[置顶] 【机器学习PAI实践一】搭建心脏病预测案例

标签： 机器学习 数据挖掘 阿里云

2016-12-13 15:06 2720人阅读 评论(10) 收藏 举报

分类：

机器学习（31） DataMining（27）

目录(?) [+]

产品地址：https://data.aliyun.com/product/learn?spm=a21gt.99266.416540.102.OwEfx2

一、背景

心脏病是人类健康的头号杀手。全世界1 / 3的人口死亡是因心脏病引起的，而我国，每年有几十万人死于心脏病。 所以，如果可以通过提取人体相关的体侧指标，通过数据挖掘的方式来分析不同特征对于心脏病的影响，对于预测和预防心脏病将起到至关重要的作用。本文将会通过真实的数据，通过阿里云机器学习平台搭建心脏病预测案例。

二、数据集介绍

数据源： UCI开源数据集heart_disease
针对美国某区域的心脏病检查患者的体测数据，共303条数据。具体字段如下表：

字段名	含义	类型	描述
age	年龄	string	对象的年龄，数字表示
sex	性别	string	对象的性别，female和male
cp	胸部疼痛类型	string	痛感由重到无typical、atypical、non-anginal、asymptomatic
trestbps	血压	string	血压数值
chol	胆固醇	string	胆固醇数值
fbs	空腹血糖	string	血糖含量大于120mg/dl为true，否则为false
restecg	心电图结果	string	是否有T波，由轻到重为norm、hyp
thalach	最大心跳数	string	最大心跳数
exang	运动时是否心绞痛	string	是否有心绞痛，true为是，false为否

oldpeak	运动相对于休息的ST depression	string	st段压数值
slop	心电图ST segment的倾斜度	string	ST segment的slope，程度分为down、flat、up
ca	透视检查看到的血管数	string	透视检查看到的血管数
thal	缺陷种类	string	并发种类，由轻到重norm、fix、rev
status	是否患病	string	是否患病，buff是健康、sick是患病

三、数据探索流程

数据挖掘流程如下：

整体实验流程：

1.数据预处理

数据预处理也叫作数据清洗，主要在数据进入算法流程前对数据进行去噪、填充缺失值、类型

开源夏令营 (13)
python (16)
git (5)
面试 (5)
SQL (3)
Hadoop (1)
分布式计算 (3)
shell (1)
机器学习 (32)

变换等操作。本次实验的输入数据包括14个特征和1个目标队列。需要解决的场景是根据用户的体检指标预测是否会患有心脏病，每个样本只有患病或不患病两种，是分类问题。因为本次分类实验选用的是线性模型逻辑回归，要求输入的特征都是double型的数据。
输入数据展示：

数据探查 - heart_disease_prediction - (仅显示前一百条)

age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slop	ca	thal	status	style
63.0	male	ang...	145.0	233.0	true	hyp	150.0	fal	2.3	down	0.0	fix	buff	H
67.0	male	asy...	160.0	286.0	fal	hyp	108.0	true	1.5	flat	3.0	norm	sick	S2
67.0	male	asy...	120.0	229.0	fal	hyp	129.0	true	2.6	flat	2.0	rev	sick	S1
37.0	male	not...	130.0	250.0	fal	norm	187.0	fal	3.5	down	0.0	norm	buff	H
41.0	fem	abn...	130.0	204.0	fal	hyp	172.0	fal	1.4	up	0.0	norm	buff	H
56.0	male	abn...	120.0	236.0	fal	norm	178.0	fal	0.8	up	0.0	norm	buff	H
62.0	fem	asy...	140.0	268.0	fal	hyp	160.0	fal	3.6	down	2.0	norm	sick	S3
57.0	fem	asy...	120.0	354.0	fal	norm	163.0	true	0.6	up	0.0	norm	buff	H
63.0	male	asy...	130.0	254.0	fal	hyp	147.0	fal	1.4	flat	1.0	rev	sick	S2
53.0	male	asy...	140.0	203.0	true	hy	155.0	true	3.1	down	0.0	rev	sick	S1

我们看到有很多数据是文字描述的，在数据预处理的过程中我们需要根据每个字段的含义将字符型转为数值。
1）二值类的数据
二值类的比较容易转换，如sex字段有两种表现形式female和male，我们可以将female表示成0，把male表示成1。
2）多值类的数据
比如cp字段，表示胸部的疼痛感，我们可以通过疼痛的由轻到重映射成0~3的数值。

