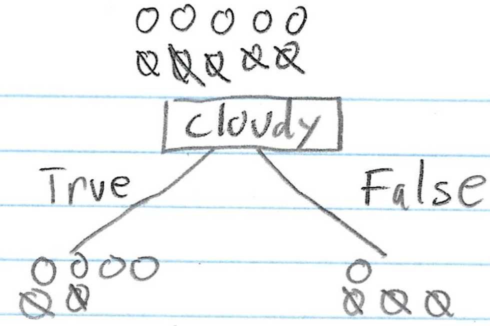
Homework 4

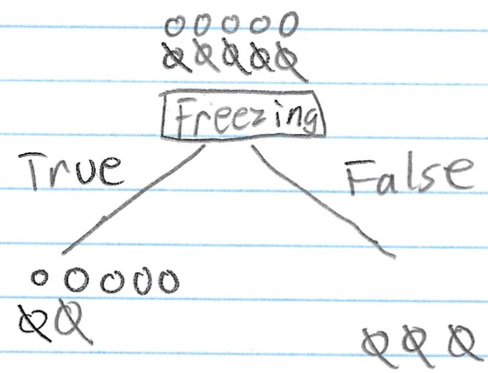
Michael Clausen

1. a)



$$\text{Gain(cloudy)} = 1 - \left[\tfrac{6}{10} I(4,2) + \tfrac{4}{10} I(1,3)\right] =$$

$$1 - [.6 \cdot .918 + .4 \cdot .811] = 1 - .8752 = .1248$$

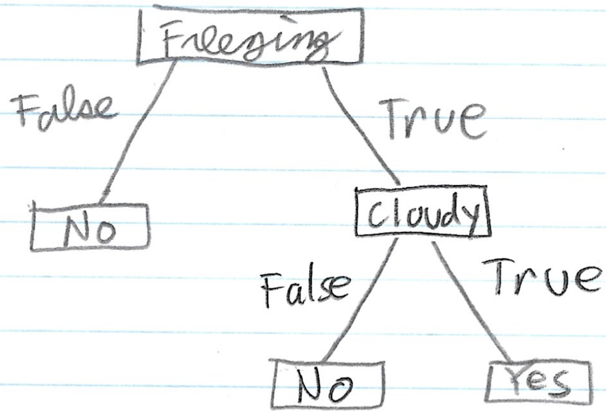


$$\text{Gain(freezing)} = 1 - \left[\tfrac{7}{10} I(5,2) + \tfrac{3}{10} I(0,3)\right] =$$

$$1 - [.7 \cdot .863 + .3 \cdot 0] = 1 - .6041 = .3959$$

$$\text{Gain(freezing)} > \text{gain(cloudy)}$$

Decision Tree



b) Classification Rules

(Freezing = False) → No

(freezing = True, cloudy = false) → No

(freezing = true, cloudy = true) → Yes

c) 90%

d) No, the combination of cloudy = false, freezing = true has 2 No's and 1 Yes. Just freezing = true or cloudy = false on their own have even lower accuracy.

2	0.918	1.000	0.971	0.918	0.863	0.811	0.764	0.722	0.684	0.650
3	0.811	0.971	1.000	0.985	0.954	0.918	0.881	0.845	0.811	0.779
4	0.722	0.918	0.985	1.000	0.991	0.971	0.946	0.918	0.890	0.863
5	0.650	0.863	0.954	0.991	1.000	0.994	0.980	0.961	0.940	0.918
6	0.592	0.811	0.918	0.971	0.994	1.000	0.996	0.985	0.971	0.954
7	0.544	0.764	0.881	0.946	0.980	0.996	1.000	0.997	0.989	0.977
8	0.503	0.722	0.845	0.918	0.961	0.985	0.997	1.000	0.998	0.991
9	0.469	0.684	0.811	0.890	0.940	0.971	0.989	0.998	1.000	0.998
10	0.439	0.650	0.779	0.863	0.918	0.954	0.977	0.991	0.998	1.000

b) Rewrite the decision tree as a list of rules.

