信用评分

李志勇/张兴敏

西南财经大学

Southwestern University of Finance and Economics

第二章 基本概念

我不想听任何统计数字，我要把它们全部拿来当烟点了！

——Mark Twain(1983)

基本概念

- **频数（count）**：重复n次独立试验，其中某一结果发生的次数是r次
- **频率（frequency）**：该结果发生次数所占的比例

$$f = r / n$$

- **概率（probability）**：系统的属性或内部结构以及所处环境所共同决定的可能性
- **比率（odds）**：一个事件发生（a:b）的比率是它发生的概率与它不发生的概率之比

$$a : b = p / (1 - p)$$

$$p = a / (a + b)$$

基本概念

$A \cup B$	**集合A和集合B的并集，元素属于A或B**
$A \cap B$	集合A和集合B的交集，元素同时属于A和B
$A \subset B$	A是B的子集，但B不一定是A的子集
A^C	A的补集
$a \in A$	a是A中的元素，不是子集
$p(A)$	事件A发生的概率，在0到1之间，也记为p_A
$p(A\|B)$	在给定条件B的情况下，A发生的条件概率
∵	因为
∴	所以

基本概念

- **特征characteristic**：描述个体的一个维度
- **属性attribute**：某特征的可能值

  案例：人的年龄（特征），>70（属性）

- Feature：CS里面的特征
- **变量variable**：模型中的输入
- **协变量covariate**、控制变量control variable
- 自变量independent variable，因变量dependent variable

基本概念

$p(G)$：也就是$p(G|G \cup B)$，即一个由好账户和坏账户构成的集合中好账户的概率。

x：一个属性或者属性向量。

$p(x)$：账户具有属性 x 的概率。

$p(G|x)$：具有属性 x 的账户是好账户的概率。

$p(x|G)$：好账户中具有属性 x 的概率。

符号	解释	符号	解释
Σ	连加或求和。	Π	连乘或求积。
α	检验假设中的显著水平。	Φ	累积标准正态分布，均值为 0，标准差为 1。
z	z 统计量，偏离均值的标准差的数量。	X^2	卡方统计量。
μ	均值或期望。	σ	标准差，σ^2 是方差。
γ	相关系数。	X_i	变量 X 的第 i 个值。
β_i	线性回归中变量 X_i 的系数。	b_i	回归系数。
$\hat{s}$	变量 s 的估计。	e	误差项，真实值和估计值之间的差$s_i - \hat{s}_i$。
λ	风险率，或死亡率。	exp(y)	指数 e^y。
ln(x)	以 e 为底的自然对数。		

基本概念

G:用好人（Good）的首字母G来表示“令人满意的表现”

B:用坏人的（Bad）的首字母B来表示“不令人满意的表现”

信用分数定义: 信用分数是描述具有属性x的借款人在贷款上表现令人满意的概率的一个充分统计量(sufficient statistic)。

$$P(G \mid \mathbf{x}) = P(G \mid s(\mathbf{x})), \mathbf{x} \in \mathbf{X}$$

充分统计量：关于信用风险的所有信息都包含在分数里了。

信用分数（好人分数）包含了预测贷款人是好人所需的全部信息。

简单评分卡案例

居住条件		年龄（岁）	
属性	得分	属性	得分
自有住房	30	18～25	5
租房	17	26～35	10
与父母同住	20	36～43	15
其他	0	44＋	20
贷款目的		现址居住时长（年）	
属性	得分	属性	得分
买新车	31	＜2	4
买二手车	9	2～5	9
房屋修缮	14	6～11	16
其他	0	12＋	18

