Notes with Partial Exam II (variant A)
Artificial Intelligence
Course code: 192140302
May 26, 2014

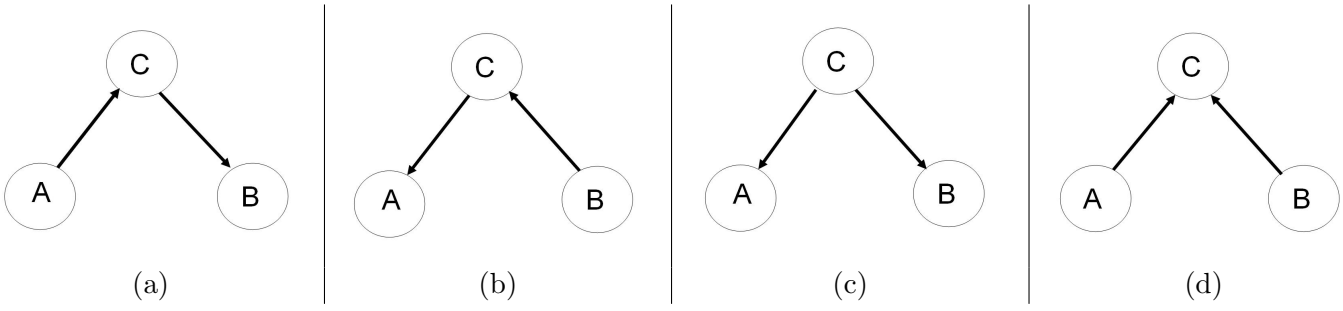
1. Figure 1 shows four Bayesian network structures on three nodes. The difference between the networks is in the direction of the arcs. For exactly one of these networks the conditional independency expressed by the equation

$$P(A|C) = P(A|C, B)$$

does NOT hold. Which one?

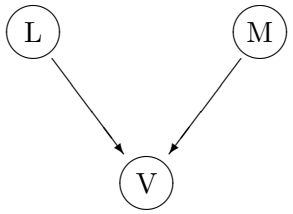
(a) a
(b) b
(c) c
(d) d

Answer: network d. This concerns the semantics of a Bayesian Network. The formula expresses the statement that "A and B are independent given C". See Ertel's book page 154, Figures 7.16 and page 155, Figure 7.17.



Figuur 1: Four Bayesian network structures.

2. In the Bayesian Network below with three boolean variables the probabilities for P and M are: $P(M = true) = 0,1$ and $P(L = true) = 0.7$ and the conditional probabilities for variable V are as shown in the table.



L	M	$P(V = true \mid L, M)$
true	true	0,9
true	false	0,5
false	true	0,3
false	false	0,05

What is the value of $P(V = false \mid L = false)$?

(a) 0.075

(b) 0.54

(c) 0.46

(d) 0.925

Answer: d (0.925).

Following the same strategy via computing the full joint probability and using the definition of conditional probability: $P(X|Y) = P(X,Y)/P(Y)$:

$P(V = f|L = f)$

= (summing out M)

$\sum_m P(V = f, M = m|L = f)$

= (definition cond. prob)

$\frac{\sum_m P(V=f,M=m,L=f)}{\sum_m \sum_v P(V=v,M=m,L=f)}$

= (use BN)

$\frac{\sum_m P(L=t)P(M=m)P(V=f|L=f,M=m)}{\sum_m \sum_v P(L=t)P(M=m)P(V=v|L=t,M=m)}$

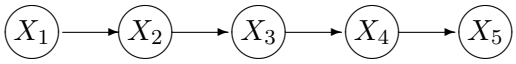
= (fill in from the CPT's in the BN)

0.925.

A short route:

$P(V = f|L = f) = \sum_m P(V = f|L = f, M = m).P(M = m) = 0.95 \times 0.9 + 0.7 \times 0.1 = 0.925$

3. Consider the following simple-chain Bayesian network, in which the five nodes represent five boolean-valued stochastic variables.



Based on the structure of the network one of the following statements is NOT generally true. Which one?

