CMSC 421 Homework 4

Solutions

1. **[15 points]**

   A cmsc421 student notices that people that drive SUVs (S) consume large amounts of gas (G) and are involved in more accidents than the national average (A). They have constructed the following Bayesian network:

   (a) Compute $P(A)$ in two ways:

   i. By generating the *entire joint distribution* over these variables and explicitly summing the appropriate entries.

S	G	A	$P(S) \cdot P(A \mid S) \cdot P(G \mid S)$	$P(S, A, G)$
t	t	t	$0.3 \cdot 0.7 \cdot 0.6$	0.126
t	t	f	$0.3 \cdot 0.7 \cdot 0.4$	0.084
t	f	t	$0.3 \cdot 0.3 \cdot 0.6$	0.054
t	f	f	$0.3 \cdot 0.3 \cdot 0.4$	0.036
f	t	t	$0.7 \cdot 0.2 \cdot 0.3$	0.042
f	t	f	$0.7 \cdot 0.2 \cdot 0.7$	0.098
f	f	t	$0.7 \cdot 0.8 \cdot 0.3$	0.168
f	f	f	$0.7 \cdot 0.8 \cdot 0.7$	0.392

$$\begin{aligned} P(a) &= P(s,a,g) + P(s,a,\bar{g}) + P(\bar{s},a,g) + P(\bar{s},a,\bar{g}) \\ &= 0.39 \\ P(\bar{a}) &= P(s,\bar{a},g) + P(s,\bar{a},\bar{g}) + P(\bar{s},\bar{a},g) + P(\bar{s},\bar{a},\bar{g}) \\ &= 0.61 \end{aligned}$$

   ii. Using the variable elimination algorithm
   One way:

$$\begin{aligned} P(A) &= \sum_{S,G} P(S) \cdot P(A \mid S) \cdot P(G \mid S) \\ &= \sum_{S} P(S) P(A \mid S) \cdot F_G(S) \end{aligned}$$

   where

$$F_G(S) = \sum_{G} P(G \mid S) =$$

S	$F_G(S)$
t	1.0
f	1.0

$$\begin{aligned} P(A) &= \sum_{S} P(S) \cdot P(A \mid S) \cdot F_G(S) \\ &= \end{aligned}$$

A	$F_S(A)$
t	$0.3 \cdot 0.6 + 0.7 \cdot 0.3 = 0.39$
f	$0.3 \cdot 0.4 + 0.7 \cdot 0.7 = 0.61$

(b) Using conditional independence, compute $P(\bar{g}, a \mid s)$ and $P(\bar{g}, a \mid \bar{s})$. Then use Bayes rule to compute $P(s|\bar{g}, a)$.

$$\begin{aligned} P(\bar{g}, a \mid s) &= P(\bar{g}|a, s) \cdot P(a|s) \\ &= P(\bar{g}|s) \cdot P(a|s) \\ &= 0.3 \cdot 0.6 = 0.18 \end{aligned}$$

similarly for $\bar{s}$: $P(\bar{g}, a \mid \bar{s}) = 0.24$

Using Bayes rule:

$$P(s \mid \bar{g}, a) = \frac{P(\bar{g}, a \mid s)P(s)}{P(\bar{g}, a)} = 0.129$$

(c) The enterprising cmsc421 student notices that there are two types of people that drive SUVs, people from California (C) and people with large families (F). After collecting some statistics, the student arrives at the following form for the Bayesian network:

Using the chain rule from probability compute the probability $P(\bar{g}, a, s, c, \bar{f})$.

$$P(\bar{g}, a, s, c, \bar{f}) = P(\bar{g} \mid s)P(a \mid s)P(s \mid c, \bar{f})P(c)P(\bar{f}) = 0.2205$$

(d) Without explicitly generating the entire probability distribution, compute $P(s)$ and $P(s \mid c)$ and $P(s \mid \bar{c})$. (Hint: consider different values of F.)

$$P(S \mid C) = \sum_F P(S \mid C, F)$$

$$P(S) = \sum_C P(S \mid C)$$

(e) Show how you can compute the value of $P(G)$ in the complete distribution without doing any additional work (i.e., based solely on the work you have already done). Explain.

$$P(G) = \sum_S P(G \mid S)P(S)$$

(f) Using the rules for determining when two variables are independent of each other given evidence, answer the following (T/F) for the BN above:

- I(C,G) is false
- I(F,A | S) is true
- I(C,F) is true
- I(A,G) is false
- I(C,F | S) is false
- I(C,F | A) is false

2. **[15 points]**

(a) Suppose we have the query P(G) and the following elimination ordering: first A, then B, C, D, E, F. Show what the variable elimination algorithm will do on this graph by listing for each variable eliminated, the new factor generated and the factors used to generate the new factor. The factors should be written in the form $f_i(X_1, \ldots, X_k)$ for some set of variables $X_1, \ldots, X_k$.

