HUAWEI

3.1 机器学习中的分类问题（上）

目录

Contents

什么是分类

一个实例：银行在做贷款业务的时候，专门有一个部门在做贷款申请的资格审核，他们需要得出一个结论：这个贷款的申请者的风险是否是安全可控的。这时候就需要构造一个模型或者分类器去将贷款申请者进行分类，比如1代表安全，0代表有风险。

数据标签：分类问题属于“监督学习”的范畴，即模型建立的必要条件是训练数据当中必须包含类别标签（label）数据，比如：

样本编号	特征1	特征2	特征3	特征4	特征5	label
1	…	…	…	…	…	1
2	…	…	…	…	…	0

1、数据标签用于在进行分类模型训练时对于label列分布的规律进行监督式的发现，准确的数据label是构建一个准确的分类模型的非常重要的因素。

2、label本身无论是数值还是字符串都没有次序、大小方面的意义，它们只是表示了类别的不同。

分类的定义与评估

定义：使用带有类别标签的训练数据通过一定的算法或规则得出一个针对类别标签的分类器，这个分类器对于不带标签的数据给出标签的预测结果。

数据定义：

训练数据：1个样本是1个n维向量X*=(x1,x2,x3,...,xn)，每1个样本拥有1个自己的标签Y*。则样本{<X1*，Y1*>, <X2*，Y2*>,...., <Xm*，Ym*>}构成了一个样本空间（m x （n+1））。

预测数据：1个样本是1个n维向量X=(x1,x2,x3,...,xn)，没有标签Y。则样本{<X1>, <X2>,...., <Xk>}构成了一个样本空间（k x n）。

过程：

第一阶段进行分类规则、分类器的的学习，称为训练阶段。即在训练数据集上使用算法或规则得到一个样本X*与标签Y*的映射：Y*=f(X*)

第二阶段进行分类规则、分类器的应用与更新，称为预测阶段。即使用训练阶段的结果映射f对于预测样本X进行处理：Y= f(X)

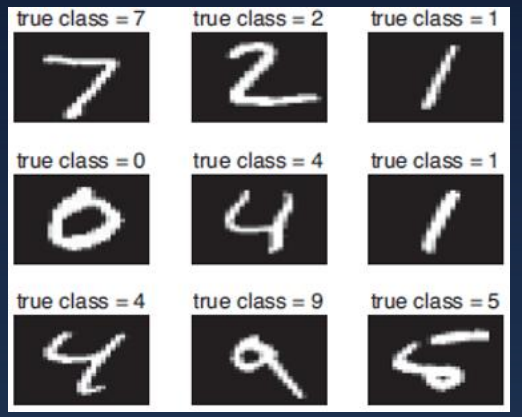
评估：

在训练阶段当中，将带有类别标签的训练数据划分为训练集和验证集，每一次使用训练集得到一个f ，即用验证集去进行验证。

分类实例&典型算法

分类问题实例：

- 文档分类、垃圾邮件过滤
- 图像分类、手写体识别
- 人脸侦测与识别
- 离网预测
- ……



典型分类算法

- Decision Tree
- Native Bayes
- Random Forest
- Logistic Regression
- Support Vector Machine
- ……

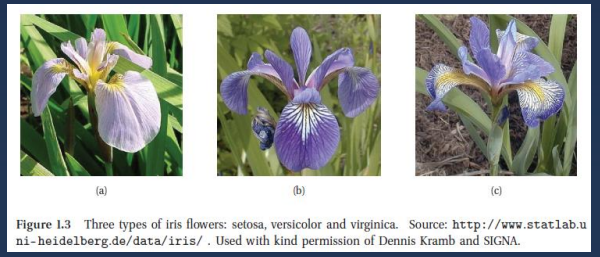


Figure 1.3 Three types of iris flowers: setosa, versicolor and virginica. Source: http://www.statlab.uni-heidelberg.de/data/iris/ . Used with kind permission of Dennis Kramb and SIGNA.