(a) Before computing $P(X_3|X_1)$ we may remove node X_5 from the network.
(b) Before computing $P(X_3|X_1)$ we may remove node X_4 as well as node X_5.
(c) $P(X_4|X_2) = P(X_4|X_1, X_2)$.
(d) $P(X_4|X_2) = P(X_4|X_5, X_2)$

Answer: d is NOT true. This question concerns the semantics of a Bayesian Network. Nothing is known about X_4 and X_5 so these nodes do not affect $P(X_3|X_1)$. This can easily be checked by working out the formula for $P(X_3|X_1)$. Formula c) says that if X_2 is given then X_4 and X_1 are independent. This correct. Formula d) says that if X_2 is given then X_4 and X_5 are independent. This is NOT correct: the values can be different; the structure does not guarantee their equality. See Ertel's book, page 154, Figures 7.16 and page 155, Figure 7.17.

4. Let D be a data set. A data element d in this set is a vector of k feature values $d =< v_0, ...v_{k-1} >$, where v_j is the value of the $j-th$ feature A_j of d. With $v_j(d)$ we denote the value of this $j-th$ component of the vector d. Let A_{k-1} be the class feature. The value of this class feature determines the class the data element belongs to.
Let A be a feature with n distinct possible values $a_i (i = 1..n)$. Let D_i be the following subset of D: $D_i = \{d \epsilon D | v_j(d) = a_i\}$. Thus D_i is the subset of D that contains those elements of D that have the feature value of feature A equal to the $i-th$ value of feature A.

For a data set D the InfoGain for feature A is defined as:

$$InfoGain(D, A) = H(D) - \sum_{i=1}^{i=n} \frac{|D_i|}{|D|}.H(D_i)$$

where $H(D)$ is the Shannon entropy of the probability distribution P that is the most likely estimator (i.e. determined by relative frequencies) of the probability distribution of the class values in the data set D. Similarly, $H(D_i)$ is the Shannon entropy of the subset D_i of D.
The Shannon Entropy of the probability distribution $p =< p_0, ...p_{n-1} >$ equals:

$$H(p) = -\sum_{i-1}^{i=n} p_i.log_2(p_i)$$

A retailer wants for marketing purposes distinguish between costumers younger than 35 (class Y) and customers older than 35 (class O). The following table summarizes the data set in the data base of the retailer in an abstract form. The relevant attributes, determined by domain knowledge, are for convenience denoted by A with values *a1*, *a2* and *a3*, B with values *b1* and *b2*, C with values *c1* and *c2* and D with values *d1* and *d2*

A	B	C	D	Number of Instances	
				Y	O
a1	b1	c1	d1	12	4
a2	b1	c1	d2	4	6
a3	b1	c1	d1	6	0
a1	b2	c1	d2	0	12
a2	b2	c1	d1	4	2
a3	b2	c1	d2	0	4
a1	b1	c2	d1	0	8
a2	b1	c2	d2	8	0
a3	b1	c2	d1	4	0
a1	b2	c2	d2	0	4
a2	b2	c2	d1	7	0
a3	b2	c2	d2	5	0

The analyst wants to learn the above classification problem using decision trees. If he uses "information gain" as selection criteria what will be the first attribute for splitting the examples?

(a) A (0.1519)

(b) B (0.0388)

(c) C (0.0244)

(d) D (0.0696)

Answer: a. Attribute A has the maximum info gain. Straightforward using the formula given.

5. Let $P(X)$ denote the probability that proposition X is true and $P(X|Y)$ the conditional probability that X is true given Y is true. Let A and B denote two arbitrary propositions. Which of the following two statements are true?

i) $(P(A|B) = 1) \Leftrightarrow (B \rightarrow P(A) = 1)$

ii) $P(A|B) = P(B \rightarrow A)$

a) only i is true

b) i as well as ii is true

c) only ii is true

d) neither i nor ii is true

Answer: d (both not true). i) is false. ii) is false. Suppose B is false, then, according the definition of $P(A|B)$ it is undefined. The right hand sides are true and 1, respectively. See also Table 7.1 on page 128 of Ertel's book.

6. Often accuracy is used as performance measure for a classifier. Which of the following statements is true?

   i. Accuracy is insufficient if we want to take into account the different costs of type I and type II errors (false negatives and false positives).

   ii. Accuracy is a more general measure than type I and type II errors, since type I and type II errors can only be used as a performance measure for binary classifications (i.e. classifications into two classes).

   (a) i and ii are both true

   (b) only i is true

   (c) only ii is true

   (d) i and ii are both false

   b) only I is true. ii) also for non binary classifications it makes sense to use type I and type II errors for each of the classes, independent how many there are, because for each of the classes one can compute how often these types of errors are made (for each class, say A, for each object, say O the classifier says: object O is in class A or it is not in class A, it is in one of the other classes).