Classification using Bayes Decision Theory

- In this approach classification is carried out using probabilities of classes.

- It is assumed that we know the *a priori* or the prior probability of each class. If we have two classes C_1 and C_2, then the prior probability for class C_1 is P_{C1} and the prior probability for class C_2 is P_{C2}.

- If the prior probability is not known, the classes are taken to be equally likely.

- If prior probability is not known and there are two classes C_1 and V_2, then it is assumed $P_{C1} = P_{C2} = 0.5$.

- If P_{C1} and P_{C2} are known, then when a new pattern x comes along, we need to calculate $P(C_1|x)$ and $P(C_2|x)$.

- The bayes theorem is used to compute $P(C_1|x)$ and $P(C_2|x)$.

- Then if $P(C_1|x) \geq P(C_2|x)$, the pattern is assigned to Class 1 and if $P(C_1|x) < P(C_2|x)$, it is assigned to Class 2. This is called the Bayes decision rule.

Bayes Rule

- If $P(C_i)$ is the prior probability of Class i, and $P(X|C_i)$ is the conditional density of X given class C_i, then the *a posteriori* or posterior probability of C_i is given by

  $P(C_i|X) = \frac{P(X|C_i)P(C_i)}{P(X)}$

- Bayes theorem provides a way of calculating the posterior probability $P(C_i \mid X)$. In other words, after observing X, the posterior probability that the class is c_i can be calculated from Bayes theorem. It is useful to convert prior probabilities to posterior probabilities.

- $P(X)$ is given by

  $P(X) = \sum_i P(X \mid C_i)P(C_i)$

- Let the probability that an elephant is black be 80% and that an elephant is white be 20%. This means P(elephant is black) = 0.8 and P(elephant is white) =0.2 . With only this information, any elephant will be classified as being black. This is because the probability of error in this case is only 0.2 as opposed to classifying the elephant as white which results in a probability of error of 0.8 . When additional information is available, it can be used along with the information above.

  If we have probability that elephant belongs to region X is 0.2 . Now if the elephant belongs to region X, we need to calculate the posterior probability that the elephant is white. i.e. P(elephant is white | elephant belongs to region) or $P(W \mid X)$. This can be calculated by using Bayes theorem. If 95% of the time when the elephant is white, it is because it belongs to the region X. Then

  $P(W \mid X) = \frac{P(X|W)*P(W)}{P(X)} = \frac{0.95*0.2}{0.2} = 0.95$

  The probability of error is 0.05 which is the probability that the elephant is not white given that it belongs to region X.

Minimum Error Rate Classifier

- If it is required to classify a pattern X, then the minimum error rate classifier classifies the pattern X to the class C which has the maximum posterior probability $P(C \mid X)$.
- If the test pattern X is classified as belonging to Class C, then the error in the classification will be (1 - $P(C \mid X)$).
- It is evident to reduce the error, X has to be classified as belonging to the class for which $P(C \mid X)$ is maximum.
- The expected probability of error is given by

$$\int_X (1 - P(C|X))P(X)dX$$

This is minimum when $P(C \mid X)$ is maximum (for a specified value of $P(X)$.

- Let us consider an example of how to use minimum error rate classifier for a classification problem. Let us consider an example with three classes *small*, *medium* and *large* with prior probability

  $P(small) = \frac{1}{3}$
  $P(medium) = \frac{1}{2}$
  $P(large) = \frac{1}{6}$

  We have a set of nails, bolts and rivets in a box and the three classes correspond to the size of these objects in the box.

  Now let us consider the class-conditional probabilities of these objects :

  For Class *small* we get

  $P(nail \mid small) = \frac{1}{4}$
  $P(bolt \mid small) = \frac{1}{2}$
  $P(rivet \mid small) = \frac{1}{4}$

  For Class *medium* we get

  $P(nail \mid medium) = \frac{1}{2}$
  $P(bolt \mid medium) = \frac{1}{6}$
  $P(rivet \mid medium) = \frac{1}{3}$

  For Class *large* we get

  $P(nail \mid large) = \frac{1}{3}$
  $P(bolt \mid large) = \frac{1}{3}$

$P(rivet \mid large) = \frac{1}{3}$

Now we can find the probability of the class labels given that it is a nail, bolt or rivet. For doing this we need to use Bayes Classifier. Once we get these probabilities, we can find the corresponding class labels of the objects.