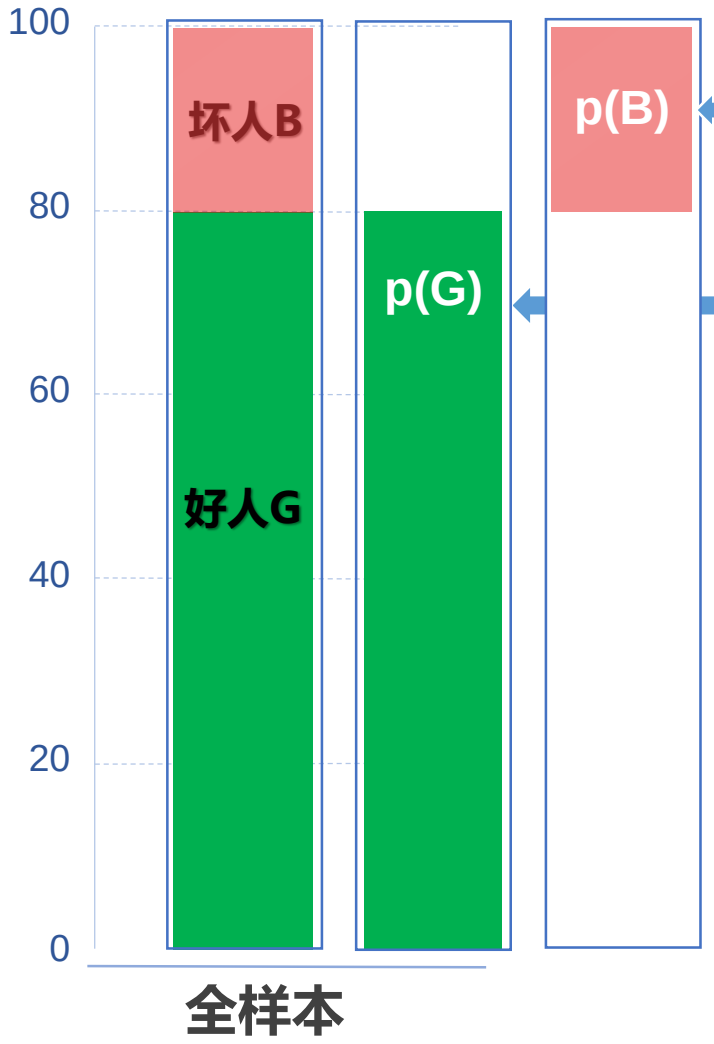
思考：

- 一个47岁、租房、在当前住址住了10 年、想借钱度假的申请者得多少分？
- 一个25岁、有自己的房产、在当前住址住了2年、想借钱买二手车的人得多少分？
- 一个38岁、与父母同住、在当前住址住了18个月、想借钱装修的人也得到多少分？

事实上，一共有七个组合可以得到53分，虽然各自情况都不一样，但对贷款机构来说，代表了同样的风险水平。

评分系统采用了补偿机制，即借款人的缺点可以用优点去弥补。

基本概念



借款人是坏人的概率，
或借款人总体中坏人的比例

借款人是好人的概率，
或借款人总体中好人的比例

性质： p(G) + p(B) = 1

好坏比率（**总体比率**、好人比率）：

$$o(G) = p(G) : p(B) = p_G : p_B$$

对数比率分数：$s(\mathbf{x}) = \ln o(G|\mathbf{x}) = \ln \frac{p(G|\mathbf{x})}{p(B|\mathbf{x})}$

例子

某银行接受了8000位贷款申请者，之后的某年，其中的7000人按时还款，1000人发生违约。如果每个好人平均带来1000元利润，每个坏人带来10000元损失。

A

需要多少个好人才能抵消一个坏人带来的损失？

B

潜在收益和潜在风险对称吗？

C

总体比率是多少？

列联表

序号	公司性质	信用等级
1	国企	AAA
2	民企	AAA
3	民企	AA
4	民企	A
5	国企	AAA
……	……	……

		信用等级		
		AAA	AA	A
公司性质	国企	2	0	0
	民企	1	1	1

列联表

	好人	P(x|G)	坏人	P(x|B)	好坏比率/总体比率
已婚	4900	0. 7	400	0. 4	4900 : 400=12. 25 : 1
未婚	2100	0. 3	600	0. 6	2100 : 600=3. 5 : 1
合计	7000	1	1000	1	7000 : 1000=7 : 1

