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：小象学院
  - 新浪微博：小象AI学院

1

申请评分卡中的数据预处理和特征衍生

目录

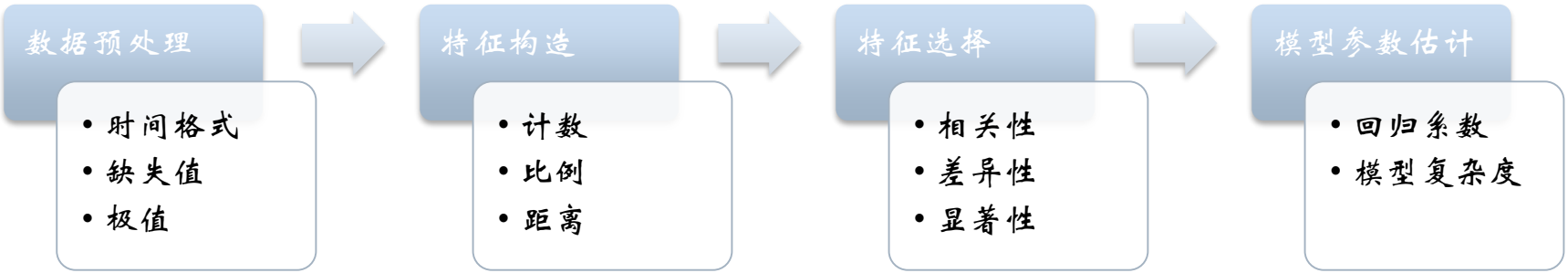
构建信用风险类型的特征

特征的分箱

WOE编码

构建信用风险类型的特征

数据预处理



构建信用风险类型的特征

数据格式的处理

原始数据带有一定的格式，需要转换成正确的格式。

例如：

- 利率

带%的百分比，需要转化成浮点数

- 日期

Nov – 17，需要转化为python的时间

- 工作年限

“<1 year”转化成0，“> 10 years”转化成11

构建信用风险类型的特征

数据格式的处理(续)

文本类的数据的处理方式

After amassing credit card debt through several years of college, I now have spending under control and a stable job. With this loan I plan to pay off all the credit cards and close them down immediately. I have a good balance between living expenses and debt repayment. Over several years in college I amassed a large debt in credit cards. Now that I have a stable job and spending under control I will use this loan to consolidate credit card debt and to close them immediately. I never miss payments and I follow a strict monthly budget of $1,350 in living expenses and a $339 car note. I have enough in savings to cover 6 months of all expenses, including this loan payment. Absolutely every other dollar is dedicated to debt repayment. Please let me know if you have any other questions. Borrower added on 12/15/11 > …

- 主题提取(NPL)

优点：提取准确、详细的信息，对风险的评估非常有效

缺点：NPL的模型较为复杂，且需要足够多的训练样本

- 编码

有点：简单

缺点：信息丢失很高

构建信用风险类型的特征

缺失值

缺失在数据分析的工作是频繁出现的。

- 缺失的种类
  - 完全随机缺失
  - 随机缺失
  - 完全非随机缺失
- 处理的方法
  - 补缺
  - 作为一种状态

构建信用风险类型的特征

- 构建特征

常用的特征衍生

- ✓ 计数：过去1年内申请贷款的总次数
- ✓ 求和：过去1年内的网店消费总额
- ✓ 比例：贷款申请额度与年收入的占比
- ✓ 时间差：第一次开户距今时长
- ✓ 波动率：过去3年内每份工作的时间的标准差

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

特征的分箱

特征的分箱(续)

