APPENDIX B. SIGNIFICANCE AND THE EXACT FORMULATION OF INFORMATION GAIN

Assume that we want to decide on the predictive power of a binary categorical attribute, a, with x being the number of training points that have value 0 and y the number of training points having value 1. Assume that the class label is also a binary attribute. Note that we can later generalize the results for any situation in which attribute a and the class label can be split into two sections in some defined way. The generalized assumption will be valid for all classification tasks in this thesis.

Equation (1) shows a contingency table in which x_0 refers to the number of items for which attribute a as well as the class label are 0, y_0 refers to the number of items for which attribute is 1 and the class label is 0; etc. The contingency table is

x_0	x_1	x	$(= x_0 + x_1)$
y_0	y_1	y	$(= y_0 + y_1)$
t_0	t_1	t	$(= t_0 + t_1)$

$$(x_1 + y_1 = t_1;\ x_0 + y_0 = t_0) \tag{1}$$

The probability of a particular value of x_1 (and thereby $y_1 = t_1 - x_1$), i.e., a particular distribution of class label values among those x items for which attribute a is 0, is

$$\frac{\binom{x}{x_1}\binom{y}{y_1}}{\binom{t}{t_1}} = p_{combination} \tag{2}$$

The total distribution for attribute a is known (x and y) as well as the total distribution of class labels (t and t_1). The probability, $p_{combination}$, is largest if attribute a

divides the set in such a way that the proportions of class label 0 and 1 are the same with or without considering the value of *a*, i.e., $\frac{x_1}{x} = \frac{y_1}{y} = \frac{t_1}{t}$

B.1. Significance

To derive significance from this quantity, the sum is taken over all contingency tables that are more extreme than the one that was encountered in the experiment. This derivation is standard and will not be discussed.

B.2. Information Gain

Information gain, as it is commonly stated (*InfoGain*), can be derived as an approximated version of the logarithm of this quantity. Writing the combinations as factorials, we get:

$$p_{combinations} = \frac{x!}{x_0!x_1!}\frac{y!}{y_0!y_1!}\frac{t_0!t_1!}{t!}$$

Let us focus on the first term and take the logarithm

$$\log\left(\frac{x!}{x_0!x_1!}\right) = \log x! - \log x_0! - \log x_1!$$

Using Stirling's approximation ($\log n! \cong n\log n - n$), we get

$$\log\left(\frac{x!}{x_0!x_1!}\right) \cong x\log x - x - x_0\log x_0 + x_0 - x_1\log x_1 + x_1$$

The expression can be simplified using $x = x_0 + x_1$ twice

$$\log\left(\frac{x!}{x_0!x_1!}\right) \cong x\log x - x_0\log x_0 - x_1\log x_1$$
$$= (x_0 + x_1)\log x - x_0\log x_0 - x_1\log x_1$$
$$= -\left(x_0\log\left(\frac{x_0}{x}\right) + x_1\log\left(\frac{x_1}{x}\right)\right)$$
$$= -x\left(\frac{x_0}{x}\log\left(\frac{x_0}{x}\right) + \frac{x_1}{x}\log\left(\frac{x_1}{x}\right)\right)$$

The fraction (x_0/x) is the probability of class label 0 among the items with a=0 and is commonly referred to as p_0. Using the standard definition of information ("Info"),

$$Info(p_0, p_1) = -(p_0\log p_0 + p_1\log p_1)$$

We can now write

$$\log p_{combination} \cong x\ Info\left(\frac{x_0}{x}, \frac{x_1}{x}\right) + y\ Info\left(\frac{y_0}{y}, \frac{y_1}{y}\right) - t\ Info\left(\frac{t_0}{t}, \frac{t_1}{t}\right)$$

Using the standard definition of information gain of attribute *a* ("*InfoGain*(*a*)"),

$$InfoGain(a) = Info\left(\frac{t_0}{t}, \frac{t_1}{t}\right) - \frac{x}{t}Info\left(\frac{x_0}{x}, \frac{x_1}{x}\right) - \frac{y}{t}Info\left(\frac{y_0}{y}, \frac{y_1}{y}\right)$$

We can now rewrite

$$InfoGain(a) \cong -\frac{1}{t}\log(p_{combination})$$

Note that Stirling's approximation was used in the derivation, and the relationship is, therefore, only expected to hold for large numbers. We will now look at the exact formula for Information ("*ExactInfo*") and Information Gain ("*ExactInfoGain*").

$$ExactInfo(x, x_0, x_1) = \frac{1}{x}(\log x! - \log x_1! - \log x_2!)$$

Note that, according to the derivation, the natural logarithm is used in this equation. Using $\log_2$ results in a constant scaling factor.

The expression for information gain in its exact form can be written analogously to the approximated version:

$$ExactInfoGain(a) = ExactInfo(t, t_0, t_1) - \frac{x}{t} ExactInfo(x, x_0, x_1) - \frac{y}{t} ExactInfo(y, y_0, y_1)$$