大米花jenni的博客

博客认证 新浪运营部技术主管
http://blog.sina.com.cn/ijenniyang ［订阅］［手机订阅］

个人资料

大米花jenni

微博

加好友 发纸条

写留言 加关注

博客地图 world map

博客等级：11
博客积分：887
博客访问：10,899
关注人气：10
获赠金笔：2
赠出金笔：0
荣誉徽章：

相关博文

北京最窄的小喇叭胡同
老莫咔嘛眼

用紧身裤勾勒曲线的潮女
原生泰

GBDT代码解读
HuoChengfu

LDA主题模型
superbear

正文

字体大小：大 中 小

[转载]【转】GBDT算法介绍

(2013-01-04 18:37:39)

转 载 ▼

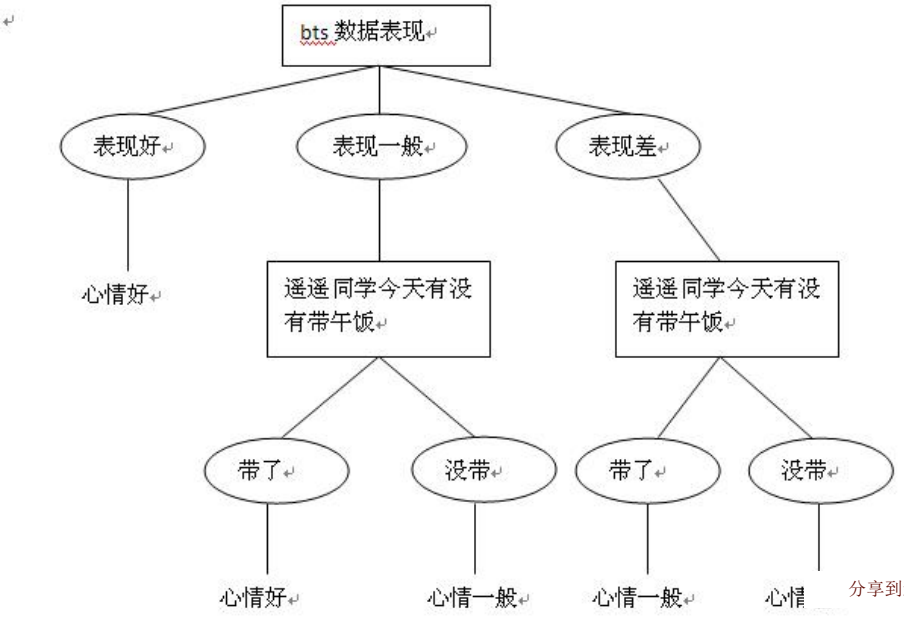
标签： 转载

原文地址：【转】GBDT算法介绍 作者：autom

http://www.searchtb.com/2010/12/an-introduction-to-treelink.html

“机器学习”这个名词对大家来说想必不是一个陌生的词汇，特别对算法组的同学来说，工作中或多或少接触使用过这种“高科技“。对于我来说，刚来淘宝工作一个月就开始接触了机器学习，当时做主搜索功夫熊猫项目，和小致飘雪一起做交易模型，正是使用了机器学习的方法，也首次接触了treelink模型。做完那个项目后对机器学习解决问题的流程有了一定的了解，但对其内部的工作原理和实现机制还是完全不知道，基本也就是在黑盒使用机器学习工具。后面也多多少少听了一些机器学习的讲座，但都是一些比较宽泛的基本概念，没有深入的原理性的介绍。也自己尝试过专研一下，但生硬晦涩的E文让人望而生畏。一直到今年做导购搜索的项目，又再次需要使用机器学习，“怀揣着对科学真理的向往”，主动请缨做模型方面的工作。经过一个多月的学习实践，算是对treelink模型有了一定的了解。下面做一些对treelink模型通俗版的介绍。都是自己的一些理解，如果有误，多指教。

在介绍treelink之前首先不得不简单介绍一下决策树算法，决策树相信大家都有所了解，任何一本机器学习书籍都会介绍这种算法，也是应用最广的归纳推理算法之一。该模型学习的结果是一棵决策树，这棵决策树可以被表示成多个if-else的规则。下图是一个典型的学习得到决策树。这棵决策树根据两个特征因素来分类“元涵今天的心情好坏”。长方形的表示特征，椭圆型的表示特征的取值，最下面的叶子节点就是最后的分类结果了。



学习得到如上这棵决策树之后，当输入一个待预测的样本实例的时候，我们就可以根据这个样本的两个特征的取值来把这个样本划分到某一个叶子节点，得到分类结果了，这就是决策树模型的预测过程，决策树的学习构建过程这里就不介绍了，大家看书吧，比较经典的有ID3算法和C4.5算法。

