

- You have approximately 3 hours.
- The exam is closed book, closed notes except a one-page crib sheet.
- Please use non-programmable calculators only.
- Mark your answers ON THE EXAM ITSELF. If you are not sure of your answer you may wish to provide a *brief* explanation. All short answer sections can be successfully answered in a few sentences AT MOST.

First name	
Last name	
SID	
EdX username	
First and last name of student to your left	
First and last name of student to your right	

For staff use only:

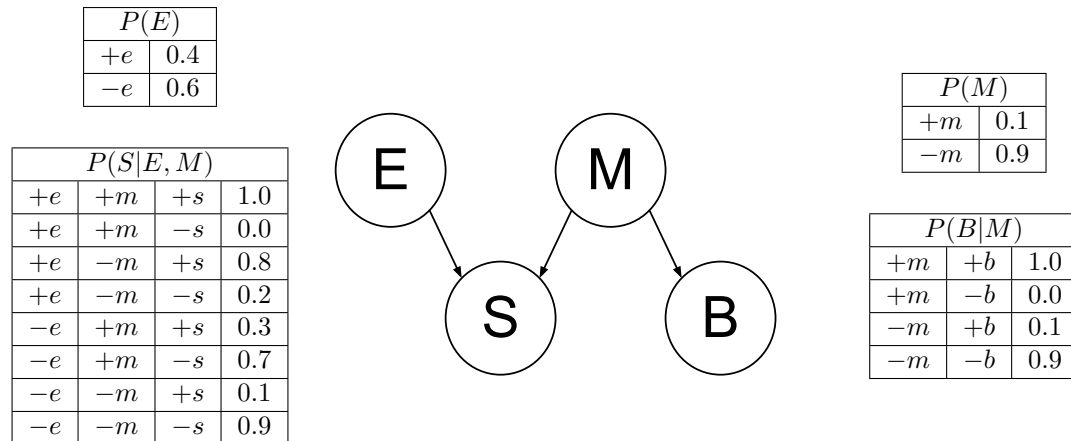
Q1. December 21, 2012	/10
Q2. Bayes' Nets Representation	/16
Q3. Variable Elimination	/13
Q4. Bayes' Nets Sampling	/10
Q5. Probability and Decision Networks	/15
Q6. Election	/12
Q7. Naïve Bayes Modeling Assumptions	/6
Q8. Model Structure and Laplace Smoothing	/7
Q9. ML: Short Question & Answer	/11
Total	/100

THIS PAGE IS INTENTIONALLY LEFT BLANK

Q1. [10 pts] December 21, 2012

A smell of sulphur (S) can be caused either by rotten eggs (E) or as a sign of the doom brought by the Mayan Apocalypse (M). The Mayan Apocalypse also causes the oceans to boil (B). The Bayesian network and corresponding conditional probability tables for this situation are shown below. For each part, you should give either a numerical answer (e.g. 0.81) or an arithmetic expression in terms of numbers from the tables below (e.g. $0.9 \cdot 0.9$).

Note: be careful of doing unnecessary computation here.



(a) [2 pts] Compute the following entry from the joint distribution:

$$P(-e, -s, -m, -b) = P(-e)P(-m)P(-s|-e, -m)P(-b|-m) = (0.6)(0.9)(0.9)(0.9) = 0.4374$$

by expanding the joint according to the chain rule of conditional probability.

(b) [2 pts] What is the probability that the oceans boil?

$$P(+b) = P(+b|+m)P(+m) + P(+b|-m)P(-m) = (1.0)(0.1) + (0.1)(0.9) = 0.19$$

by marginalizing out m according to the law of total probability.

(c) [2 pts] What is the probability that the Mayan Apocalypse is occurring, given that the oceans are boiling?

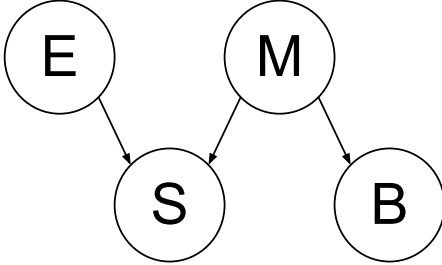
$$P(+m|+b) = \frac{P(+b|+m)P(+m)}{P(+b)} = \frac{(1.0)(0.1)}{0.19} \approx .5263$$

by the definition of conditional probability.

The figures and table below are identical to the ones on the previous page and are repeated here for your convenience.

$P(E)$	
$+e$	0.4
$-e$	0.6

$P(S E, M)$			
$+e$	$+m$	$+s$	1.0
$+e$	$+m$	$-s$	0.0
$+e$	$-m$	$+s$	0.8
$+e$	$-m$	$-s$	0.2
$-e$	$+m$	$+s$	0.3
$-e$	$+m$	$-s$	0.7
$-e$	$-m$	$+s$	0.1
$-e$	$-m$	$-s$	0.9



$P(M)$	
$+m$	0.1
$-m$	0.9

$P(B M)$		
$+m$	$+b$	1.0
$+m$	$-b$	0.0
$-m$	$+b$	0.1
$-m$	$-b$	0.9

- (d) [2 pts] What is the probability that the Mayan Apocalypse is occurring, given that there is a smell of sulphur, the oceans are boiling, and there are rotten eggs?

$$P(+m | +s, +b, +e) =$$

$$\begin{aligned}
 \frac{P(+m, +s, +b, +e)}{\sum_m P(m, +s, +b, +e)} &= \frac{P(+e)P(+m)P(+s | +e, +m)P(+b | +m)}{\sum_m P(+e)P(m)P(+s | +e, m)P(+b | m)} \\
 &= \frac{(0.4)(0.1)(1.0)(1.0)}{(0.4)(0.1)(1.0)(1.0) + (0.4)(0.9)(0.8)(0.1)} \\
 &= \frac{0.04}{0.04 + 0.0288} \approx .5814
 \end{aligned}$$

- (e) [2 pts] What is the probability that rotten eggs are present, given that the Mayan Apocalypse is occurring?

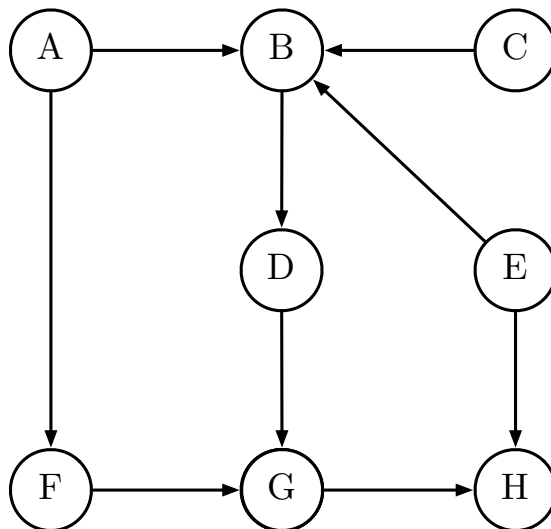
$$P(+e | +m) = P(+e) = 0.4$$

The first equality holds true as we have $E \perp\!\!\!\perp M$ (E is independent of M), which can be inferred from the graph of the Bayes' net.

