法律声明

- 本课件包括：演示文稿，示例，代码，题库，视频和声音等，小象学院拥有完全知识产权的权利；只限于善意学习者在本课程使用，不得在课程范围外向任何第三方散播。任何其他人或机构不得盗版、复制、仿造其中的创意，我们将保留一切通过法律手段追究违反者的权利。

- 课程详情请咨询
  - 微信公众号：大数据分析挖掘
  - 新浪微博：ChinaHadoop

申请评分卡中的数据预处理和特征衍生

目录

构建信用风险类型的特征

已加工成型的信息

Master　　idx：每一笔贷款的unique key，可以与另外2个文件里的idx相匹配。

UserInfo_*：借款人特征字段

WeblogInfo_*：Info网络行为字段

Education_Info*：学历学籍字段

ThirdParty_Info_PeriodN_*：第三方数据时间段N字段

SocialNetwork_*：社交网络字段

ListingInfo：借款成交时间

Target：违约标签(1 = 贷款违约，0 = 正常还款)

构建信用风险类型的特征

需要衍生的信息

借款人的登陆信息

ListingInfo：借款成交时间

LogInfo1：操作代码

LogInfo2：操作类别

LogInfo3：登陆时间

idx：每一笔贷款的unique key

Idx	Listinginfo1	LogInfo1	LogInfo2	LogInfo3
3	2013/11/5	4	1	2013/8/30
3	2013/11/5	-4	6	2013/8/31
3	2013/11/5	-4	6	2013/9/3
3	2013/11/5	-4	6	2013/9/4
3	2013/11/5	-4	6	2013/10/23
3	2013/11/5	1	2	2013/10/23

- 有多个操作日期
- 每个日期有多个操作
- 有多种操作

构建信用风险类型的特征

需要衍生的信息(续)

时间切片：

两个时刻间的跨度

例： 申请日期之前30天内的登录次数

申请日期之前第30天至第59天内的登录次数

基于时间切片的衍生

- 申请日期之前180天内，平均每月(30天)的登录次数

常用的时间切片

- (1、2个)月，(1、2个)季度，半年，1年，1年半，2年

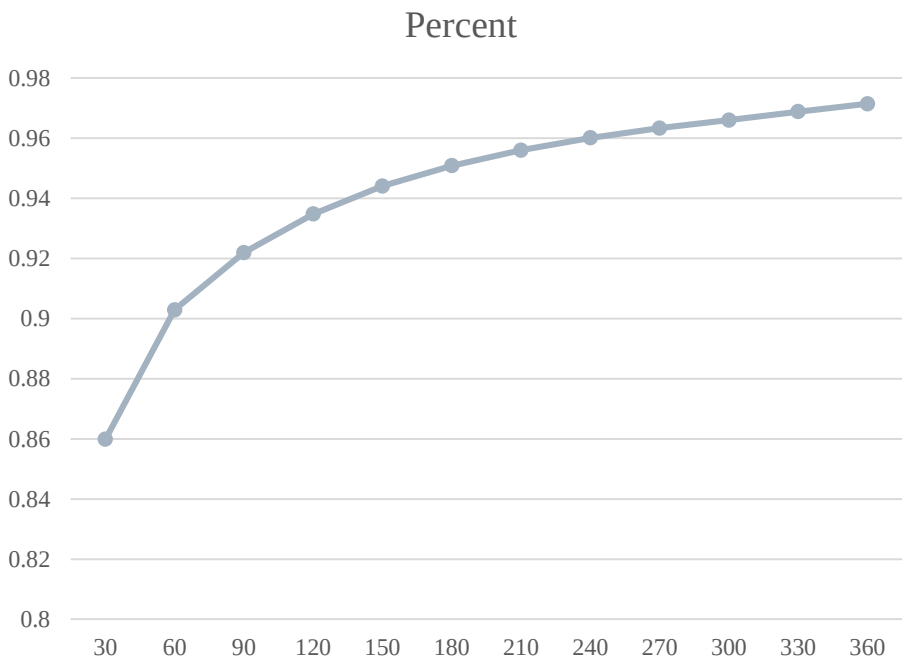
时间切片的选择

- 不能太长：保证大多数样本都能覆盖到
- 不能太短：丢失信息

构建信用风险类型的特征

需要衍生的信息(续)