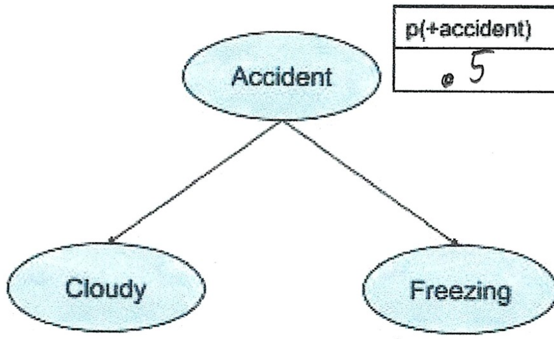
c) For your decision tree, what is the accuracy on the given dataset? (Hint: predict every case and compare predictions with the given Accident column)

d) Does there exist a decision tree (that only uses Cloudy and Freezing as input for its decision) that gives 100% accuracy on the given training data? Justify your answer (1-2 sentences).

2. Using the same data in Problem 1:

(a) Fill the conditional probability tables for the Bayesian Network shown below.

p(+accident)
.5



Cloudy	A	P
+	+	.4
+	-	.2
-	+	.1
-	-	.3

Accident	p(+cloudy|Accident)	-cloudy
True	.4\.5=.8	.2
False	.2\.5=.4	.6

Accident	p(+freezing|Accident)	
True	.5\.5=1	0
False	.2\.5=.4	.6

F	A	P
+	+	.5
+	-	.2
-	+	0
-	-	.3

2) b)

Cloudy	Freezing	Accident	P(C, F, A)
T	T	T	$.8(1)(.5) = .4$
T	T	F	$.4(.4)(.5) = .08$
T	F	T	0
T	F	F	$.4(.6)(.5) = .12$
F	T	T	$.2(1)(.5) = .1$
F	T	F	$.6(.4)(.5) = .12$
F	F	T	0
F	F	F	$.6(.6)(.5) = .18$

	P(-cloudy | Accident)
T	$.3 \backslash .5 = .2$
F	$.1 \backslash .5 = .6$

	P(-freezing | accident)
T	0
F	$.3 \backslash .5 = .6$

Instance

1. $P(-a|-c,-f) > P(+a|-c,-f)$?

$P(-a,-c,-f) = .18 > P(+a,-c,-f) = 0$ True

2. $P(-a|+c,-f) > P(+a|+c,-f)$?

$.12 > 0$ true

3. $P(-a|+f,-c) > P(+a|+f,-c)$?

$.04 \not> .1$ false

$\frac{7}{10} =$ 70% accuracy

4. $P(+a|+f,+c) > P(-a|+f,+c)$?

$.4 > 0$ true

5. same as 4, true

6. $P(-a|-f,+c) > P(+a|-f,+c)$?

$.12 > 0$ true

7. same as 3, false

8. $P(+a|+f,-c) > P(-a|+f,-c)$?

$.1 > .12$ false

9. same as 4, true

10. same as 4, true

3.

A	B	C	D	P(A, B, C, D)	
T	T	T	T	.4 (.3) (.1) (.95)	= .0114
T	T	T	F	.4 (.3) (.1) (.05)	= .0006
T	T	F	T	.4 (.3) (.9) (.95)	= .1026
T	T	F	F	.4 (.3) (.9) (.05)	= .0054
T	F	T	T	.4 (.7) (.6) (.98)	= .16464
T	F	T	F	.4 (.7) (.6) (.02)	= .00336
T	F	F	T	.4 (.7) (.4) (.98)	= .10976
T	F	F	F	.4 (.7) (.4) (.02)	= .00224
F	T	T	T	.6 (.99) (.1) (.95)	= .05643
F	T	T	F	.6 (.99) (.1) (.05)	= .00297
F	T	F	T	.6 (.99) (.9) (.95)	= .50787
F	T	F	F	.6 (.99) (.9) (.05)	= .02673
F	F	T	T	.6 (.01) (.6) (.98)	= .003528
F	F	T	F	.6 (.01) (.6) (.02)	= .000072
F	F	F	T	.6 (.01) (.4) (.98)	= .002352
F	F	F	F	.6 (.01) (.4) (.02)	= .000048

