1 Bayesian Network Inference

1.1 Factor

A factor is a function $\{0,1\}^n \to R$. For example, every CPT in the Bayesian networks are factor.

Multiply two factors.

$$f \cdot g(\mathbf{x}, \mathbf{y}, \mathbf{z}) = f(\mathbf{x}, \mathbf{y}) \cdot g(\mathbf{y}, \mathbf{z})$$

A	B	f
F	F	$f(F,F)$
F	T	$f(F,T)$
T	F	$f(T,F)$
T	T	$f(T,T)$

B	C	g
F	F	$g(F,F)$
F	T	$g(F,T)$
T	F	$g(T,F)$
T	T	$g(T,T)$

A	B	C	$f \cdot g$
F	F	F	$f(F,F) \cdot g(F,F)$
F	F	T	$f(F,F) \cdot g(F,T)$
F	T	F	$f(F,T) \cdot g(T,F)$
F	T	T	$f(F,T) \cdot g(T,T)$
T	F	F	$f(T,F) \cdot g(F,F)$
T	F	T	$f(T,F) \cdot g(F,T)$
T	T	F	$f(T,T) \cdot g(T,F)$
T	T	T	$f(T,T) \cdot g(T,T)$

1.2 Factor Sum-out

Given a factor f, we can sum out a variable

$$\sum_X f(\mathbf{y}) = f(x, \mathbf{y}) + f(\bar{x}, \mathbf{y})$$

Example:

A	B	f
F	F	$f(F,F)$
F	T	$f(F,T)$
T	F	$f(T,F)$
T	T	$f(T,T)$

B	$\sum_A f$
F	$f(F,F)+f(T,F)$
T	$f(F,T)+f(T,T)$

Observations: Say we have a BN over variables $X_1, ..., X_n$, whose CPTs (parameters) are $f_{X_1|P_1}, ...f_{X_n|P_n}$. The joint distribution is the factor multiplication of all the parameters.

$$Pr(x_1, ..., x_n) = f_{X_1|P_1}(x_1 \mid \mathbf{p}_1) \cdot f_{X_2|P_2}(x_2 \mid \mathbf{p}_2)... \cdot f_{X_n|P_n}(x_n \mid \mathbf{p}_n)$$

In BN $X_1 \to X_2 \to X_3$, we will have

$$f_{X_1|P_1} = Pr(A), f_{X_2|P_2} = Pr(B|A), f_{X_3|P_3} = Pr(C|B).$$

Say we have the joint distribution, $Pr(X_1, X_2, X_3)$, we can get a marginal on variables X_1, X_2 by sum-out X_3. In another word,

$$Pr(X_1, X_2) = \sum_{X_3} Pr(X_1, X_2, X_3)$$

1.3 Probability of an evidence in BN

Evidence is a partial assignment of variables, i.e. assignment of leaf variables.

1.3.1 Methods to calculate the probability of the evidence

- Use Truth table, as given in the last discussion. $O(2^n)$
- Use sum out operator.

1.3.2 How to use variable elimination order

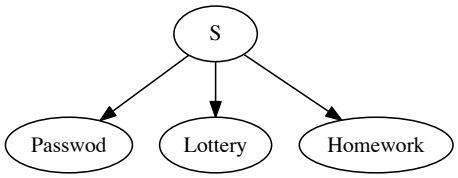
- Remove leave variables which is not appeared in the evidence, call it BN'.
- Let $f_1, f_2, ...$ be the CPTs of the BN'. For every CPT, if the input violates the evidence, we set the mapped value to 0.
- Decide a variable elimination order.
- For every variable X in the order, we find all the CPTs containing that variables. Multiply all them together and sum out X to form a new factor f_{new}. Add the f_{new} to the CPTs. All the factors now do not contain variable X.
- In the end, multiply all the constant factors together, and it will be the probability of the query.

1.4 How to find the marginal of a variable under evidence

We first calculate the $Pr(X = t, \mathbf{e})$ and $Pr(X = f, \mathbf{e})$. Then we just normalize the probability to get the marginal over X conditioned on evidence $\mathbf{e}$.

2 Naive Bayes

Class variable C and Feature variables $\mathbf{F} = \{F_1, ..., F_n\}$ What's the Markov Assumptions, Markovian Blankets.



S	$Pr(S)$
F	0.9
T	0.1

S	P	$Pr(P\|S)$
F	F	0.8
F	T	0.2
T	F	0.6
T	T	0.4

S	L	$Pr(L\|S)$
F	F	0.95
F	T	0.05
T	F	0.4
T	T	0.6

S	H	$Pr(H\|S)$
F	F	0.7
F	T	0.3
T	F	0.05
T	T	0.95

Say we know the email contains words Homework, but not Password, and we don't know whether it contains word Lottery. What is the probability of this event.

