信用评分

李志勇\张兴敏

西南财经大学

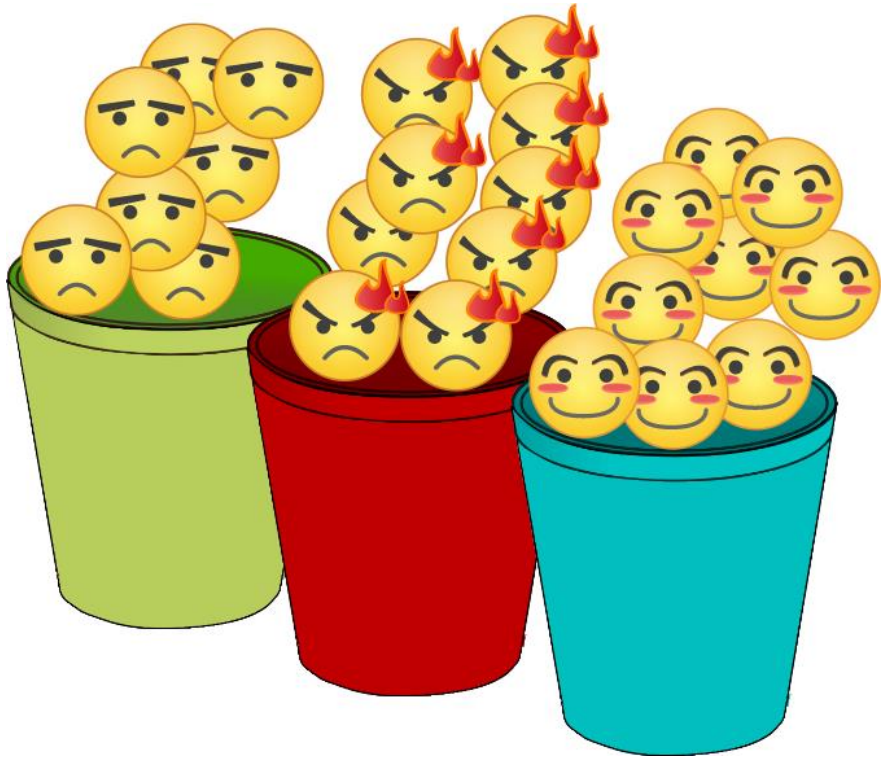
Southwestern University of Finance and Economics

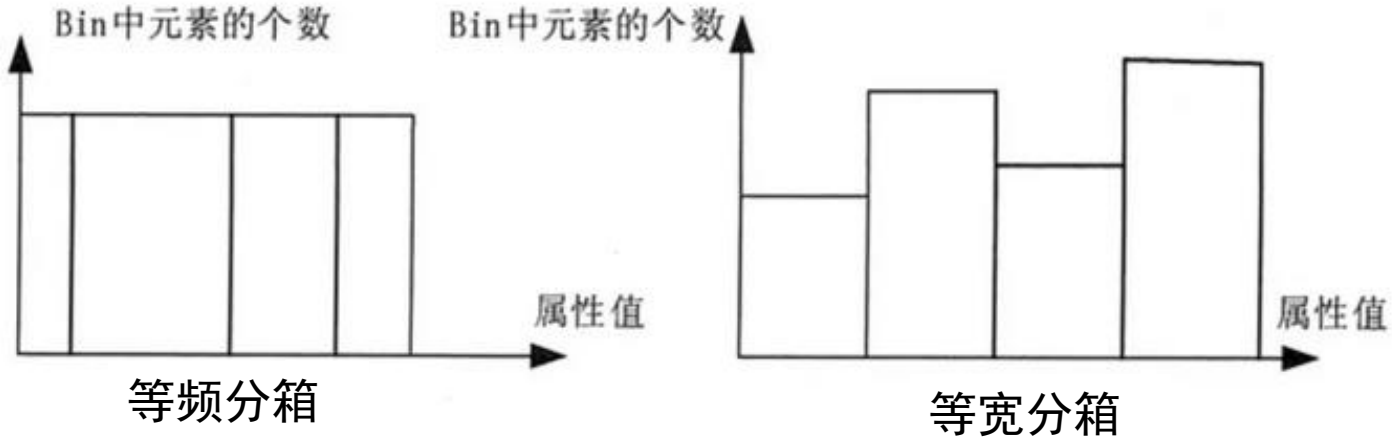
数据处理与特征工程的Python实现

- 数据分类-分箱
- 数据分类-粗分类
- 数据转换-哑变量
- 数据转换-证据权重
- 数据转换-信息值

数据分类-分箱

- 分箱是一种将数据排序并分组的方法，分为等宽分箱和等频分箱。
- 所谓**等宽分箱**，是用同等大小的格子来将数据范围分成N个间隔。等宽分箱比较直观和容易操作，但是对于偏态分布的数据，等宽分箱并不是太好，因为可能出现许多箱中没有样本点的情况。
- 所谓**等频分箱**是将数据分成N个间隔，每个间隔包含大致相同的数据样本个数，这种分箱方法有着比较好的可扩展性。
- 将数据分箱后，可以用箱均值、箱中位数和箱边界来对数据进行平滑，平滑可以在一定程度上削弱离群点对数据的影响。