Q2. [16 pts] Bayes' Nets Representation

(a) [6 pts] Graph Structure: Conditional Independence

Consider the Bayes' net given below.



Remember that $X \perp\!\!\!\perp Y$ reads as “ X is independent of Y given nothing”, and $X \perp\!\!\!\perp Y \mid \{Z, W\}$ reads as “ X is independent of Y given Z and W .”

For each expression, fill in the corresponding circle to indicate whether it is True or False.

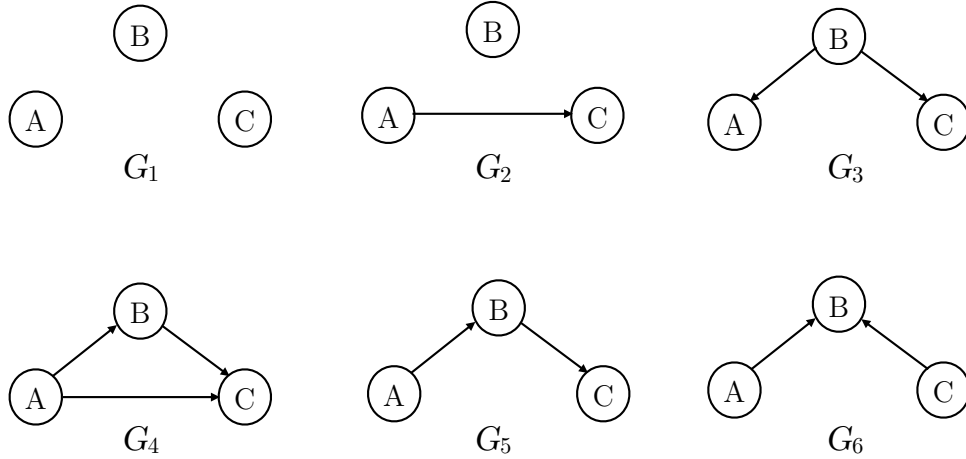
- (i) ☐ True ☒ False It is guaranteed that $A \perp\!\!\!\perp B$
 The edge between A and B implies that independency is not guaranteed.
- (ii) ☒ True ☐ False It is guaranteed that $A \perp\!\!\!\perp C$
 No active paths.
- (iii) ☐ True ☒ False It is guaranteed that $A \perp\!\!\!\perp D \mid \{B, H\}$
 An active path: $A \rightarrow F \rightarrow G$ (descendant H observed) $\leftarrow D$.
- (iv) ☒ True ☐ False It is guaranteed that $A \perp\!\!\!\perp E \mid F$
 No active paths.
- (v) ☐ True ☒ False It is guaranteed that $G \perp\!\!\!\perp E \mid B$
 An active path: $G \leftarrow F \leftarrow A \rightarrow B$ (observed) $\leftarrow E$.
- (vi) ☐ True ☒ False It is guaranteed that $F \perp\!\!\!\perp C \mid D$
 An active path: $F \leftarrow A \rightarrow B$ (descendant D observed) $\leftarrow C$.
- (vii) ☒ True ☐ False It is guaranteed that $E \perp\!\!\!\perp D \mid B$
 No active paths.
- (viii) ☐ True ☒ False It is guaranteed that $C \perp\!\!\!\perp H \mid G$
 An active path: $C \rightarrow B$ (descendant G observed) $\leftarrow E \rightarrow H$.

(b) Graph structure: Representational Power

Recall that any directed acyclic graph G has an associated family of probability distributions, which consists of all probability distributions that can be represented by a Bayes' net with structure G .

For the following questions, consider the following six directed acyclic graphs:

In general, the absence of an edge implies independence but the presence of an edge does not guarantee dependence. For a Bayes' net to represent a joint distribution, it can only make a subset of the conditional independence assumptions given by the joint. If a Bayes' net makes more independence assumptions than the joint, its family of distributions is not guaranteed to include the joint distribution because the Bayes' net family is constrained by more independence relationships. For instance G_1 can only represent the completely independent joint $P(A, B, C) = P(A)P(B)P(C)$.



- (i) [2 pts] Assume all we know about the joint distribution $P(A, B, C)$ is that it can be represented by the product $P(A|B, C)P(B|C)P(C)$. Mark each graph for which the associated family of probability distributions is guaranteed to include $P(A, B, C)$.

☐ G_1

☐ G_2

☐ G_3

☒ G_4

☐ G_5

☐ G_6

G_4 is fully connected, and is therefore able to represent any joint distribution. The others cannot represent $P(A|B, C)P(B|C)P(C)$ because they make more independence assumptions, which you can verify. For example, G_3 assumes $A \perp\!\!\!\perp C|B$ but this is not given by the joint.

- (ii) [2 pts] Now assume all we know about the joint distribution $P(A, B, C)$ is that it can be represented by the product $P(C|B)P(B|A)P(A)$. Mark each graph for which the associated family of probability distributions is guaranteed to include $P(A, B, C)$.

☐ G_1

☐ G_2

☒ G_3

☒ G_4

☒ G_5

☐ G_6

G_1 assumes all variables are independent, G_2 B is independent of the others, and G_6 assumes $A \perp\!\!\!\perp C$.

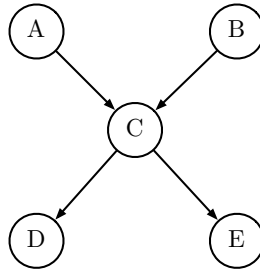
(c) **Marginalization and Conditioning**

Consider a Bayes' net over the random variables A, B, C, D, E with the structure shown below, with full joint distribution $P(A, B, C, D, E)$.

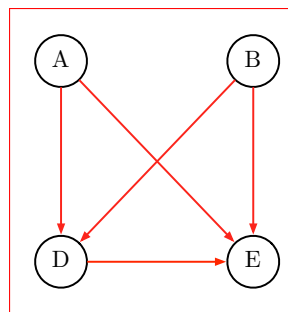
The following three questions describe different, unrelated situations (your answers to one question should not influence your answer to other questions).

Marginalization renders the neighbors of the marginalized out variable dependent.

Conditioning fixes the observed variables and renders their ancestors dependent according to the rules of d-separation.



- (i) [2 pts] Consider the marginal distribution $P(A, B, D, E) = \sum_c P(A, B, c, D, E)$, where C was eliminated. On the diagram below, draw the minimal number of arrows that results in a Bayes' net structure that is able to represent this marginal distribution. If no arrows are needed write "No arrows needed."