var eliminated	factors used	factor generated
A	$P(C \mid A), P(D \mid A)P(A)$	$f_1(C, D)$
B	$P(D \mid B), P(B)$	$f_2(D)$
C	$P(E \mid C, D), f_1(C, D)$	$f_3(D, E)$
D	$P(D \mid G)f_2(D), f_3(D, E)$	$f_4(G, E)$
E	$P(F \mid E, G)f_4(G, E)$	$f_5(F, G)$
F	$f_5(F, G)$	$f_6(G)$

(b) Elimination ordering matters to the complexity of inference. Assume all variables have the same domain size $m > 2$. Find an elimination ordering for the same query whose time complexity as a function of m is worse than the ordering given above. As in the above problem, show what the factors generated at each step.

var eliminated	factors used	factor generated
E	$P(E \mid C, D), P(F \mid E, G)$	$f_1(C, D, F, G)$
B	$P(B)P(D \mid A, B)$	$f_2(A, D)$
A	$P(A)P(C \mid A)f_2(A, D)$	$f_3(C, D)$
D	$f_1(C, D, F, G), f_3(C, D), P(G \mid D)$	$f_4(C, F, G)$
C	$f_4(C, F, G)$	$f_5(F, G)$
F	$f_5(F, G)$	$f_6(G)$

There are a number of other possibilities.

3. **[15 points]** The expected monetary value of the lottery L is

$$EMV(L) = \frac{1}{50} \times \$10 + \frac{1}{2,000,000} \times \$1,000,000 = \$0.70$$

Although \$0.70 < \$1,it is not necessarily irrational to buy the ticket. First we will consider just the utilities of the monetary outcomes,ignoring the utility of actually playing the lottery game. Using U(S_{k+n}) to represent the utility to the agent of having n dollars more than the current state,and assuming that utility is linear for small values of money(i.e., U(S_{k+n})=n(U(S_{k+1}) -U(S_k)) for $-10 \leq n \leq 10$),the utility of the lottery is:

$$U(L) = \frac{1}{50}U(S_{k+10}) + \frac{1}{2,000,000}U(S_{k+1,000,000})$$

$$\approx \frac{1}{5}U(S_{k+1}) + \frac{1}{2,000,000}U(S_{k+1,000,000})$$

This is more than U(S_{k+1}) when U($S_{k+1,000,000}$) > $1,6000,000U(\$1)$. Thus,for a purchase to be rational(when only money is considered),the agent must be quite risk-seeking. This would be unusual for low-income individuals,for whom the price of a ticket is non-trivial. It is possible that some buyers do not internalize the magnitude of the very low probability of winning–to imagine an event is to assign it a "non-trivial" probability, in effect. Apparently,these buyers are better at internalizing the large magnitude of the prize. Such buyers are clearly acting irrationally.

Some people may feel their current situation is intolerable,that is, U(S_k) ≈U(S_{k+1}) ≈ $u_\perp$. Therefore the situation of having one dollar less would be equally intolerable,and it would be rational to gamble on a high payoff,even if one that has low probability.

Gamblers also derive pleasure from the excitement of the lottery and the temporary possession of at least a non-zero chance of wealth. So we should add to the utility of playing the lottery the term t to represent the thrill of participation. Seen this way,the lottery is just another form of entertainment,and buying a lottery ticket is no more irrational than buying a movie ticket. Either way,you pay your money,you get a small thrill t,and(most likely) you walk away empty-handed.(Note that it could be argued that doing this kind of decision- theoretic computation decreases the value of t. It is not clear if this is a good thing or a bad thing.)

4. **[15 points]**

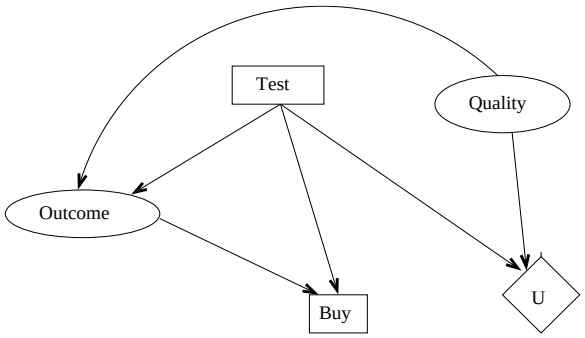
   (a) The Decision Network is described in the figure.