$P(small \mid nail) = \frac{P(nail)|small)P(small)}{P(nail|small).P(small)+P(nail|medium).P(medium)+P(nail|large).P(large)}$

This will give

$P(small \mid nail) = \frac{\frac{1}{4}\cdot\frac{1}{3}}{\frac{1}{4}\cdot\frac{1}{3}+\frac{1}{2}\cdot\frac{1}{2}+\frac{1}{3}\cdot\frac{1}{6}} = 0.2143$

Similarly, we calculate $P(medium \mid nail)$ and we get

$P(medium \mid nail) = \frac{\frac{1}{2}.frac12}{\frac{1}{4}\cdot\frac{1}{3}+\frac{1}{2}\cdot\frac{1}{2}+\frac{1}{3}\cdot\frac{1}{6}} = 0.6429$

and also $P(large \mid nail)$

$P(large \mid nail) = \frac{\frac{1}{3}\cdot\frac{1}{6}}{\frac{1}{4}\cdot\frac{1}{3}+\frac{1}{2}\cdot\frac{1}{2}+\frac{1}{3}\cdot\frac{1}{6}} = 0.1429$

Since $P(medium \mid nail) > P(small \mid nail)$ and
$P(medium \mid nail) > P(large \mid nail)$

we classify nail as belonging to the class *medium*. The probability of error $P(error \mid nail) = 1 - 0.6429 = 0.3571$

In a similar way, we can find the posterior probability for bolt

$P(small \mid bolt) = \frac{\frac{1}{2}\cdot\frac{1}{3}}{\frac{1}{2}\cdot\frac{1}{3}+\frac{1}{6}\cdot\frac{1}{2}+\frac{1}{3}\cdot\frac{1}{6}} = 0.5455$

$P(medium \mid bolt) = \frac{\frac{1}{6}\cdot\frac{1}{2}}{\frac{1}{2}\cdot\frac{1}{3}+\frac{1}{6}\cdot\frac{1}{2}+\frac{1}{3}\cdot\frac{1}{6}} = 0.2727$

$P(large \mid bolt) = \frac{\frac{1}{3} \cdot \frac{1}{6}}{\frac{1}{2} \cdot \frac{1}{3} + \frac{1}{6} \cdot \frac{1}{2} + \frac{1}{3} \cdot \frac{1}{6}} = 0.1818$

Since $P(small \mid bolt) > P(medium \mid bolt)$ and
$P(small \mid bolt) > P(large \mid bolt)$

we classify bolt as belonging to the class *small* and the probability of error $P(error \mid bolt) = 1 - 0.5455 = 0.4545$

In a similar way, we can find the posterior probability for rivet

$P(small \mid rivet) = \frac{\frac{1}{4} \cdot \frac{1}{3}}{\frac{1}{4} \cdot \frac{1}{3} + \frac{1}{3} \cdot \frac{1}{2} + \frac{1}{3} \cdot \frac{1}{6}} = 0.2727$

$P(medium \mid rivet) = \frac{\frac{1}{3} \cdot \frac{1}{2}}{\frac{1}{4} \cdot \frac{1}{3} + \frac{1}{3} \cdot \frac{1}{2} + \frac{1}{3} \cdot \frac{1}{6}} = 0.5455$

$P(large \mid rivet) = \frac{\frac{1}{3} \cdot \frac{1}{6}}{\frac{1}{4} \cdot \frac{1}{3} + \frac{1}{3} \cdot \frac{1}{2} + \frac{1}{3} \cdot \frac{1}{6}} = 0.1818$

Since $P(medium \mid rivet) > P(small \mid rivet)$ and
$P(medium \mid rivet) > P(large \mid rivet)$

we classify bolt as belonging to the class *medium* and the probability of error $P(error \mid rivet) = 1 - 0.5455 = 0.4545$

