Problem 1

In this problem we relate the performance of three learning procedures (Procedure W, Identification Trees and Nearest Neighbor) on a data set of Arch descriptions. The training data is attached on a separate page at the end of the quiz – feel free to tear it off for handy reference. It consists of eight samples: four arches and four near misses.

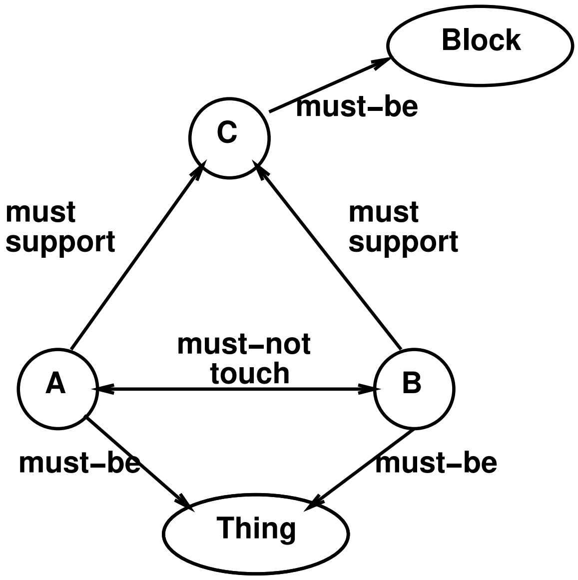
Procedure W operates on models represented as nets and both Identification Trees and Nearest Neighbors operate on feature vectors To enable comparisons, we have devised a simple way of encoding the Procedure W nets into binary (yes/no) feature vectors. Each of the features denotes the presence of a particular link in the corresponding Procedure W model. So, for example, the fact that A supports B in the sample is indicated by placing a yes in the `A-supports-B?` feature. Note that, in contrast to the text, we have assumed a symmetric `touch` relation and we don't have a `left-of` relation. We only show those features that are relevant to this particular data set; in general, many more features would be required.

Note that we have features indicating the classes, e.g. brick, cylinder, etc., of the objects in the description. There would be a feature corresponding to each class in the AKO tree and a yes entry for the feature corresponding to **every** class that the object belongs to. We have not shown them all since they are not all relevant to this data set. Throughout this problem we assume the following AKO tree:

```
          Thing
         /     \
     Block    Curved
    /    \    /    \
Brick  Wedge Cone Cylinder
```

Part A

(a) Choose a sequence of samples from the data set given above so that Procedure W learns the following concept for an Arch. For each sample, indicate the induction heuristic used, drawing from the following list as given in Chapter 16: forbid-link, require-link, climb-tree, enlarge-set, drop-link, and close-interval. You may or may not have to use all the entries in the data set; the number of spaces below is not intended as a hint.



Sample number	Induction heuristic
	initial model

(b) In general, is there a unique order in which the samples must be presented for Procedure W to learn a particular concept? Briefly, justify your answer.

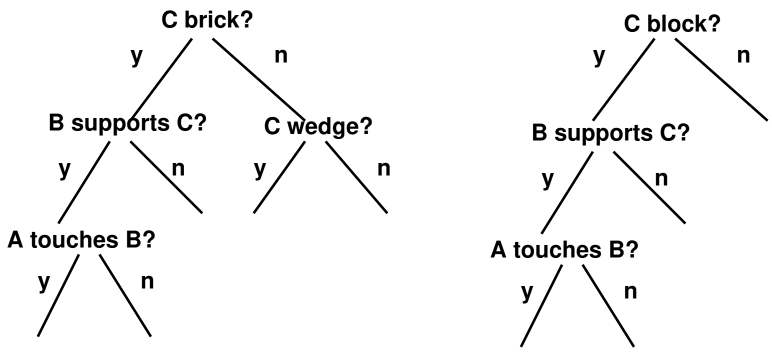
Part B

Now we proceed to build an Identification Tree for this data. In this problem, we will use the following simple measure of disorder for a set (assuming only two classes, e.g. arch and non-arch). Identify the most common class in the set (if it's balanced, pick randomly). Then, the disorder is the following ratio:

$$\frac{\text{number in less common class}}{\text{number in most common class}}$$

. Note that this measure is zero for a homogeneous set and one for an evenly split set, which is precisely what we want. We will further assume that the disorder for a test will be the simple average of the disorder for the branches. In this question, you should be able to make most of the disorder comparisons by inspection.

