

信息熵在决策树分类中的简单应用



j13hu (/u/cd591e266c5d) [+关注](#)

2015.11.12 22:36* 字数 855 阅读 737 评论 2 喜欢 5

(/u/cd591e266c5d)

（一）信息熵

信息熵是信息的期望值，描述信息的不确定度。熵越大，表明集合信息的混乱程度越高，换句话说，集合信息混沌，其包含信息价值少。

信息熵计算公式：

$$H(U) = - \sum_{i=1}^n p_i \log p_i$$

信息熵公式

（二）信息增益

信息增益是对信息前后变化量的描述。

计算公式：

```
infoGain = baseEntropy - EntropyAfter
```

当 $\text{infoGain} > 0$ ，表明集合信息熵减小，包含的信息更纯更有序，价值得到提高。

当 $\text{infoGain} < 0$ ，信息变得混沌。

$\text{infoGain} = 0$ ，信息量没有变化，但不表明信息没有变化。

我们要依据样本特征来分割数据集以使数据变得更加有序的过程就是要求一个使得

$\text{infoGain} > 0$ 的分类特征，而且是在所有的特征中使得 infoGain 值最大的特征。

在决策树分类中，数据集包含的分类特征不止一个，那么，在第一次划分数据集时如何从整个特征集中抽出一个特征？第二次呢。信息增益便是对这种划分数据集前后信息是更有序了，还是更混乱的度量。

选出取得信息增益最大的特征，以该特征对数据集分类，如果该特征共有 n 个特征值，那么划分数据集后会生成 n 个节点。这些 n 个节点再计算信息增益，并选出最大的特征，以此往复，直到最终节点仅包含同一类数据为止。

（三）Python代码实现

1，划分数据

`splitDataSet` 方法接收待分类的数据集、预选中的分类特征 `featIndex` 及特征的值 `featValue`。如果样本特征 `featIndex` 的值与 `featValue` 值相等，则归集该样本且移除该分类特征。返回的 `retDataSet` 即为移除分类特征的子数据集。

```
def splitDataSet(dataSet, featIndex, featValue):
    retDataSet = []
    for singleData in dataSet:
        if singleData[featIndex] == featValue:
            reducedFeatVect = singleData[:featIndex]
            reducedFeatVect.extend(single[featIndex + 1:])
            retDataSet.append(reducedFeatVect)
    return retDataSet
```

2，选出最佳分类特征

试探性选出最佳分类特征的过程是：

第一步，先从数据集的第一列作为分割数据的特征，接着第二列，第三列直到最后一列。取出第一列的所有值，并存入集合以去重复。

第二步，遍历集合中的分类特征的值，把每一个值传入`splitDataSet`方法来判断该样本是否应该化为该特征的值代表的同质集，换句话说，与该特征的值相等的归集到一类。

第三步，对同一个分类特征而言，可能具有不同的值。特征的值有多少个，对应节点分



支有多少支。因此，在对一个分类特征计算信息熵大小时，必须把各个特征的值的的信息熵相加，这才代表该分类特征所具有的信息熵。

第四步，原信息熵 $baseEntropy$ 与分类后的信息熵 $newEntropy$ 相减即求得信息增益。该次迭代求出的信息增益与上次相比较，若大于上次的信息增益，说明集合信息变得更纯些（熵减），则把此次信息增益赋给变量 $bestInfoGain$ 并把该特征所在列表索引赋给变量 $bestFeature$ ，最后返回最佳的分类特征

```
def chooseBestFeature(dataSet):
    numFeatures = len(dataSet[0]) - 1
    baseEntropy = calcShannonEnt(dataSet)
    bestInfoGain = 0.0
    bestFeature = -1
    for i in range(numFeatures):
        featList = [ds[i] for ds in dataSet]
        uniqueVals = set(featList)
        newEntropy = 0.0
        for value in uniqueVals:
            subDataSet = splitDataSet(dataSet, i, value)
            prob = float(len(subDataSet)) / len(dataSet)
            newEntropy += prob * calcShannonEnt(subDataSet)
            infoGain = baseEntropy - newEntropy
        if infoGain > bestInfoGain:
            bestInfoGain = infoGain
            bestFeature = i
    return bestFeature
```