Naive Bayes Classifier

- A naive bayes classifier is based on applying Bayes theorem to find the class of a pattern.
- The assumption made here is that every feature is class conditionally independent.
- Due to this assumption, the probabilistic classifier is simple.
- In other words, it is assumed that the effect of each feature on a given class is independent of the value of other features.
- Since this simplifies the computation, though it may not be always true, it is considered to be a naive classifier.
- Even though this assumption is made, the Naive Bayes Classifier is found to give results comparable in performance to other classifiers like neural network classifiers and classification trees.
- Since the calculations are simple, this classifier can be used for large databases where the results are obtained fast with reasonable accuracy.
- Using the minimum error rate classifier, we classify the pattern X to the class with the maximum posterior probability $P(c \mid X)$. In the naive bayes classifier, this can be written as

  $P(C \mid f_1, ..., f_d)$.

  where $f_1, ..., f_d$ are the features.
- Using Bayes theorem, this can be written as

  $P(C \mid f_1, ..., f_d) = \frac{P(C)\ P(f_1,...,f_d)|C}{p(f_1,...,f_d)}$

  Since every feature f_i is independent of every other feature f_j, for $j \neq i$, given the class

$P(f_i, f_j \mid C) = P(f_i \mid C)P(f_j \mid C)$

So we get,

$P(C, f_1, ..., f_d) = P(C)\ P(f_1 \mid C)\ P(f_2 \mid C)\ \cdots p(f_d \mid C)$

$=$

$$p(C) \prod_{i=1}^{d} p(f_i|C).$$

The conditional distribution over the class variable C is

$$p(C|f_1, \dots, f_n) = \frac{1}{Z} p(C) \prod_{i=1}^{n} p(f_i|C)$$

where Z is a scaling factor.

- The Naive Bayes classification uses only the prior probabilities of classes P(C) and the independent probability distributions $p(f_i \mid C)$.

Parameter Estimation

- In supervised learning, a training set is given. Using the training set, all the parameters of the bayes model can be computed.
- If n_C of the training examples out of n belong to Class C, then the prior probability of Class C will be

  $P(C) = \frac{n_C}{n}$

- In a class C, if n_1 samples take a range of values (or a single value) out of a total of n_C samples in the class, then the prior probability of the feature being in this range in this class will be

$P(f_1$ is in range (a,b)$) = \frac{n_1}{n_C}$

In case of the feature taking a small number of integer values, this can be calculated for each of these values. For example, it would be

$P(f_1$ is 6$) = \frac{n_2}{n_C}$

if n_2 of the n_C patterns of Class c take on the value 6.

- If some class and feature never occur together, then that probability will be zero. When this is multiplied by other probabilities, it may make some probabilities zero. To prevent this, it is necessary to give a small value of probability to every probability estimate.

- Let us estimate the parameters for a training set which has 100 patterns of Class 1, 90 patterns of Class 2, 140 patterns of Class 3 and 100 patterns of Class 4. The prior probability of each class can be calculated.
  The prior probability of Class 1 is

  $P(C_1) = \frac{100}{100+90+140+100} = 0.233$

  The prior probability of Class 2 is

  $P(C_2) = \frac{90}{100+90+140+100} = 0.210$

  The prior probability of Class 3 is

  $P(C_2) = \frac{140}{100+90+140+100} = 0.326$

  The prior probability of Class 4 is

$P(C_2) = \frac{100}{100+90+140+100} = 0.233$

Out of the 100 examples of Class 1, if we consider a particular feature f_1 and if 30 patterns take on the value 0, 45 take on the value 1 and 25 take on the value 2, then the prior probability that in Class 1 the feature f_1 is 0 is

$P(f_1 \text{ is } 0) = \frac{30}{100} = 0.03$

The prior probability that in Class 1 the feature f_1 is 1 is

$P(f_1 \text{ is } 1) = \frac{45}{100} = 0.45$

The prior probability that in Class 1 the feature f_1 is 2 is

$P(f_1 \text{ is } 2) = \frac{25}{100} = 0.25$

Example for Naive Bayes Classifier

Let us take an example dataset.

