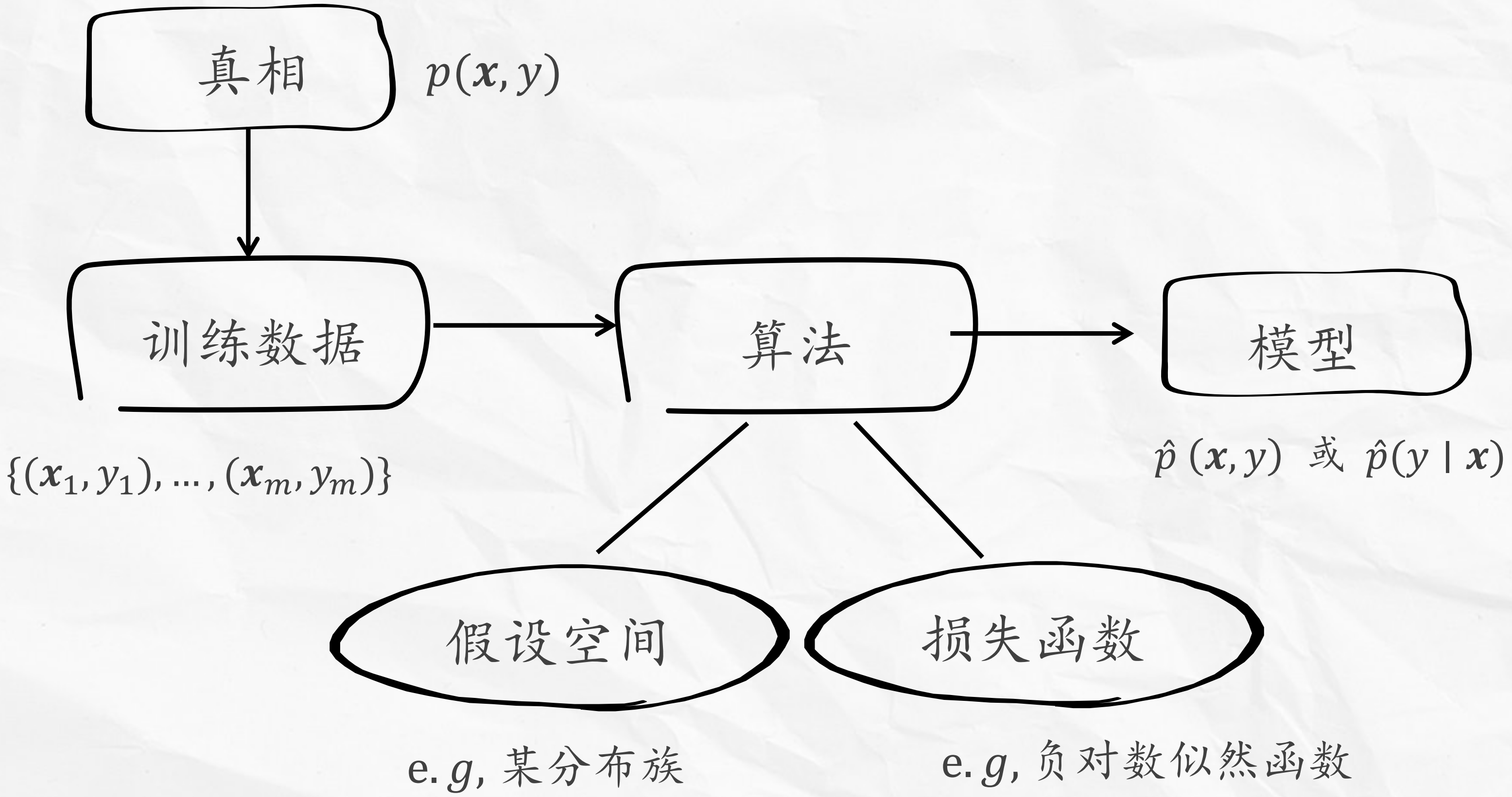
机器学习

线性分类

涂文婷
tu.wenting@mail.shufe.edu.cn

回顾：概率论角度下的线性回归

- 机器学习框架 (概率论角度)



