UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN

CS440/ECE448 Artificial Intelligence

Exam 1

Spring 2024

Exam 1 will be February 26-28, 2024 at CBTF

Question 1 ***(0 points)***

Consider three binary events, A, B, and C, with probabilities given by $\Pr(A) = 0.7$, $\Pr(B) = 0.4$, and $\Pr(C) = 0.3$.

(a) What's the largest possible $\Pr(B \wedge C)$?

Solution:

$$\max \Pr(B \wedge C) = \min(\Pr(B), \Pr(C)) = 0.3$$

(b) If A and B are independent, what's $\Pr(A \wedge B)$?

Solution:

$$\Pr(A)\Pr(B) = 0.28$$

(c) If B and C are mutually exclusive, what's $\Pr(B \wedge (\neg C))$?

Solution: If B and C are mutually exclusive, then $\Pr(B \wedge C) = 0$, so

$$\Pr(B \wedge (\neg C)) = \Pr(B) = 0.4$$

(d) What's the largest possible $P(\neg(B \wedge C))$?

Solution: $\neg(B \wedge C)$ is the same as $(\neg B \vee \neg C)$.

$$\Pr(\neg B \vee \neg C) = \Pr(\neg B) + \Pr(\neg C) - \Pr(\neg B \wedge \neg C) \leq \Pr(\neg B) + \Pr(\neg C) = 1.3$$

Since $1.3 > 1$, we conclude that the largest possible $P(\neg(B \wedge C)) = 1$, i.e., this would occur if B and C are mutually exclusive.

Question 2 ***(0 points)***

Let A and B be independent binary random variables with $P(A=1)=0.1$, $P(B=1)=0.4$. Let C denote the event that at least one of them is 1, and let D denote the event that exactly one of them is 1.

(a) What is $\Pr(C)$?

Solution:

$$\begin{aligned}\Pr(C) &= P(A=1,B=1)+P(A=1,B=0)+P(A=0,B=1)\\ &= (0.1)(0.4)+(0.1)(0.6)+(0.9)(0.4)=0.46\end{aligned}$$

where the last line follows from the independence of A and B.

(b) What is $\Pr(D)$?

Solution:

$$\begin{aligned}\Pr(D) &= P(A=1,B=0)+P(A=0,B=1)\\ &= (0.1)(0.6)+(0.9)(0.4)=0.42\end{aligned}$$

(c) What is $\Pr(D|A=1)$?

Solution:

$$\begin{aligned}\Pr(D|A=1) &= \Pr(D,A=1)/P(A=1)\\ &= P(A=1,B=0)/P(A=1)\\ &= \frac{0.06}{0.1}=0.6\end{aligned}$$

(d) Are A and D independent? Why?

Solution: No. $\Pr(D|A=1)\neq\Pr(D)$.

Question 3 ***(0 points)***

Use the axioms of probability to prove that $\Pr(\neg A) = 1 - \Pr(A)$.

Solution:

- From the third axiom, $\Pr(A \vee \neg A) = \Pr(A) + \Pr(\neg A) - \Pr(A \wedge \neg A)$.
- The event $(A \vee \neg A)$ is always true, so from the second axiom, $\Pr(A \vee \neg A) = 1$. The event $(A \wedge \neg A)$ is always false, so from the second axiom, $\Pr(A \wedge \neg A) = 0$.
- Combining the two statements above, $1 = \Pr(A) + \Pr(\neg A)$. Q.E.D.

Question 4 ***(0 points)***

20% of students at U of I are Malazan secret agents. Amongst these students, 10% study engineering. Furthermore, 15% of the entire student body studies engineering. Given that we know that a student studies engineering, what is the probability that the student is not a Malazan secret agent?