分类模型的评估

- 准确率（accuracy）
- 精确率（查准率，precision）
- 召回率（查全率，recall）
- F1 score
- ROC曲线
- AUC曲线
- 混淆矩阵（Confusion matrix）

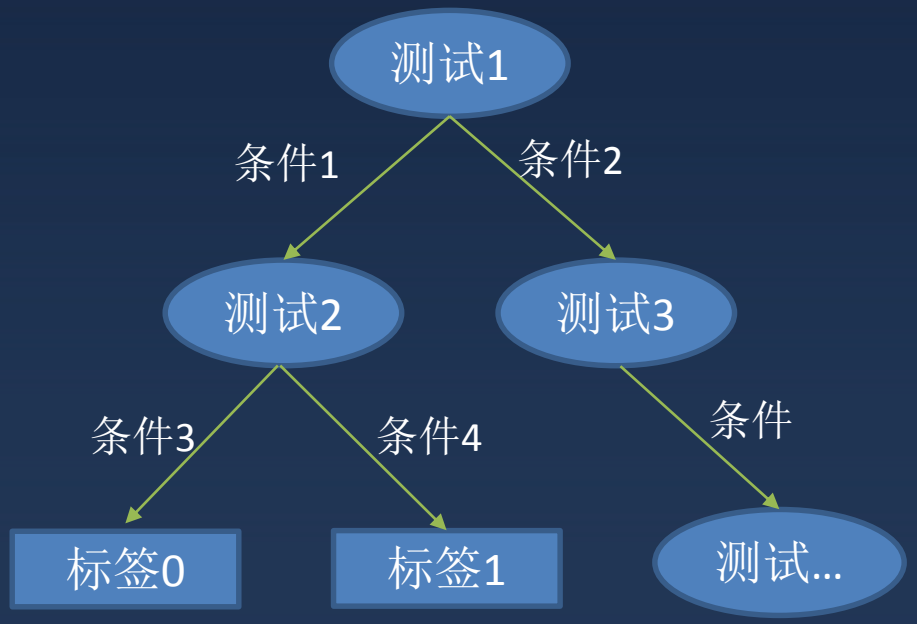
目录

Contents

决策树是什么

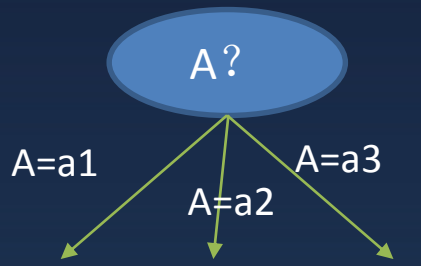
决策树是一种类似于流程图的树形结构，其中每个非叶结点都表示在某一个特征或者特征组合上的测试，每一个分支表示该测试的一个条件输出，而每个叶结点存放一个标签。

决策树的典型形式

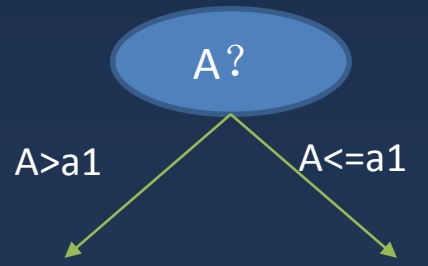


3种形式的测试与条件输出的形式

A是离散值



A是连续值



A是离散值
且只能产生
二叉分裂

决策树分裂准则

决策树的关键在于给出一种方法可以科学的计算哪一种分裂的方式是合理的。

信息增益：基于信息论，定义数据D中的期望信息为$Info(D) = -\sum_{i=1}^{m} p_i log_2(p_i)$，$p_i$是D中任意样本属于类别i的非零概率。$Info(D)$又称为D的熵。某个特征A的期望信息为$Info_A(D) = \sum_{j=1}^{v} \frac{|Dj|}{|D|} \times Info(Dj)$。 $|D|$是指D中的样本个数。