逻辑回归

- 广义线性模型

$$y = g^{-1}(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x} + b)$$

$\uparrow$

“联系函数”(link function)

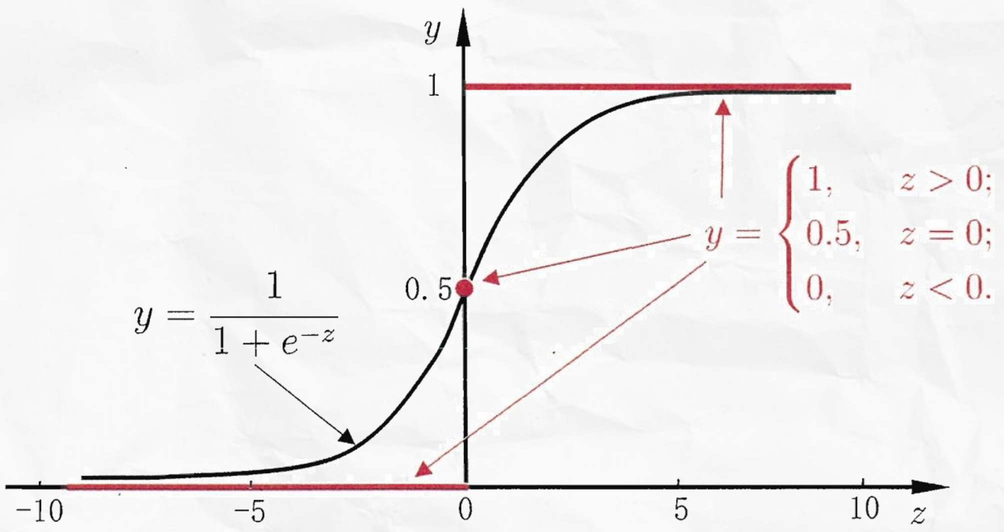
逻辑回归

- sigmoid函数

$$p(y=1\mid \boldsymbol{x})$$

$$=g^{-1}\left(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}+b\right)$$

$$=\frac{1}{1+e^{-(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}+b)}}$$



逻辑回归

- 对数几率回归

$$p(y=1 \mid \boldsymbol{x}) = \frac{1}{1+e^{-(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}+b)}}$$

$$\ln \frac{p(y=1 \mid \boldsymbol{x})}{1-p(y=1 \mid \boldsymbol{x})} = \boldsymbol{w}^{\mathrm{T}}\boldsymbol{x} + b$$

$$\underbrace{\ln \underbrace{\frac{p(y=1 \mid \boldsymbol{x})}{p(y=0 \mid \boldsymbol{x})}}_{\text{几率}}}_{\text{对数几率}}$$

回归

逻辑回归

- $\boldsymbol{w}^*, b^*$的求解

$$\ell(\boldsymbol{w}, b) = \sum_{i=1}^{m} \ln p(y_i \mid \boldsymbol{x}_i; \boldsymbol{w}, b)\,, \boldsymbol{\beta}=(\boldsymbol{w}; b), \hat{\boldsymbol{x}}=(\boldsymbol{x}; 1) \rightarrow \boldsymbol{w}^{\mathrm{T}}\boldsymbol{x} + b = \boldsymbol{\beta}^{\mathrm{T}}\hat{\boldsymbol{x}}$$

$$p(y_i \mid \boldsymbol{x}_i; \boldsymbol{w}, b) = y_i p_1(\hat{\boldsymbol{x}}_i; \boldsymbol{\beta}) + (1 - y_i) p_0(\hat{\boldsymbol{x}}_i; \boldsymbol{\beta}),\ p_1(\hat{\boldsymbol{x}}_i; \boldsymbol{\beta}) = p(y = 1 \mid \hat{\boldsymbol{x}}; \boldsymbol{\beta})$$

$$\ell(\boldsymbol{\beta}) = \sum_{i=1}^{m} (-y_i \boldsymbol{\beta}^{\mathrm{T}} \hat{\boldsymbol{x}}_i + \ln(1 + e^{\boldsymbol{\beta}^{\mathrm{T}} \hat{\boldsymbol{x}}_i}))$$

$$\boldsymbol{\beta}^* = \arg\min_{\boldsymbol{\beta}} \ell(\boldsymbol{\beta})$$