1. $P(+a, -b \mid -c, +d) = \dfrac{P(+a, -b, -c, +d)}{P(-c, +d)} =$

$\dfrac{.10976}{(.1026 + .10976 + .50787 + .002352)} =$

$\dfrac{.10976}{.722582} \approx .1519$

2. $P(+b \mid -c, +d) = \dfrac{P(+b, -c, +d)}{P(-c, +d)} =$

$\dfrac{(.1026 + .50787)}{.722582} \approx .8448$

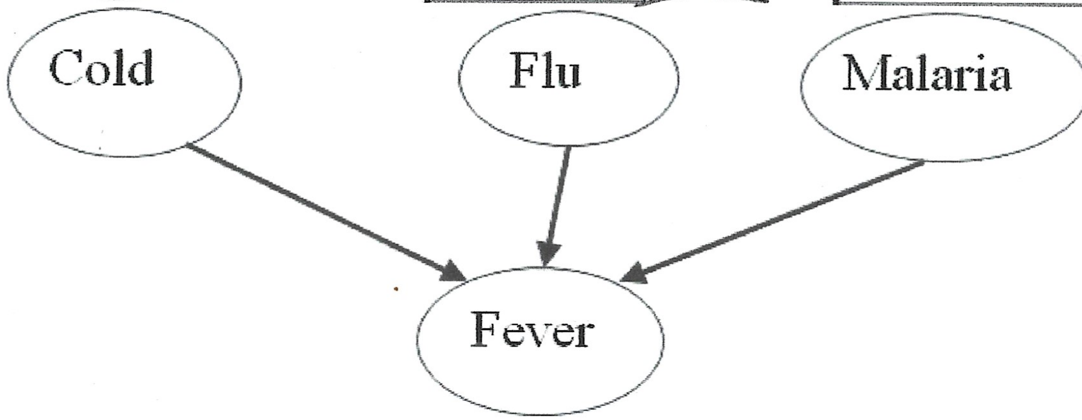
1) 1. false 2. false 3. true
4. false 5. false 6. true
7. true 8. false 9. true

5. Consider the Bayesian network shown below where the Conditional Probability table for the Fever variable is represented as a Noisy-OR (same network as in class work).

P(+cold)	-cold
0.4	.6

P(+flu)	-flu
0.2	.8

P(+malaria)	-malaria
0.1	.9



Inhibition probabilities for Fever:

- P(~Fever | cold, ~flu, ~malaria) = 0.6
- P(~Fever | ~cold, flu, ~malaria) = 0.2
- P(~Fever | ~cold, ~flu, malaria) = 0.1

(a) Complete the Conditional Probability table at Fever:

Cold	Flu	Malaria	P(+Fever | Cold,Flu,Malaria
False	False	False	0
False	False	True	0.9
False	True	False	1-.2=.8
False	True	True	1-(.2*.1)=.98
True	False	False	1-.6=.4
True	False	True	1-(.6*.1)=.94
True	True	False	1-(.6*.2)=.88
True	True	True	1-(.6*.1)=.94

(b) Calculate the probability of having Malaria given that Fever is observed.

(c) Calculate the probability of having Fever given that Malaria is observed (and it is not known if the person has Cold or Flu).

You may find it simpler to first complete the full Joint Probability table represented by the Bayesian network even if every row is ***not*** required to answer the two questions.

Cold	**Flu**	**Malaria**	**Fever**	**P(Cold,Flu,Malaria,Fever)**
False	False	False	False	0.432
False	False	False	True	0
False	False	True	False	.6(.8)(.1)(1-.9) = .0048
False	False	True	True	.6(.8)(.1)(.9) = .0432
False	True	False	False	.6(.2)(.9)(1-.8) = .0216
False	True	False	True	.6(.2)(.9)(.8) = .0864
False	True	True	False	.6(.2)(.1)(1-.98) = .00024
False	True	True	True	.6(.2)(.1)(.98) = .01176
True	False	False	False	.4(.8)(.9)(1-.4) = .1728
True	False	False	True	.4(.8)(.9)(.4) = .1152
True	False	True	False	.4(.8)(.1)(1-.94) = .00192
True	False	True	True	.4(.8)(.1)(.94) = .03008
True	True	False	False	.4(.2)(.9)(1-.88) = .0086
True	True	False	True	.4(.2)(.9)(.88) = .06336
True	True	True	False	.4(.2)(.1)(1-.94) = .00048
True	True	True	True	.4(.2)(.1)(.94) = .00752