Solution: Define M=student is a Malazan secret agent, E=student studies engineering. We are given that $\Pr(M) = 0.2$ and $\Pr(E|M) = 0.1$, from which we may infer that $\Pr(E,M) = 0.02$. We are also given that $\Pr(E) = 0.15$, from which we may infer that

$$\Pr(\neg M, E) = \Pr(E) - \Pr(E, M) = 0.13$$

$$\Pr(\neg M|E) = \frac{\Pr(\neg M, E)}{\Pr(E)} = \frac{0.13}{0.15} = \frac{13}{15}$$

Question 5 ***(0 points)***

Consider the following joint probability distribution:

$$\Pr(A,B) = 0.12$$
$$\Pr(A,\neg B) = 0.18$$
$$\Pr(\neg A,B) = 0.28$$
$$\Pr(\neg A,\neg B) = 0.42$$

What are the marginal distributions of A and B? Are A and B independent and why?

Solution: $\Pr(A) = 0.3, \Pr(\neg A) = 0.7, \Pr(B) = 0.4, \Pr(\neg B) = 0.6$. They are independent, because $\Pr(A)\Pr(B) = \Pr(A,B) = 0.12, \Pr(A)\Pr(\neg B) = \Pr(A,\neg B) = 0.18$, and so on.

Question 6 ***(0 points)***

A friend who works in a big city owns two cars, one small and one large. Three-quarters of the time he drives the small car to work, and one-quarter of the time he drives the large car. If he takes the small car, he usually has little trouble parking, and so is at work on time with probability 0.9. If he takes the large car, he is at work on time with probability 0.6. Given that he was on time on a particular morning, what is the probability that he drove the small car?

Solution: Let S be the event "takes the small car," and T is the event "arrives on time." Then

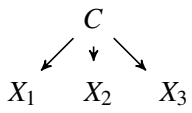
$$\Pr(S|T) = \frac{\Pr(T|S)\Pr(S)}{\Pr(T)} = \frac{\Pr(T|S)\Pr(S)}{\Pr(T|S)\Pr(S) + \Pr(T|\neg S)\Pr(\neg S)} = \frac{0.9(3/4)}{0.9(3/4) + 0.6(1/4)} = \frac{27}{33}$$

Question 7 ***(0 points)***

We have a bag of three biased coins, a, b, and c, with probabilities of coming up heads of 20%, 60%, and 80%, respectively. One coin is drawn randomly from the bag (with equal likelihood of drawing each of the three coins), and then the coin is flipped three times to generate the outcomes X_1, X_2, and X_3.

(a) Draw the Bayesian network corresponding to this setup and define the necessary conditional probability tables (CPTs).

Solution: You need an intermediate variable, $C \in \{a, b, c\}$, to specify which coin is drawn, then the graph is



and the CPTs are

C	$P(C=c)$	$P(X_1=H\|C=c)$	$P(X_2=H\|C=c)$	$P(X_3=H\|C=c)$
a	1/3	0.2	0.2	0.2
b	1/3	0.6	0.6	0.6
c	1/3	0.8	0.8	0.8

(b) Calculate which coin was most likely to have been drawn from the bag if the observed flips come out heads twice and tails once.

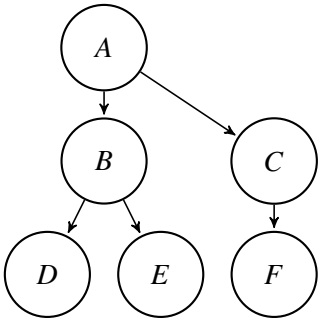
Solution:

$$P(C=a, HHT) = (0.2)(0.2)(0.8)/3 = 32/3000$$
$$P(C=b, HHT) = (0.6)(0.6)(0.4)/3 = 144/3000$$
$$P(C=c, HHT) = (0.8)(0.8)(0.2)/3 = 128/3000$$

The maximum-posterior-probability event is also the maximum-joint-probability event, which is the event $C=b$.

Question 8 ***(0 points)***

Consider the following Bayes network (all variables are binary):



$$P(A=T)=0.8$$

a	$P(B=T\|A=a)$	$P(C=T\|A=a)$
0	0.2	0.6
1	0.5	0.8

b	$P(D=T\|B=b)$	$P(E=T\|B=b)$
0	0.5	0.8
1	0.5	0.8

c	$P(F=T\|C=c)$	
0	0.01	
1	0.2	

(a) Are D and E independent?