特征A上的信息增益定义为：Gain(A)=$Info(D)$-$Info_A(D)$。

每一次进行数据分裂时，计算当前数据上每个特征的信息增益，选择信息增益最大的特征作为分裂的测试。

基尼指数：定义数据D中的不纯度信息为$Gini(D) = 1 - \sum_{i=1}^{m} {p_i}^2$，$p_i$是D中任意样本属于类别$i$的非零概率。基尼指数考虑对一个特征A进行二叉分裂，假设给定A一个二叉分裂sp_1，则这种分裂的不纯度信息为

$$Gini_{A\_sp1}(D) = \frac{|D_1|}{|D|} \times Gini(D_1) + \frac{|D_2|}{|D|} \times Gini(D_2)$$

则这种分裂导致的不纯度的降低为$\Delta\ Gini(A_sp1) = Gini(D) - Gini_{A_sp1}(D)$

每一次进行数据分裂时，计算当前数据上每个特征的每一种二叉分裂的$\Delta\ Gini$ ，选择$\Delta\ Gini$最大的特征的分裂作为分裂的测试。

决策树分裂计算

银行客户的信息如下

id	income	age	working	history_credit	label
1	h	y	0	0	0
2	h	y	0	1	0
3	h	m	0	0	1
4	m	s	0	0	1
5	l	s	1	0	1
6	l	s	1	1	0
7	l	m	1	1	1
8	m	y	0	0	0
9	l	y	1	0	1
10	m	s	1	0	1
11	m	y	1	1	1
12	m	m	0	1	1
13	h	m	1	0	1
14	m	s	0	1	0

1、请尝试计算在特征age上的信息增益。

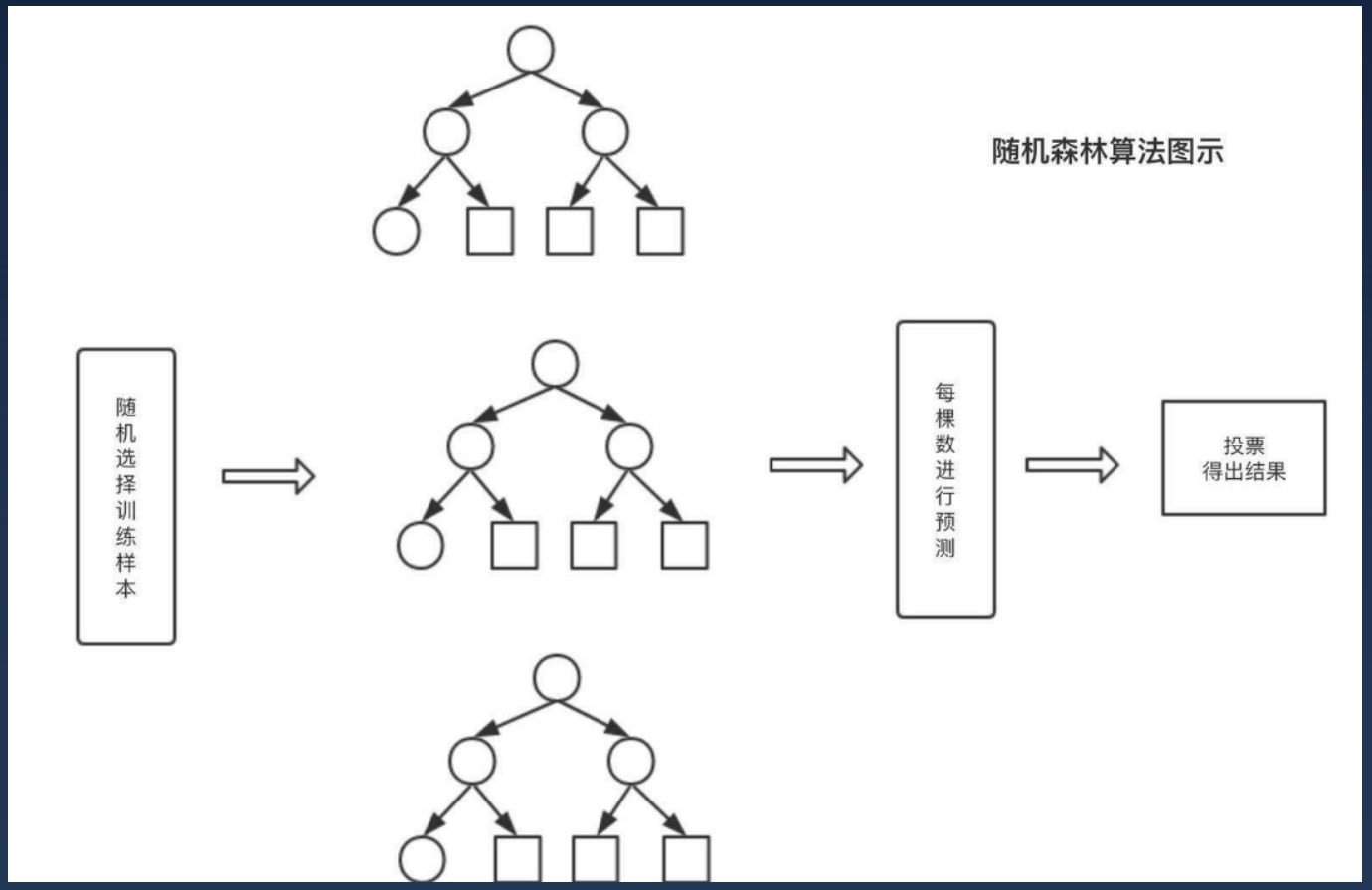
2、请尝试计算在特征income上 $\Delta\ Gini(A_sp)$最大的分裂

