Re: 回复: 11月7日 结果

"xiazhenyu@weiresearch.com" <xiazhenyu@weiresearch.com>

收件人:"段翠婷" <duancuiting@126.com>

抄　送: "周像金" <zhouxiangjin@weiresearch.com>

时　间: 2014-11-12 3:08:51

附　件:

我加你QQ了，你微信是？如果不给你添麻烦的话，我希望加上你最直接的联系方式。尽量不影响你的生活圈子，哈哈哈。

因为咱们有很多算法和程序转换的地方。例如：做decision tree时，不清楚是ID3，CART，C4.5那个更合适现在的数据。咱们要实验
咱们需要具有
1. 数据处理能力，可能是SQL+Java能力。要能至少脚本编程。比如现在DB的原始数据，可能还不能完成建模。
2. 算法比较能力。可能咱们选取某个算法，进行数据分析，比较效果。比如：几种算法的比较。仅仅是距离就有多种算法。欧氏距离，曼哈顿距离，契比雪夫距离。咱们得选择效果最好的。
3. 算法确定后，咱们要编写程序改进算法。因为咱们分析数据量不小。最后一定用程序实现整个过程。
　例如：咱们建模用K-means，计算很慢，最后需要想法把它变成hadoop下的算法，或者直接用R的分布式lib(这个我还不会)

我觉得工作挑战可能是目前国内首屈一指，但是也需要你能跟上。尤其是训练一个月左右，需要你跟上咱们中科院预测研究所的同事们。共勉吧，希望你能有所收获，微瑞也能培养出改变中国的一群新人。
如果18年前，互联网黄金一代开始酝酿一样，现在是大数据黄金一代酝酿期。

Regards,

Ryan

发件人: 段翠婷
发送时间: 2014-11-11 23:34
收件人: 夏振宇
主题: 回复: Re: 答复: 11月7日 结果

1054083631 其实我用微信更多 可以加我微信～

段翠婷

在2014年11月11日 11:54，夏振宇写道：

> 你的QQ是？方便沟通。
> 微瑞原来用的RapidMiner，其实就是weka。

现在可能不得不用R来做。这次优酷你慢慢会和中科院预测研究所的伙伴一起完成。

---------- Original ----------

From: "段翠婷"<duancuiting@126.com>;
Date: Tue, Nov 11, 2014 11:51 AM
To: "xiazhenyu@weiresearch.com"<xiazhenyu@weiresearch.com>;
Subject: Re:Re: 答复: 11月7日 结果

领导好:

我再学习下～谢谢～

祝好
段翠婷

在 2014-11-09 12:05:33, "xiazhenyu@weiresearch.com" <xiazhenyu@weiresearch.com> 写道:

翠婷, 你好

我叫夏振宇, 咱们见过面。没来得及多聊两句, 只是我上飞机前, 跟你电话说过一些情况。非常高兴能跟你合作, 咱们完成当前的项目。
为了便于交流, 总结经验
1. 我给你开了邮件箱 duancuiting@weiresearch.com , 密码: welcome1 是腾讯的企业邮箱 http://exmail.qq.com/
2. 并且用这个企业邮箱注册了咱们研发使用的wiki
http://www.web30core.org/trac/web30core/wiki/MoviePredicationDB
用户名:duancuiting, 密码: welcome1, 首次登录用duancuiting@weiresearch.com激活获得权限

>我觉得可以按这个逻辑来做, 但是好像ID3和C4.5都是通过计算信息增益来自动的添加比较重要的维度的, 所以我不太明白为何要"**咱们讨论一下, 第一次选取那些纬度吧**"。

用ID3, 咱们要确定输入那些数据, 用户观看的电影类型, 用户地域, 用户支付方式?那些属性作为输入, 就是提的纬度。
例如:wiki描述 http://www.web30core.org/trac/web30core/wiki/MoviePredicationDB。

> 还有一点, 我不太明白邮件里, 分区统计的意思。原来的数据里总共有4个月18周的数据, 每周都有购买记录的用户只有一个, 每月都有购买记录的应该不少, 还没有分析。

原来我认为当前数据是随机给出，首先统计不同购买区间的人多寡。根据统计获得区间。例如：每周付费的区间是否合理？还是两周付费的区间更合理？
如果每周都购买的只有一个，咱们得考虑
1. 是不是这样的用户样本不足？是不是需要补充。怎么补充才能保证抽样不偏差太大？用年龄，用地域，还是随机在抽取 2万个付费用户，从中看区间？

> 是不是说比如考察每月都有购买记录的用户，用决策树算法，取出主要决定因素，然后再用这些主要因素分别做一下聚类和回归？

我原来考虑，是不是分了区间后，再用决策时树分析重点因素。然后再对比不分区间得出的形成购买的重点因素。例如：有冒险电影是每周付费的决定因素？
找到决定因素，用决定因素对用户聚类。然后回归。是这样