The high level overview for these types of problems is that the resultant graph must be able to encode the same conditional independence assumptions from the initial Bayes' net we have. For example, let's look at the BN above. We see the following independence assumptions:

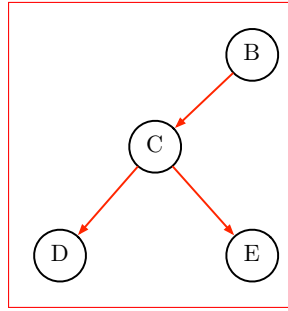
- $A \perp B$
- $A \perp D|C$
- $B \perp D|C$
- $A \perp E|C$
- $B \perp E|C$
- $D \perp E|C$

When we marginalize out C , we remove C from the graph. The conditional independence assumptions involving C no longer matter, so we just need to preserve:

$$A \perp B$$

To do this, we cannot have an edge between A and B . A and B must also be D-separated in the resultant BN, which it is in the solution. Every other edge is fair game because we don't make any other conditional independence assumptions.

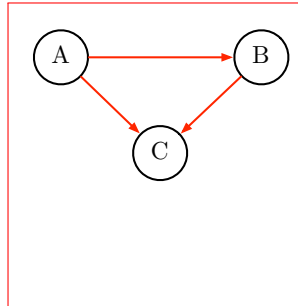
If you think about it, having $E \rightarrow D$ or $D \rightarrow E$ will fit the requirement above (it's also a valid point to say that the BN is symmetrical so the direction should not matter.) However, the arrow between A and D matters because we want ADB to be a common effect triple, which is an inactive triple if the middle node is unobserved, hence preserving the $A \perp B$ requirement.



- (ii) [2 pts] Assume we are given an observation: $A = a$. On the diagram below, draw the minimal number of arrows that results in a Bayes' net structure that is able to represent the conditional distribution $P(B, C, D, E \mid A = a)$. If no arrows are needed write "No arrows needed."

Observing A fixes its value and removes it from the Bayes' net. By d-separation no further dependence is introduced.

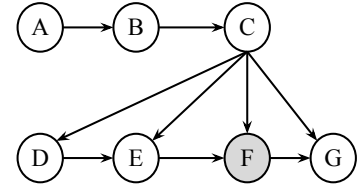
- (iii) [2 pts] Assume we are given two observations: $D = d, E = e$. On the diagram below, draw the minimal number of arrows that results in a Bayes' net structure that is able to represent the conditional distribution $P(A, B, C \mid D = d, E = e)$. If no arrows are needed write "No arrows needed." Observing D and E



makes an active path to their parent C , which in turn activates the common effect triple, and renders A and B dependent.

Q3. [13 pts] Variable Elimination

For the Bayes' net shown on the right, we are given the query $P(B, D \mid +f)$. All variables have binary domains. Assume we run variable elimination to compute the answer to this query, with the following variable elimination ordering: A, C, E, G .



(a) Complete the following description of the factors generated in this process:

After inserting evidence, we have the following factors to start out with:

$$P(A), P(B|A), P(C|B), P(D|C), P(E|C, D), P(+f|C, E), P(G|C, +f)$$

When eliminating A we generate a new factor f_1 as follows:

$$f_1(B) = \sum_a P(a)P(B|a)$$

This leaves us with the factors:

$$P(C|B), P(D|C), P(E|C, D), P(+f|C, E), P(G|C, +f), f_1(B)$$

(i) [2 pts] When eliminating C we generate a new factor f_2 as follows:

$$f_2(B, D, E, +f, G) = \sum_c P(c|B)P(D|c)P(E|c, D)P(+f|c, E)P(G|c, +f)$$

This leaves us with the factors:

$$f_1(B), f_2(B, D, E, +f, G)$$

(ii) [2 pts] When eliminating E we generate a new factor f_3 as follows:

$$f_3(B, D, +f, G) = \sum_e f_2(B, D, E, +f, G)$$

This leaves us with the factors:

$$f_1(B), f_3(B, D, +f, G)$$

(iii) [2 pts] When eliminating G we generate a new factor f_4 as follows:

$$f_4(B, D, +f) = \sum_g f_3(B, D, +f, g)$$

This leaves us with the factors:

$$f_1(B), f_4(B, D, +f)$$

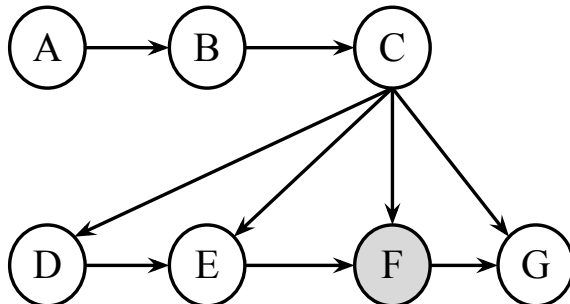
(b) [2 pts] Explain in one sentence how $P(B, D \mid +f)$ can be computed from the factors left in part (iii) of (a)?

Join $f_1 f_4$ to obtain $P(B, D, +f)$ and normalize it to get $P(B, D \mid +f)$ Concretely, $P(b, d \mid +f) = \frac{f_1(b)f_4(b, d, +f)}{\sum_{b', d'} f_1(b')f_4(b', d', +f)}$.

- (c) [1 pt] Among f_1, f_2, \dots, f_4 , which is the largest factor generated, and how large is it? Assume all variables have binary domains and measure the size of each factor by the number of rows in the table that would represent the factor.

$f_2(B, D, E, +f, G)$ is the largest factor generated. It has 4 variables, hence $2^4 = 16$ entries.

For your convenience, the Bayes' net from the previous page is shown again below.



- (d) [4 pts] Find a variable elimination ordering for the same query, i.e., for $P(B, D \mid +f)$, for which the maximum size factor generated along the way is smallest. Hint: the maximum size factor generated in your solution should have only 2 variables, for a table size of $2^2 = 4$. Fill in the variable elimination ordering and the factors generated into the table below.

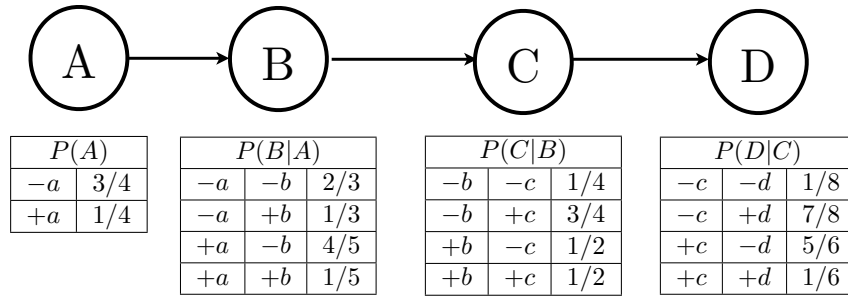
Variable Eliminated	Factor Generated
A	$f_1(B)$
G	$f_2(C, +f)$
E	$f_3(C, D)$
C	$f_4(B, D, +f)$

For example, in the naive ordering we used earlier, the first line in this table would have had the following two entries: $A, f_1(B)$. For this question there is no need to include how each factor is computed, i.e., no need to include expressions of the type $= \sum_a P(a)P(B|a)$.

Note: multiple orderings are possible. In particular in this case all orderings with E and G before C are correct.

