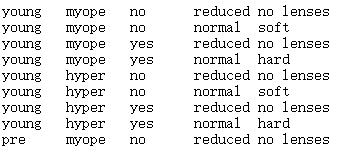
编程作业说明：决策树

**任务一：使用决策树预测隐形眼镜类型**

**问题**：眼科医生是如何判断患者需要佩戴的镜片类型的？

隐形眼镜数据集是非常著名的数据集，它包含了很多患者眼部状况的观察条件以及医生推荐的隐形眼镜类型。隐形眼镜类型包括硬材质（hard）、软材质（soft）以及不适合佩戴隐形眼镜（no lenses）。以下为该数据集的部分数据，包括年龄、近视or远视类型，是否散光，是否容易流泪，最后1列为应佩戴眼镜类型：



**准备数据**：用Python解析文本文件，解析tab键分割的数据行；

**分析数据**：快速检查数据，确保正确地解析数据内容；

**手撕算法**：不调用现有的机器学习库，自己实现决策树分类算法，获得预测隐形眼镜类型的决策树。

所需提交材料：任务一需要编程画出决策树（例如可以在Python中使用Matplotlib），编写实验报告进行简述其原理，编程思路等。

**任务二：根据用户采集的WiFi信息采用决策树预测用户所在房间**

**（1）数据集讲解**

**数据集**：训练集存于TrainDT.csv中；测试集存于 TestDT.csv中。

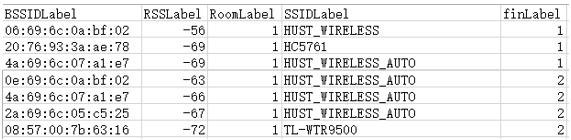
BSSIDLabel： BSSID标识符，每个AP（接入点，如路由器）拥有1个或多个不同的BSSID，但1个BSSID只属于1个AP；

RSSLabel：该BSSID的信号强度，单位dbm；

RoomLabel: 该BSSID被采集时所属的房间号，为类标签，测试集中也含该标签，主要用于计算预测准确度；

SSIDLabel: 该BSSID的名称，不唯一；

finLabel：finLabel标号相同，表示这部分BSSID在同一时刻被采集到；我们将在同一时刻采集的所有BSSID及其相应RSS构成的矢量称为一个指纹；由于BSSID的RSS在不同位置大小不同，因此指纹可以唯一的标识一个位置。