Regards,

Ryan

发件人： 段翠婷
发送时间： 2014-11-08 15:02
收件人： xiazhenyu@weiresearch.com; '周像金'
主题： 答复: 11月7日 结果

领导好，

不好意思这么晚才给您回邮件。我是段翠婷，SPSS，SAS和R，还有Clementine，我都会用一点。

我觉得可以按这个逻辑来做，但是好像ID3和C4.5都是通过计算信息增益来自动的添加比较重要的维度的，所以我不太明白为何要“**咱们讨论一下，第一次选取那些纬度吧**”。

第2是为了对所有的影响因素降维，然后选取主要因素，第3是聚类，第4是回归。

还有一点，我不太明白邮件里，分区统计的意思。原来的数据里总共有4个月18周的数据，每周都有购买记录的用户只有一个，每月都有购买记录的应该不少，还没有分析。

是不是说比如考察每月都有购买记录的用户，用决策树算法，取出主要决定因素，然后再用这些主要因素分别做一下聚类和回归？

祝好

段翠婷

发件人: xiazhenyu@weiresearch.com [mailto:xiazhenyu@weiresearch.com]

发送时间: 2014年11月8日 0:04

收件人: 月牙湾; 周像金

主题: Re: 11月7日 结果

翠婷, 你好

我是夏振宇, 看了你统计的信息, 你看咱们用以下逻辑开展下边的工作进行如何?我也想了解一下你当前使用的分析工具什么。SPSS, SAS还是啥

1. 根据用户观看频度进行分区统计。例如:每周观看的用户, 每月观看。咱们得看那个区间的用户多。
2. 根据区间分析, 使用ID3, C4.5我不知道那个效果好, 分析属于这个区间的决定因素, 把影响因素减少。例如:分析每周观看这个区间的记录, 能否得出那类影片, 那个地区, 那种支付是决定因素

可以先取电影类型, 支付类型, 观众地区, **咱们讨论一下, 第一次选取那些纬度吧**

电影类型参照, 这是系统内标准分类, 可以建立一个32 dimension的vector, 电影有这个类型就是1。例如:{1,1,0,0,1,0,0,,,,,} 表示类型 喜剧/动作/动画

喜剧 comedy

动作 action

剧情 drama

悬疑 suspense

动画 carton

科幻 fiction

纪录片 record

爱情 romance

家庭 family

儿童 child

历史 history

惊悚 horror

短片 short

奇幻 fantasy

歌舞/音乐 musical

武侠 Swordplay

恐怖/惊栗 horror

战争 war

西部 western

古装 ancient_costume

犯罪 criminal

运动 sport

冒险 adventure

舞台 stage

戏曲 chinese_opera

传记 biography

情色 adult

成人 adult

真人秀 trueman show

同性 lesbian_gay

鬼怪 ghost

黑色 noir

也可以用你现在聚合的大类别

3. 确定相对少的影响因素后，例如：类型，支付方式是决定因素。再用kmeans根据这些纬度进行聚类。按照区间分成，一周看的，几个月看的。

4. 最后根据逻辑回归，给出一个用户成为每周看片用户的可能性是多大，每月是多大

Regards,

Ryan

发件人： 月牙湾

发送时间： 2014-11-07 18:49

收件人： 周像金

抄送： 夏振宇

主题： 11月7日 结果

统计工作如下：

这两天的任务主要集中在数据统计分析阶段，主要在当前170万条记录的基础上，

完成以下内容的

√基础统计工作:交易日期统计,节目类型统计,交易来源统计,支付手段统计,支付设备统计,用户地域统计

会员统计工作:

•√ 沉睡会员统计:已是会员,近6个月无观看行为(享受会员服务)会员用户。

•× 沉睡用户统计:曾经是会员,但目前过期状态,近6个月无购买会员。(没完成原因:需要其他数据来做)

•√ 活跃会员统计:每周有1以上次观看优酷内容。

---------- 原始邮件 ----------

发件人: "周像金";<zhouxiangjin@weiresearch.com>;

发送时间: 2014年11月6日(星期四) 上午10:11

收件人: "段翠婷"<1054083631@qq.com>;

抄送: "夏振宇"<xiazhenyu@weiresearch.com>;

主题: 会员购买记录分析任务

翠婷,

目前已经导入到数据表中的数据有约170多万条,附件Excel文件是数据库中的一个样例解释,你可以根据这个来看具体每个字段的含义。

这两天的任务主要集中在数据统计分析阶段,主要在当前170万条记录的基础上,完成以下内容的

基础统计工作:交易日期统计,交易来源统计,节目类型统计,支付手段统计,支付设备统计,用户地域统计

会员统计工作:

沉睡会员统计:已是会员,近6个月无观看行为(享受会员服务)会员用户。

沉睡用户统计:曾经是会员,但目前过期状态,近6个月无购买会员。

活跃会员统计:每周有1以上次观看优酷内容。

你有疑问的,随时过来和我沟通。

周像金