Q4. [10 pts] Bayes' Nets Sampling

Assume the following Bayes' net, and the corresponding distributions over the variables in the Bayes' net:



(a) You are given the following samples:

$+a$	$+b$	$-c$	$-d$	$+a$	$-b$	$-c$	$+d$
$+a$	$-b$	$+c$	$-d$	$+a$	$+b$	$+c$	$-d$
$-a$	$+b$	$+c$	$-d$	$-a$	$+b$	$-c$	$+d$
$-a$	$-b$	$+c$	$-d$	$-a$	$-b$	$+c$	$-d$

(i) [1 pt] Assume that these samples came from performing Prior Sampling, and calculate the sample estimate of $P(+c)$.

$5/8$

(ii) [2 pts] Now we will estimate $P(+c \mid +a, -d)$. Above, clearly cross out the samples that would **not** be used when doing Rejection Sampling for this task, and write down the sample estimate of $P(+c \mid +a, -d)$ below.

$2/3$

(b) [2 pts] Using Likelihood Weighting Sampling to estimate $P(-a \mid +b, -d)$, the following samples were obtained. Fill in the weight of each sample in the corresponding row.

Sample	Weight
$-a \quad +b \quad +c \quad -d$	$P(+b \mid -a)P(-d \mid +c) = 1/3 * 5/6 = 5/18 = 0.277$
$+a \quad +b \quad +c \quad -d$	$P(+b \mid +a)P(-d \mid +c) = 1/5 * 5/6 = 5/30 = 1/6 = 0.17$
$+a \quad +b \quad -c \quad -d$	$P(+b \mid +a)P(-d \mid -c) = 1/5 * 1/8 = 1/40 = 0.025$
$-a \quad +b \quad -c \quad -d$	$P(+b \mid -a)P(-d \mid -c) = 1/3 * 1/8 = 1/24 = 0.042$

(c) [1 pt] From the weighted samples in the previous question, estimate $P(-a \mid +b, -d)$.

$\frac{5/18 + 1/24}{5/18 + 5/30 + 1/40 + 1/24} = 0.625$

(d) [2 pts] Which query is better suited for likelihood weighting, $P(D \mid A)$ or $P(A \mid D)$? Justify your answer in one sentence.

$P(D \mid A)$ is better suited for likelihood weighting sampling, because likelihood weighting conditions only on upstream evidence.

(e) [2 pts] Recall that during Gibbs Sampling, samples are generated through an iterative process.

Assume that the only evidence that is available is $A = +a$. Clearly fill in the circle(s) of the sequence(s) below that could have been generated by Gibbs Sampling.

☒ Sequence 1

1 :	$+a$	$-b$	$-c$	$+d$
2 :	$+a$	$-b$	$-c$	$+d$
3 :	$+a$	$-b$	$+c$	$+d$

☐ Sequence 2

1 :	$+a$	$-b$	$-c$	$+d$
2 :	$+a$	$-b$	$-c$	$-d$
3 :	$-a$	$-b$	$-c$	$+d$

☒ Sequence 3

1 :	$+a$	$-b$	$-c$	$+d$
2 :	$+a$	$-b$	$-c$	$-d$
3 :	$+a$	$+b$	$-c$	$-d$

☐ Sequence 4

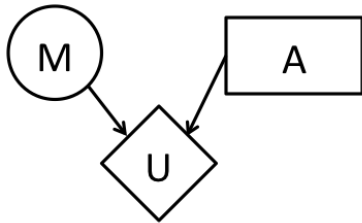
1 :	$+a$	$-b$	$-c$	$+d$
2 :	$+a$	$-b$	$-c$	$-d$
3 :	$+a$	$+b$	$-c$	$+d$

Gibbs sampling updates one variable at a time and never changes the evidence.

The first and third sequences have at most one variable change per row, and hence could have been generated from Gibbs sampling. In sequence 2, the evidence variable is changed. In sequence 4, the second and third samples have both B and D changing.

Q5. [15 pts] Probability and Decision Networks

The new Josh Bond Movie (M), Skyrise, is premiering later this week. Skyrise will either be great ($+m$) or horrendous ($-m$); there are no other possible outcomes for its quality. Since you are going to watch the movie no matter what, your primary choice is between going to the theater (*theater*) or renting (*rent*) the movie later. Your utility of enjoyment is only affected by these two variables as shown below:



M	P(M)
+m	0.5
-m	0.5

M	A	U(M,A)
+m	<i>theater</i>	100
-m	<i>theater</i>	10
+m	<i>rent</i>	80
-m	<i>rent</i>	40

(a) [3 pts] Maximum Expected Utility

Compute the following quantities:

$$EU(\textit{theater}) = P(+m)U(+m, \textit{theater}) + P(-m)U(-m, \textit{theater}) = 0.5 * 100 + 0.5 * 10 = 55$$

$$EU(\textit{rent}) = P(+m)U(+m, \textit{rent}) + P(-m)U(-m, \textit{rent}) = 0.5 * 80 + 0.5 * 40 = 60$$

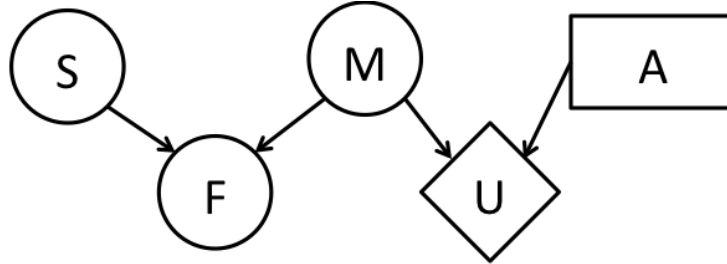
$$MEU(\{\}) = 60$$

Which action achieves $MEU(\{\}) = \textit{rent}$

(b) [3 pts] **Fish and Chips**

Skyrise is being released two weeks earlier in the U.K. than the U.S., which gives you the perfect opportunity to predict the movie's quality. Unfortunately, you don't have access to many sources of information in the U.K., so a little creativity is in order.

You realize that a reasonable assumption to make is that if the movie (M) is great, citizens in the U.K. will celebrate by eating fish and chips (F). Unfortunately the consumption of fish and chips is also affected by a possible food shortage (S), as denoted in the below diagram.



The consumption of fish and chips (F) and the food shortage (S) are both binary variables. The relevant conditional probability tables are listed below:

S	M	F	$P(F S, M)$
+s	+m	+f	0.6
+s	+m	-f	0.4
+s	-m	+f	0.0
+s	-m	-f	1.0

S	M	F	$P(F S, M)$
-s	+m	+f	1.0
-s	+m	-f	0.0
-s	-m	+f	0.3
-s	-m	-f	0.7

S	$P(S)$
+s	0.2
-s	0.8

You are interested in the value of revealing the food shortage node (S). Answer the following queries:

$$EU(theater| +s) =$$

The shortage variable is independent of the parents of the utility node when no additional evidence is present; thus, the same values hold:

$$EU(theater| +s) = EU(theater) = 55$$

$$EU(rent| +s) = EU(rent) = 60$$

$$MEU(\{+s\}) = 60$$

$$\text{Optimal Action Under } \{+s\} = r \quad (Rent)$$

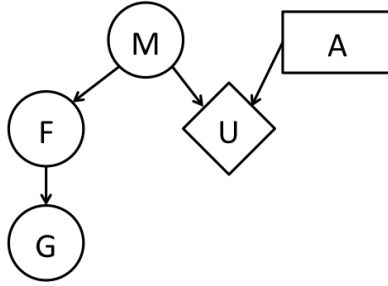
$$MEU(\{-s\}) = 60$$

$$\text{Optimal Action Under } \{-s\} = r \quad (Rent)$$

$VPI(S) = 0$, since the Value of Perfect Information is the expected difference in MEU given the evidence vs. without the evidence and here the evidence is uninformative.

