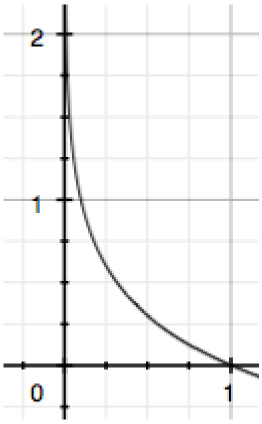
信息量

越不可能发生的事件发生了，获取到的信息量就越大。

越可能发生的事件发生了，获取到的信息量就越小。

信息量和事件发生的概率有关

$$I(x_0) = -log(p(x_0))$$



熵是对平均不确定性的度量

$$H(X) = -\sum_{x \in X} P(x) \cdot \log P(x)$$

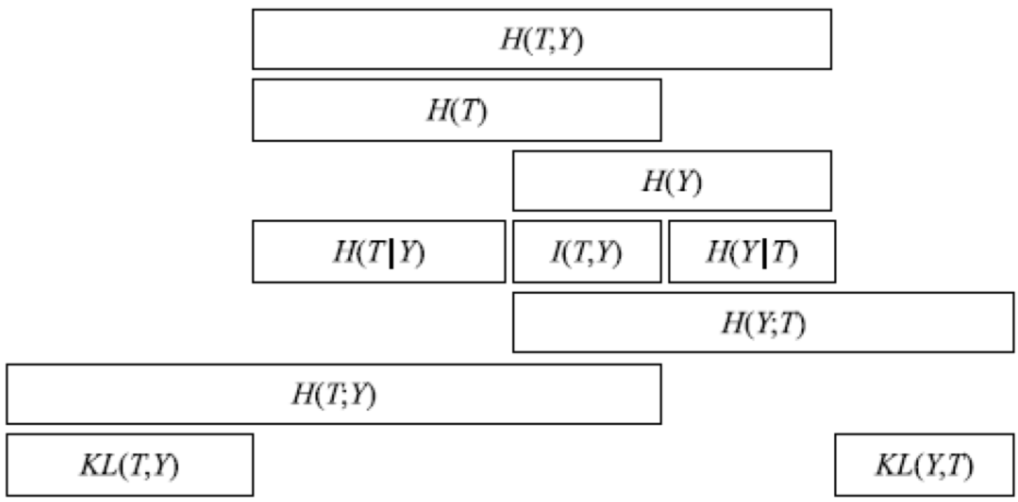
二项分布问题，对于这类问题，熵的计算可以简化为：

$$\begin{aligned} H(X) &= -\sum_{i=1}^{n} p(x_i) log(p(x_i)) \\ &= -p(x) log(p(x)) - (1 - p(x)) log(1 - p(x)) \end{aligned}$$

互信息：

$$I(X;Y) = \sum_{x \in X, y \in Y} P(x,y) \log \frac{P(x,y)}{P(x)P(y)}$$

Name	Formula	(Dis)similarity	(A)symmetry
Joint Information	$H(T,Y) = -\sum_t \sum_y p(t,y) \log_2 p(t,y)$	Inapplicable	Symmetry
Mutual Information	$I(T,Y) = \sum_t \sum_y p(t,y) \log_2 \frac{p(t,y)}{p(t)p(y)}$	Similarity	Symmetry
Conditional Entropy	$H(Y\|T) = -\sum_t \sum_y p(t,y) \log_2 p(y\|t)$	Dissimilarity	Asymmetry
Cross Entropy	$H(T;Y) = -\sum_z p_t(z) \log_2 p_y(z)$	Dissimilarity	Asymmetry
KL Divergence	$KL(T,Y) = \sum_z p_t(z) \log_2 \frac{p_t(z)}{p_y(z)}$	Dissimilarity	Asymmetry



熵 H(T) = 条件熵 H(T|Y)+互信息 I(T,Y)

相对熵又称 KL 散度：

$$D_{KL}(p||q) = \sum_{i=1}^{n} p(x_i) log(\frac{p(x_i)}{q(x_i)})$$

n 为事件的所有可能性

交叉熵 H(T,Y) ：

$$D_{KL}(p||q) = \sum_{i=1}^{n} p(x_i) log(p(x_i)) - \sum_{i=1}^{n} p(x_i) log(q(x_i))$$

$$= -H(p(x)) + [-\sum_{i=1}^{n} p(x_i) log(q(x_i))]$$

等式的前一部分恰巧就是 p 的熵，等式的后一部分就是交叉熵：

$$H(p,q) = -\sum_{i=1}^{n} p(x_i) log(q(x_i))$$

信息增益：

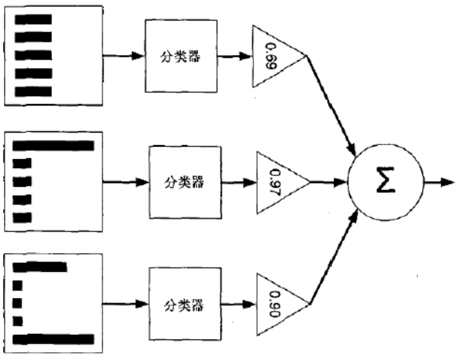
- 特征A对训练数据集D的信息增益g(D,A)，定义为集合D的经验熵H(D)与特征A给定条件下D的经验条件熵H(D|A)之差，即：
  - g(D,A)=H(D) – H(D|A)
  - 显然，这即为训练数据集D和特征A的互信息。
- 遍历所有特征，选择信息增益最大的特征作为当前的分裂特征