   (b) The expected net gain in \$ is

$$P(q^+)(200 - 1500) + P(q^-)(2000 - 2200) = 0.7 \times 500 + 0.3 \times -200 = 290$$

   (c) The question could also have been stated: Bayes theorem is used to compute P($p^+ \mid Pass$)) , etc., whereas conditionilization is sufficient to compute P(Pall)

$$P(Pass) = P(Pass \mid q^+))P(q^+) + P(Pass \mid q^- = 0.665$$



(d) If the car passes the test, the expected value of buying is

$$500P(q^+ \mid Pass) - 200P(q^- \mid Pass) = 378.92$$

If the car fails the test, the expected value of buying it is :

$$500P(q^+ \mid \neg Pass) - 200P(q^- \mid \neg Pass) = 92.53$$

Buying again is the best decision.

(e) Since the action is the same for both outcomes of the test, the test itself is worthless (if it is the only possible test) and the optimal plan is to buy the car without the test.

5. [**15 points**] Decision tree learning

(a) Information gain at root First we compute the information in the original data set. There are nine positive ("Play") instances and five negative ("Don't Play") instances, so we have:

$$I(T) = -9/14\log_2(9/14) - 5/14\log_2(5/14) = 0.94 \text{ bits} \quad (1)$$

If we use Outlook to subdivide T, the data set is divided into three sets. The Sunny set has two positive and three negative instances; Overcast has four positives; and Rain has three positive and two negative instances. The information of this split is:

$$\begin{aligned} I(T, Outlook) &= 5/14(-2/5\log_2(2/5) - 3/5\log_2(3/5)) \\ &+ \quad 4/14(-4/4\log_2(4/4) - 0/4log_2(0/4)) \\ &+ \quad 5/14(-3/5\log_2(3/5) - 2/5\log_2(2/5)) \\ &= \quad 0.69 \text{ bits} \end{aligned}$$

Applying the same technique to the Humidity attribute, we get:

$$\begin{aligned} I(T, Humidity) &= 9/14(-5/9\log_2(5/9) - 4/9\log_2(4/9)) \\ &+ \quad 5/14(-4/5\log_2(4/5) - 1/5\log_2(4/5)) \\ &= \quad 0.90 \text{ bits} \end{aligned}$$

The information gain associated with each attribute is the original data set's information minus the attribute's information:

$$Gain(T, Outlook) = I(T) - I(T, Outlook) = .25 \text{ bits}$$
$$Gain(T, Humidity) = I(T) - I(T, Outlook) = .045 \text{ bits}$$

(b) Gain ratio at root
To calculate the gain ratio, we have to compute the split information associated with each attribute. The Outlook attribute produces three subsets containing five, four, and five cases, so we have

$$\begin{aligned} & SplitInfo(T, Outlook) \\ &= -5/14 \log_2(5/14) - 4/14 \log_2(4/14) - 5/14 \log_2(5/14) \\ &= 1.58 \text{ bits} \end{aligned}$$

Computing SplitInfo for the Humidity attribute, we have:

$$\begin{aligned} & SplitInfo(T, Humidity) \\ &= -8/14 \log_2(8/14) - 6/14 \log_2(6/14) \\ &= .98 \text{ bits} \end{aligned}$$

The gain ratio is defined as

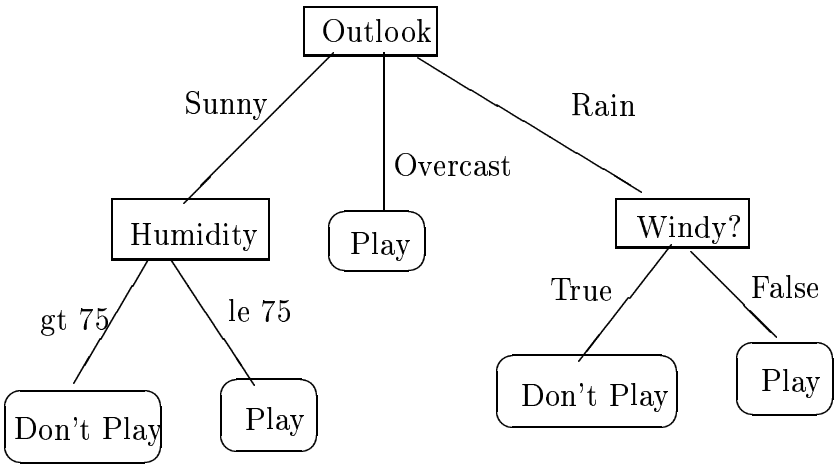
$$GainRatio(T, Attr) = \frac{Gain(T, attr)}{SplitInfo(T, attr)}$$

so we get

$$GainRatio(T, Outlook) = .25/1.58 = .158$$
$$GainRatio(T, Humidity) = .04/.98 = .046$$

Despite the "penalty" to Outlook of the higher split information (due to its larger domain size), it has a high enough gain that it also ends up having the highest gain ratio – so would be considered the best attribute to split on by either criterion.