(c) [5 pts] **Greasy Waters**

You are no longer concerned with the food shortage variable. Instead, you realize that you can determine whether the runoff waters are greasy (G) in the U.K., which is a variable that indicates whether or not fish and chips have been consumed. The prior on M and utility tables are unchanged. Given this different model of the problem:



G	F	$P(G F)$
+g	+f	0.8
-g	+f	0.2
+g	-f	0.3
-g	-f	0.7

M	$P(M)$
+m	0.5
-m	0.5

F	M	$P(F M)$
+f	+m	0.92
-f	+m	0.08
+f	-m	0.24
-f	-m	0.76

M	A	$U(M,A)$
+m	theater	100
-m	theater	10
+m	rent	80
-m	rent	40

[Decision network]

[Tables that define the model]

F	$P(F)$
+f	0.58
-f	0.42

G	$P(G)$
+g	0.59
-g	0.41

M	G	$P(M G)$
+m	+g	0.644
-m	+g	0.356
+m	-g	0.293
-m	-g	0.707

G	M	$P(G M)$
+g	+m	0.760
-g	+m	0.240
+g	-m	0.420
-g	-m	0.580

M	F	$P(M F)$
+m	+f	0.793
-m	+f	0.207
+m	-f	0.095
-m	-f	0.905

[Tables computed from the first set of tables. Some of them might be convenient to answer the questions below]

Answer the following queries:

$$MEU(+g) = \max(EU(theater|+g), EU(rent|+g))$$

$$EU(theater|+g) = P(+m|+g) * U(+m, theater) + P(-m|+g) * U(-m, theater) = (0.644) * 100 + (0.356) * 10 = 67.96$$

$$EU(rent|+g) = P(+m|+g) * U(+m, rent) + P(-m|+g) * U(-m, rent) = (0.644) * 80 + (0.356) * 40 = 65.76$$

$$\max(EU(theater|+g), EU(rent|+g)) = 67.96$$

$$MEU(-g) = \max(EU(theater|-g), EU(rent|-g))$$

$$EU(theater|-g) = P(+m|-g) * U(+m, theater) + P(-m|-g) * U(-m, theater) = (0.293) * 100 + (0.707) * 10 = 36.37$$

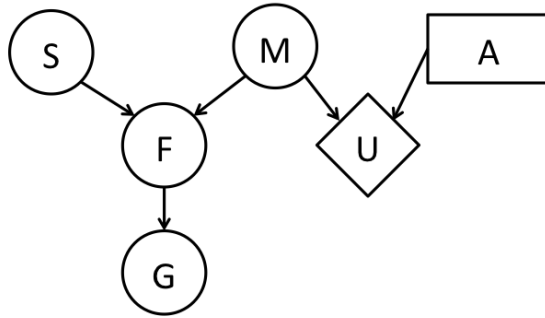
$$EU(rent|-g) = P(+m|-g) * U(+m, rent) + P(-m|-g) * U(-m, rent) = (0.293) * 80 + (0.707) * 40 = 51.72$$

$$\max(EU(theater|-g), EU(rent|-g)) = 51.72$$

$$VPI(G) = P(+g) * MEU(+g) + P(-g) * MEU(-g) - MEU(\{\}) = 0.59 * 67.96 + 0.41 * 51.72 - MEU(\{\}) = 61.3 - 60 = 1.3$$

(d) **VPI Comparisons**

We consider the shortage variable (S) again, resulting in the decision network shown below. The (conditional) probability tables for $P(S)$, $P(M)$, $P(F|S, M)$ and $P(G|F)$ are the ones provided above. The utility function is still the one shown in part (a). Circle all statements that are true, and provide a brief justification (no credit without justification).



(i) [1 pt] $VPI(S)$:

$$VPI(S) < 0 \quad \textcolor{red}{VPI(S) = 0} \quad VPI(S) > 0 \quad VPI(S) = VPI(F) \quad VPI(S) = VPI(G)$$

Justify:

With no evidence, $VPI(S)$ is zero because it is conditionally independent of the parents of the utility node.

(ii) [1 pt] $VPI(S|G)$:

$$VPI(S|G) < 0 \quad \textcolor{red}{VPI(S|G) = 0} \quad VPI(S|G) > 0 \quad VPI(S|G) = VPI(F) \quad VPI(S|G) = VPI(G)$$

Justify:

We accepted several solutions for this question.

Observing G turns $S- > F < -M$ into an active triple, which means S is no longer conditionally independent from the parents of the utility node (M). This introduces the possibility for $VPI(S|G)$ to be strictly positive.

It is possible for $VPI(S|G)$ to be zero due to the presence of conditional independencies that are not exposed by the graph structure (e.g., if F were a coin flip that was independent of S due to a very special choice of CPT entries). It is clear from the CPTs there are no such conditional independencies; consequently, this is not a reason why $VPI(S|G)$ could be 0.

It is possible for $VPI(S|G)$ to be zero if observing S does not change the optimal action for any possible value of S or G (essentially knowing S when G is observed does not allow you to improve your expected performance). Determining this requires substantial computation. We didn't expect you to do so, and accepted justifications that reflect (most of) the above reasoning pattern.

For your interest, we did compute the optimal actions for each situation, and *theater* is optimal for $(+s, +g)$ and for $(-s, +g)$ and *rent* is optimal for $(+s, -g)$ and for $(-s, -g)$ and hence it happens to be the case that $VPI(S|G) = 0$.

(iii) [1 pt] $VPI(G|F)$:

$$VPI(G|F) < 0 \quad \textcolor{red}{VPI(G|F) = 0} \quad VPI(G|F) > 0 \quad VPI(G|F) = VPI(F) \quad VPI(G|F) = VPI(G)$$

Justify:

G is independent of the parents of the utility node if F is observed, so $VPI(G|F) = 0$

(iv) [1 pt] $VPI(G)$:

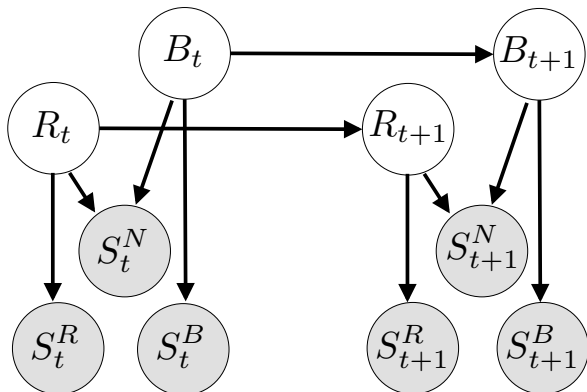
$$VPI(G) = 0 \quad \textcolor{red}{VPI(G) > 0} \quad VPI(G) > VPI(F) \quad \textcolor{red}{VPI(G) < VPI(F)} \quad VPI(G) = VPI(F)$$

Justify:

Since G is a noisy indicator of F , it still has a positive VPI (because VPI of F is positive), but its value is less than the VPI of F . Only if G was an exact representation of F would their VPI's be equal

Q6. [12 pts] Election

The country of Purplestan is preparing to vote on its next President! In this election, the incumbent President Purple is being challenged by the ambitious upstart Governor Fuschia. Purplestan is divided into two states of equal population, Redexas and Blue York, and the Blue York Times has recruited you to help track the election.



