## 信息增益

1. 比特编码

X 有如下4种取值情况

|  |  |  |  |
| --- | --- | --- | --- |
| P(X=A)=1/4 | P(X=B)=1/4 | P(X=C)=1/4 | P(X=D)=1/4 |

假设我们要编码的序列为：BAACBADCDADDDA…

可以将(e.g. A = 00, B = 01, C = 10, D = 11),，每个字符使用2个比特位。

0100001001001110110011111100…

1. 更佳的比特编码方式

也许有人会告诉你每个字符的概率不一定相等,如下表

|  |  |  |  |
| --- | --- | --- | --- |
| P(X=A)=1/2 | P(X=B)=1/4 | P(X=C)=1/8 | P(X=D)=1/8 |

那有没有一种编码方式，能将平均每个字符花1.75bits, 如果有的话，那应该如何编码？可能大部分的人第一反应就是哈夫曼编码(Huffman Coding)。

|  |  |
| --- | --- |
| A | 0 |
| B | 10 |
| C | 110 |
| D | 111 |

当然这只是众多编码中的一种编码方法。

Suppose there are three equally likely values…

|  |  |  |
| --- | --- | --- |
| P(X=A)=1/3 | P(X=B)=1/3 | P(X=C)=1/3 |

最简单方便的编码方法是每个字符花费2 bits，那有没有一种编码方式，能将平均每个字符花1.6bits。理论上，每只字符只要1.58496 bits，就可以完成编码方法。

一般情况，假如随机变量X有m种取值可能，V1, V2, … Vm

|  |  |  |  |
| --- | --- | --- | --- |
| P(X=V1)=P1 | P(X=V2)=P2 | …. | P(X=Vm)=Pm |