CAD中模型与布局的设置
Nicole

MapReduce的Shuffle阶段
枝叶飞扬

DISC模型
江边渔火

C的|、||、&、&&、异或、~、！运
金色雨露

血凝四项及D-二聚体临床意义
检验之星

短文本的topicmodel
liushengbing

源码安装MongoDB的过程
pzghost

C++中运算的溢出检查
FollowKobeBryant

更多>>

推荐博文

【2016右腿摄】村子里发了1

北美崔哥三八节檄文：千万别和女

高三男生给女生下药，我们只看到

摇孩子会不会把孩子摇傻

为何换下开胡的马丁内斯？球队问

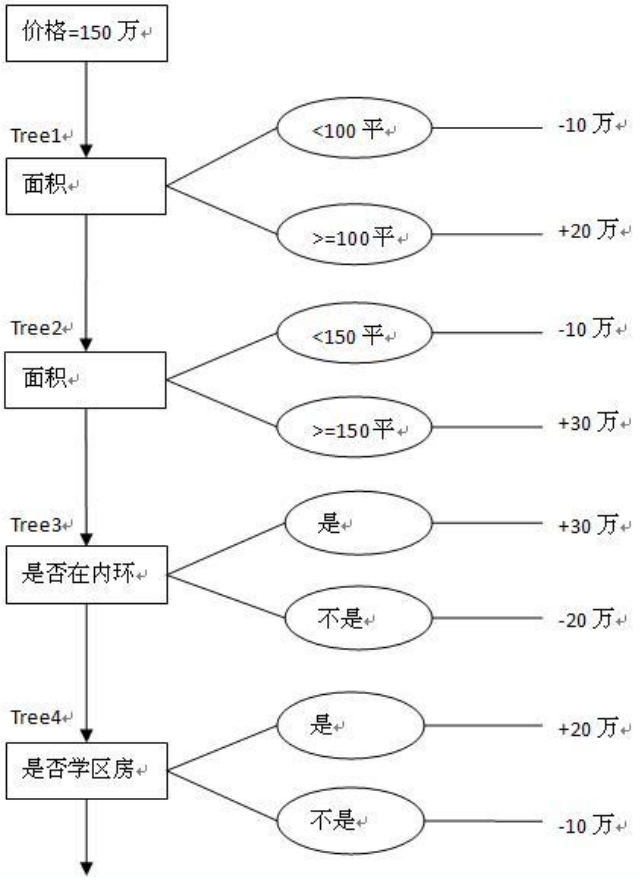
延迟退休应由谁做主？

切入正题下面开始说treelink算法，treelink这个名字其实只是我们阿里集团内部的叫法，学术上的名称叫GBDT（Gradient boost decision tree）。treelink模型不像决策树模型那样仅由一棵决策树构成，而是由多棵决策树构成，通常都是上百棵树，而且每棵树规模都较小（即树的深度会比较浅）。模型预测的时候，对于输入的一个样本实例，首先会赋予一个初值，然后会遍历每一棵决策树，每棵树都会对预测值进行调整修正，最后得到预测的结果。

$$F(X) = F_0 + \beta_1 T_1(X) + \beta_2 T_2(X) + \ldots + \beta_M T_M(X)$$

F0是设置的初值， Ti是一棵一棵的决策树。对于不同的问题（回归问题或者分类问题）和选择不同的损失函数，初值的设定是不同的。比如回归问题并且选择高斯损失函数，那么这个初值就是训练样本的目标的均值。

下面是一个简单的treelink模型示意图。模型的目标是上海一套普通商品房的价格，特征有三个：房子的面积（连续特征），是否在内环（分类特征），是否学区房（分类特征）。模型由四棵决策树构成，每棵决策树只进行了一次分裂，即树的深度为一（这种树被称为Decision Stump）。实际应用中通常较复杂，深度不会为一。



初值设定为上海普通商品房价格的均值150万，每经过一棵决策树，都会对根据相应特征的取值对预测价格进行调整。比如一个面积为120平的内环非学区房的价格预测值为：150+20-10+30-10=180万。

那为什么要用多棵决策树，一棵决策树为什么不好呢？使用单棵决策树，最大的问题就是会因为过度分裂，而造成过拟合，失去泛化能力。试想一下，对于给定的一批训练数据，完全可以只构造一棵树，不断分裂下去，直到每个叶子节点包含的样本的目标值都一样，然后把这节点的预测值设定成这个目标值，这样构造出来的这棵树就可以在这批训练数据上达到100%的准确性。但这样一棵过度分裂的决策树，对于新的样本基本没有什么预测能力。而如果分裂太少，又会造成学习不够充分。Treelink使用多棵决策树正是希望能够在训练精度和泛化能力两个方面都能达到较好的效果。作为一种boosting算法，Treelink自然包含了boosting的思想：将一系列弱分类器组合起来，构成一个强分类器。它不要求每棵树学到太多的东西，每颗树都学一点点知识，然后将这些学到的知识累加起来构成一个强大的模型。举个现实生活中的例子，电视里的那种益智类节目，如开心词典，答题者有三次请求帮组的机会，其中一个就是请求现场所有观众，通过他们的选择来给出答案。我们可以把每个观众当做一个弱的分类器，他们各个单独的准确率都不高，但把他们的知识综合起来这个准确率会大大提升。也许上面这个例子不太能说服你，我们来把这个例子量化。假如我们有三个观众，他们各自的准确率为60%（非常弱的分类器，只比随机分类器好一点点），如果这三位观众中有大于等于两位的答案是正确的，那么则认为我们正确了，反之则错误。那么我们正确的概率是多少呢？preciseness = p(三个人都正确) + p(三个人中有两个人正确) = 0.6*0.6*0.6 + 3*0.6*0.6*0.4 = 0.648，比单个人0.6的正确率有所提升，验了中国那句老话“三个臭皮匠顶个诸葛亮”。随着人数的增加，这个正确率还会提升。