Drift and Error Models

x	$D(x)$	$E_R(x)$	$E_B(x)$	$E_N(x)$
5	.01	.00	.04	.00
4	.03	.01	.06	.00
3	.07	.04	.09	.01
2	.12	.12	.11	.05
1	.17	.18	.13	.24
0	.20	.30	.14	.40
-1	.17	.18	.13	.24
-2	.12	.12	.11	.05
-3	.07	.04	.09	.01
-4	.03	.01	.06	.00
-5	.01	.00	.04	.00

To begin, you draw the dynamic Bayes net given above, which includes the President's true support in Redexas and Blue York (denoted R_t and B_t respectively) as well as weekly survey results. Every week there is a survey of each state, S_t^R and S_t^B , and also a national survey S_t^N whose sample includes equal representation from both states.

The model's transition probabilities are given in terms of the random drift model $D(x)$ specified in the table above:

$$P(R_{t+1}|R_t) = D(R_{t+1} - R_t)$$

$$P(B_{t+1}|B_t) = D(B_{t+1} - B_t)$$

Here $D(x)$ gives the probability that the support in each state shifts by x between one week and the next. Similarly, the observation probabilities are defined in terms of error models $E_R(x)$, $E_B(x)$, and $E_N(x)$:

$$P(S_t^R|R_t) = E_R(S_t^R - R_t)$$

$$P(S_t^B|B_t) = E_B(S_t^B - B_t)$$

$$P(S_t^N|R_t, B_t) = E_N\left(S_t^N - \frac{R_t + B_t}{2}\right)$$

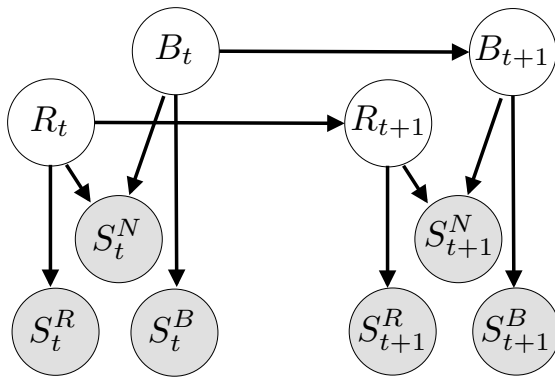
where the error model for each survey gives the probability that it differs by x from the true support; the different error models represent the surveys' differing polling methodologies and sample sizes. Note that S_t^N depends on both R_t and B_t , since the national survey gives a noisy average of the President's support across both states.

(a) Particle Filtering. First we'll consider using particle filtering to track the state of the electorate. Throughout this problem, you may give answers either as unevaluated numeric expressions (e.g. $0.4 \cdot 0.9$) or as numeric values (e.g. 0.36).

(i) [2 pts] Suppose we begin in week 1 with the two particles listed below. Now we observe the first week's surveys: $S_1^R = 51$, $S_1^B = 45$, and $S_1^N = 50$. Write the weight of each particle given this evidence:

Particle	Weight
$(r = 49, b = 47)$	$E_R(51 - 49) \cdot E_B(45 - 47) \cdot E_N(50 - 48) = .12 \cdot .11 \cdot .05 = 0.00066$
$(r = 52, b = 48)$	$E_R(51 - 52) \cdot E_B(45 - 48) \cdot E_N(50 - 50) = .18 \cdot .09 \cdot .40 = 0.00648$

The figures and table below are identical to the ones on the previous page and are repeated here for your convenience.



Drift and Error Models

x	$D(x)$	$E_R(x)$	$E_B(x)$	$E_N(x)$
5	.01	.00	.04	.00
4	.03	.01	.06	.00
3	.07	.04	.09	.01
2	.12	.12	.11	.05
1	.17	.18	.13	.24
0	.20	.30	.14	.40
-1	.17	.18	.13	.24
-2	.12	.12	.11	.05
-3	.07	.04	.09	.01
-4	.03	.01	.06	.00
-5	.01	.00	.04	.00

- (ii) [2 pts] Now we resample the particles based on their weights; suppose our resulting particle set turns out to be $\{(r = 52, b = 48), (r = 52, b = 48)\}$. Now we pass the first particle through the transition model to produce a hypothesis for week 2. What's the probability that the first particle becomes $(r = 50, b = 48)$?

$$D(50 - 52) \cdot D(48 - 48) = .12 \cdot .20 = 0.024$$

- (iii) [2 pts] In week 2, disaster strikes! A hurricane knocks out the offices of the company performing the Blue York state survey, so you can only observe $S_2^R = 48$ and $S_2^N = 50$ (the national survey still incorporates data from voters in Blue York). Based on these observations, compute the weight for each of the two particles:

Particle	Weight
$(r = 50, b = 48)$	$E_R(48 - 50) \cdot E_N(50 - 49) = .12 \cdot .24 = 0.0288$
$(r = 49, b = 53)$	$E_R(48 - 49) \cdot E_N(50 - 51) = .18 \cdot .24 = 0.0432$

- (iv) [4 pts] Your editor at the Times asks you for a “now-cast” prediction of the election if it were held today. The election directly measures the true support in both states, so R_2 would be the election result in Redexas and B_t the result in Blue York.

