Basic Probability

T/F For all random variables X, Y and Z and distributions over them,

$$\frac{P(Y|X)P(Z|X)P(X)}{P(Y,Z)} = P(X|Y,Z)$$

this indicates ~~$\not\perp\!\!\!\perp$~~ $Y \perp\!\!\!\perp Z \mid X$.

Question #1 — 20 points

Indicate whether the following statement is true and briefly justify your answer.

(i) (2 points) If we add more training data, the number of support vectors will always increase.

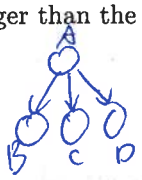
(ii) (2 points) For a given training set, support vectors are invariant to the kernel choice.

(iii) (2 points) There is a distribution over five variables that can be represented by any five node Bayesian network topology.

Yes, when the variables are in dependent

(iv) (2 points) The largest factor created by the variable elimination algorithm is no bigger than the largest initial factor.

No.



$p(A,B,C,D) = p(A)\,p(B|A)\,p(C|A)\,p(D|A) \Rightarrow$ the biggest one have two variables involved

$p(B,C,D) = \sum_A p(A)\,p(B|A)\,p(C|A)\,p(D|A) \Rightarrow f(B,C,D)$ this factor has 3 variables involved.

(v) (2 points) The variable elimination algorithm can require time exponential in the size of the Bayesian network.

Yes, for example, when all the variables (nodes) are connected.

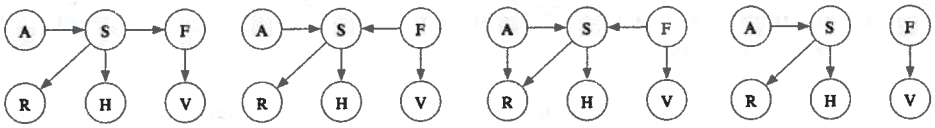
(vi) (2 points) Observing an additional node (i.e. making the node an evidence node) in a Bayesian network can only increase the number of variable pairs which are conditionally independent.

$A \perp\!\!\!\perp B$, but $A \not\perp\!\!\!\perp B \mid E$

Question #3 — 20 points

Assume there are two types of conditions: (S)inus congestion and (F)lu. Sinus congestion is caused by (A)llergy or the flu.

There are three observed symptoms for these conditions: (H)eadache, (R)unny nose, and fe(V)er. Runny nose and headaches are directly caused by sinus congestion (only), while fever comes from having the flu (only). For example, allergies only cause runny noses indirectly. Assume each variable is binary (i.e. true/false).

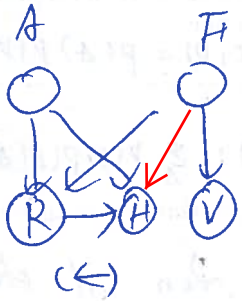


(i) (ii) (iii) (iv)

(a) (5 points) Consider the four Bayesian networks shown. Indicate which one models the domain (as described above) best.

Best network : i / (ii) / iii / iv

(b) (5 points) Assume we wanted to remove the Sinus congestion (S) node. Using the least number of edges, draw a Bayesian network over the remaining variables which can encode the original model's marginal distribution over the remaining variables.



The following samples were drawn from the correct Bayesian network using prior sampling:

Sample #	A	S	R	H	F	V
1	true	true	true	false	false	false
2	true	true	false	true	true	false
3	true	false	false	false	false	false
4	true	false	false	true	true	false
5	true	true	false	true	false	false

(c) (2 points) Give the sample estimate of $P(F = true)$ or state why it cannot be computed.

$\frac{2}{5}$

(d) (2 points) Give the sample estimate of $P(F = true|H = true)$ or state why it cannot be computed.

$\frac{2}{3}$

(e) (2 points) Give the sample estimate of $P(F = true|V = true)$ or state why it cannot be computed.

No, because no V=true is observed.

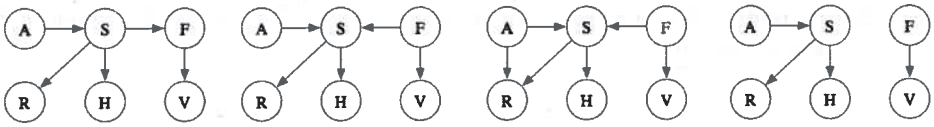
(f) (4 points) For rejection sampling in general (not necessarily on these samples), which query will require more samples to compute a certain degree of accuracy, $P(F|H)$ or $P(F|H, A)$? Briefly justify.