Consider the example given in decision trees given in the Table 1. We have a new pattern

money = 90, has-exams=yes, and weather=fine

We need to classify this pattern as either belonging to goes-to-movie=yes or goes-to-movie=no.
There are four examples out of 11 belonging to goes-to-movie=yes.
The prior probability of P(goes-to-movie=yes)= $\frac{4}{11}$= 0.364

The prior probability of P(goes-to-movie=no) = $\frac{7}{11}$ = 0.636

There are 4 examples with $money 50 - 150$ and goes-to-movie=no and 1 examples with $money < 50$ and goes-to-movie=yes. Therefore,

Money	Has-exams	weather	Goes-to-movie
25	no	fine	no
200	no	hot	yes
100	no	rainy	no
125	yes	rainy	no
30	yes	rainy	no
300	yes	fine	yes
55	yes	hot	no
140	no	hot	no
20	yes	fine	no
175	yes	fine	yes
110	no	fine	yes

Table 1: Example training data set

$P(money50 - 150 \mid goes - to - movie = yes) = \frac{1}{4} = 0.25$ and

$P(money50 - 150 \mid goes - to - movie = no) = \frac{4}{7} = 0.429$

There are 4 examples with has-exams=yes and goes-to-movie=no and 2 examples with has-exams=yes and goes-to-movie=yes. Therefore,

$P(has - exams \mid goes - to - movie = yes) = \frac{2}{4} = 0.5$

$P(has - exams \mid goes - to - movie = no) = \frac{4}{7} = 0.429$

There are 2 examples with weather=fine and goes-to-movie=no and 2 examples with weather=fine and goes-to-movie=yes. Therefore,

$P(weather = fine \mid goes - to - movie = yes) = \frac{2}{4} = 0.5$

$P(weather = fine \mid goes - to - movie = no) = \frac{2}{7} = 0.286$

Therefore

$$P(goes - to - movie = yes \mid X) = 0.364 * 0.25 * 0.5 * 0.5 = 0.023$$

$$P(goes - to - movie = no \mid X) = 0.636 * 0.429 * 0.429 * 0.286 = 0.033$$

Since $P(goes - to - movie = no \mid X)$ is larger, the new pattern is classified as belonging to the class goes-to-movie=no.

Bayesian Belief Network

- A Bayesian network is a graphical model of a situation which represents a set of variables and the dependencies between them by using probability.

- The nodes in a Bayesian network represent the variables and the directional arcs represent the dependencies between the variables. The direction of the arrows show the direction of the dependency.

- Each variable is associated with a conditional probability table which gives the probability of this variable for different values of the variables on which this node depends.

- Using this model, it is possible to perform inference and learning.

- Bayesian networks that model a sequence of variables varying in time are called dynamic Bayesian networks.

- Bayesian networks with decision nodes which solve decision problems under uncertainly are Influence diagrams.

- The graphical model will be a directed acyclic graph with the nodes representing the variables. The arcs will represent the dependencies between nodes. If there is a directed arc between A and B, then A is called the parent of B and B is a child of A.

- A variable which has no parent is a variable which does not depend on any other variable. It is a variable which is independent and is not conditioned on any other variable.

- The conditional probability table associated with each node gives the probability of this variable for different values of its parent nodes.

- If a node does not have any parents, the conditional probability table is very simple as there are no dependencies.

- The joint distribution of the variable values can be written as the product of the local distribution of each node and its parents. In other words, if $f_1, f_2, ..., f_d$ are the variables, the joint distribution

$P(f_1, f_2, ..., f_d)$ is given by

$P(f_1, f_2, ..., f_d) = \prod_{j=1}^{d} P(f_j \mid parents(f_j))$

Using the above equation, it is possible to get the joint distribution of the variables for different values.

- As as example, let us consider the following scenario.

  Lakshman travels by air if he is on an official visit. If he is on a personal visit, he travels by air if he has money. If he does not travel by plane, he travels by train but sometimes also takes a bus.