First, remove leaves which are not contained in the evidence. The leave Lottery is removed.

After filtering the CPTs, we update all the factors Let $f_1 = Pr(S), f_2 = Pr(P|S), f_3 = Pr(H|S)$.

S	$f_1 = Pr(S)$
F	0.9
T	0.1

S	P	$f_2 = Pr(P\|S)$
F	F	0.8
F	T	0
T	F	0.6
T	T	0

S	H	$f_3 = Pr(H\|S)$
F	F	0
F	T	0.3
T	F	0
T	T	0.95

Assert a elimination order H, P, S Eliminating H Factors containing H : f_3 So we sum-out H on factor f_3, and $f_4 =$

S	f_4
F	0.3
T	0.95

Now we have factors f_1, f_2, f_4 as f_3 already processed.
Eliminating P Factors containing $P : f_2$ so we sum-out P on factor f_2, and $f_5 =$

S	f_5
F	0.8
T	0.6

Now we have factors f_1, f_4, f_5 as f_5 is processed.
Eliminating S Factors containing $S : f_1, f_4, f_5$

S	$f_1 \cdot f_4 \cdot f_6$
F	$0.9 \cdot 0.3 \cdot 0.8 = 0.216$
T	$0.1 \cdot 0.95 \cdot 0.6 = 0.057$

Therefore $f_6 = \sum_S f_1 \cdot f_4 \cdot f_5$, f_6 is a constant factor with $f_6 = 0.216 + 0.057 = 0.273$.
After eliminating all variable, we only have one constant factor f_6. Therefore, the probability of the evidence is 0.273.

Sometimes, we want to know the probability of the email is a spam given the evidence. How? By asserting $\{\mathbf{e}, S = T\}$ and $\{\mathbf{e}, S = F\}$. It then gives the marginal probability $Pr(S|\mathbf{e})$.

3 Decision Tree Learning

3.1 Entropy

Given a distribution, entropy measures the uncertainty.

$$\mathsf{H}(X) = \sum_{\mathbf{x}} -f(\mathbf{x}) \log(f(\mathbf{x}))$$

It is the expected value of log probability.

For example,

A	B	$Pr(A, B)$
F	F	$Pr(F, F)$
F	T	$Pr(F, T)$
T	F	$Pr(T, F)$
T	T	$Pr(T, T)$

$$\begin{aligned}\mathsf{H}(A,B) = &-Pr(F,F)\cdot\log(Pr(F,F))\\ &-Pr(F,T)\cdot\log(Pr(F,T))\\ &-Pr(T,F)\cdot\log(Pr(T,F))\\ &-Pr(T,T)\cdot\log(Pr(T,T))\end{aligned}$$

If there is no uncertainty, i.e. $Pr(F,F)=1$ and all other world has $Pr=0$, then

$$\mathsf{H}(A,B) = -1\cdot 0 - 0\cdot\log(0) - ... = 0$$

If there is extreme uncertainty, i.e. Pr is an uniform distribution,

$$\mathsf{H}(A,B) = -1/4\cdot\log(1/4) - 1/4\cdot\log(1/4) - 1/4\cdot\log(1/4) - 1/4\cdot\log(1/4) = 2$$

When uncertainty increases, entropy will also increases.

3.2 Conditional Entropy

It measures the information gain knowing a new random variable.

$$H(A|B) = \sum_b Pr(b)\cdot H(A|B=b)$$

A	$Pr(A\|B=T)$
F	$\frac{Pr(F,T)}{Pr(F,T+Pr(T,T))}$
T	$\frac{Pr(T,T)}{Pr(F,T+Pr(T,T))}$

$$\begin{aligned}H(A|B=T) = &-\frac{Pr(F,T)}{Pr(F,T+Pr(T,T))}\cdot\log(\frac{Pr(F,T)}{Pr(F,T+Pr(T,T))})\\ &-\frac{Pr(T,T)}{Pr(F,T+Pr(T,T))}\cdot\log(\frac{Pr(T,T)}{Pr(F,T+Pr(T,T))})\end{aligned}$$

A	$Pr(A\|B=F)$
F	$\frac{Pr(F,F)}{Pr(F,F+Pr(T,F))}$
T	$\frac{Pr(T,F)}{Pr(F,F+Pr(T,F))}$

$$H(A|B=F) = -\frac{Pr(F,F)}{Pr(F,F+Pr(T,F))} \cdot \log(\frac{Pr(F,F)}{Pr(F,F+Pr(T,F))}) - \frac{Pr(T,F)}{Pr(F,F+Pr(T,F))} \cdot \log(\frac{Pr(T,F)}{Pr(F,F+Pr(T,F))})$$

Then, we will get $Pr(B=T) = Pr(F,T) + Pr(F,F)$, and $Pr(B=F) = Pr(F,F) + Pr(T,F)$. As a result, we have

$$H(A|B) = Pr(B=T)H(A|B=T) + Pr(B=F)H(A|B=F)$$

Note this property

$$H(X|Y) <= H(X)$$

3.3 Decision Tree

We have discrete variable $F_1, F_2, ..., F_n$ to represent features of an object, and we have a class variable C to represent the class of the corresponding object.