P(F|H,A) needs more samples because samples that have (H, A) cannot be more common than those that have (H).

Question #3 — 20 points

Assume there are two types of conditions: (S)inus congestion and (F)lu. Sinus congestion is caused by (A)llergy or the flu.

There are three observed symptoms for these conditions: (H)eadache, (R)unny nose, and fe(V)er. Runny nose and headaches are directly caused by sinus congestion (only), while fever comes from having the flu (only). For example, allergies only cause runny noses indirectly. Assume each variable is binary (i.e. true/false).

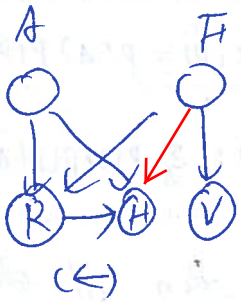


(i) (ii) (iii) (iv)

(a) (5 points) Consider the four Bayesian networks shown. Indicate which one models the domain (as described above) best.

Best network : i / (ii) / iii / iv

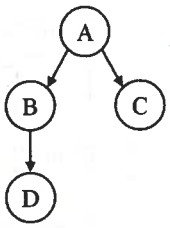
(b) (5 points) Assume we wanted to remove the Sinus congestion (S) node. Using the least number of edges, draw a Bayesian network over the remaining variables which can encode the original model's marginal distribution over the remaining variables.



Question #5

20 points

The next parts involve computing various quantities in the network below. These questions are designed so that they can be answered with minimum computation. If you find yourself doing copious amount of computation for each part, step back and consider whether there is a simpler way to deduce the answer.



A	P(A)
true	0.1
false	0.9

B	A	P(B|A)
true	true	0.5
false	true	0.5
true	false	0.8
false	false	0.2

C	A	P(C|A)
true	true	0.4
false	true	0.6
true	false	0.7
false	false	0.3

D	B	P(D|B)
true	true	0.9
false	true	0.1
true	false	0.2
false	false	0.8

(a) (2 points) P(A=true, B=false, C=true, D=false)

$= P(A=true) \times P(B=false|A=true) \times P(C=true|A=true) \times P(D=false|B=false)$

$= 0.1 \times 0.5 \times 0.4 \times 0.8$

(b) (2 points) P(A=true, B=true)

$= \sum_{CD} P(A=true) P(B=true|A=true) \times P(C|A=true) \times P(D|B=true)$

$= P(A=true) \times P(B=true|A) \times \sum_{C} P(C|A=true) \times \sum_{D} P(D|B=true)$

$= 0.1 \times 0.5$ (1) (1)

(c) (2 points) P(B=true)

$P(B=true) = \sum_{A} P(A, B=true) = 0.1 \times 0.5 + 0.9 \times 0.8 = 0.05 + 0.72 = 0.77$

(d) (2 points) P(A=true|B=true)

$\frac{P(A=true, B=true)}{P(B=true)} = \frac{0.05}{0.77}$

(e) (2 points) P(D=true|A=true)

$P(D=true, A=true) = \sum_{BC} P(A=true) P(B|A=true) P(C|A=true) \cdot P(D=true|B)$

$= P(A=true) \times \sum_{C} P(C|A=true) \sum_{B} P(B|A=true) \cdot P(D=true|B)$

(1) (0.55)

$= 0.1 \times 0.55 = 0.055$

$P(D=false, A=true) = 0.045$

$\therefore P(D=true|A=true) = \frac{0.055}{0.1}$

(f) (2 points) P(D=true|A=true,C=true)

P(D = true, A = true, C = true)
= sum_{B} P(A = true, B, C = true, D = true)
= 0.022

P(D = false, A = true, C = true)
= sum_{B} P(A = true, B, C = true, D = false)
= 0.018

P(D = false |A = true, C = true) = 0.022/(0.022+0.018)

Another faster approach to solving (f) will be:
D is independent from C given A; therefore,
P(D | A, C) = P(D|A). As a reaults, the Answer to (f) is the same as the one to (e).