Treelink模型的学习过程，就是多颗树的构建过程。和决策树模型一样，在树的构建过程中，最重要的就是寻找分裂点（某个特征的某个取值）。我们希望选择的这个分裂点是最能区分样本的。那么如何衡量一个分裂点

对样本的区分能力？在treelink算法我们通过Loss（衡量样本预测值与目标值的差异）的减小程度用来衡量这个区分能力，Loss减小得越多，这个分裂点就越好。即以某个分裂点划分，把样本分成两部分，使得分裂后样本的损失函数（Loss Function）值减小的最多。好像有点不太通俗了，没办法。

训练流程：

1 估计初值

2 按如下方式构造M颗树

2.1 随机选取部分样本作为本颗树的训练数据

2.2 按如下方式寻找最优分裂点，进行N次叶子节点的分裂

2.2.1 对当前所有叶子节点

2.2.1.1 计算该叶子节点的最优划分以及其增益（损失函数减少量）

2.2.1.2 选择增益最大的叶子节点及其划分点，进行分裂，将样本划分到子节点中

2.2.1.3 更新样本估计值

集团开发的mllib机器学习工具包中treelink是最重要的一个模型。对于如何使用这个工具包， mllib user manual里面已经写的非常详细了。下面说一下其中一些重要参数的意义及如何设置。

tree_count：前面提到的决策树的个数，这个数设的越大学习就越充分，但太大也会造成过度拟合，而且也消耗训练和预测的时间。可以先选择比较大的树个数，然后观察训练过程中的损失减少趋势，损失减少比较平缓时，树个数就比较合适了。tree_count和shrinkage也有关系，shrinkage越大，学习越快，需要的树越少。

shrinkage：步长，它代表的是学习的速度，越小表示学习越保守（慢），而越大则表示学习越冒进（快）。通常我们可以把Shrinkage设小一点，把树的个数设大一点。

sample_rate：样本采样率，一次学习使用全部的样本是浪费，为了构造出具有不同倾向性的模型，需要使用样本的子集来进行训练，而过多的样本对简单的模型无益，只会造成更多的过拟合和局部极小问题。这个采样的比例一般选择50%-70%比较合适。

variable_sample_rate：特征采样率，和上面的样本采样率不同，这个采样率是指从样本的所有特征中选取部分的特征来学习，而不使用全部特征。当你发现训练出来的模型，某一两个特征非常强势，重要性很大，而造成其他特征基本学不到的时候，可以考虑设置一下把这个参数设置成<1的数。

最后是对于使用机器学习的一些常见问题的一些理解。

1 机器学习是万能的？

当然不是，如果你认为可以简单地把一堆特征样本扔给机器学习，就期望它给一个好的模型，那是不可能。在使用机器学习之前，一定要对所使用的模型有个基本的了解，最好能够知道它的计算原理。如果你都不知道treelink为何物，那些参数大概是个什么意思，你最好就不要用了，用也是在拼人品。

2 使用机器学习的同时做好数据分析工作

就算使用机器学习，数据的分析工作也是省不了的，做好了分析工作，你会发现自己都可以像机器一样找到那些规律。其实机器学习本身也就是统计，帮你找数据之间的规律，并把这些规律做成规则。好莱坞奥斯卡金像奖电影《美丽心灵》中的男主人翁，诺贝尔经济学奖得主数学家约翰纳什在电影中被称为最厉害的人肉密码破译者，充分向我们展现了how people learning beats machine learning，当然是有些夸张的成份。下面是一些可以做的最基本的数据分析工作：

1） 特征的分布：按特征的取值分段，每一段包含的样本数量，特征均值，方差。

2） 目标分布同上

3） 特征目标关系：特征分段，每段中包含的样本的目标取值。

4） 目标特征关系：目标分段，每段中包含的样本的特征取值

3 模型在训练数据上效果不错，但做Cross-validation效果不佳

主要原因有两个：

1） 选取的样本数据太少，覆盖度不够，考虑增加训练样本

2） 样本特征过多，可以考虑减少一些特征，只留下重要的特征

4 模型在类似Cross-validation这样的封闭测试上效果不错，但在开放测试上效果不佳

1） 选取的训练数据覆盖度不够，不具备代表性，不能体现真实数据的分布。

2） 模型迁移（Model drift），随着时间变化，特征数据也随之变化。比如3个月前做的模型对现在的特征可能不会有好的效果。

分享：

前一篇：机器学习ppt
后一篇：今日头条 app

〈 前一篇
机器学习ppt

后一篇 〉
今日头条 app