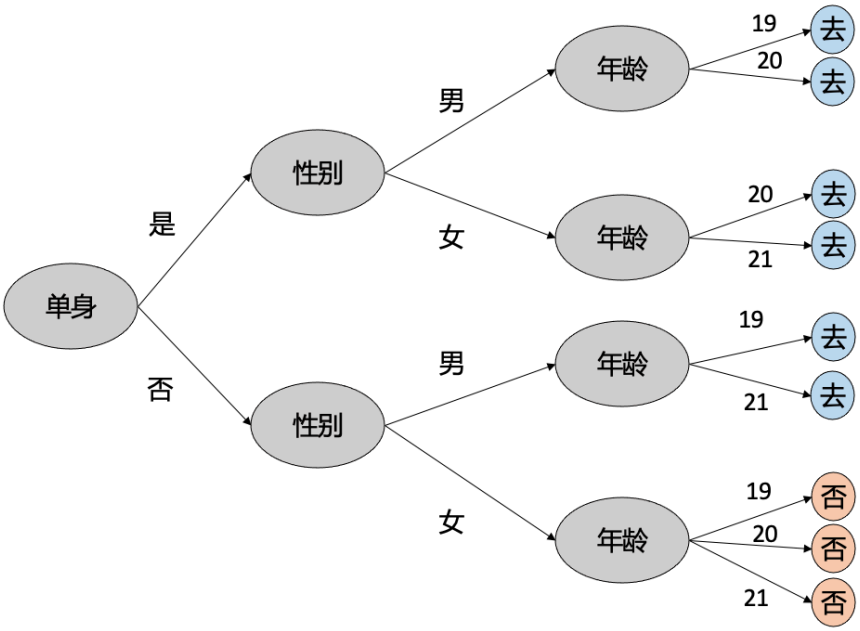
# 1 决策树

在机器学习中，最常见的一类问题是分类问题，就是通过观察数据的特点（特征），构造一个函数，将数据分为特定的类别。比如说其中一个最直观的应用“图像分类”，就是通过观察图像中每个像素的值，来分类图像（图像中是猫还是狗）。在我们这次上机作业中，我们需要通过观察学生的“年龄”、“性别”和“单身情况”，来判断学生是否“周日会去图书馆”。

在上次上机作业中，我们发现，当选择那个让熵减少最多的特征用于分类时，分类得最准确。即选取“单身情况”作为分类依据，可以达到 80% 的准确率。这时，我们的分类函数为，“如果学生单身，则判断周末会去图书馆，如果不单身则判断周末不去图书馆”。

在本次上机作业中，我们将建立一个决策树算法，综合考虑所有特征，建立复杂的分类函数。决策树是机器学习中应用最广泛的非线性分类器，在很多实际问题中基于决策树的算法都是当前效果最好的算法，例如 XGBOOST 等，很多时候会优于深度学习算法。

决策树是一个树状的决策器（分类器），比如我们上机作业的例子中，决策器的分类方法是：先看学生是否单身，如果单身，再看学生性别，如果是男，再看年龄，如果 19 岁则判断周末会去图书馆。如下图



本次上机作业需要构造一个可以自动适应任意类似的数据的决策树。构造决策树需要以下的步骤：

1. 像上次作业一样，找出使熵下降最多的特征，下称“最优特征”。
2. 对最优特征的每一个取值可能，在剩下的特征中找到“最优特征”。
3. 以此递归。
4. 如果只剩一个特征，那么就不再需要找最优特征了，这时，对该特征的每个取值对应的决策器的输出，是该取值下最多的那一类。

我们从一个具体例子开始，1. 像上次上机一样，选择“single”作为最优特征。

2. ”single” 特征有两个取值“True”和“False”,我们先看“True”。在”single=True” 的这一支，只剩下

{”age”: 19, ”male”: True, ”single”: True, ”visit_library_in_Sunday”: True},
{”age”: 20, ”male”: True, ”single”: True, ”visit_library_in_Sunday”: True},
{”age”: 20, ”male”: False, ”single”: True, ”visit_library_in_Sunday”: True},
{”age”: 20, ”male”: False, ”single”: True, ”visit_library_in_Sunday”: True},
{”age”: 21, ”male”: False, ”single”: True, ”visit_library_in_Sunday”: True},
{”age”: 21, ”male”: False, ”single”: True, ”visit_library_in_Sunday”: True},

这些可能，此时我们来到决策树的第二层，在这些数据里选择“最优特征”，作为这一层的节点。例如，通过计算，我们发现 “male” 是此子数据集下的最优特征，那么第二层的节点就是 “male”。3. 然后我们考虑”male” 的所有取值。“male” 有两个取值 “True” 和 “False” 我们先看 “False”。在”male=False” 的这一支，只剩下

{”age”: 20, ”male”: False, ”single”: True, ”visit_library_in_Sunday”: True},
{”age”: 20, ”male”: False, ”single”: True, ”visit_library_in_Sunday”: True},
{”age”: 21, ”male”: False, ”single”: True, ”visit_library_in_Sunday”: True},
{”age”: 21, ”male”: False, ”single”: True, ”visit_library_in_Sunday”:

True},
这些数据，此时我们来到决策树的第三层，这时只有一个特征了，就是“age”，那”age” 作为这一层的节点。4. 此时”age” 有两种可能“20” 和“20”，当”age=20” 时，”visit_library_in_Sunday” 最多的取值是 True（因为 True 有 2 个，False 有 0 个）。5. 然后我们依次计算剩下的所有的取值可能，最终构建出整个决策树。

上机作业中需要完成的”predict” 函数，需要对任意给定的一行数据，给出学生是否周日会去图书馆的判断。**注意，该决策树不应只针对这一个数据集，换成另一个完全不同的数据集也应该可以使用**。例如，下面这种判断是否生病的例子
{”height”: 170, ”weight”: 60, ”age”: 20, ”exercise”: True, ”eat_breakfast”: True, ”disease”: False},