概率和比率

某银行有1000个历史借款人的样本，每个借款人有三个特征：年龄、居住条件和信用卡持有状况。每个借款人都已确定为好人或坏人。数据中有900个好借款人，100个坏借款人。现在只考虑居住条件，包括三个属性值：自有、租房和其他。

居住条件	好人数量	坏人数量
自有	570	30
租房	150	50
其他	180	20
总数	900	100

计算：自有住房、租房和其他的好人概率P（G|x）和好人比率O(G|x)。

概率和比率

$$p(\mathrm{G}\,|\,\mathrm{owner}) = \frac{570}{570+30} = 0.95 \qquad o(\mathrm{G}\,|\,\mathrm{owner}) = \frac{570}{30} = 19.0$$

$$p(\mathrm{G}\,|\,\mathrm{renter}) = \frac{150}{150+50} = 0.75 \qquad o(\mathrm{G}\,|\,\mathrm{renter}) = \frac{150}{50} = 3.0$$

$$p(\mathrm{G}\,|\,\mathrm{others}) = \frac{180}{180+20} = 0.90 \qquad o(\mathrm{G}\,|\,\mathrm{others}) = \frac{180}{20} = 9.0$$

贝叶斯定理

•**源于**贝叶斯关于“逆概”问题的文章，而这篇文章是在他死后才由他的一位朋友发表出来的。

•**正概率问题**：如“假设袋子里面有N个白球，M个黑球，你伸手进去摸一把，摸出黑球的概率是多大”。

•**逆概率问题**：“如果我们事先并不知道袋子里面黑白球的比例，而是闭着眼睛摸出一个（或好几个）球，观察这些取出来的球的颜色之后，那么我们可以就此对袋子里面的黑白球的比例作出什么样的推测”。

•**贝叶斯论文的工作**：直接求解逆概率问题，并没有意识到这里面所包含的深刻思想。

Thomas Bayes (1701 –1761) was an English statistician, philosopher and Presbyterian minister who is known for formulating a specific case of the theorem that bears his name: Bayes' theorem.

贝叶斯定理

Thomas Bayes (1701 –1761) was an English statistician, philosopher and Presbyterian minister who is known for formulating a specific case of the theorem that bears his name: Bayes' theorem.

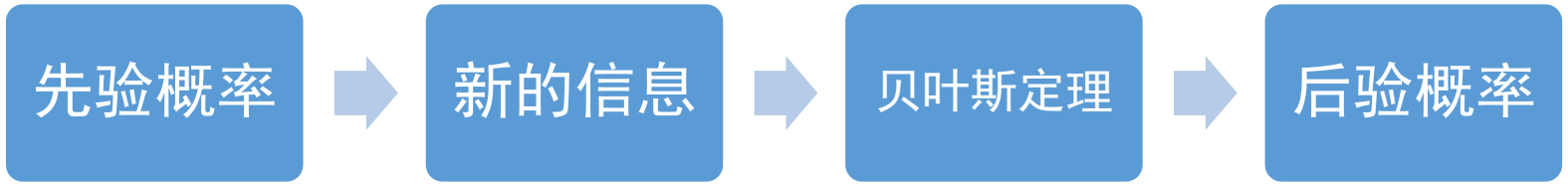
•贝叶斯方法发展：席卷概率论，广泛应用到各领域，是机器学习的核心方法之一。

•背后原因：现实世界本身就是不确定的，人类的观察能力是有局限性的，我们日常所观察到的只是事物表面上的结果。此时，需要提供一个猜测（hypothesis，更为严格的说法是“假设”，这里用“猜测”更通俗易懂一点），所谓猜测，当然就是不确定的（很可能有好多种乃至无数种猜测都能满足目前的观测），但也绝对不是两眼一抹黑瞎蒙。

需要做两件事情：1. 算出各种不同猜测的可能性大小。2. 算出最靠谱的猜测是什么。第一个就是计算特定猜测的后验概率，对于连续的猜测空间则是计算猜测的概率密度函数。第二个则是所谓的模型比较，模型比较如果不考虑先验概率的话就是最大似然方法。