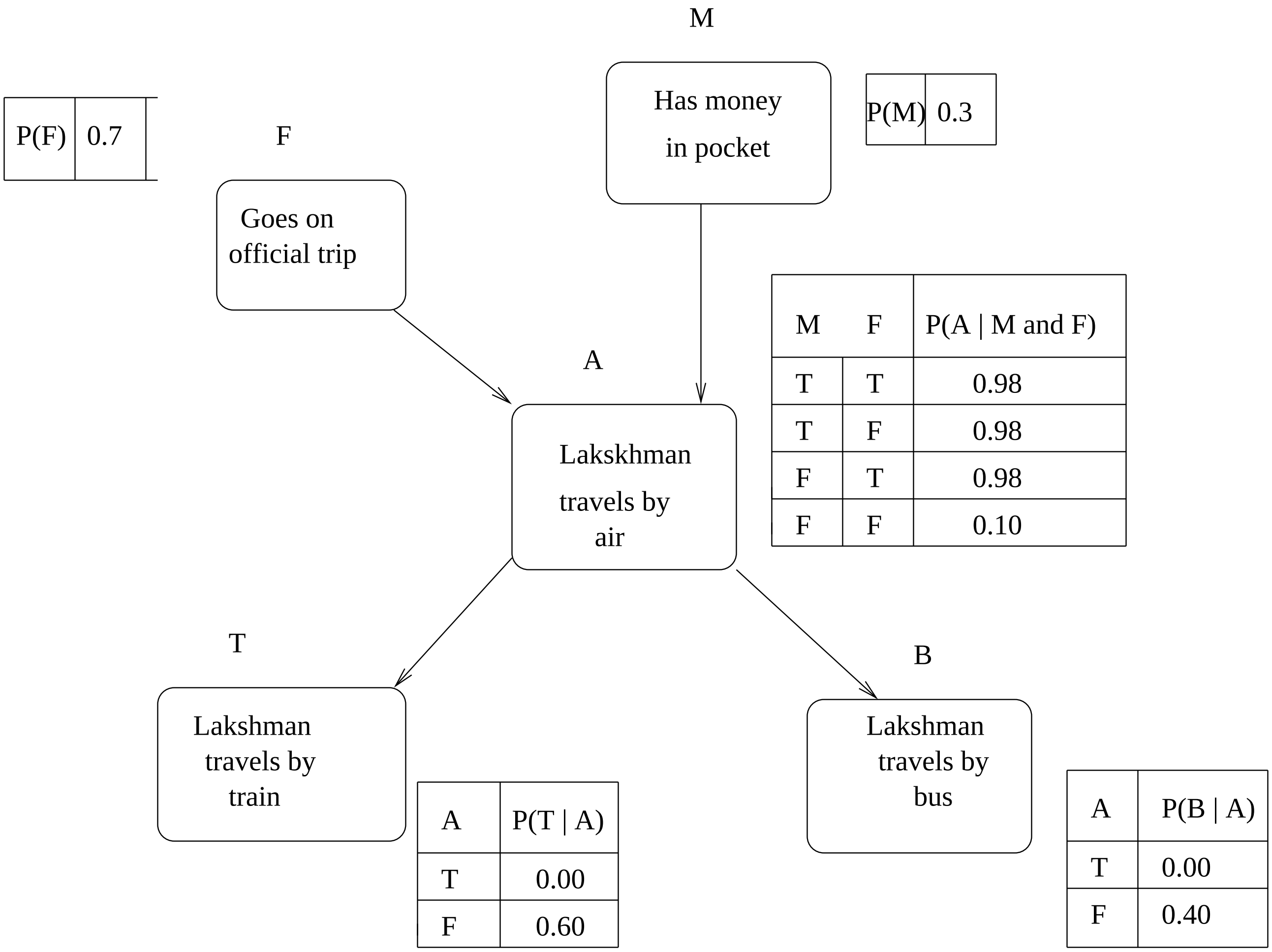
  The variables involved are :

1. Lakshman travels by air(A)
2. Goes on official visit(F)
3. Lakshman has money(M)
4. Lakshman travels by train(T)
5. Lakshman travels by bus(B)

This situation is converted into a belief network as shown in Figure 1.

In the graph, we can see the dependencies with respect to the variables. The probability values at a variable are dependent on the value of its parents. In this case, the variable A is dependent on F and M. The variable T is dependent on A and variable B is dependent on A. The variables F and M are independent variables which do not have any parent node. So their probabilities are not dependent on any other variable. Node A has the biggest conditional probability table as A depends on F and M. T and B depend on A.

Once the graph is drawn which is a directed acyclic graph (DAG), the probabilities of the variables depending on the values of their parent node has to be entered. This requires us to know the problem at hand and estimate



P(F)	0.7

P(M)	0.3

M	F	P(A | M and F)
T	T	0.98
T	F	0.98
F	T	0.98
F	F	0.10

A	P(T | A)
T	0.00
F	0.60

A	P(B | A)
T	0.00
F	0.40

Figure 1: Bayesian Belief Network

these probabilities. So for each node the Conditional Probability Table has to be entered.