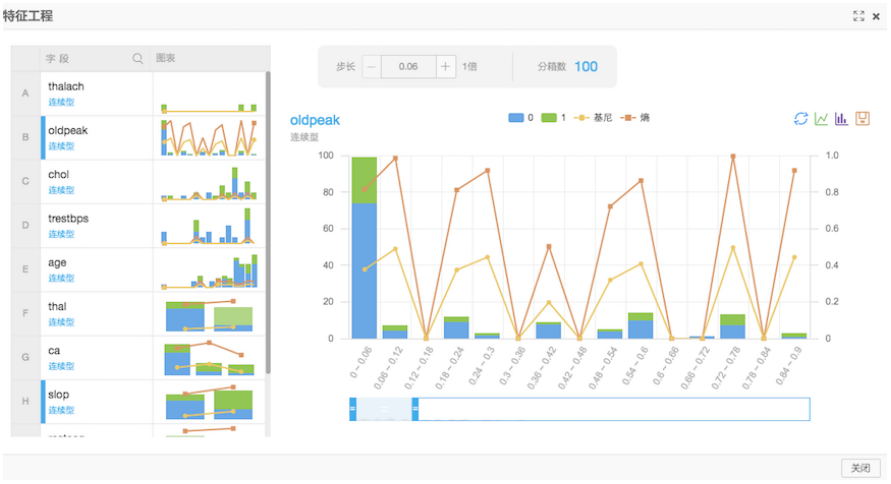
数据的预处理通过sql脚本来实现，具体请参考SQL脚本-1组件，

```
select age,
(case sex when 'male' then 1 else 0 end) as sex,
(case cp when 'angina' then 0  when 'notang' then 1 else 2 end) as cp,
trestbps,
chol,
(case fbs when 'true' then 1 else 0 end) as fbs,
(case restecg when 'norm' then 0  when 'abn' then 1 else 2 end) as restecg,
thalach,
(case exang when 'true' then 1 else 0 end) as exang,
oldpeak,
(case slop when 'up' then 0  when 'flat' then 1 else 2 end) as slop,
ca,
(case thal when 'norm' then 0  when 'fix' then 1 else 2 end) as thal,
(case status  when 'sick' then 1 else 0 end) as ifHealth
from  ${t1};
```

2.特征工程

特征工程主要是包括特征的衍生、尺度变化等。本例中有两个组件负责特征工程的部分。

1）过滤式特征选择
主要是通过这个组件判断每个特征对于结果的影响，通过信息熵和基尼系数来表示，可以通过查看评估报告来显示最终的结果。



2）归一化

因为本次实验选择的是通过逻辑回归二分类来进行模型训练，需要每个特征去除量纲的影响。归一化的作用是将每个特征的数值范围变为0到1之间。归一化的公式为result=(val-min)/(max-min)。

归一化结果：

数据探查 - pai_temp_2954_36756_1 - (仅显示前一百条)

sex	cp	fbs	restecg	exang	slop	thal	ifhealth	age	trestbps	chol	thalach	oldpeak
1	0	1	1	0	1	0.5	0	0.70...	0.4811320...	0.244...	0.603053...	0.370967...
1	1	0	1	1	0.5	0	1	0.79...	0.6226415...	0.365...	0.282442...	0.241935...
1	1	0	1	1	0.5	1	1	0.79...	0.2452830...	0.235...	0.442748...	0.419354...
1	0.5	0	0	0	1	0	0	0.16...	0.3396226...	0.283...	0.885496...	0.564516...
0	1	0	1	0	0	0	0	0.25	0.3396226...	0.178...	0.770992...	0.225806...
1	1	0	0	0	0	0	0	0.5625	0.2452830...	0.251...	0.816793...	0.129032...
0	1	0	1	0	1	0	1	0.6875	0.4339622...	0.324...	0.679389...	0.580645...
0	1	0	0	1	0	0	0	0.58...	0.2452830...	0.520...	0.702290...	0.096774...
1	1	0	1	0	0.5	1	1	0.70...	0.3396226...	0.292...	0.580152...	0.225806...
1	1	1	1	1	1	1	1	0.5	0.4339622...	0.175...	0.641221...	0.5
1	1	0	0	0	0.5	0	0	0.58...	0.4339622...	0.150...	0.587786...	0.064516...