**（2）注意：**

**1. 连续值处理**：一方面，可以将每个特征划分为两个属性，未接收到RSS用0表示，接收到RSS用1表示，则一个样本可表示为

另一方面可采用二分法对连续属性进行处理，计算每个划分点的信息增益；

**2. 特征构造**：不同样本中BSSID集合不尽相同，因此可以采用所有样本BSSID集合的并集作为特征，如指纹的BSSID集合为，则特征可表示为。

**3. 缺失值处理：**本身缺失值也可以作为特征属性；若采用功能二分法则可以填补特殊值-100等。

**4. 举例说明**：

则本身只接收到、共3个本身只接收到、共3个；特征为所有样本的并集 接收到的其值用1表示，缺失值用0填充；最后一列表示样本类标签，属于房间0，属于房间1。

（上述只是提供一种思路，采用其它方法构造均可以）

**所需提交材料**：采用训练集对决策树进行训练，使用测试集进行测试，计算精度：预测正确样本数/样本总数。编写实验报告，报告中需要说明数据处理、编程思路等，需要在报告中写明自己模型在测试集中所达到的精度，可以将测试结果部分截图展示在报告中。

**任务三：预测哪些乘客在泰坦尼克号沉船中幸存**

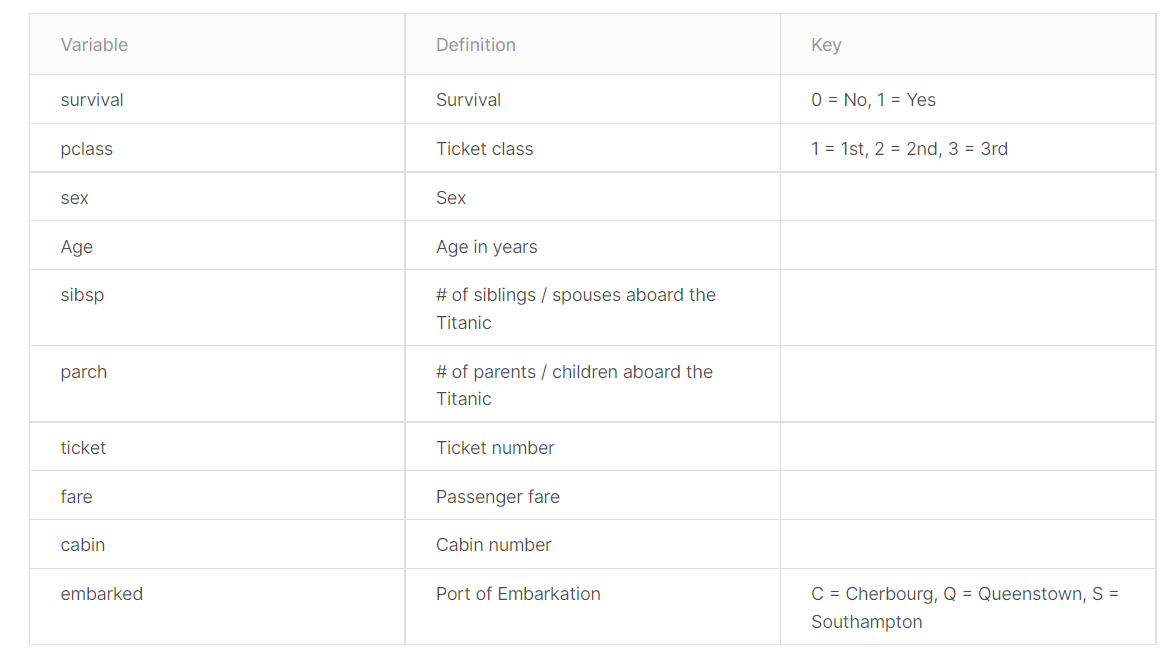
1. **任务介绍**

泰坦尼克号的沉没是历史上最臭名昭著的沉船事故之一。1912年4月15日，在她的处女航中，被广泛认为“永不沉没”的“泰坦尼克号”在与冰山相撞后沉没。不幸的是，没有足够的救生艇供船上所有人使用，导致2224名乘客和船员中有1502人死亡。虽然生存中有一些运气因素，但似乎有些群体比其他群体更有可能生存下来。在本次任务中，请使用乘客数据建立一个决策树预测模型，以回答以下问题：“什么样的人更有可能生存？”。

1. **数据集介绍**

train.csv为训练集，test.csv为测试集，训练集和测试集的每一行都表示了一名乘客的特征。不同的是，训练集给出了真值（Ground truth），即该名乘客是否存活（Survived这列），用来训练机器学习的模型；测试集没有给出真值，需要用模型进行预测，根据预测的精度可以评估模型的性能。

**数据字典：**



|  |  |
| --- | --- |
| **变量名** | **变量解释** |
| PassengerId | 乘客编号 |
| Survived | 乘客是否存活(0=NO, 1=Yes) |
| Pclass | 乘客所在的船舱等级(1=1st, 2=2nd, 3=3rd;  1st = Upper, 2nd = Middle, 3rd = Lower) |
| Name | 乘客姓名 |
| Sex | 乘客性别 |
| Age | 乘客年龄(Age is fractional if less than 1. If the age is estimated, is it in the form of xx.5)[有缺失值] |
| SibSp | 乘客的兄弟姐妹/配偶数量 |
| Parch | 乘客的父母/子女数量 |
| Ticket | 票的编号 |
| Fare | 票价 |
| Cabin | 座位号[有缺失值] |
| Embarked | 乘客登船码头[有缺失值] |