Consider computing the following distributions in the above network using various methods:

(i) P(A|B=true,C=true,D=true)

(ii) P(C|D=true)

(iii) P(D|A=true)

(iv) P(D)

(g) (2 points) Which query is least expensive using inference by enumeration? Briefly justify.

(i) only needs to go through A.

(h) (2 points) Which query is most improved by using likelihood weighting instead of rejection sampling (in terms of number of samples required)? Briefly justify.

(i) because it requires samples to match three variables.
(evidence)

(i) (2 points) When computing P(D) with variable elimination, what factor(s) is/are eliminated by first eliminating variable A from the active list?

$P(A) \cdot P(B|A)\, P(C|A)$

(j) (2 points) When computing P(D) with variable elimination, show the new factor (in tabular form) that is created by first eliminating variable A from the active list.

$\sum_A P(A)\, P(B|A)\, P(C|A) = P(A=\text{true})\, P(B|A=\text{true}) \cdot P(C|A=\text{true}) + P(A=\text{false})\, P(B|A=\text{false})\, P(C|A=\text{false})$

B	C	
T	T	0.524
T	F	0.246
F	T	0.146
F	F	0.084

MDP: Walk or Jump?

Consider an MDP with states 4, 3, 2, 1, 0, where 4 is the starting state. In states $k \geq 1$, you can *walk* (W) and $T(k, W, k-1) = 1$. In states $k \geq 2$, you can also *jump* (J) and $T(k, J, k-2) = T(k, J, k) = 1/2$. State 0 is a terminal state. The reward $R(s, a, s') = (s - s')^2$ for all (s, a, s'). Use a discount of $\gamma = 1/2$.

Compute $V*(2)$

$V^*(0) = 0$

$V^*(1) = \max\{1 + \gamma V^*(0)\}$

$= 1$

$V^*(2) = \max\{1 + \gamma V^*(1), \frac{1}{2}(4 + \gamma V^*(0)) + \frac{1}{2}\gamma V^*(2)\}$

$= \max\{1, 2 + \frac{1}{4}V^*(2)\}$ $\leftarrow 2 + \frac{1}{4}V^*(2) = V^*(2)$

$\therefore \frac{3}{4}V^*(2) = 2$, $V^*(2) = \frac{8}{3}$

Compute $Q*(4, W)$ $= \frac{8}{3}$

$V^*(3) = \frac{1}{2}(4 + \frac{1}{2}V^*(1)) + \frac{1}{2} \times \frac{1}{2} V^*(3)$

$= \frac{9}{4} + \frac{1}{4}V^*(3) \Rightarrow V^*(3) = 3$

$Q^*(4, W) = 1 + \frac{1}{2}V^*(3) = \frac{5}{2}$

Now consider the same MDP, but with infinite states 4, 3, 2, 1, 0, -1, ... and no terminal states. Like before, $T(k, J, k-2) = T(k, J, k) = 1/2$ and $T(k, W, k-1) = 1$. $R(s, a, s') = (s - s')^2$. Compute $V*(2)$.

By symmetry, $V^*(s)$ is consistent, this implies:

$V^*(2) = \max\{1 + \gamma V^*(1), \frac{1}{2}(4 + \gamma V^*(0)) + \frac{1}{2}\gamma V^*(2)\}$

$= \max\{1 + \gamma V^*(2), 2 + \gamma V^*(2)\}$

$= 2 + \gamma V^*(2)$

$\therefore V^*(2) = \frac{2}{1-\gamma} = 4$.

6. (10 points.) MDPs and Reinforcement Learning

Consider an autonomous robot which can either move FAST or SLOW in any time step. Moving FAST generally gives a reward of +2, while moving SLOW gives a reward of only +1. However, the robot must also take into account its internal temperature, which can be either HOT or OK. Driving SLOW tends to lower the temperature, while driving FAST tends to raise it. If the robot is HOT, there is a danger if it overheating, at which point it must stop, cool down, and make repairs. The MDP transitions and rewards are specified as follows:

s	a	s'	$T(s,a,s')$	$R(s,a,s')$
OK	SLOW	OK	1.0	+1
OK	FAST	OK	0.5	+2
OK	FAST	HOT	0.5	+2
HOT	SLOW	OK	1.0	+1
HOT	FAST	HOT	0.5	+2
HOT	FAST	OK	0.5	-10

Note that while repairs are costly, the robot is OK afterwards (the last row in the table).

