## Escolha da melhor palavra

Em termos formais, para um conjunto O de observações e uma classe J de n objetos, Gini é calculado da seguinte forma:

$$I(O)=1-S$$

onde $S=\sum_{i=1}^{n} P(j_i|O)^2$ para $j_i \in J$ e $P(j_i|O)$ é a probabilidade da ocorrência de objetos $j_i$ em $O$.

Por exemplo, considere a palavra “deadline” e a resposta padrão C. Suponha que em 1596 mensagens, 230 devem ser respondidas com C. A Tabela 1 resume estes números. Na tabela, A representa qualquer outra resposta que não C.

| **Tabela 1: Distribuição da resposta “C” entre mensagens** | | |
|---|---|---|
| **resposta** | **# mensagens** | **percentagem** |
| C | 230 | 14,41% |
| A | 1366 | 85,59% |
| Total | 1596 | 100% |

Gini = $1 – (0{,}1441^2 + 0{,}8559^2) = 0{,}2467$

| **Tabela 2: Mensagens com a palavra “deadline”** | | |
|---|---|---|
| **resposta** | **# mensagens** | **percentagem** |
| C | 184 | 89,32% |
| A | 22 | 10,68% |
| Total | 206 | 100% |

Gini = $1 – (0{,}8932^2 + 0{,}1068^2) = 0{,}1908$

| **Tabela 3: Mensagens sem a palavra “deadline”** | | |
|---|---|---|
| **resposta** | **# mensagens** | **percentagem** |
| C | 46 | 3,31% |
| A | 1344 | 96,69% |
| Total | 1390 | 100% |

Gini = $1 – (0{,}0331^2 + 0{,}9669^2) = 0{,}0640$

Finalmente, vamos checar o quanto a presença ou ausência da palavra “deadline” em uma mensagem indica a ocorrência da resposta C ou outras respostas.

Resultado:

$$\Delta I = 0.2467 - \left(0.1908\, x\, \frac{206}{1596} + 0.0640\, x\, \frac{1392}{1596}\right)$$

Em termos formais, para um conjunto O de mensagens e suas respectivas respostas padrões, o poder discriminativo de uma palavra w é dada por:

$$\Delta I(O) = I(O) - \left(I(O_w) * p_w + I(O_{\not w}) * p_{\not w}\right)$$

onde

$I(O)$ = Gini de O (i.e, distribuição de uma resposta C em todas as mensagens, tabela 5 no exemplo considerado).

$I(O_w)$ = Gini do subconjunto de elementos de O contendo a palavra w (tabela 6 no exemplo)

$I(O_{\not w})$ = Gini do subconjunto de elementos de O sem a palavra w (tabela 7 no exemplo)

$p_w$ e $p_{(\not w)}$ são as proporções dos elementos em $I(O_w)$ e $I(O_{\not w})$ .