由于此数据集包含的特征较多，有的特征对于预测是否存活并不一定有用，并且缺失值和连续值的处理方法也十分多样，没有统一结论。因此，本任务不限定同学们特征工程的处理方法，仅提出**最低要求**和**一种思路**：

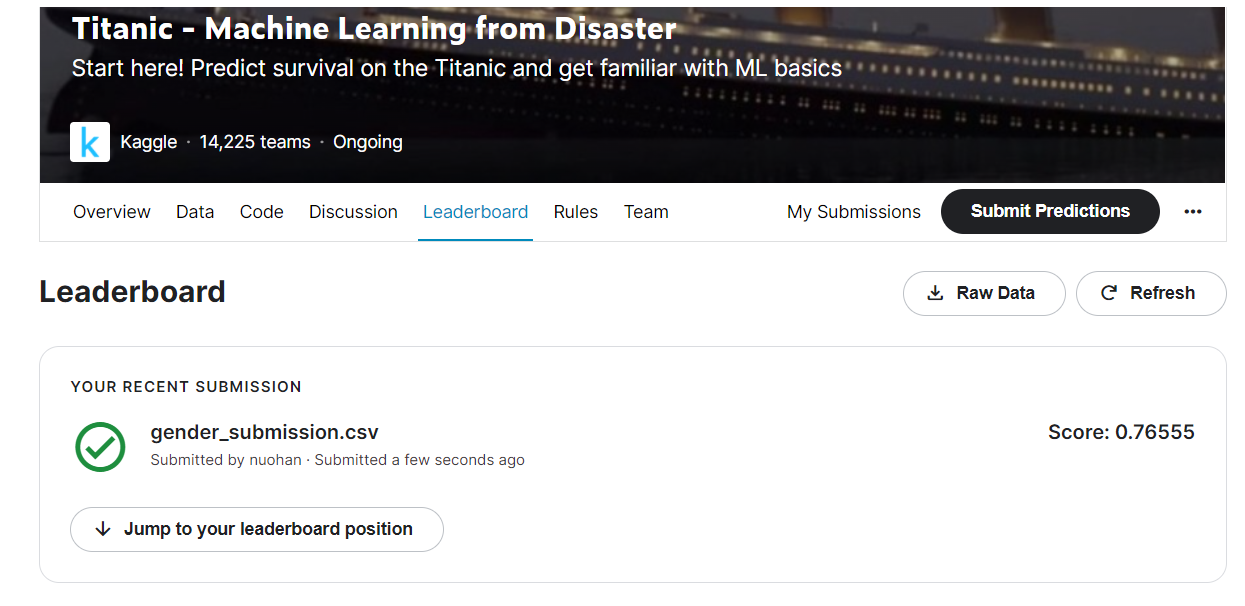
1. 选取特征时，Pclass、Sex、Age为必选项，其余为非必选项，同学们可以自行决定。
2. 缺失值处理：本身缺失值也可以作为特征属性，若采用功能二分法则可以填补特殊值-100等。
3. 连续值处理：可以将连续值转化为离散值。

**举例说明：**将Age分为5类：缺失值-100，未成年人[0-18)，青壮年人[18-45)，中年人[45-65），老年人[65-)分别用0,1,2,3,4表示。

**（3）具体要求**

（a）采用训练集对决策树进行训练，使用测试集进行预测，将预测结果存为gender\_submission.csv中示例的形式。

将预测的结果上传到网站https://www.kaggle.com/c/titanic/leaderboard，可以计算模型的预测精度，截图附在实验报告中。截图示例如下：



（b）编写实验报告，报告中需要说明：特征工程（选取了哪些特征，选取这些特征的理由，如何处理这些特征），编程思路，自己模型在测试集中所达到的精度，回答以下问题：“什么样的人更有可能生存？”。

鼓励同学们实验多种特征工程，比较它们的性能差异。

三个任务的实验报告编写在一份报告中，需要提交电子档。

（希望同学们得出理想的实验结果，学习到更多的知识！）