分箱的优势

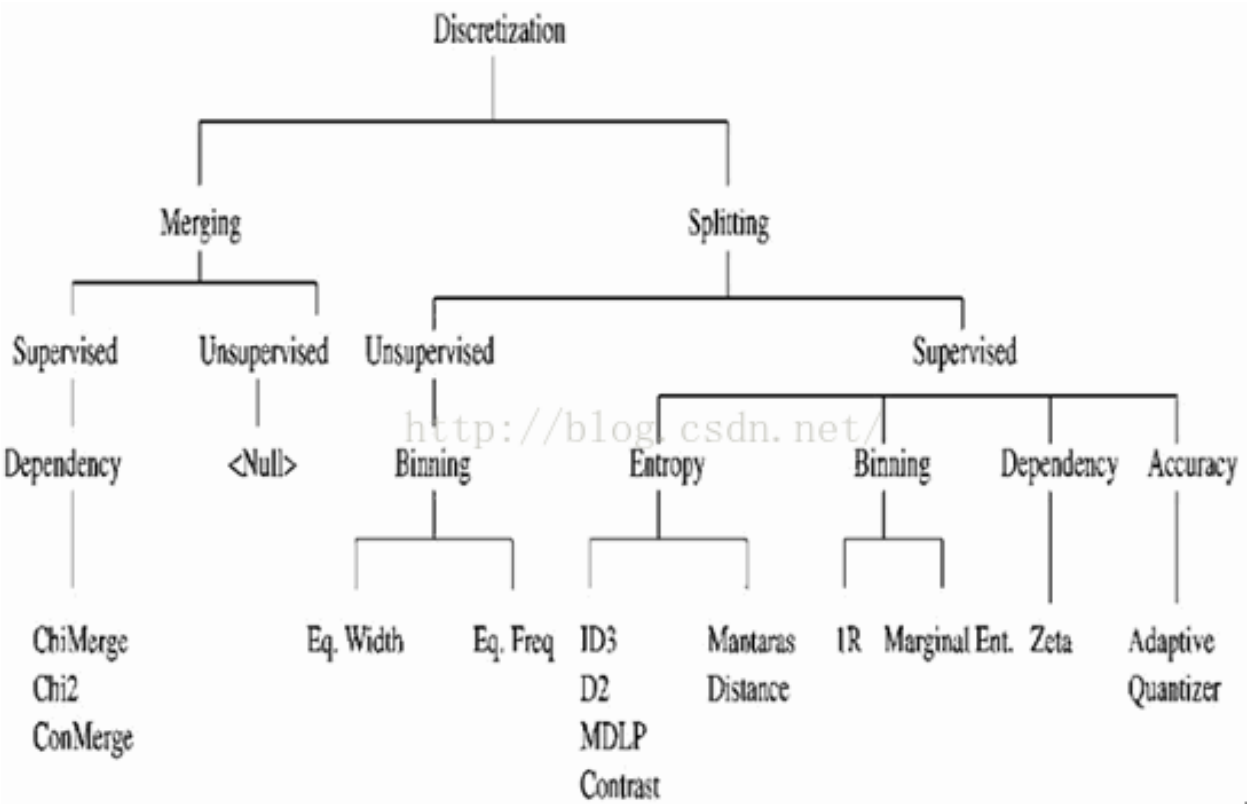
- 可以将缺失作为独立的一个箱带入模型中
- 将所有变量变换到相似的尺度上

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。

和Best－KS相比，ChiMerge可以应用在multi-Class的情形下。

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。

特征的分箱

卡方分箱法(续)

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

分箱的注意点

对于连续型变量，

- 使用ChiMerge进行分箱(默认分成5个箱)
- 检查分箱后的bad rate单调性；倘若不满足，需要进行相邻两箱的合并，直到bad rate为止
- 上述过程是收敛的，因为当箱数为2时，bad rate自然单调
- 分箱必须覆盖所有训练样本外可能存在的值！
- 原始值很多时，为了减小时间的开销，通常选取较少(例如50个)初始切分点。但是要注意分布不均匀！

特征的分箱

分箱的注意点(续)

对于类别型变量，

- 当类别数较少时，原则上不需要分箱
- 当某个或者几个类别的bad rate为0时，需要和最小的非0的bad rate的箱进行合并
- 当该变量可以完全区分目标变量时，需要认真检查该变量的合理性

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。

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(续)

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

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	$G_{total} = \sum G_i$	$B_{total} = \sum B_i$		

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

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

- 小象问答官网
  - http://wenda.chinahadoop.cn

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：小象学院
- 新浪微博：小象AI学院