贝叶斯定理

- 概率分析中，已知**先验概率**
- 然后，通过抽样或者试验，获得**额外信息**
- 通过这些额外信息，计算更新**后验概率**
- 贝叶斯定理（Bayes' theorem）是一种更新先验概率的方法
- 用途：通过已知的三个概率而推出第四个概率



贝叶斯定理

似然概率　　先验概率

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} = \frac{P(A \cap B)}{P(B)}$$

后验概率　　边际似然概率

贝叶斯定理

- 设$\mathbf{X} = (X_1, X_2, \ldots, X_n)$ 是借款人的特征，如年龄、婚姻、住房等；
- 设$\mathbf{x} = (x_1, x_2, \ldots, x_n)$ 是借款人特征的属性值，如年龄的属性有：18-25岁，26-35岁，36-43岁,>43岁等；
- p(G) 和 p(B) 是先验概率；
- 后验概率p(G|x)是给定某些属性值时借款人是好人的概率
- 后验概率 p(B|x)是给定某些属性值时借款人是坏人的概率
- p(x|G)和 p(x|B) 是在好人或坏人总体中，属性值x的似然值，也表示为 $f(x|.)$
- 根据贝叶斯定理：

$$O(\mathrm{G}\,|\,\mathbf{x}) = \frac{P(\mathrm{G}\,|\,\mathbf{x})}{P(\mathrm{B}\,|\,\mathbf{x})} = \frac{P(\mathbf{x}\,|\,G)\times P(G)/P(\mathbf{x})}{P(\mathbf{x}\,|\,B)\times P(B)/P(\mathbf{x})} = I(\mathbf{x})\times O_{Pop}$$

信息比率

列联表

	好人	P(married|G)	坏人	P(married|B)	边际比率
已婚	4900	0. 7	400	0. 4	4900 : 400=12. 25 : 1
未婚	2100	0. 3	600	0. 6	2100 : 600=3. 5 : 1
合计	7000	1	1000	1	

已婚边际比率：$0.7:0.4\times 7:1=12.25$

未婚边际比率：$0.3:0.6\times 7:1=3.5$

信息比率

O（G|x）=边际比率 = 信息比率 × 总体比率

计算自有住房、租房和其他的信息比率、边际比率

居住条件	好人数量	坏人数量	总数
自有住房(owner)	570	30	600
租房(renter)	150	50	200
其他(other)	180	20	200
总数(Total)	900	100	1000

- 好人占比：p(G) = p_G= 900/1000 = 0.9
- 坏人占比：p(B) = p_B= 100/1000 = 0.1
- p(自有住房owner) = 600/1000 = 0.6
- p(租房renter) = 200/1000 = 0.2
- p(其他other) = 200 / 1000 = 0.2

p(owner|G) = 570/900 = 0.633

p(owner|B) = 30/100 = 0.3

$$P(G \mid owner) = \frac{p(owner \mid G) * p(G)}{p(owner)}$$

$$= \frac{0.633 * 0.9}{0.6} = 0.95$$

$$s(owner) = \ln(\frac{p(G \mid owner)}{p(B \mid owner)})$$

$$= \ln(\frac{p_G}{p_B}) + \ln(\frac{p(owner \mid G)}{p(owner \mid B)})$$

$$= \ln(\frac{0.9}{0.1}) + \ln(\frac{0.633}{0.3})$$

$$= \ln(9) + \ln(2.11)$$

两个特征

- 婚姻状况

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

- 工作经验

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**	

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

- 未婚和三年以上工作时间的好人比率？

未婚和三年以上工作经验的好人比率=7/1 ×0.3/0.6 ×0.33/0.11=10.5

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 ）

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

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**	

风险决策

好坏比率	婚姻状况	工作经验
36.75:1	已婚	3年以上工作经验
24.5:1	已婚	6m-3y工作经验
11.76:1	已婚	0-6m工作经验
10.5:1	未婚	3年以上工作经验
7:1	未婚	6m-3y工作经验
3.675:1	已婚	无工作经验
3.36:1	未婚	0-6m工作经验
1.05:1	未婚	无工作经验