

## Tutorial: The Spanning Tree Algorithm

This tutorial is the same material as the coursework. You may wish to use it for revision.

A data warehouse contains the following vast data set connecting three variables A B and C:

A	B	C
a1	b1	c1
a1	b1	c2
a1	b2	c2
a1	b2	c1
a2	b2	c1
a2	b1	c1
a2	b2	c1
a2	b2	c1

We will find a spanning tree and check its accuracy.

1. Construct co-occurrence matrices for the three possible pairings AB, BC and AC:

	a1	a2
b1		
b2		

etc

2. From the co-occurrences construct the joint probability table for each pair, and the marginalisations, using the following format:

	a1	a2	P(B)
b1			
b2			
P(A)			

etc.

3. Calculate the L1 metric for each possible pair of nodes

$$\text{Dep}(A,B) = \sum_{A \times B} |P(a_i \& b_j) - P(a_i)P(b_j)|$$

4. Given that A is the root node construct the tree.

5. Calculate the conditional probability matrices  $P(B|A)$  and  $P(C|A)$

6. Calculate the joint probabilities  $P(a1 \& b1 \& c1)$  and  $P(a2 \& b2 \& c1)$ , using first the data and secondly the tree you have constructed.

7. If it is known from other sources that the prior probability of A is  $P(A) = \{0.4, 0.6\}$  and not the value  $\{0.5, 0.5\}$  displayed in the data, what result will the tree give for  $P(a1 \& b1 \& c1)$  and  $P(a2 \& b2 \& c1)$ ?

8. Would using the weighted L1 metric make any difference to your result?