3，计算信息熵

```
def calcShannonEnt(dataSet):
    numberEnt = len(dataSet)
    labelCounts = {}
    for feat in dataSet:
        currentLabel = feat[-1]
        if currentLabel not in labelCounts.keys():
            labelCounts[currentLabel] = 0
        labelCounts[currentLabel] += 1
    shannonEnt = 0.0
    for key in labelCounts:
        prob = float(labelCounts[key]) / numberEntries
        shannonEnt -= prob * log(prob, 2)
    return shannonEnt
```

小礼物走一走，来简书关注我

赞赏支持

机器学习 (/nb/2431546)

举报文章 © 著作权归作者所有



j13hu (/u/cd591e266c5d)

写了 8397 字，被 63 人关注，获得了 69 个喜欢
(/u/cd591e266c5d)

+ 关注

♡ 喜欢 (/sign_in?utm_source=desktop&utm_medium=not-signed-in-like-button) | 5



更多分享

(http://cwb.assets.jianshu.io/notes/images/238865)



下载简书 App ▶
随时随地发现和创作内容



(/apps/download?utm_source=nbc)





登录 (/sign-in?utm_source=desktop&utm_medium=not-signed-in-comment-form)

2条评论

只看作者

按喜欢排序 按时间正序 按时间倒序



小武子 (/u/f4nqA1)

2楼 · 2015.11.15 10:39

(/u/f4nqA1)
正是我需要的 😊

👍 赞 💬 回复



j13hu (/u/cd591e266c5d) 作者

3楼 · 2015.11.15 10:49

(/u/cd591e266c5d)



共同学习

👍 赞 💬 回复

被以下专题收入，发现更多相似内容



机器学习与模式识别 (/c/1395428608b4?

utm_source=desktop&utm_medium=notes-included-collection)



机器学习 (/c/e6b06fcb9cad?utm_source=desktop&utm_medium=notes-

included-collection)



首页投稿 (/c/bDHhpK?utm_source=desktop&utm_medium=notes-included-

collection)



机器学习 (/c/b46b3137e014?utm_source=desktop&utm_medium=notes-

included-collection)

推荐阅读

更多精彩内容 > (/)

应用朴素贝叶斯分类器对文本简单分类 (/p/f18047f24e8c?utm_campaign=...

朴素贝叶斯分类器 一，生成词向量（词集模型）第一，假设这里有两个参数vocabList, inputSet. vocabList代表着包含很多无重复的词，词量足够大，inputSet代表着我们预转换的词列表。第二，创建一个与...

j13hu (/u/cd591e266c5d?

utm_campaign=maleskine&utm_content=user&utm_medium=pc_all_hots&utm_source=recommendation)

开始简书之旅 (/p/31c3ebf9e603?utm_campaign=maleskine&utm_conte...

听说百度空间要关闭了，故转至简书。慢慢地学起了python，决定朝着技术方向发展。前几天在捣弄BeautifulSoup一些基础的东西指定获取某标签的内容，如 简书 获取到“简书”元素。我使用的方法是把文档...

j13hu (/u/cd591e266c5d?

utm_campaign=maleskine&utm_content=user&utm_medium=pc_all_hots&utm_source=recommendation)

李小璐出轨，苍井空结婚，关于爱情你怎么看？ (/p/dd... (/p/dd003bad28d9?

01 今天娱乐圈爆出的消息，信息量有点大。总结一下就是，有人出道多年，终于结婚；有人结婚多年，终于出轨。2017年的最后一天，消失很久的卓伟放...

衷曲无闻 (/u/deeea9e09cbc?

utm_campaign=maleskine&utm_content=user&utm_medium=pc_all_hots&utm_source=recommendation)



写作最重要的天赋是？ (/p/29df29fd0ed5?utm_campaign=maleskine&utm_content=note&utm_medium=seo_notes&utm_source=recommendation) (/p/29df29fd0ed5?utm_campaign=maleskine&utm_content=note&utm_medium=seo_notes&utm_source=recommendation)

很多人说他喜欢写作，但是他一个字都没有写，因为觉得自己没有天赋。呃，貌似这是一个十分充足的理由，无懈可击的样子。那么什么是天赋呢？词典里...

断鹞 (/u/65096740cbfc?utm_campaign=maleskine&utm_content=user&utm_medium=pc_all_hots&utm_source=recommendation)

二十多岁的我们，拥有多少存款？ (/p/c12d523ca005?utm_campaign=maleskine&utm_content=note&utm_medium=seo_notes&utm_source=recommendation) (/p/c12d523ca005?utm_campaign=maleskine&utm_content=note&utm_medium=seo_notes&utm_source=recommendation)

你是不是有过这样的经历？看到一件很喜欢的大衣，看了看吊牌价，记下尺码和款型，在某宝的里开始搜索，收藏，等待着重大节点再下单。北漂上漂的...

