Upper bound of the Information Value (IV)

Let T be a $k \times 2$ table of N observations: a cross-tabulation of a categorical variable C with k categories, by a binary outcome variable which is split into 'Good' and 'Bad', denoted by G and B respectively.

Category	Good	Bad
C_1	G_1	B_1
C_2	G_2	B_2
$\vdots$	$\vdots$	$\vdots$
C_k	G_k	B_k
Total	N_G	N_B

$\Longrightarrow$

Category	prop(Good)	prop(Bad)
C_1	g_1	b_1
C_2	g_2	b_2
$\vdots$	$\vdots$	$\vdots$
C_k	g_k	b_k
Total	1	1

So $N_G + N_B = N$, the number of observations in the table, and for all categories, $G_i > 0$ and $B_i > 0$, where $i = 1..k$. [1]

Then $g_i = G_i/N_G$, the proportion of goods in category i, and similarly, $b_i = B_i/N_B$, the proportion of bads in category i. Clearly:

$$\sum_{i=1}^{k} g_i = 1 \quad , \quad \sum_{i=1}^{k} b_i = 1 \quad , \quad 0 < g_i, b_i < 1 \tag{1}$$

The **Information Value** (IV) is defined as:

$$\text{IV} = \sum_{i=1}^{k} (g_i - b_i) \log_e(g_i/b_i) \tag{2}$$

Note that if g_i or b_i were to equal zero, then by the convention that $0 \times \log_a(0) = 0$, the summand at i would be zero. In other words, if a category is composed of only goods or bads, then that category does not contribute towards the IV, and is ignored.

From here onwards, we will drop the e subscript to log, and we only refer to the natural logarithm.

Equation (2) can be re-written as:

$$\begin{aligned} \text{IV} &= \sum_{i=1}^{k} (g_i - b_i)(\log(g_i) - \log(b_i)) & (3) \\ &= \sum_{i=1}^{k} g_i \log(g_i) - \sum_{i=1}^{k} g_i \log(b_i) - \sum_{i=1}^{k} b_i \log(g_i) + \sum_{i=1}^{k} b_i \log(b_i) & (4) \end{aligned}$$

and hence the IV is at a maximum when the 1st and 4th terms are maximised, and the 2nd and 3rd terms are minimised.

[1] If the categorical variable C had one or more categories with zero goods or bads, then this variable would not be suitable for inclusion in our model. Either it should be a policy rule, or the distributions in each category need re-assessing.

The 1st and 4th terms in equation (4) have the same form:

$$U = \sum_i u_i \log(u_i) \quad , \quad 0 < u_i < 1 \tag{5}$$

The upper bound for $u_i \log(u_i)$ is clearly zero, as $\log(x)$ grows more slowly than x. Hence, an upper bound for U is also zero.

(Not needed here, but $u_i \log(u_i)$ is a minimum when $u_i = (1/e)$, with a value of $(-1/e)$ hence a lower bound for U is $(-k/e)$.)

The 2nd and 3rd terms of equation (4) have the form:

$$Z = \sum_i v_i \log(w_i) \quad , \quad 0 < v_i, w_i < 1 \tag{6}$$

As w_i approaches 1, $v_i \log(w_i)$ clearly tends to zero, hence an upper bound for Z is zero. If w_i could be any real number, then there is no lower bound, as $\log(x)$ goes to $-\infty$ as x approaches zero. However, in our case, the smallest w_i could be is $1/\max(N_G, N_B)$.

Re-write

$$\sum_{i=1}^{k} g_i \log b_i \tag{7}$$

as

$$\sum_{i=1}^{k} \frac{G_i}{N_G} \log\left(\frac{B_i}{N_B}\right) \tag{8}$$

We can order the G_i such that $G_1 = N_G - (k-1)$, and $G_i = 1, i = 2..k$. Similarly, let $B_k = N_B - (k-1)$, and $B_i = 1, i = 1..(k-1)$.

Then

$$\begin{aligned}
\sum_{i=1}^{k} \frac{G_i}{N_G} \log\left(\frac{B_i}{N_B}\right) \quad &= \quad \frac{N_G - (k-1)}{N_G} \log\left(\frac{1}{N_B}\right) \\
&+ \quad \sum_{i=2}^{k-1} \frac{1}{N_G} \log\left(\frac{1}{N_B}\right) \\
&+ \quad \frac{1}{N_G} \log\left(\frac{N_B - (k-1)}{N_B}\right) \qquad (9) \\
&= \quad \frac{N_G - 1}{N_G} \log\left(\frac{1}{N_B}\right) + \frac{1}{N_G} \log\left(\frac{N_B - (k-1)}{N_B}\right) \qquad (10)
\end{aligned}$$

As N_G and N_B increase, the second term in equation (10) approaches zero. The first term approaches $\log(1/N_B)$, which is hence the lower bound.

Re-applying this result to equation (4), it is clear that as N_G and N_B grow larger, the upper bound for the IV approaches $\log(N_G) + \log(N_B)$.