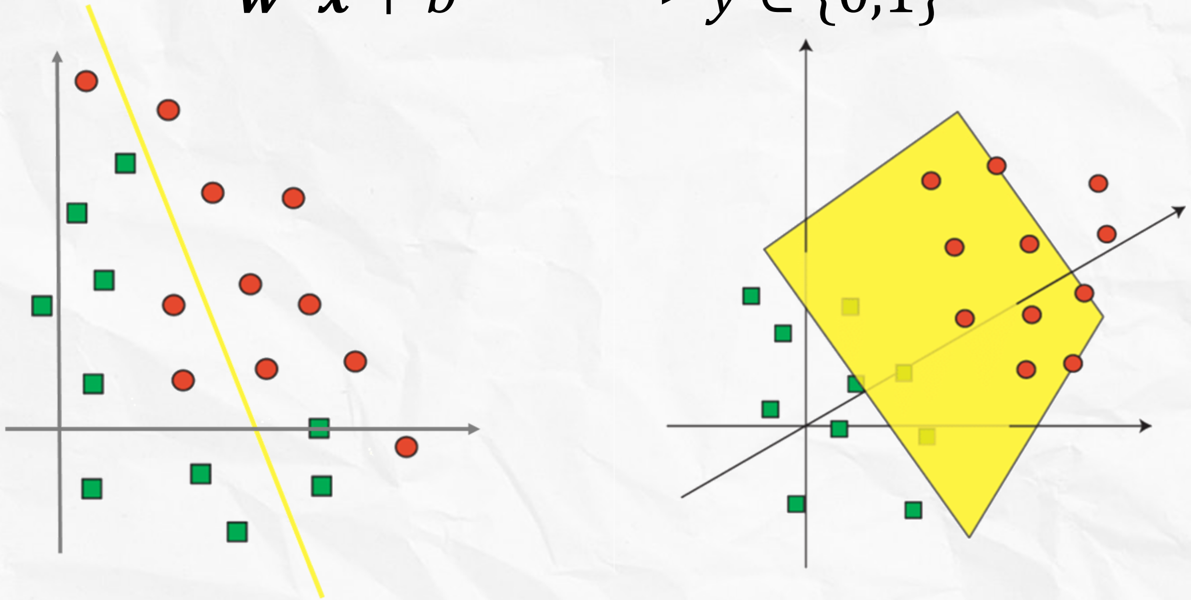
$$\frac{\partial \ell(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = -\sum_{i=1}^{m} \hat{\boldsymbol{x}}_i (y_i - p_1(\hat{\boldsymbol{x}}_i; \boldsymbol{\beta}))$$

$$\frac{\partial^2 \ell(\boldsymbol{\beta})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^{\mathrm{T}}} = \sum_{i=1}^{m} \hat{\boldsymbol{x}}_{i_i} \hat{\boldsymbol{x}}_{i_i}^{\mathrm{T}} p_1(\hat{\boldsymbol{x}}_i; \boldsymbol{\beta})(1 - p_1(\hat{\boldsymbol{x}}_i; \boldsymbol{\beta}))$$

线性分类

- 线性分类超平面

$$\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}+b \xrightarrow{\text{决定}} y\in\{0,1\}$$



$$y=\begin{cases}0, & \boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}+b<0\\ 1, & \boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}+b\geq 0\end{cases}$$

扩展：多分类的实现

- Softmax回归

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix} = \text{softmax}\left(\begin{bmatrix} W_{1,1} & W_{1,2} & W_{1,3} \\ W_{2,1} & W_{2,2} & W_{2,3} \\ W_{3,1} & W_{3,2} & W_{3,3} \end{bmatrix} \cdot \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} + \begin{bmatrix} b_1 \\ b_2 \\ b_3 \end{bmatrix} \right) \begin{matrix} \longleftrightarrow \boldsymbol{\beta}_1^\top \hat{\boldsymbol{x}} \\ \longleftrightarrow \boldsymbol{\beta}_2^\top \hat{\boldsymbol{x}} \\ \longleftrightarrow \boldsymbol{\beta}_3^\top \hat{\boldsymbol{x}} \end{matrix}$$

$$p(y = c \mid \boldsymbol{x}) = \text{softmax}(\boldsymbol{\beta}_c^\top \hat{\boldsymbol{x}}) = \frac{\exp(\boldsymbol{\beta}_c^\top \hat{\boldsymbol{x}})}{\sum_{c=1}^{C} \exp(\boldsymbol{\beta}_c^\top \hat{\boldsymbol{x}})}$$