To simplify notation, let $I_2 = (S_1^R, S_1^B, S_1^N, S_2^R, S_2^N)$ denote all of the information you observed in weeks 1 and 2, and also let the variable W_i indicate whether President Purple would win an election in week i :

$$W_i = \begin{cases} 1 & \text{if } \frac{R_i + B_i}{2} > 50 \\ 0 & \text{otherwise.} \end{cases}$$

For improved accuracy we will work with the *weighted* particles rather than resampling. Normally we would build on top of step (iii), but to decouple errors, let's assume that after step (iii) you ended up with the following weights:

Particle	Weight
$(r = 50, b = 48)$.12
$(r = 49, b = 53)$.18

Note this is not actually what you were supposed to end up with! Using the weights from this table, estimate the following quantities:

- The current probability that the President would win:

$$P(W_2 = 1 | I_2) \approx \frac{\text{weight of winning particles}}{\text{total particle weight}} = \frac{.18}{.12 + .18} = 3/5.$$

- Expected support for President Purple in Blue York:

$$\mathbb{E}[B_2 | I_2] \approx \frac{.12}{.12 + .18} \cdot 48 + \frac{.18}{.12 + .18} \cdot 53 = 51.$$

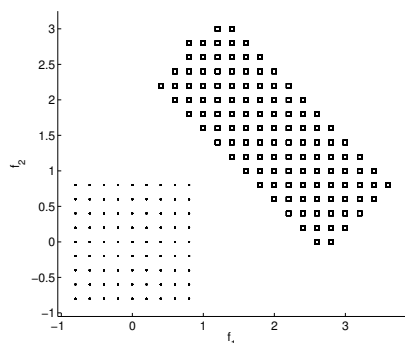
- (v) [2 pts] The real election is being held next week (week 3). Suppose you are representing the current joint belief distribution $P(R_2, B_2 | I_2)$ with a large number of *unweighted* particles. Explain using no more than two sentences how you would use these particles to forecast the national election (i.e. how you would estimate $P(W_3 = 1 | I_2)$, the probability that the President wins in week 3, given your observations from weeks 1 and 2).

Sample a new version of each particle from the transition model. Return the fraction of the new particles in which $\frac{1}{2}(r + b) > 50$.

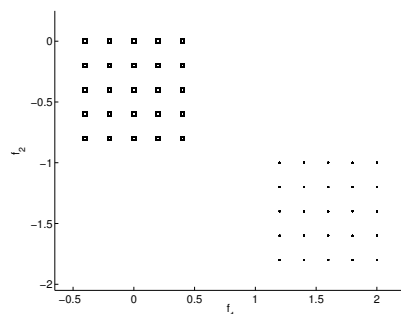
Q7. [6 pts] Naïve Bayes Modeling Assumptions

You are given points from 2 classes, shown as rectangles and dots. For each of the following sets of points, mark if they satisfy all the Naïve Bayes modelling assumptions, or they do not satisfy all the Naïve Bayes modelling assumptions. Note that in (c), 4 rectangles overlap with 4 dots.

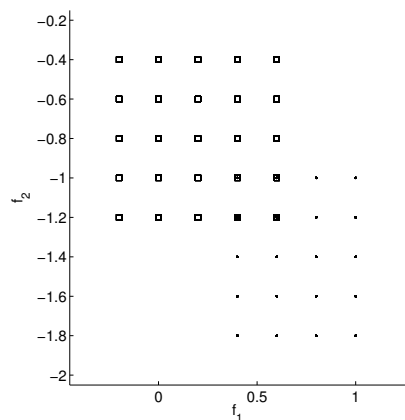
The conditional independence assumptions made by the Naïve Bayes model are that features are conditionally independent when given the class. Features being independent once the class label is known means that for a fixed class the distribution for f_1 cannot depend on f_2 , and the other way around. Concretely, for discrete-valued features as shown below, this means each class needs to have a distribution that corresponds to an axis-aligned rectangle. No other assumption is made by the Naïve Bayes model. Note that linear separability is not an assumption of the Naïve Bayes model—what is true is that for a Naïve Bayes model with all binary variables the decision boundary between the two classes is a hyperplane (i.e., it's a linear classifier). That, however, wasn't relevant to the question as the question examined which probability distribution a Naïve Bayes model can represent, not which decision boundaries.



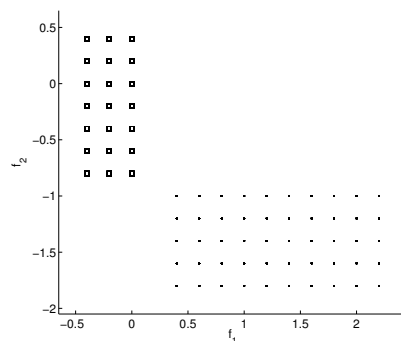
(a) ☐ Satisfies ☒ Does not Satisfy



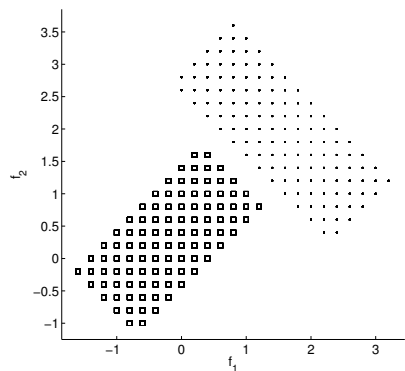
(b) ☒ Satisfies ☐ Does not Satisfy



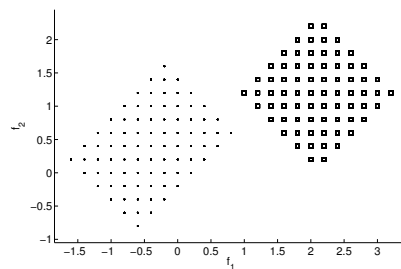
(c) ☒ Satisfies ☐ Does not Satisfy



(d) ☒ Satisfies ☐ Does not Satisfy



(e) ☐ Satisfies ☒ Does not Satisfy

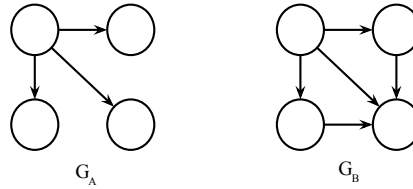


(f) ☐ Satisfies ☒ Does not Satisfy

A note about feature independence: The Naïve Bayes model assumes features are conditionally independent given the class. Why does this result in axis-aligned rectangles for discrete feature distributions? Intuitively, this is because fixing one value is uninformative about the other: within a class, the values of one feature are constant across the other. For instance, the dark square class in (b) has $f_1 \in [-0.5, 0.5]$ and $f_2 \in [-1, 0]$ and fixing one has no impact on the domain of the other. However, when the features of a class are not axis-aligned then fixing one limits the domain of the other, inducing dependence. In (e), fixing $f_2 = 1.5$ restricts f_1 to the two points at the top, whereas fixing $f_2 = 0$ gives a larger domain.

Q8. [7 pts] Model Structure and Laplace Smoothing

We are estimating parameters for a Bayes' net with structure G_A and for a Bayes' net with structure G_B . To estimate the parameters we use Laplace smoothing with $k = 0$ (which is the same as maximum likelihood), $k = 5$, and $k = \infty$.



Let for a given Bayes' net BN the corresponding joint distribution over all variables in the Bayes' net be P_{BN} then the likelihood of the training data for the Bayes' net BN is given by

$$\prod_{x_i \in \text{Training Set}} P_{BN}(x_i)$$

Let \mathcal{L}_A^0 denote the likelihood of the training data for the Bayes' net with structure G_A and parameters learned with Laplace smoothing with $k = 0$.

Let \mathcal{L}_A^5 denote the likelihood of the training data for the Bayes' net with structure G_A and parameters learned with Laplace smoothing with $k = 5$.

Let \mathcal{L}_A^∞ denote the likelihood of the training data for the Bayes' net with structure G_A and parameters learned with Laplace smoothing with $k = \infty$.