(1) (5 pts): Run two rounds of value iteration in the table below, using a discount of 0.8. You may skip the greyed-out square.

$V_{i+1}(s) \leftarrow \max_a \{ \sum_{s'} T(s,a,s') [R(s,a,s') + \gamma V_i(s')]$

s	V_0	V_1	V_2
OK	0	2	3.2
HOT	0	1	

(1) (5 pts): Run Q-learning with a discount of 0.8 and a learning rate of 0.5, using the transition samples below. Do not copy over q-values which have not changed in a given step.

Assume the agent experiences the samples:

OK, FAST, HOT, reward +2, calculate Q_1
HOT, FAST OK, reward -10, calculate Q_2
OK, SLOW, OK, reward +1, calculate Q_3

$Q(s,a) \leftarrow Q(s,a) + 0.5 [R(s,a,s') + 0.8 \max_{a'} Q(s',a') - Q(s,a)]$

s	a	Q_0	Q_1	Q_2	Q_3
OK	SLOW	0			0.9
OK	FAST	0	1.0		
HOT	SLOW	0			
HOT	FAST	0		-4.6	

End of Exam

6 Perceptron (8 points)

The following table shows a data set and the number of times each point is misclassified during a run of the perceptron algorithm, starting with zero weights. What is the equation of the separating line found by the algorithm, as a function of x_1, x_2, and x_3? Assume that the learning rate is 1 and the initial weights are all zero.

x_1	x_2	x_3	y	times misclassified
2	3	1	+1	12
2	4	0	+1	0
3	1	1	-1	3
1	1	0	-1	6
1	2	1	-1	11

m ⇒ # samples.

$$W = \eta \sum_{i=1}^{m} \alpha_i y^i \bar{X}_i$$

$$= (12) \times 1 \times (1,2,3,1) + (3)(-1)(1,3,1,1) + 6(-1)(1,1,1,0) + (11) \times (-1) \times (1,1,2,1) = (-8, -2, -5, -2)$$

So the equation of the separating line B:

$$-2X_1 + 5X_2 - 2X_3 - 8 = 0$$

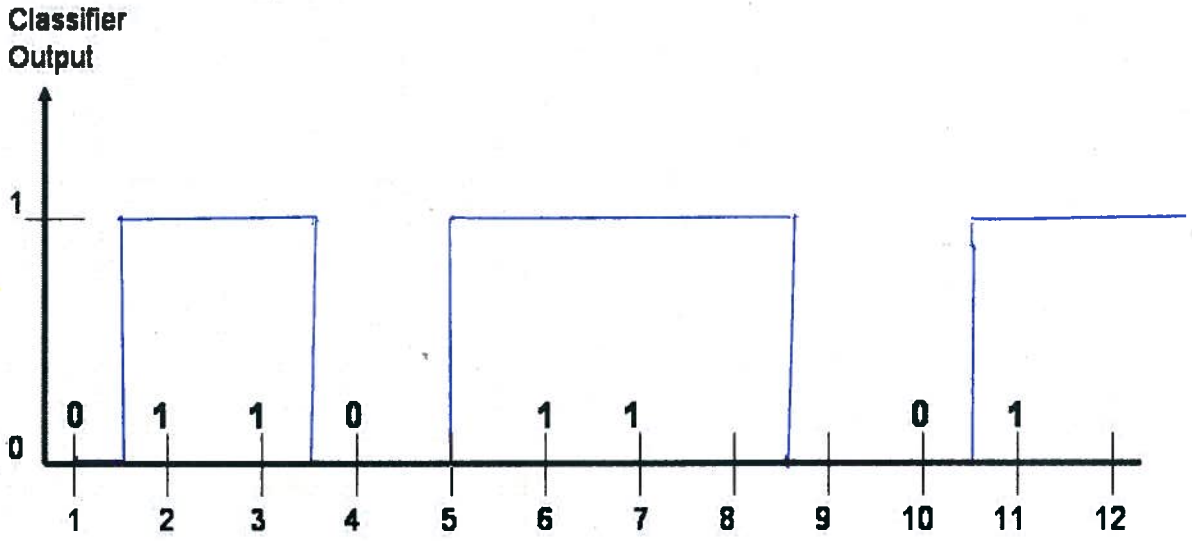
4 Machine Learning — Continuous Features (20 points)

In all the parts of this problem we will be dealing with one-dimensional data, that is, a set of points (x^i) with only one feature (called simply x). The points are in two classes given by the value of y^i. We will show you the points on the x axis, labeled by their class values; we also give you a table of values.