目录

Contents

随机森林是什么

决策树产出的是一棵树去进行分类的判定，随机森林是使用抽样有放回的方式产生多个训练集，每个训练集产出一棵决策树，最终形成一个决策森林。



最终的分类判定由决策森林中的每棵树投票产生。投票方式可以有如下形式：

1、平等投票，即每棵树的票权重相同。最终哪个类别得到的票数多，则这个类别作为模型的判定类别。

2、权重投票，权重与产生决策树的数据量相关，数据量越大权重越大。最终哪个类别得到的权重票越高，则这个类别作为模型的判定类别。

一般采用方式1。

随机森林中的随机过程

I、随机森林的第一步是随机产生多个数据集，需要提供的参数是森林中树的个数，随机的方式有以下两种：

设森林中树的个数为k，原始训练集为D，一共m个样本，每个样本n个特征。

1、随机样本：对于D中的样本进行有放回的抽样，抽样k次，产生k个小训练集，每个小训练集的样本数小于m，每个样本的特征数依然是n。

2、随机特征：对于D中的特征进行有放回的抽样，抽样k次，产生k个小训练集，每个小训练集的样本数等于m，每个样本的特征数小于n。

3、随机特征组合：随机产生一个数字L，L<n，随机选择L个特征，随机从[-1，1]中产出系数，L个特征和系数产生F个线性组合。

一般3种方式组合进行使用，最终产生的k个小训练集中的每一个的样本数小于m、特征数小于n

II、随机森林的第二步是随机进行分裂规则选择，每个小训练集进行决策树产出的时候，可以随机的选择是使用信息增益还是基尼指数等，

III、随机森林的第三步是随机进行分裂特征的指定，每个小训练集进行决策树产出的时候，进行到每一步的分裂计算时，随机产生一个小于n的数t，且t小于这个小训练集的特征数，然后在当前分裂上随机选择t个特征进行分裂规则的计算。

随机森林的优点

1、准确率比单棵决策树高。

2、对于离群点更加鲁棒。

3、随着森林中树个数的增加，森林的泛化误差收敛。

4、每棵树由于样本数和特征数都比原始训练集小，计算简单。

5、随机森林使用随机过程保持了每棵树的独立性。

6、随机森林在大型数据上非常有效。

7、随机森林具有强可解释性，可以给出特征重要性的内在估计。

Thank You.

华为云机器学习服务MLS
www.huaweicloud.com/product/mls.html