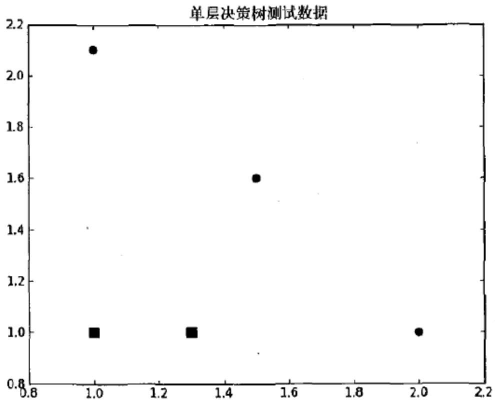
- 信息增益率：gr(D,A) = g(D,A) / H(A)
  - C4.5
- Gini系数：
  - CART

$$Gini(p) = \sum_{k=1}^{K} p_k(1-p_k) = 1 - \sum_{k=1}^{K} p_k^2$$

$$= 1 - \sum_{k=1}^{K} \left(\frac{|C_k|}{|D|} \right)^2$$

AdaBoost，是英文"Adaptive Boosting"（自适应增强）的缩写，由Yoav Freund和Robert Schapire在1995年提出。它的自适应在于：前一个基本分类器分错的样本会得到加强，加权后的全体样本再次被用来训练下一个基本分类器。同时，在每一轮中加入一个新的弱分类器，直到达到某个预定的足够小的错误率或达到预先指定的最大迭代次数。





给定一个训练数据集T={(x1,y1), (x2,y2)…(xN,yN)}，其中实例$x \in \mathcal{X}$，而实例空间$\mathcal{X} \subset \mathbb{R}^n$，yi属于标记集合{-1,+1}，Adaboost的目的就是从训练数据中学习一系列弱分类器或基本分类器，然后将这些弱分类器组合成一个强分类器。

Adaboost的算法流程如下：

步骤1. 首先，初始化训练数据的权值分布。每一个训练样本最开始时都被赋予相同的权重：1/N。

$$D_1=(w_{11},w_{12}\cdots w_{1i}\cdots,w_{1N}),\ w_{1i}=\frac{1}{N},\ i=1,2,\cdots,N$$

步骤2. 进行多轮迭代，用m = 1,2, ..., M表示迭代的第多少轮

a. 使用具有权值分布Dm的训练数据集学习，得到基本分类器：

$$G_m(x):\ \chi \to \{-1,+1\}$$

b. 计算Gm(x)在训练数据集上的分类误差率

$$e_m=P(G_m(x_i)\neq y_i)=\sum_{i=1}^{N}w_{mi}I(G_m(x_i)\neq y_i)$$

由上述式子可知，Gm(x)在训练数据集上的**误差率**em就是被Gm(x)误分类样本的权值之和。

c. 计算Gm(x)的系数，am表示Gm(x)在最终分类器中的重要程度（目的：得到基本分类器在最终分类器中所占的权重）：

$$\alpha_m=\frac{1}{2}\log\frac{1-e_m}{e_m}$$

由上述式子可知，em <= 1/2时，am >= 0，且am随着em的减小而增大，意味着分类误差率越小的基本分类器在最终分类器中的作用越大。

d. 更新训练数据集的权值分布（目的：得到样本的新的权值分布），用于下一轮迭代

$$D_{m+1}=(w_{m+1,1},w_{m+1,2}\cdots w_{m+1,i}\cdots,w_{m+1,N}),$$

$$w_{m+1,i}=\frac{w_{mi}}{Z_m}\exp(-\alpha_m y_i G_m(x_i)),\ i=1,2,\cdots,N$$

使得被基本分类器Gm(x)误分类样本的权值增大，而被正确分类样本的权值减小。就这样，通过这样的方式，AdaBoost方法能“聚焦于”那些较难分的样本上。

其中，Zm是规范化因子，使得Dm+1成为一个概率分布：

$$Z_m=\sum_{i=1}^{N}w_{mi}\exp(-\alpha_m y_i G_m(x_i))$$

步骤3. 组合各个弱分类器

$$f(x)=\sum_{m=1}^{M}\alpha_m G_m(x)$$

从而得到最终分类器，如下：

$$G(x)=sign(f(x))=sign\left(\sum_{m=1}^{M}\alpha_m G_m(x)\right)$$

X	0	1	2	3	4	5	6	7	8	9
Y	1	1	1	−1	−1	−1	1	1	1	−1

迭代过程1

对于m=1，在权值分布为**D1**（10个数据，每个数据的权值皆初始化为0.1）的训练数据上，经过计算可得：

1. **阈值v取2.5**时误差率为0.3（x < 2.5时取1，x > 2.5时取-1，**则6 7 8分错**，误差率为0.3），
2. 阈值v取5.5时误差率最低为0.4（x < 5.5时取1，x > 5.5时取-1，则3 4 5 6 7 8皆分错，误差率0.6大于0.5，不可取。故令x > 5.5时取1，x < 5.5时取-1，则0 1 2 9分错，误差率为0.4），
3. 阈值v取8.5时误差率为0.3（x < 8.5时取1，x > 8.5时取-1，则3 4 5分错，误差率为0.3）。