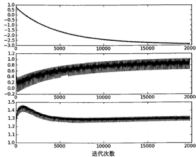
羊达令 (/u/ce94d617e045?utm_campaign=maleskine&utm_content=user&utm_medium=pc_all_hots&utm_source=recommendation)

分类算法-决策树 (/p/0d03b1e01487?utm_campaign=maleskine&utm_content=note&utm_medium=seo_notes&utm_source=recommendation) (/p/0d03b1e01487?utm_campaign=maleskine&utm_content=note&utm_medium=seo_notes&utm_source=recommendation)

决策树理论在决策树理论中，有这样一句话，“用较少的东西，照样可以做很好的事情。越是小的决策树，越优于大的决策树”。数据分类是一个两阶段过程，包括模型学习阶段（构建分类模型）和分类预测阶段（使...

制杖灶灶 (/u/75d6088120ca?utm_campaign=maleskine&utm_content=user&utm_medium=seo_notes&utm_source=recommendation)

(/p/ed9ae5385b89?utm_campaign=maleskine&utm_content=note&utm_medium=seo_notes&utm_source=recommendation) (/p/ed9ae5385b89?utm_campaign=maleskine&utm_content=note&utm_medium=seo_notes&utm_source=recommendation)



注：题中所指的『机器学习』不包括『深度学习』。本篇文章以理论推导为主，不涉及代码实现。前些日子定下了未来三年左右的计划，其中很重要的一点是成为一名出色的人工智能产品经理，说是要每月至少读...

我偏笑_NSNirvana (/u/2293f85dc197?utm_campaign=maleskine&utm_content=user&utm_medium=seo_notes&utm_source=recommendation)

(/p/43267a5b61ce?utm_campaign=maleskine&utm_content=note&utm_medium=seo_notes&utm_source=recommendation) (/p/43267a5b61ce?utm_campaign=maleskine&utm_content=note&utm_medium=seo_notes&utm_source=recommendation)

序号	问题描述 (描述)	问题类型 (类别)	问题难度 (难度)	问题来源 (来源)	问题状态 (状态)
1	问题描述	类别	难度	来源	状态
2	问题描述	类别	难度	来源	状态
3	问题描述	类别	难度	来源	状态
4	问题描述	类别	难度	来源	状态
5	问题描述	类别	难度	来源	状态
6	问题描述	类别	难度	来源	状态
7	问题描述	类别	难度	来源	状态
8	问题描述	类别	难度	来源	状态
9	问题描述	类别	难度	来源	状态
10	问题描述	类别	难度	来源	状态

分类与预测 餐饮企业经常会碰到下面的问题：如何预测未来一段时间内，哪些顾客会流失，哪些顾客最有可能成为VIP客户？如何预测一种产品的销售量，以及在何种类型的客户中会较受欢迎？除此之外，餐厅...

Skye_kh (/u/70b4fd000153?utm_campaign=maleskine&utm_content=user&utm_medium=seo_notes&utm_source=recommendation)

(/p/07cede509d90?utm_campaign=maleskine&utm_content=note&utm_medium=seo_notes&utm_source=recommendation) (/p/07cede509d90?utm_campaign=maleskine&utm_content=note&utm_medium=seo_notes&utm_source=recommendation)

$$\sum_{i=1}^n \frac{1}{x_i} \leq \frac{1}{x} \leq \sum_{i=1}^n \frac{1}{x_i^2}$$

是当第 1 个元素的和，并记第 1 个元素 (即，A 中为 a1) 中的第 1 个元素为 a1，子集划分的情况如下：注意，对于第 1 个元素 a1，

$$P(a_1, a_2, \dots, a_n) = \frac{1}{n!} \cdot \frac{1}{n!}$$

是 A 中的样本属于 C 的概率。

每获得的信息熵是

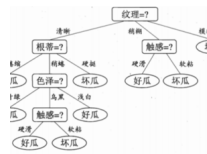
$$H(A) = -\sum_{i=1}^n P(a_i) \log_2 P(a_i)$$

用判定树归纳分类 (/p/07cede509d90?utm_campaign=maleskine&utm_content=note&utm_medium=seo_notes&utm_source=recommendation) (/p/07cede509d90?utm_campaign=maleskine&utm_content=note&utm_medium=seo_notes&utm_source=recommendation)

“什么是判定树？”判定树是一个类似于流程图的树结构；其中，每个内部结点表示在一个属性上的测试，每个分枝代表一个测试输出，而每个树叶结点代表类或类分布。树的最顶层结点是根结点。一棵典型的判定树...