4.1 Nearest Neighbors

i	x^i	y^i
1	1	0
2	2	1
3	3	1
4	4	0
5	6	1
6	7	1
7	10	0
8	11	1

1. In the figure below, draw the output of a 1-Nearest-Neighbor classifier over the range indicated in the figure.



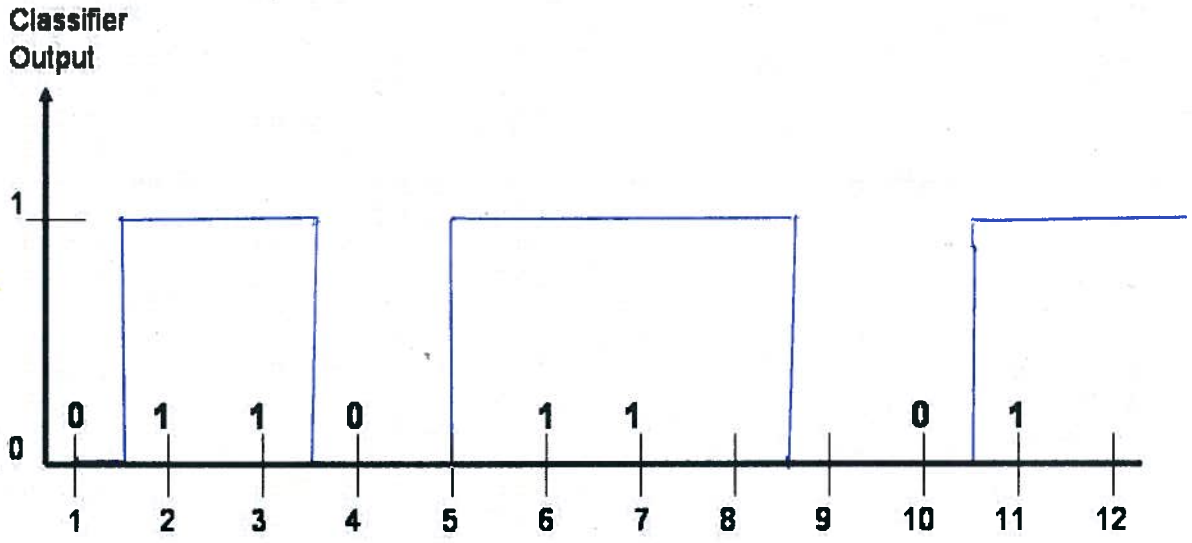
4 Machine Learning — Continuous Features (20 points)

In all the parts of this problem we will be dealing with one-dimensional data, that is, a set of points (x^i) with only one feature (called simply x). The points are in two classes given by the value of y^i. We will show you the points on the x axis, labeled by their class values; we also give you a table of values.