所以无论阈值v取2.5，还是8.5，总得分错3个样本，故可任取其中任意一个如2.5，弄成第一个基本分类器为：

$$G_1(x)=\begin{cases}1, & x<2.5\\ -1, & x>2.5\end{cases}$$

上面说阈值v取2.5时则6 7 8分错，所以误差率为0.3

从而得到G1(x)在训练数据集上的误差率（被G1(x)误分类样本“6 7 8”的权值之和）**e1=P(G1(xi)≠yi) = 3*0.1 = 0.3**。

然后根据误差率e1计算G1的系数：

$$\alpha_1=\frac{1}{2}\log\frac{1-e_1}{e_1}=0.4236$$

这个a1代表G1(x)在最终的分类函数中所占的权重，为0.4236。

接着更新训练数据的权值分布，用于下一轮迭代：

$$D_{m+1}=\left(w_{m+1,1},w_{m+1,2}\cdots w_{m+1,i}\cdots,w_{m+1,N}\right),$$

$$w_{m+i}=\frac{w_{mi}}{Z_m}\exp\left(-\alpha_m y_i G_m(x_i)\right),\ i=1,2,\cdots,N$$

第一轮迭代后，最后得到各个数据**新**的权值分布**D2** = (0.0715, 0.0715, 0.0715, 0.0715, 0.0715, 0.0715, 0.1666, 0.1666, 0.1666, 0.0715)。由此可以看出，因为样本中是数据“6 7 8”被G1(x)分错了，所以它们的权值由之前的0.1增大到0.1666，反之，其它数据皆被分正确，所以它们的权值皆由之前的0.1减小到0.0715。

分类函数f1(x)= a1*G1(x) = 0.4236G1(x)。

此时，得到的第一个基本分类器sign(f1(x))在训练数据集上有3个误分类点（即6 7 8）。

从上述第一轮的整个迭代过程可以看出：被误分类样本的权值之和影响误差率，误差率影响基本分类器在最终分类器中所占的权重。

迭代过程2

对于m=2，在权值分布为**D2** = (0.0715, 0.0715, 0.0715, 0.0715, 0.0715, 0.0715, 0.1666, 0.1666, 0.1666, 0.0715)的训练数据上，经过计算可得：

1. 阈值v取2.5时误差率为0.1666*3（x < 2.5时取1，x > 2.5时取-1，则6 7 8分错，误差率为0.1666*3），
2. 阈值v取5.5时误差率最低为0.0715*4（x > 5.5时取1，x < 5.5时取-1，则0 1 2 9分错，误差率为0.0715*3 + 0.0715），
3. **阈值v取8.5**时误差率为0.0715*3（x < 8.5时取1，x > 8.5时取-1，**则3 4 5分错**，误差率为0.0715*3）。

所以，阈值v取8.5时误差率最低，故第二个基本分类器为：

$$G_2(x)=\begin{cases}1, & x<8.5\\ -1, & x>8.5\end{cases}$$

计算G2的系数：

$$\alpha_2=\frac{1}{2}\log\frac{1-e_2}{e_2}=0.6496$$

更新训练数据的权值分布：

$$D_{m+1}=(w_{m+1,1},w_{m+1,2}\cdots w_{m+1,i}\cdots,w_{m+1,N})$$

$$w_{m+i}=\frac{w_{mi}}{Z_m}\exp(-\alpha_m y_i G_m(x_i)),\quad i=1,2,\cdots,N$$

D3 = (0.0455, 0.0455, 0.0455, 0.1667, 0.1667, 0.01667, 0.1060, 0.1060, 0.1060, 0.0455)。

迭代过程3

$$G_3(x)=\begin{cases}1, & x<5.5\\ -1, & x>5.5\end{cases}$$

$$D_{m+1}=(w_{m+1,1},w_{m+1,2}\cdots w_{m+1,i}\cdots,w_{m+1,N})$$

$$w_{m+i}=\frac{w_{mi}}{Z_m}\exp(-\alpha_m y_i G_m(x_i)),\quad i=1,2,\cdots,N$$

D4 = (0.125, 0.125, 0.125, 0.102, 0.102, 0.102, 0.065, 0.065, 0.065, 0.125)。

$$\alpha_3=\frac{1}{2}\log\frac{1-e_3}{e_3}=0.7514$$

G(x) = sign[f3(x)] = sign[a1 * G1(x) + a2 * G2(x) + a3 * G3(x)]

G(x) n[f3(x)] = sign[0.4236G1(x) + 0.6496G2(x)+0.7514G3(x)]