# 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：

```
young   myope   no      reduced no lenses
young   myope   no      normal  soft
young   myope   yes     reduced no lenses
young   myope   yes     normal  hard
young   hyper   no      reduced no lenses
young   hyper   no      normal  soft
young   hyper   yes     reduced no lenses
young   hyper   yes     normal  hard
pre     myope   no      reduced no lenses
```

**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；

**训练算法**：采用决策树分类算法，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，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$f_i = [BSSID_1\colon\ RSS_1, BSSID_2\colon\ RSS_2, \dots, \text{RoomLabel}]$；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。

| BSSIDLabel | RSSLabel | RoomLabel | SSIDLabel | finLabel |
|---|---|---|---|---|
| 06:69:6c:0a:bf:02 | -56 | 1 | HUST_WIRELESS | 1 |
| 20:76:93:3a:ae:78 | -69 | 1 | HC5761 | 1 |
| 4a:69:6c:07:a1:e7 | -69 | 1 | HUST_WIRELESS_AUTO | 1 |
| 0e:69:6c:0a:bf:02 | -63 | 1 | HUST_WIRELESS_AUTO | 2 |
| 4a:69:6c:07:a1:e7 | -66 | 1 | HUST_WIRELESS_AUTO | 2 |
| 2a:69:6c:05:c5:25 | -67 | 1 | HUST_WIRELESS_AUTO | 2 |
| 08:57:00:7b:63:16 | -72 | 1 | TL-WTR9500 | 2 |

**（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

$$f_i = [BSSID_1\colon\ 1, BSSID_2\colon\ 0, \dots]$$

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$f_i$的 BSSID 集合为$B_i = \{BSSID_j | BSSID_j \in f_i\}$，则特征可表示为$B_u = \bigcup_{i=1}^{N} B_i$。

**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**：

$$f_1 = [BSSID_1\colon 1, BSSID_2\colon 0, BSSID_3\colon 1, BSSID_4\colon 1,0]$$

$$f_2 = [BSSID_1\colon 1, BSSID_2\colon 1, BSSID_3\colon 1, BSSID_4\colon 0,1]$$

则$f_1$本身只接收到$BSSID_1$、$BSSID_3$和$BSSID_4$共 3 个$BSSID$；$f_2$本身只接收到$BSSID_1$、$BSSID_2$和$BSSID_3$共 3 个$BSSID$；特征为所有样本$BSSID$的并集$B_u = \{BSSID_1，BSSID_2，BSSID_3，BSSID_4\}$；接收到的$BSSID$其值用 1 表示，缺失值用 0 填充；最后一列表示样本类标签，$f_1$属于房间 0，$f_2$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：IMDB 数据集电影评测分类（二分类问题）

**（1）数据集讲解：**

该数据集是 IMDB 电影数据集的一个子集，已经划分好了测试集和训练集，训练集包括 25000 条电影评论，测试集也有 25000 条，该数据集已经经过预处理，将每条评论的具体单词序列转化为词库里的整数序列，其中每个整数代表该单词在词库里的位置。例如，整数 104 代表该单词是词库的第 104 个单词。为实验简单，词库仅仅保留了 10000 个最常出现的单词，低频词汇被舍弃。每条评论都具有一个标签，0 表示为负面评论，1 表示为正面评论。

训练数据在 train_data.txt 文件下，每一行为一条评论，训练集标签在 train_labels.txt 文件下，每一行为一条评论的标签；测试数据在 test_data.txt 文件下，测试数据标签未给出。

**（2）思路：**

这里提供一个最简单的思路，还有其它思路同学们可以自行思考。

首先，需要将每条评论转换为特征向量，这里可以采用 one-hot 编码，举个例子：这里采用的词库大小为 10000，因此转换的 one-hot 编码也是 10000 维的，如某条评论为[3, 5]，则转换得到的 one-hot 编码的 10000 维向量，只有索引为 3 和 5 的元素为 1，其余全部为 0。

将每条评论都转换为 one-hot 编码后，再采用决策树算法进行分类。

**（3）具体要求：**

将测试数据预测结果，与训练数据标签存储方式相同，存储为 txt 文件，每一行为一条评论的标签。将测试集预测结果的 txt 文件发送到邮箱 zhangzizhuo@hust.edu.cn。实验报告中需要写明具体实验流程，思路等。

三个任务的实验报告编写在一份报告中，需要提交纸质和电子档。

（希望同学们得出理想的实验结果，学习到更多的知识！）