等频分箱　　等宽分箱

数据处理和变换

- 虚拟变量设置

序号	学历	X1(本科)	X2（硕士）	X3（博士）
1	本科	1	0	0
2	硕士	0	1	0
3	本科	1	0	0
4	本科	1	0	0
5	高中	0	0	0
6	硕士	0	1	0
7	博士	0	0	1

4类

3个虚拟变量

数据处理和变换

风险变量——**证据权重**weights of evidence

- 证据权重（WOE）用在某一分类特征变量X的每个类别上**x**

$$w(\mathbf{x}) = \ln(I(\mathbf{x})) = \ln\left(\frac{p(G\mid\mathbf{x})/p(B\mid\mathbf{x})}{p_G/p_B}\right) = \ln\left(\frac{g_i/b_i}{n_G/n_B}\right) = \ln\left(\frac{g_i n_B}{b_i n_G}\right)$$

- 其中，

  b_i 和 g_i 是该特征落在某类别 i 中坏人和好人的数量

  n_B 和 n_G 是坏人和好人的总数量

- 证据权重将某个分类特征转换成数值量化的变量，可以直接放入模型

数据处理和变换

信息值Information Value（IV）

- FICO公司采用信息值来度量某个特征的预测能力，信息值IV又叫Kullback散度。

$$\mathrm{IV}=\sum_{i=1}^{n}\left[\left(\frac{g_i}{n_G}-\frac{b_i}{n_B}\right)\times \mathrm{WoE}_i\right]$$

- 某特征的IV值小于0.1效果不太好，大于0.5存疑。

收入	好人数量	坏人数量	好坏比率	好人比例	坏人比例	证据权重
低	5000	2000	2.5	14.3	33.3	−0.847
中等	10000	2000	5.0	28.6	33.3	−0.154
高	20000	2000	10.0	57.1	33.3	0.539
合计	35000	6000	5.8	100.0	100.0	信息值

多个特征

- 如果有两个特征，需要一个三维的列联表：

$$O(\mathrm{G}\mid x_1,x_2)=\frac{P(\mathrm{G}\mid x_1,x_2)}{P(\mathrm{B}\mid x_1,x_2)}=\frac{p_{\mathrm{G}}P(x_1,x_2\mid G)}{p_{\mathrm{B}}P(x_1,x_2\mid B)}=\frac{p_{\mathrm{G}}}{p_B}\times\frac{P(x_1\mid G)}{P(x_1\mid B)}\times\frac{P(x_2\mid G,x_1)}{P(x_2\mid B,x_1)}$$

- 如果两个特征独立，那么根据乘法法则

$$P(E\cap F)=P(E)\times P(F)$$

$$p(\mathbf{x}\mid G)=p(x_1\mid G)\times p(x_2\mid G)\ldots p(x_n\mid G)$$

- 但如果有很多特征，怎么办？
- n个独立特征的发生比率=总体比率×信息比率(X1) × …×信息比率(Xn)

$$O(\mathrm{G}\mid x_1,x_2)=\frac{P(\mathrm{G}\mid x_1,x_2)}{P(\mathrm{B}\mid x_1,x_2)}=\frac{p_{\mathrm{G}}P(x_1,x_2\mid G)}{p_{\mathrm{B}}P(x_1,x_2\mid B)}=\frac{p_{\mathrm{G}}P(x_1\mid G)P(x_2\mid G)}{p_{\mathrm{B}}P(x_1\mid B)P(x_2\mid B)}=O_{Pop}\times I(x_1)\times I(x_2)$$

多个特征

如果婚姻状况和工作时间相互独立

已婚和无工作的好人比率=7/1 ×0.7/0.4 ×0.15/0.5=7 ×1.75 ×0.3=3.675

	Good	$P(x|G)$	Bad	$P(x|B)$
Married	4900	0.7	400	0.4
Not married	2100	0.3	600	0.6

未婚和三年以上工作时间的好人比率？

0	1050	0.15	500	0.5
up to 6 m	1680	0.24	250	0.25
6m - 3y	1960	0.28	140	0.14
3y+	2310	0.33	110	0.11
Total	**7000**		**1000**	

对以上等式取对数：

Log odds score = ln(7) + ln(1.75)+ln(0.3) = ln(3.675)

= 1.3

s(X)　　证据权重（ weights of evidence ）