扩展：多分类的实现

- 交叉熵损失函数

模型判断的样本
为第c类的概率

$$\mathcal{L}(\mathbf{y},\mathbf{x},\boldsymbol{\beta}) = -\sum_{c=1}^{C} \boxed{y_c} \log \boxed{f_c(\mathbf{x},\boldsymbol{\beta}_c)} = -\log LL(\mathbf{y},\mathbf{x},\boldsymbol{\beta})$$

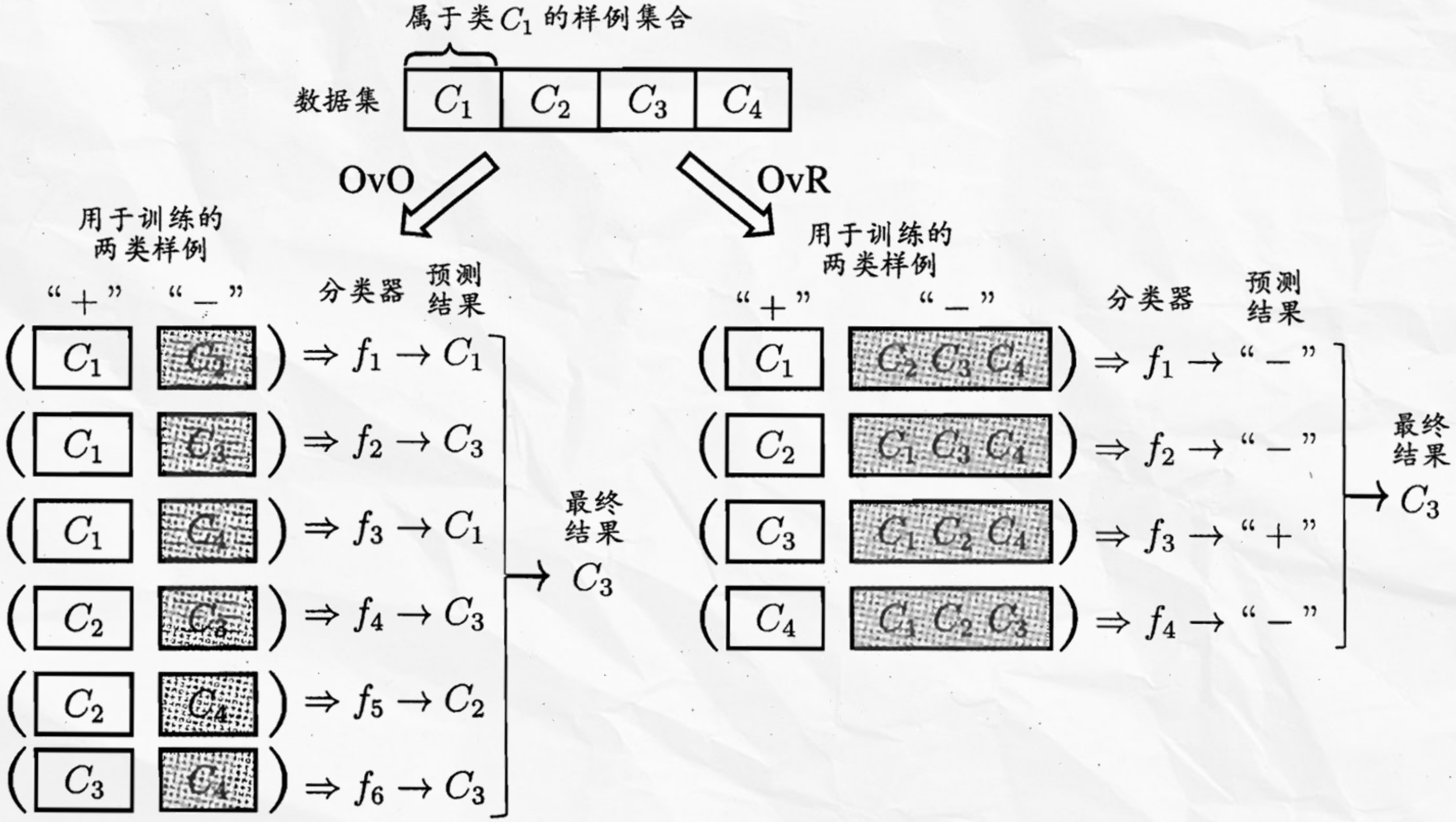
$\mathbf{y}$是一个C维的向量，用来标注样本的多
类标签。假设样本的标签为c，那么它只
有c维是1，其余维都为0
y_c是$\mathbf{y}$的第c维