We similarly define $\mathcal{L}_B^0, \mathcal{L}_B^5, \mathcal{L}_B^\infty$ for structure G_B .

For a given Bayes' net structure, maximum likelihood parameters would give them maximum likelihood on the training data. As you add more and more smoothing you tend to move away from the MLE and get lesser and lesser likelihood on the training data. Hence parts (a), (b), and (c).

Across models, G_B can represent a larger family of models and hence has a higher MLE estimate than we get with using G_A . Hence, we have (d). In (e), in the case of infinite smoothing, all variables become independent and equally probable to take any value for both G_A and G_B , hence giving equal training likelihood. (f) follows from (a) and (d). A priori, we can not say anything about (g) because the going from \mathcal{L}_A^0 to \mathcal{L}_B^5 we increase model power (which would increase likelihood) and increase smoothing (which would decrease likelihood) and a priori we don't know which effect would dominate.

For each of the questions below, mark which one is the correct option.

(a) [1 pt] Consider \mathcal{L}_A^0 and \mathcal{L}_A^5

- ☐ $\mathcal{L}_A^0 \leq \mathcal{L}_A^5$
☒ $\mathcal{L}_A^0 \geq \mathcal{L}_A^5$
☐ $\mathcal{L}_A^0 = \mathcal{L}_A^5$
☐ Insufficient information to determine the ordering.

(b) [1 pt] Consider \mathcal{L}_A^5 and \mathcal{L}_A^∞

- ☐ $\mathcal{L}_A^5 \leq \mathcal{L}_A^\infty$
☒ $\mathcal{L}_A^5 \geq \mathcal{L}_A^\infty$
☐ $\mathcal{L}_A^5 = \mathcal{L}_A^\infty$
☐ Insufficient information to determine the ordering.

(c) [1 pt] Consider \mathcal{L}_B^0 and \mathcal{L}_B^∞

- ☐ $\mathcal{L}_B^0 \leq \mathcal{L}_B^\infty$
☒ $\mathcal{L}_B^0 \geq \mathcal{L}_B^\infty$
☐ $\mathcal{L}_B^0 = \mathcal{L}_B^\infty$
☐ Insufficient information to determine the ordering.

(d) [1 pt] Consider \mathcal{L}_A^0 and \mathcal{L}_B^0

- ☒ $\mathcal{L}_A^0 \leq \mathcal{L}_B^0$
☐ $\mathcal{L}_A^0 \geq \mathcal{L}_B^0$
☐ $\mathcal{L}_A^0 = \mathcal{L}_B^0$
☐ Insufficient information to determine the ordering.

(e) [1 pt] Consider \mathcal{L}_A^∞ and \mathcal{L}_B^∞

☐ $\mathcal{L}_A^\infty \leq \mathcal{L}_B^\infty$
☐ $\mathcal{L}_A^\infty \geq \mathcal{L}_B^\infty$
☒ $\mathcal{L}_A^\infty = \mathcal{L}_B^\infty$
☐ Insufficient information to determine the ordering.

(f) [1 pt] Consider \mathcal{L}_A^5 and \mathcal{L}_B^0

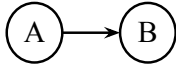
☒ $\mathcal{L}_A^5 \leq \mathcal{L}_B^0$
☐ $\mathcal{L}_A^5 \geq \mathcal{L}_B^0$
☐ $\mathcal{L}_A^5 = \mathcal{L}_B^0$
☐ Insufficient information to determine the ordering.

(g) [1 pt] Consider \mathcal{L}_A^0 and \mathcal{L}_B^5

☐ $\mathcal{L}_A^0 \leq \mathcal{L}_B^5$
☐ $\mathcal{L}_A^0 \geq \mathcal{L}_B^5$
☐ $\mathcal{L}_A^0 = \mathcal{L}_B^5$
☒ Insufficient information to determine the ordering.

Q9. [11 pts] ML: Short Question & Answer

- (a) **Parameter Estimation and Smoothing.** For the Bayes' net drawn on the left, A can take on values $+a$, $-a$, and B can take values $+b$ and $-b$. We are given samples (on the right), and we want to use them to estimate $P(A)$ and $P(B|A)$.



$(-a, +b)$ $(-a, +b)$ $(-a, -b)$ $(-a, -b)$
 $(-a, -b)$ $(-a, -b)$ $(-a, +b)$
 $(-a, +b)$ $(-a, -b)$ $(+a, +b)$

- (i) [3 pts] Compute the maximum likelihood estimates for $P(A)$ and $P(B|A)$, and fill them in the 2 tables on the right.

A	$P(A)$
+a	1/10
-a	9/10

A	B	$P(B A)$
+a	+b	1/1
+a	-b	0/1
-a	+b	4/9
-a	-b	5/9

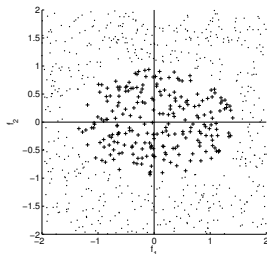
- (ii) [3 pts] Compute the estimates for $P(A)$ and $P(B|A)$ using Laplace smoothing with strength $k = 2$, and fill them in the 2 tables on the right.

A	$P(A)$
+a	3/14
-a	11/14

A	B	$P(B A)$
+a	+b	3/5
+a	-b	2/5
-a	+b	6/13
-a	-b	7/13

- (b) [2 pts] **Linear Separability.** You are given samples from 2 classes ($.$ and $+$) with each sample being described by 2 features f_1 and f_2 . These samples are plotted in the following figure. You observe that these samples are not *linearly* separable using just these 2 features. Circle the minimal set of features below that you could use alongside f_1 and f_2 , to *linearly* separate samples from the 2 classes.

The samples are elliptically separable, so consider the equation for an ellipse.



- ☐ $f_1 < 1.5$ ☐ $f_1 > -1.5$ ☐ $f_2 < -1$ ☒ f_2^2
☐ $f_1 > 1.5$ ☐ $f_2 < 1$ ☐ $f_2 > -1$ ☐ $|f_1 + f_2|$
☐ $f_1 < -1.5$ ☐ $f_2 > 1$ ☒ f_1^2

☐ Even using all these features alongside f_1 and f_2 will not make the samples linearly separable.

- (c) [3 pts] **Perceptrons.** In this question you will perform perceptron updates. You have 2 classes, $+1$ and -1 , and 3 features f_0, f_1, f_2 for each training point. The $+1$ class is predicted if $w \cdot f > 0$ and the -1 class is predicted otherwise.

You start with the weight vector, $w = [1 \quad 0 \quad 0]$. In the table below, do a perceptron update for each of the given samples. If the w vector does not change, write *No Change*, otherwise write down the new w vector.

f_0	f_1	f_2	Class	Updated w
1	7	8	-1	$[0 \quad -7 \quad -8]$
1	6	8	-1	Unchanged (no misclassification)
1	9	6	+1	$[1 \quad 2 \quad -2]$