借款人的登陆信息中的时间



- 180天的时间切片能覆盖到95%的事件
- 选取[30,60,90,120,150,180]做为不同的切片，衍生变量

构建信用风险类型的特征

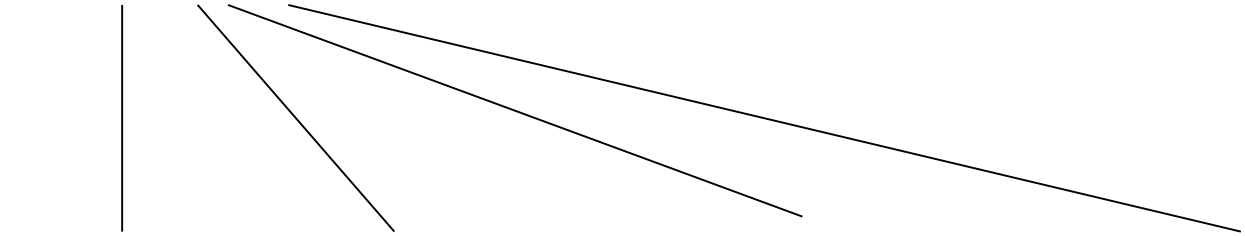
需要衍生的信息(续)

在同一个时间切片内，可以衍生的特征

(注意到LogInfo1、LogInfo2是类别和代码，不能进行数值运算)

- 操作的次数
- 不同类别／代码的个数
- 同一类别／代码的平均操作次数

共计6*(1+2+2)=30个变量



6个时间切片

同一个时间切片内的操作次数

LogInfo1、LogInfo2的不同值的个数

LogInfo1、LogInfo2的不同值的平均操作次数

构建信用风险类型的特征

需要衍生的信息(续)

对Userupdate_Info的变量衍生

- 时间切片的选取方式如前所述
- 特别地，需要做数据预处理
  - 统一大小写
  - 统一Phone, Mobilephone
- 需要关注几个特殊的变量
  - 是否修改IDNumber
  - 是否修改Mobilephone
  - 是否修改*HASBUYCAR*
  - 是否修改*MARRIAGESTATUSID*

构建信用风险类型的特征

需要衍生的信息(续)

数据清洗

- 对于类别型变量
  - 删除缺失率超过50%的变量
  - 剩余变量中的缺失做为一种状态
- 对于连续型变量
  - 删除缺失率超过70%的变量
  - 利用随机抽样法对剩余变量中的缺失进行补缺