1想得美 (/u/ffeea433b895?utm_campaign=maleskine&utm_content=user&utm_medium=seo_notes&utm_source=recommendation)

(/p/8c4a3ef74589?utm_campaign=maleskine&utm_content=note&utm_medium=seo_notes&utm_source=recommendation)



utm_campaign=maleskine&utm_content=note&utm_medium=seo_notes&utm_source=recommendation)
机器学习-决策树 (/p/8c4a3ef74589?utm_campaign=maleskine&utm_co...

1、引言 决策树 (Decision Tree) 是数据挖掘中一种基本的分类和回归方法，它呈树形结构，在分类问题中，表示基于特征对实例进行分类的过程，可以认为是if-then规则的集合，也可认为是定义在特征空间与...

 wwlovet (/u/c5df9e229a67?)


utm_campaign=maleskine&utm_content=user&utm_medium=seo_notes&utm_source=recommendation)

(/p/1909d53cdb94?



utm_campaign=maleskine&utm_content=note&utm_medium=seo_notes&utm_source=recommendation)
iOS 1分钟超简单配置pod环境 (/p/1909d53cdb94?utm_campaign=males...

1. 查看源:sudo gem sources -l 2. 删除源:sudo gem sources -r https://rubygems.org/ 3. 添加源:sudo gem sources -a https://ruby.taobao.org 4. 安装cocoa...

 翻滚的炒勺2013 (/u/884a67907187?)


utm_campaign=maleskine&utm_content=user&utm_medium=seo_notes&utm_source=recommendation)

(/p/86d7ff89d87f?



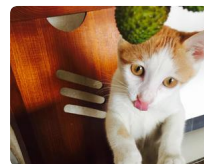
utm_campaign=maleskine&utm_content=note&utm_medium=seo_notes&utm_source=recommendation)
2017-07-30 (/p/86d7ff89d87f?utm_campaign=maleskine&utm_content=...

我真的无法拒绝：一个“喝”出来的事业机会！！ 风靡全球50多个国家！ 全球独一无二的爆品---- XS能量饮料！ 你拒绝得了吗？？ [太阳][太阳] 全国招商中 [握手][握手] 15922939939/13368158338

 安利龍蝶语XS全国招代理 (/u/8834b15f4d0b?)


utm_campaign=maleskine&utm_content=user&utm_medium=seo_notes&utm_source=recommendation)

(/p/f4acc3b2e5c9?



utm_campaign=maleskine&utm_content=note&utm_medium=seo_notes&utm_source=recommendation)
0603感谢29 (/p/f4acc3b2e5c9?utm_campaign=maleskine&utm_content=...


0603感谢29 稳住，我们能赢！——《王者荣耀》进入六月，高考季。忙碌着高考考场的工作，看到高三师生的紧张，感受着学校的高考氛围。儿子仍然淡定，情绪反复，有时明白，有时糊涂。 感谢儿子前几天的...

 文放wf (/u/a63abc76e52a?)

utm_campaign=maleskine&utm_content=user&utm_medium=seo_notes&utm_source=recommendation)

雨季不再来 (/p/8ea4527fc609?utm_campaign=maleskine&utm_content=...

在辽阔的草原上，一阵热浪席卷而来，将土地龟裂，染青草枯黄。原上的生灵祈求着一场甘霖，但他们不知道在这被人们遗忘的地方，雨季不再来。 寝室小王走的时候，我帮他吧行李搬下楼，他去与友人交谈，我...


 牧笛少年 (/u/17dd72f00293?)

utm_campaign=maleskine&utm_content=user&utm_medium=seo_notes&utm_source=recommendation)

眠 (/p/2997f5b0c017?utm_campaign=maleskine&utm_content=note&u...

想睡好觉 一夜无梦 醒来清明 充满朝气



 Wuooo (/u/ca24a9d60c91?
utm_campaign=maleskine&utm_content=user&utm_medium=seo_notes&utm_source=recommendation)