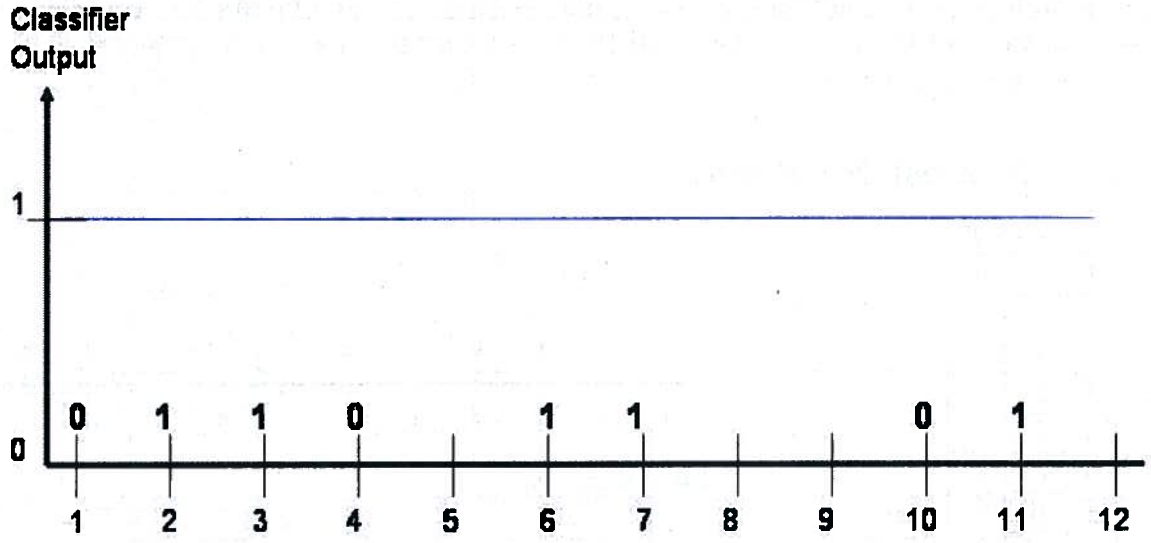
4.1 Nearest Neighbors

i	x^i	y^i
1	1	0
2	2	1
3	3	1
4	4	0
5	6	1
6	7	1
7	10	0
8	11	1

1. In the figure below, draw the output of a 1-Nearest-Neighbor classifier over the range indicated in the figure.



2. In the figure below, draw the output of a 5-Nearest-Neighbor classifier over the range indicated in the figure.



4.2 Decision Trees

Answer this problem using the same data as in the Nearest Neighbor problem above.

Which of the following three tests would be chosen as the top node in a decision tree?

$$x \leq 1.5 \qquad x \leq 5 \qquad x \leq 10.5$$

Justify your answer.

Recall that entropy for each side of a split is:

$$H = -p \log p - (1-p) \cdot \log(1-p)$$

so for $X \leq 1.5$:

$$H = \frac{1(0) + 7\left(-\frac{5}{7}\log_2\frac{5}{7} - \frac{2}{7}\log_2\frac{2}{7}\right)}{8} = 0.761$$

You may find this table useful.

x	y	-(x/y)*lg(x/y)		x	y	-(x/y)*lg(x/y)
1	2	0.50		1	8	0.38
1	3	0.53		3	8	0.53
2	3	0.39		5	8	0.42
1	4	0.50		7	8	0.17
3	4	0.31		1	9	0.35
1	5	0.46		2	9	0.48
2	5	0.53		4	9	0.52
3	5	0.44		5	9	0.47
4	5	0.26		7	9	0.28
1	6	0.43		8	9	0.15
2	6	0.53		1	10	0.33
5	6	0.22		3	10	0.52
1	7	0.40		7	10	0.36
2	7	0.52		9	10	0.14
3	7	0.52				
4	7	0.46				
5	7	0.35				
6	7	0.19				

$-\frac{1}{4}\log_2\frac{1}{4}$

$X \leq 5$

$$H = \frac{4\left(-\frac{2}{4}\log_2\frac{2}{4} - \frac{2}{4}\log_2\frac{2}{4}\right) + 4\left(-\frac{3}{4}\log_2\frac{3}{4} - \frac{1}{4}\log_2\frac{1}{4}\right)}{8}$$

$$= 0.905$$

$X \leq 10.5$

$$H = \frac{1(0) + 7\left(-\frac{4}{7}\log_2\frac{4}{7} - \frac{3}{7}\log_2\frac{3}{7}\right)}{8}$$

$$= 0.85$$

∴ we should choose $X \leq 1.5$.

Problem 2: Overfitting (20 points)

For each of the supervised learning methods that we have studied, indicate how the method could overfit the training data (consider both your design choices as well as the training) and what you can do to minimize this possibility. There may be more than one mechanism for overfitting, make sure that you identify them all.

Part A: Nearest Neighbors (5 Points)

1. How does it overfit?

Every point in dataset defines its own decision boundary.

2. How can you reduce overfitting?

Use k-NN for larger k

but not too large; otherwise it may cause underfitting.

Use cross-validation to choose k.

Part B: Decision Trees (5 Points)

1. How does it overfit?

By adding new tests to the tree to correctly classify every data point in the training set.

2. How can you reduce overfitting?

By pruning the resulting tree based on performance on a validation (development) set.