注：连续变量中的缺失也可以当成一种状态

目录

特征的分箱

特征的分箱

分箱的定义

- 将连续变量离散化
- 将多状态的离散变量合并成少状态

分箱的重要性

- 稳定性：避免特征中无意义的波动对评分带来的波动
- 健壮性：避免了极端值的影响

分箱的优势

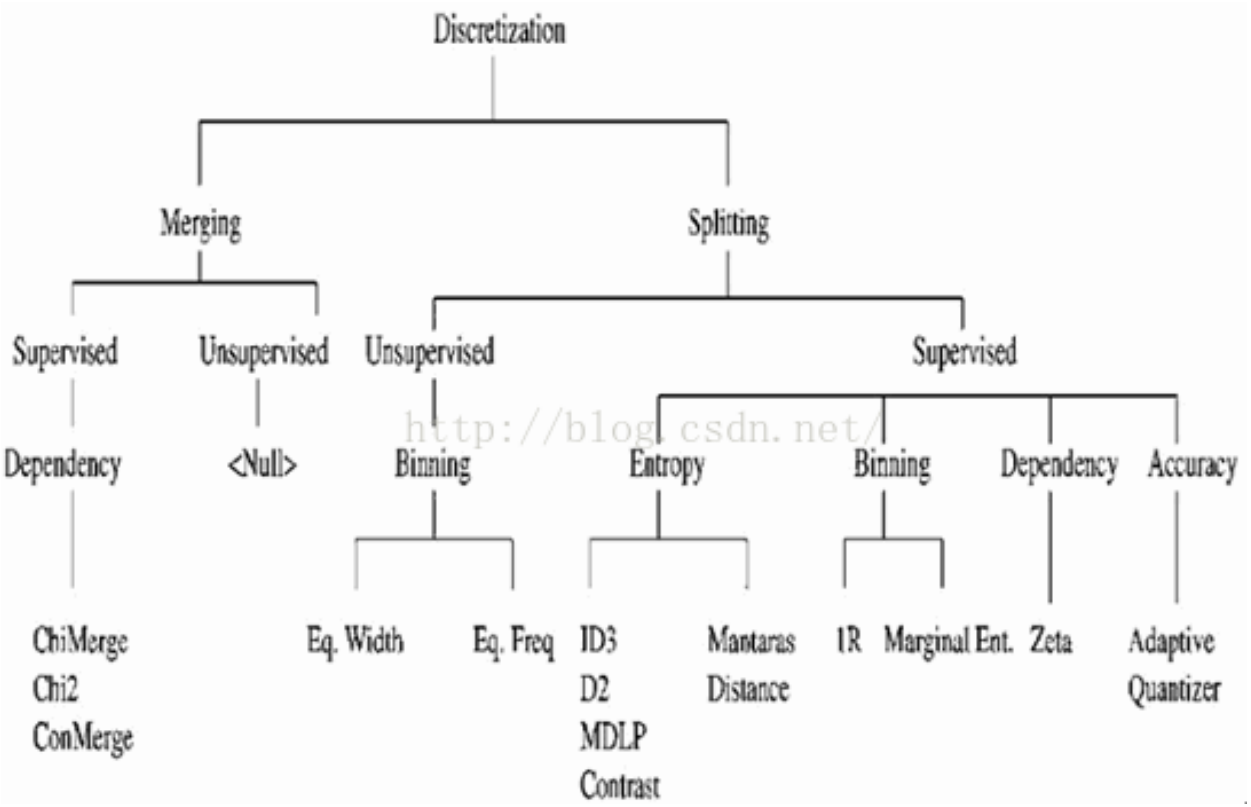
- 可以将缺失作为独立的一个箱带入模型中
- 将所有变量变换到相似的尺度上

分箱的限制

- 计算量大　　分箱后需要编码

特征的分箱

- 分箱的方法



常用的方法

有监督

- Best – KS
- ChiMerge

无监督

- 等频
- 等距
- 聚类

特征的分箱

分箱的方法(续)

监督式分箱法：　　Best－KS

原理：让分箱后组别的分布的差异最大化

- 对于连续变量

1，排序，x＝ $\{x_1, x_2, \ldots, x_k\}$

2，计算每一点的KS值

3，选取最大的KS对应的特征值x_m，将x分为$\{x_i \leq x_m\}$与$\{x_i > x_m\}$两部

对于每一部分，重复2-3，直到满足终止条件之一

- 终止条件

1，下一步分箱后，最小的箱的占比低于设定的阈值(常用0.05)

2，下一步分箱后，该箱对应的y类别全部为0或者1

3，下一步分箱后，bad rate不单调

- 对于离散度很高的变量

1，编码

2，依据连续变量的方式进行分箱

特征的分箱

卡方分箱法(ChiMerge)

监督式分箱法：　　卡方分箱法(ChiMerge)

自底向上的(即基于合并的)数据离散化方法。它依赖于卡方检验：具有最小卡方值的相邻区间合并在一起，直到满足确定的停止准则。

基本思想：对于精确的离散化，相对类频率在一个区间内应当完全一致。因此，如果两个相邻的区间具有非常类似的类分布，则这两个区间可以合并；否则，它们应当保持分开。而低卡方值表明它们具有相似的类分布。