6. Assume that the observed data between Smoking and Cancer can be summarized by the following joint probability table:

5) b) $P(+\text{malaria} \mid +\text{fever}) = \dfrac{P(+\text{malaria}, +\text{fever})}{P(+\text{fever})} =$

$$\frac{(.0432 + .01176 + .00192 + .00752)}{(0 + .0432 + .08644 + .01176 + .1152 + .03008 + .06336 + .00752)} =$$

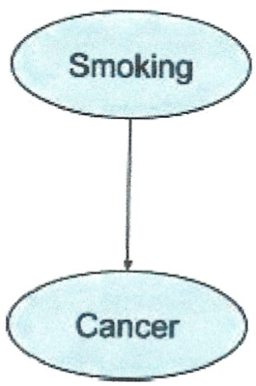
$$\frac{.0644}{.42544} \approx .1514$$

c) $P(+\text{fever} \mid +\text{malaria}) = \dfrac{P(+f, +m)}{P(+m)} =$

$$\frac{.0644}{.1} \approx .644$$

Smoking	Cancer	P(Smoking, Cancer)
True	True	0.08
True	False	0.10
False	True	0.02
False	False	0.80

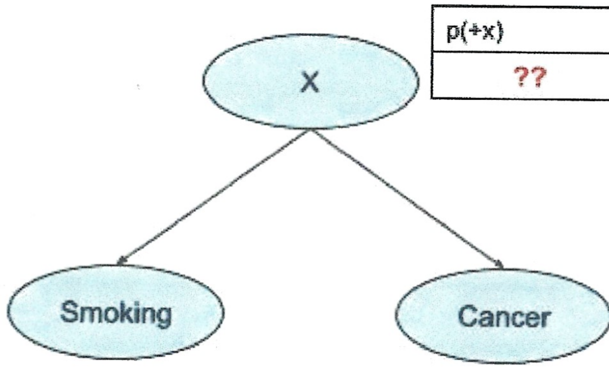
This can be modeled using the following Bayesian network (can verify that the probability distribution matches the above joint distribution).



p(+smoking)
0.18

Smoking	p(+cancer | Smoking)
True	0.4444
False	0.0244

The tobacco industry proposes an alternate model that includes an unknown/hidden "X" factor (perhaps genetic or environment-related):



p(+x)
??

.875

X	p(+smoking | X)
True	?? .0914
False	?? .8

~~Accident~~ X	p(+cancer | X)
True	0
False	0.8

6)

Smoking	Cancer	X	P(smoking, cancer, x)
+	+	+	$x \cdot s1 \cdot 0$
+	+	−	$(1-x) \cdot s2 \cdot .8 = .08$
+	−	+	$x \cdot s1 \cdot 1$
+	−	−	$(1-x) \cdot s2 \cdot .2 = .1$
−	+	+	$x \cdot (1-s1) \cdot 0$
−	+	−	$(1-x) \cdot (1-s2) \cdot .8 = .02$
−	−	+	$x \cdot (1-s1) \cdot 1$
−	−	−	$(1-x) \cdot (1-s2) \cdot .2 = .8$

1. $\cancel{(x \cdot s1 \cdot 0)} + ((1-x) \cdot s2 \cdot .8) = .08$
2. $(x \cdot s1 \cdot 1) + ((1-x) \cdot s2 \cdot .2) = .1$
3. $\cancel{(x \cdot s3 \cdot 0)} + ((1-x) \cdot (1-s2) \cdot .8) = .02$
4. $(x \cdot (1-s1)) + ((1-x) \cdot (1-s2) \cdot .2) = .8$

1. $S2 = \frac{.08}{.8(1-x)} = \frac{.1}{1-x}$

3. $(1-S2)(1-x) = .025$

$(1-x) - S2 \cdot (1-x) = .025$

$(1-x) - \frac{.1}{\cancel{(1-x)}} \cancel{(1-x)} = .025$

$1 - x - .1 = .025$

$.9 - x = .025$

$x = .875$

1. $S2 = \frac{.08}{.8(1-.875)} = .8$

2. $.875(S1) + (.8(1-.875) \cdot .2) = .1$

$.875(S1) + .02 = .1$

$\frac{.875(S1) = .08}{.875}$

$S1 \approx .0914$