3.模型训练和预测

本次实验是监督学习，因为我们已经知道每个样本是否患有心脏病，所谓监督学习就是已知结果来训练模型。解决的问题是预测一组用户是否患有心脏病。

1）拆分

首先通过拆分组件将数据分为两部分，本次实验按照训练集和预测集7：3的比例拆分。训练集数据流入逻辑回归二分类组件用来训练模型，预测集数据进入预测组件。

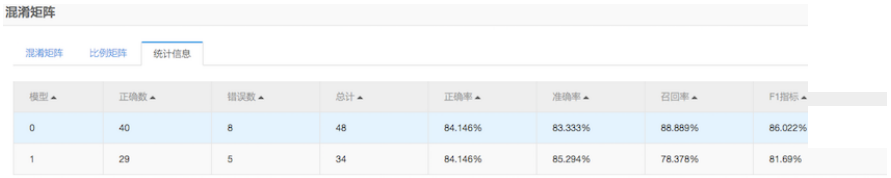
2）逻辑回归二分类

逻辑回归是一个线性模型，在这里通过计算结果的阈值实现分类。具体的算法详情推荐大家在网上或者书籍中自行了解。逻辑回归训练好的模型可以在模型页签中查看。

3)预测
预测组件的两个输入分别是模型和预测集。预测结果展示的是预测数据、真实数据、每组数据不同结果的概率。

4.评估

通过混淆矩阵组件可以评估模型的准确率等参数，



通过此组件可以方便的通过预测的准确性来评估模型。

四.总结

通过以上数据探索的流程我们可以得到以下的结论。

1）特征权重
我们可以通过过滤式特征选择得到每个特征对于结果的权重。

featname	weight
thalach	0.16569171224597157
oldpeak	0.14640697618779352
thal	0.13769166559906015
ca	0.11467097546217575
chol	0.10267709576600859
age	0.07876430484527841
trestbps	0.0772599125640569
slop	0.07702762609078306
restecg	0.015246832497405105
cp	0.0037507283721422424
exang	0
fbs	0
sex	0

-可以看出thalach(心跳数)对于是否发生心脏病影响最大。
-性别对于心脏病没有影响
2）模型效果
通过上文提供的14个特征，可以达到百分之八十多的心脏病预测准确率。模型可以用来做预测，辅助医生预防和治疗心脏病。

五、其它

免费体验：阿里云数加机器学习平台

- 上一篇 Google Java编程风格指南
- 下一篇 【机器学习PAI实践二】人口普查统计

相关文章推荐

- hive表信息查询：查看表结构、表操作等（转）
- MySQL在微信支付下的高可用运营--莫晓东
- 转：Confusion Matrix(混淆矩阵) 解释最全的一个
- 容器技术在58同城的实践--姚远
- 【机器学习PAI实践六】金融贷款发放预测
- SDCC 2017之容器技术实战线上峰会
- 【机器学习PAI实践十】深度学习Caffe框架实现图...
- SDCC 2017之数据库技术实战线上峰会
- 【机器学习PAI实践二】人口普查统计
- 腾讯云容器服务架构实现介绍--董晓杰
- 【机器学习PAI实践四】如何实现金融风控
- 微博热点事件背后的数据库运维心得--张冬洪
- 【机器学习PAI实践十二】机器学习算法基于信用卡..
- 【机器学习PAI实践十二】机器学习实现双十一购物..
- 【机器学习PAI实践十二】机器学习实现双十一购物..
- 【机器学习PAI实践五】机器学习眼中的《人民的名..

查看评论

lyuchen65 4楼 4天前 21:24发表

那个数据集是在data folder里面吗？不知道要在哪里下载数据？

河南家里蹲大学 3楼 2017-04-11 11:33发表

大神，你给的群号是不是错了啊？

李博Garvin Re: 2017-04-11 14:53发表

回复河南家里蹲大学：钉钉群，不是QQ

河南家里蹲大学 Re: 2017-04-12 09:01发表

回复李博Garvin：哈哈 没用过这个 下载一个试玩

河南家里蹲大学 Re: 2017-04-12 09:21发表

回复河南家里蹲大学：博主 试玩一下 没有找到如何加入群 有没有别的联系方式联系到你呢？

河南家里蹲大学 2楼 2017-04-09 10:09发表

大神 你好 我最近正在体验阿里的机器学习平台， 能否加个联系方式：QQ：593956670

李博Garvin Re: 2017-04-10 13:25发表

回复河南家里蹲大学：请加入我们的钉钉群11768691，或者在我们的贴吧交流：https://yq.aliyun.com/teams/47/type_ask?spm=5176.100239.0.0.vo2cLY

Kevin_zhai 1楼 2016-12-16 22:14发表

大神又开始写博客啦

李博Garvin Re: 2016-12-18 14:30发表

回复Kevin_zhai：哈哈，给自家产品宣传宣传