(a) Which of the following identification trees is consistent with this disorder measure and the procedure for growing trees described in the text? Explain briefly how you arrived at the answer. Hint: You should not need to do the disorder computations in detail.



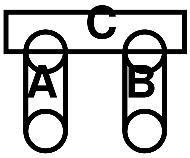
(b) Can we always build an Identification Tree from one of Procedure W's models? If so, explain the relationship between the models learnt by Procedure W and an Identification Tree. In particular, indicate what corresponds to a required (must) link and a forbidden (must not) link in the Identification Tree.

(c) Can we always build a Procedure W model from an Identification Tree built from the type of feature data we have been considering? If yes, explain why. If not, explain why not.

Part C

Suppose that we wanted to use Nearest Neighbor classification on the same Arch feature data. A natural metric to use is the, so called, Hamming distance — a count of the number of "bits" (yes/no features) that are different between two complete feature vectors.

(a) Consider the following sample:

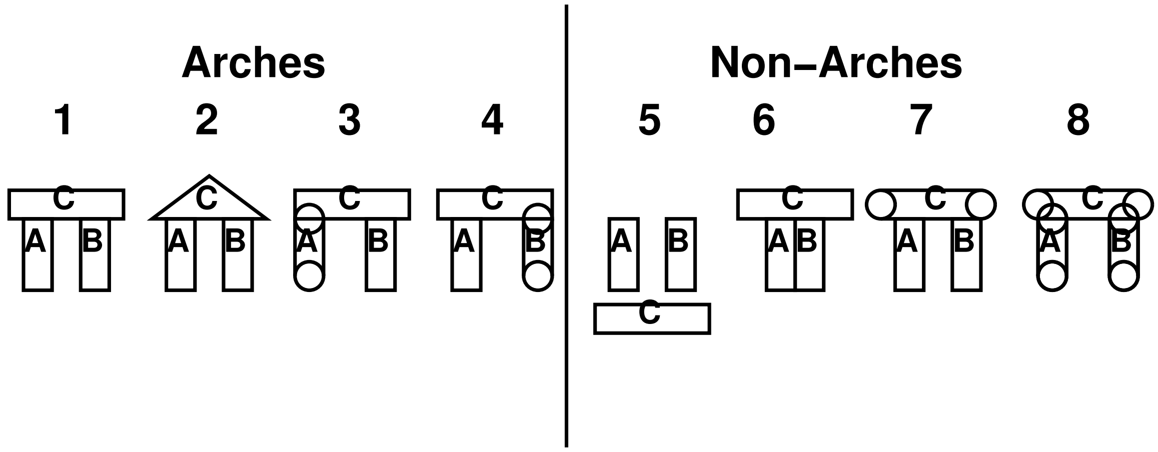


How would this sample be classified by a Nearest Neighbor (using single nearest neighbor) procedure using Hamming distance: Arch or Not-Arch. Is it well defined or is there a tie? Which sample or samples in the data set are nearest?

(b) Is the answer from Nearest Neighbor using this metric consistent with what Procedure W or the Identification Tree in Part B would give? Explain why the difference, if any.

(c) Based on the Identification Trees of Part B, suggest a variation to the Nearest Neighbor metric suggested above that might improve the performance of Nearest Neighbors in this domain.

Training Data for Problem 4



	1	2	3	4	5	6	7	8
A supports C?	y	y	y	y		y	y	y
B supports C?	y	y	y	y		y	y	y
A touches B?						y		
A brick?	y	y		y	y	y	y	
A block?	y	y		y	y	y	y	
A cylinder?			y					y
A curved?			y					y
B brick?	y	y	y		y	y	y	
B block?	y	y	y		y	y	y	
B cylinder?				y				y
B curved?				y				y
C brick?	y		y	y	y	y		
C wedge?		y						
C block?	y	y	y	y	y	y		
C cylinder?							y	y
C curved?							y	y