特征的分箱

卡方分箱法(ChiMerge)

第零步：预先设定一个卡方的阈值

第一步：初始化

根据要离散的属性对实例进行排序：每个实例属于一个区间

第二步：合并区间：

(1) 计算每一对相邻区间的卡方值

(2) 将卡方值最小的一对区间合并

$$X^2 = \sum_{i=1}^{2}\sum_{j=1}^{2}\frac{(A_{ij}-E_{ij})^2}{E_{ij}}$$

A_{ij}:第i区间第j类的实例的数量

E_{ij}: A_{ij}的期望频率, $=\frac{N_i \times C_j}{N}$, N是总样本数，$N_i$是第i组的样本数，$C_j$是第j类样本在全体中的比例

特征的分箱

卡方分箱法(ChiMerge)

卡方阈值的确定

根据显著性水平和自由度得到卡方值

自由度比类别数量小1。例如，有3类，自由度为2，则90%置信度（10%显著性水平)下，卡方的值为4.6。

阈值的意义

类别和属性独立时，有90%的可能性，计算得到的卡方值会小于4.6，这样，大于阈值的卡方值就说明属性和类不是相互独立的，不能合并。如果阈值选的大，区间合并就会进行很多次，离散后的区间数量少、区间大。

注：

1，ChiMerge算法推荐使用0.90、0.95、0.99置信度，最大区间数取10到15之间.

2，也可以不考虑卡方阈值，此时可以考虑最小区间数或者最大区间数。指定区间数量的上限和下限，最多几个区间，最少几个区间。

3，对于类别型变量，需要分箱时需要按照某种方式进行排序

特征的分箱

分箱的方法(续)

无监督分箱法：　　等距划分、等频划分

等距分箱

从最小值到最大值之间，均分为 N 等份， 这样， 如果 A,B 为最小最大值， 则每个区间的长度为 W=(B−A)/N , 则区间边界值为 A+W,A+2W,….A+(N−1)W .

等频分箱

区间的边界值要经过选择，使得每个区间包含大致相等的实例数量。比如说 N=10 ，每个区间应该包含大约10%的实例。

以上两种算法的弊端

比如，等宽区间划分，划分为5区间，最高工资为50000，则所有工资低于10000的人都被划分到同一区间。等频区间可能正好相反，所有工资高于50000的人都会被划分到50000这一区间中。这两种算法都忽略了实例所属的类型，落在正确区间里的偶然性很大。

目录

构建信用风险类型的特征

特征的分箱

WOE编码

WOE编码

WOE编码

WOE(weight of evidence, 证据权重)

一种有监督的编码方式，将预测类别的集中度的属性作为编码的数值

优势

- 将特征的值规范到相近的尺度上

(经验上讲，WOE的绝对值波动范围在0.1～3之间)

- 具有业务含义

缺点

- 需要每箱中同时包含好、坏两个类别

WOE编码

WOE编码(续)

WOE计算公式

	Good	Bad	Good Percent	Bad Percent
Group 1	G_1	B_1	G_1/G_{total}	B_1/B_{total}
Group 2	G_2	B_2	G_2/G_{total}	B_2/B_{total}
…	…	…	…	…
Group N	G_N	B_N	G_N/G_{total}	B_N/B_{total}
Total	$G_{total} = \sum G_i$	$B_{total} = \sum B_i$		

$$WOE = \sum \log(\frac{G_i/G_{total}}{B_i/B_{total}})$$

WOE编码

WOE编码(续)

WOE编码的意义

- 符号与好样本比例相关
- 要求回归模型的系数为负

疑问

- 问题答疑： http://www.xxwenda.com/
  - 可邀请老师或者其他人回答问题

联系我们

小象学院：互联网新技术在线教育领航者

- 微信公众号：大数据分析挖掘
- 新浪微博：ChinaHadoop