(c) Decision tree There are three children at the first level. In order to create leaf nodes with instances belonging to a single class, the Sunny and Rain nodes must be further subdivided. The Sunny node can be split on Humidity, and the Rain node on Windy?, giving the following decision tree.



6. [**15 points**] Version Spaces

   (a) Initial G and S sets

   The G set is initialized to { `[any-size any-color any-shape]` }. (I'll use `[any any any]` as shorthand.)

   The S set contains all of the conjunctive concepts that use the leaf-level attribute values. There are 5 leaf values for `size`, 4 for `color`, and 4 for `shape`, for a total of $5 \cdot 4 \cdot 4 = 80$ hypotheses in the initial S set. `[jumbo blue cube]` is a representative member of the S set.

   (b) Processing I1

   After processing I1, a positive example with attribute values `[jumbo yellow ovoid]`, the G set is unchanged, since `[any any any]` covers I1. The only member of the initial S set that covers I1 is `[jumbo yellow ovoid]`, so the other elements are deleted from the S set. The resulting G and S sets are:

$$\begin{aligned} G &= \{[\texttt{any any any}]\} \\ S &= \{[\texttt{jumbo yellow ovoid}]\} \end{aligned}$$

   (c) Processing I2

   Processing I2, a negative example with attribute values `[jumbo red pyramid]`, has the effect of specializing the G set. After finding the set of most-general specializations of G, some elements are removed because they don't cover any members of S (indicated as a strike-through and a ** in the list below), and others need to specialized again in order to exclude I2 (indicated as a strike-through and a * in the list below). Note that in this process, several most-general specializations don't appear at all because they include I2 and can't be specialized any further to exclude I2 (e.g., `[jumbo any any]`.

   The S set is unchanged, since the single member of S doesn't cover I2.

   The resulting G and S sets are:

$$\begin{aligned} G = \{ \quad & [\texttt{medium any any}] ** \\ & [\texttt{small any any}] ** \\ & [\texttt{large any any}] * \end{aligned}$$

$$
\begin{aligned}
& [\texttt{big any any}] * * \\
& [\texttt{any dark any}] * \\
& [\texttt{any blue any}] * * \\
& [\texttt{any light any}] \\
& [\texttt{any any angular}] * \\
& [\texttt{any any cube}] * * \\
& [\texttt{any any curved}]\} \\
S = \{ \quad & [\texttt{jumbo yellow ovoid}]\}
\end{aligned}
$$

(d) Resulting version space

The version space includes 24 concepts:

```
jumbo yellow ovoid
large yellow ovoid
jumbo light ovoid
jumbo yellow curved
any yellow ovoid
large light ovoid
large yellow curved
any light ovoid
jumbo any ovoid
jumbo light curved
any yellow curved
jumbo yellow any
large any ovoid
large light curved
jumbo any curved
large yellow any
any any ovoid
any light curved
jumbo light any
large any curved
any yellow any
large light any
any light any
any any curved
```

Notice that there are a number of generalizations of `[jumbo yellow ovoid]`, such as `[large any any]`, that are not in the version space because they are not less general than either member of the G set.

(e) After I1 and I2 are processed,

(1) Any positive instance that is not covered by any member of the G set would cause the version space to collapse. An example is `[petite blue cube]`. Other instances are possible.

(2) Any negative instance that is covered by every member of the S set would cause the version space to collapse. An example is `[jumbo yellow ovoid]`. This

is the *only* negative instance that would cause a version space collapse at this point.

(3) Any positive instance that causes S to be generalized to exactly one member of G would cause the version space to be reduced to a single concept. An example is `[medium blue sphere]`. The set of examples with this property can be completely described by the expression `[medium|petite|tiny blue|red sphere]` $\vee$ `[medium|petite|tiny white cube|pyramid]`.

(4) To reduce the size of the version space to a single concept, a negative instance would have to specialize the G set down to a single member of S. In this case, since the S set only contains a single concept, the G set would have to be specialized to `[jumbo yellow ovoid]`. This is impossible with just one instance, since several specialization steps are required to move from the G set down to the S set in this case.

The following sequence of three negative examples would reduce the S set to `[jumbo yellow ovoid]`:

```
jumbo yellow sphere
jumbo white ovoid
big yellow ovoid
```

These instances are called *near miss* instances because they are each only "one feature away" from a positive instance.

7. [**15 points**] Statistical Learning

The Bayesian approach would be to take both drugs. The maximum likelihood approach would be to take the anti-B drug. In the case where there are two versions of B, the Bayesian still recommends taking both drugs, while the maximum likelihood is now to take the anti-A drug, since it has a 40% chance of being correct, versus 30% for each of the B cases. This is of course a toy example.