熵角度：当我们将$\mathbf{y}$看做是样本标签的真实概率分布，$f_c(\mathbf{x},\boldsymbol{\beta}_c), c = 1 \ldots C$看作是类别标签的条件概率分布（模型估计的），那么$\mathcal{L}(\mathbf{y},\mathbf{x},\boldsymbol{\beta})$就对应于信息论里面交叉熵的概念，是一种衡量两个分布差距的度量，其公式为$H(p,q) = -\sum_x p(x)\log q(x)$

最大似然角度：交叉熵实际上对应于负的对数似然函数，最小化交叉熵就对应于最大似然估计

扩展：多分类的实现

- 一对一OVO/一对其余OVR



逻辑回归应用：信用风险评分卡

- 信用评分卡

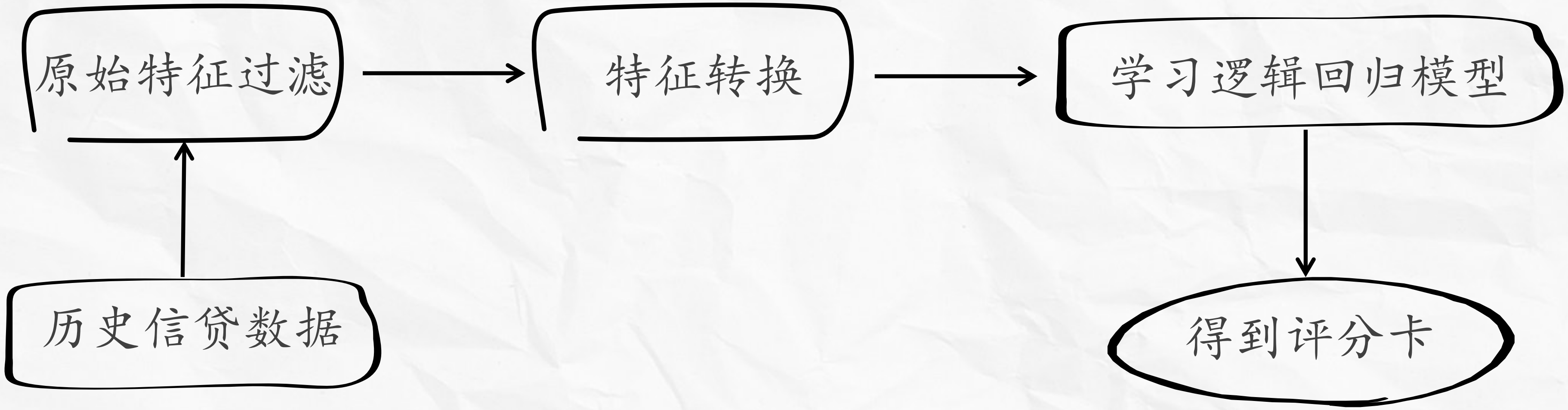
age.in.years	housing	present.employment.since
35	rent	1 <= ... < 4 years

age.in.years	housing	present.employment.since	score
201	122	148	471

variable	bin	points
age.in.years	[-Inf,26)	119
age.in.years	[26,35)	146
age.in.years	[35,37)	201
age.in.years	[37, Inf)	162
housing	rent	122
housing	own	163
housing	for free	117
present.employment.since	unemployed%,%... < 1 year	125
present.employment.since	1 <= ... < 4 years	148
present.employment.since	4 <= ... < 7 years%,%... >= 7 years	167

