特征选择与分类

——隐形眼镜数据集

特征选择

定义：我们能用很多属性描述一个西瓜，例如色泽、根蒂、敲声、纹理、触感等，但有经验的人往往只需看看根蒂、听听敲声就知道是否好瓜。换言之，对一个学习任务来说，给定属性集，其中有些属性可能很关键、很有用，另一些属性则可能没什么用。

对此，我们将属性称为“特征”(feature)，对当前学习任务有用的属性称为“**相关特征**”(relevant feature)、没什么用的属性称为“**无关特征**”(irrelevantfeature)。从给定的特征集合中选择出相关特征子集的过程，称为“**特征选择**”(feature selection)。

特征选择的两个关键环节

1.如何根据评价结果获取下一个候选特征子集?

2.如何评价候选特征子集的好坏?

第一个环节是“子集搜索" (subset search)问题。给定特征集合{a1,a2...,ad}，我们可将每个特征看作一个候选子集，对这d个候选单特征子集进行评价，假定{a2}最优，于是将{a2}作为第一轮的选定集；然后，在上一轮的选定集中加入一个特征，构成包含两个特征的候选子集，假定在这d-1个候选两特征子集中{a2，a4} 最优，且优于{a2},于是将{a2,a4}作为本轮的选定集；...假定在第k+1轮时，最优的候选(k+1)特征子集不如上一轮的选定集，则停止生成候选子集，并将上一轮选定的k特征集合作为特征选择结果。这样逐渐增加相关特征的策略称为**“前向”(forward)搜索**.

第二个环节是“子集评价”(subset evaluation)问题。给定数据集D，假定D中第i类样本所占的比例为pi (i= 1,2....1)。为便于讨论，假定样本属性均为离散型。对属性子集A，假定根据其取值将D分成了V个子集$\{D^1, D^2 \ldots D^V\}$，每个子集中的样本在A上取值相同，于是我们可计算属性子集A的**信息增益**:

$$Gain(A) = Ent(D) - \sum_{v-1}^{V} \frac{|D^v|}{|D|} \mathrm{Ent}(D^v),$$

其中信息熵定义为:

$$Ent(D) = - \sum_{i=1}^{|y|} p_k \, log_2 p_k,$$

信息增益Gain(A)越大，意味着特征子集A包含的有助于分类的信息越多。于是,对每个候选特征子集，我们可基于训练数据集D来计算其信息增益，以此作为评价准则。

数据集介绍

数据集：隐形眼镜数据集/lenses_data.txt

共有 24个样本，4个输入变量(数据集中第一列为样本编号)，1个输出变量(数据集中最后一列)

数据集信息：

Attribute Information:
-- 3 Classes
1 : the patient should be fitted with hard contact lenses,
2 : the patient should be fitted with soft contact lenses,
3 : the patient should not be fitted with contact lenses.

1. age of the patient（年龄）： (1) young, (2) pre-presbyopic, (3) presbyopic
2. spectacle prescription（症状）： (1) myope（近视）, (2) hypermetrope（远视）
3. astigmatic（是否散光）： (1) no, (2) yes
4. tear production rate（眼泪数量）： (1) reduced, (2) normal

实验内容与流程

本实验旨在对隐形眼镜数据集实现分类任务，要求采用**前向搜索算法**来选择最优特征集合，使用**信息增益**标准来衡量每个特征的重要性，实验总体流程可分为三步：

1.导入隐形眼镜数据。

2.特征选择：采用前向搜索算法，选择最优特征集合。

3.模型训练和评估：将最优特征集合中的特征作为输入，使用朴素贝叶斯算法来训练模型，并对测试集进行预测。

实验环境

Python

编辑器：Jupyter Notebook、Pycharm

可使用numpy、pandas、matplotlib等基础扩展包，建议使用anaconda安装

不可使用sklearn、pytorch等机器学习包

Jupyter Notebook使用教程：

https://zhuanlan.zhihu.com/p/33105153

Pycharm使用教程：

https://blog.csdn.net/m0_73479194/article/details/126584118

实验要求

1. 将数据集拆分成训练集（前20个）和测试集（后4个）。

2.用**前向搜索**算法来选择最优特征集合。

3.使用最优特征集合训练模型并在测试集上进行测试，输出测试集的准确率。

4. 6月5日晚上12:00之前将代码（.py或者.ipynb文件）、实验报告（doc或pdf文件）一并打包上传至邮箱。

邮箱：mlspring2023@163.com

压缩包和实验报告命名方式：

实验序号_学号_姓名，例如：实验1_111702xxxxx_王xx