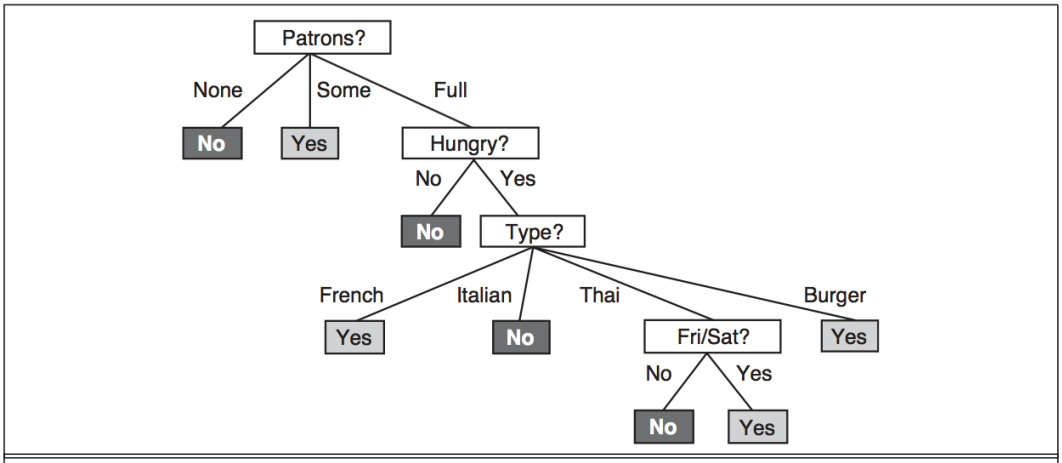


Figure 18.6 The decision tree induced from the 12-example training set.

3.4 Algorithm

3.5 Example

Data:

function DECISION-TREE-LEARNING($examples, attributes, parent\_examples$) **returns** a tree

if $examples$ is empty **then return** PLURALITY-VALUE($parent\_examples$)
else if all $examples$ have the same classification **then return** the classification
else if $attributes$ is empty **then return** PLURALITY-VALUE($examples$)
else
  $A \leftarrow \text{argmax}_{a \in attributes}$ IMPORTANCE($a, examples$)
  $tree \leftarrow$ a new decision tree with root test A
  for each value v_k of A **do**
    $exs \leftarrow \{e : e \in examples$ **and** $e.A = v_k\}$
    $subtree \leftarrow$ DECISION-TREE-LEARNING($exs, attributes - A, examples$)
    add a branch to $tree$ with label $(A = v_k)$ and subtree $subtree$
  return $tree$

F_1	F_2	C
F	F	F
F	F	T
F	T	F
T	T	T
T	F	T

We have five data points, and which is the best feature F_1 or F_2

C	$Pr(C\|F_1 = F)$
T	1/3
F	2/3

$$H(C|F_1 = F) = -1/3 \log(1/3) - 2/3 \log(2/3) = 0.92$$

C	$Pr(C\|F_1 = T)$
T	1
F	0

$$H(C|F_1 = T) = 0$$

$$H(C|F_1) = Pr(F_1 = T) \cdot H(C|F_1 = T) + Pr(F_1 = F) \cdot H(C|F_1 = F) = 2/5 * 0 + 3/5 * 0.92 = 0.552$$

What about $H(C|F_2)$

C	$Pr(C\|F_2 = F)$
T	2/3
F	1/3

$$H(C|F_2 = F) = -2/3 \log(2/3) - 1/3 \log(1/3) = 0.92$$

F_1	$Pr(C\|F_1 = T)$
T	1/2
F	1/2

$$H(C|F_2 = T) = -1/2 \log(1/2) - 1/2 \log(1/2) = 1$$

$$H(C|F_2) = Pr(F_2 = T) \cdot H(C|F_2 = T) + Pr(F_2 = F) \cdot H(C|F_2 = F) = 3/5 * 0.92 + 2/5 * 1 = 0.952$$

Which feature should we decide? F_1 as it has lower conditional entropy.