逻辑回归应用：信用风险评分卡

- 评分卡构建流程



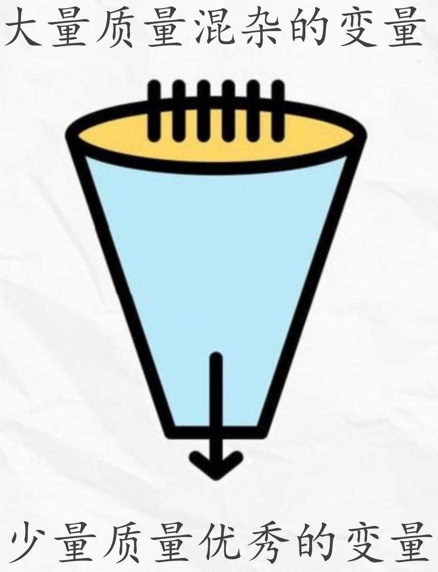
逻辑回归应用：信用风险评分卡

◦ 变量过滤

- 按照缺失值比例过滤

- 按照特征价值过滤

－信息值（Information Value)；
－其他衡量特征与目标相关性的指标



逻辑回归应用：信用风险评分卡

变量分箱

- 非监督分箱
  - 等宽、等频、聚类；
  - 优点：计算简单
  - 缺点：未利用目标变量

- 监督分箱
  - 卡方分箱法、决策树分箱法、bestKS；
  - 优点：考虑了目标变量
  - 缺点：计算量大

Income
13495
16500
18920
41315
5151
6295
…

→

Income	Income-binned
13495	Low
16500	Low
18920	Medium
41315	High
5151	Low
6295	Low
…	…

* 对于有序类别型变量，可以直接用卡方等分箱法来进行分箱
* 对于无序类别型变量：取值数量较少，一般无需分箱（但有可能需要合并优化）；取值数量较大，若要用卡方分箱，需要转换成有序型变量。常用的方法是利用每个取值的坏样本率进行数值编码。
* 对于特殊值（例如缺失值），可以将特殊值看成单独的一箱，并不考虑在单调性检查之列

逻辑回归应用：信用风险评分卡

- WOE编码

WOE（Weight of Evidence）是一个常常用于变量分箱后的编码方法，可以实现用数值代替非数值的操作，目的是为了让模型能够对其进行数学运算。

Bin	Bad Count	Good Count	Bad Percent	Good Percent	WOE
1	B_1	G_1	B_1/B	G_1/G	$\ln(\frac{G_1/G}{B_1/B})$
2	B_2	G_2	B_2/B	G_2/G	$\ln(\frac{G_2/G}{B_2/B})$
...					
N	B_N	G_N	B_N/B	G_N/G	$\ln(\frac{G_N/G}{B_N/B})$
Total	$B=\sum B_i$	$G=\sum G_i$	WOE公式的另一种表达 $\ln\left(\frac{G_i/G}{B_i/B}\right)=\ln\left(\frac{G_i}{B_i}\right)-\ln\left(\frac{G}{B}\right)$		

逻辑回归应用：信用风险评分卡

- 从后验概率到评分

－ 逻辑回归模型的输出：

$p(y = \mathrm{bad} \mid \boldsymbol{x})$

－ 转换$p(y = \mathrm{bad} \mid \boldsymbol{x})$到$credit_score(\boldsymbol{x})$

Step 1. 设定PDO (point to double odds)，涵义为当好坏比上升1倍时，分数上升PDO个单位。

Step 2. 设定$\mathrm{odds_bad}$和S_0，得到$\mathrm{S}_{\mathrm{base}}$：

$\mathrm{S}_{\mathrm{base}} = \mathrm{S}_0 + PDO/\ln(2) \times \ln(\mathrm{odds\_bad})$

Step 3. 得到$\mathrm{S}(\boldsymbol{x})$

$\mathrm{S}(\boldsymbol{x}) = \mathrm{S}_{\mathrm{base}} - PDO/\ln(2) \times \ln\frac{p(y=\mathrm{bad}|\boldsymbol{x})}{1-p(y=\mathrm{bad}|\boldsymbol{x})}$