First we take the independent nodes. Node F has a probability of $P(F) = 0.7$. Node M has a probability of $P(M) = 0.3$.

We next come to node A. The conditional probability table for this node can be represented as

F	M	P(A | F,M and P)
T	T	0.98
T	F	0.98
F	T	0.90
F	F	0.10

The conditional probability table for T can be represented as

A	P | A
T	0.0
F	0.6

The conditional probability table for B is

A	P | A
T	0.0
F	0.40

Using the bayesian belief network, we can get the probability of a combination of these variables. For example, we can get the probability that Lakshman travels by train, does not travel by air, goes on an official trip and has money. In other words, we are finding $P(T, \neg A, F, M)$. The probability of each variable given its parent is found and multiplied together to give the probability.

$P(T, \neg A, M, P) = P(T \mid \neg A) * P(\neg A \mid F \; and \; M) * P(F) * P(M)$
= 0.6 * 0.98 * 0.7 * 0.3 = 0.123

Assignment

1. Let the probability that a road is wet P(w) = 0.3. Let probability of rain, P(R) = 0.3. Given that 90% of the time when the roads are wet, it is because it has rained, and it has rained, calculate the *posterior* probability that the roads are wet.

2. Let *blue, green, and red* be three classes of objects with prior probabilities given by P(blue) = 0.3, P(green) = 0.4, P(red) = 0.3. Let there be three types of objects: *pencils, pens, and paper.* Let the class-conditional probabilities of these objects be given as follows. Use Bayes classifier to classify pencil, pen, and paper.

P(pencil|green) = 0.3	P(pen|green) = 0.5	P(paper|green) = 0.2
P(pencil|blue) = 0.5	P(pen|blue) = 0.2	P(paper|blue) = 0.3
P(pencil|red) = 0.2	P(pen|red) = 0.3	P(paper|red) = 0.5

3. Consider a two-class (Tasty or nonTasty) problem with the following training data. Use Naive Bayes classifier to classify
$Cook = Asha,\ Health - Status = Bad,\ Cuisine = Continental$

Cook	Health-Status	Cuisine	Tasty
Asha	Bad	Indian	Yes
Asha	Good	Continental	Yes
Sita	Bad	Indian	No
Sita	Good	Indian	Yes
Usha	Bad	Indian	Yes
Usha	Bad	Continental	No
Sita	Bad	Continental	No
Sita	Good	Continental	Yes
Usha	Good	Indian	Yes
Usha	Good	Continental	No

4. Consider the following dataset with three features f_1, f_2, and f_3. Consider the test pattern $f_1 = a$, $f_2 = c$, $f_3 = f$. Classify it using NNC and Naive Bayes Classifier.

f_1	f_2	f_3	Class Label
a	c	e	No
b	c	f	Yes
b	c	e	No
b	d	f	Yes
a	d	f	Yes
a	d	f	No

5. The profit a businessman makes depends on how fresh the provisions are. Further, if there a festival approaching, his profit increases. On the other hand, towards the end of the month, his sales come down. If he makes enough profit, he celebrates Diwali in a big way. Draw the belief network and suggest the likely conditional probability tables for all variables. Using this data, find the probability that the businessman celebrates Diwali big given that the provisions are fresh.

References

V. S. Devi and M. N. Murty (2011) *Pattern Recognition: An Introduction* Universities Press, Hyderabad.

R.O. Duda, P.E. Hart and D.G. Stork (2001) *Pattern Classification*, Second Edition, Wiley-Interscience.

S. Russell and P. Norvig (2003) *Artificial Intelligence : A Modern Approach*, Pearson India.

P. Domingos and M. Pazzani (1997) On the optimality of the simple Bayesian classifier under zero-one loss, *Machine Learning*, Vol. 29, pp. 103-130.

J. Pearl, (1988) *Probabilistic reasoning in intelligent systems, Morgan Kauffman.*

I. Rish *(2001) An empirical study of the naive Bayes classifier, IJCAI workshop on Empirical Methods in Artificial Intelligence*

D. Hackerman *(1996) Bayesian networks for knowledge discovery, In U.M. Fayyad, G.P. Shapiro, P. Smyth and R. Uthurusamy, editors Advances in Knowledge Discovery and Data Mining, MIT Press.*
R.E. Neopolitan, *(2003) Learning Bayesian Networks, Prentice Hall.*