Solution: Yes. This is a trick question. The structure of the Bayes net shows them to be conditionally independent given B, but not independent. However, in the probability table, notice that $\Pr(D|B)=\Pr(D|\neg B)$, therefore D is independent of B, despite the arrow shown in the Bayes net. Similarly, $\Pr(E|B)=\Pr(E|\neg B)$, therefore E is independent of B, despite the arrow shown in the Bayes net. Since there is no other path connecting D to E except the one going through B, they are independent.

(b) Are D and E conditionally independent given B?

Solution: Yes. This is not a trick question. The structure of the Bayes net shows that they are conditionally independent given B.

(c) If you did not know the Bayesian network, how many numbers would you need to represent the full joint probability table?

Solution: There are 2^6 possible combinations of 6 binary variables, so you'd need $2^6-1=63$ numbers.

(d) If you knew the Bayes network as shown above, but the variables were ternary instead of binary, how many values would you need to represent the full joint probability table and the conditional probability tables, respectively?

Solution: Conditional probability tables: For each variable, the number of trainable parameters is (# possible values of the variable, minus 1)$\times$(# possible values of its parents). $P(A)$ would need 2 trainable parameters, each of the other five variables would need $2\times 3=6$ trainable parameters, for a total of $2+5\times 2\times 3=32$ trainable parameters.

Full joint probability table: there are 3^6 possible combinations of the variables, so you would need to store 3^6-1 parameters.

(e) Write down the expression for the joint probability of all the variables in the network, in terms of the model parameters given above.

Solution:

$$P(A,B,C,D,E,F) = P(A)P(B|A)P(C|A)P(D|B)P(E|B)P(F|C)$$

(f) Find $P(A=0,B=1,C=1,D=0)$.

Solution:

$$P(A=0,B=1,C=1,D=0) = (0.2)(0.2)(0.6)(0.5) = \frac{3}{250}$$

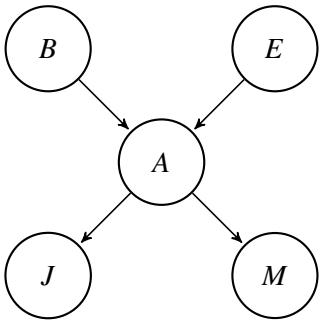
(g) Find $P(B=1|A=1,D=0)$.

Solution:

$$\begin{aligned} P(B=1|A=1,D=0) &= \frac{P(A=1,B=1,D=0)}{P(A=1,B=1,D=0)+P(A=1,B=0,D=0)} \\ &= \frac{(0.8)(0.5)(0.5)}{(0.8)(0.5)(0.5)+(0.8)(0.5)(0.5)} \\ &= \frac{1}{2} \end{aligned}$$

Question 9 ***(0 points)***

Consider the "Burglary" Bayesian network:



(a) How many independent parameters does this network have? How many entries does the full joint distribution table have?

Solution: There are five binary variables: two with no parents (B and E, one parameter each), one with two parents (A, four parameters), and two with one parent each (J and M, two parameters each), for a total of ten independent parameters. The full joint distribution table has $2^5 - 1 = 31$ parameters.

(b) If no evidence is observed, are B and E independent?

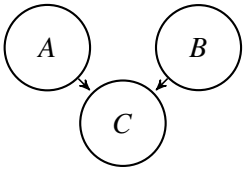
Solution: Yes, because they have no common ancestors.

(c) Are B and E conditionally independent given the observation that $A =$True?

Solution: No. Knowing that Earthquake=True makes Burglary less probable.

Question 10 ***(0 points)***

Consider the following Bayes network (all variables are binary):



$$P(A=T)=0.4, P(B=T)=0.1$$

a,b	$P(C=T\|A=a,B=b)$
False,False	0.7
False,True	0.7
True,False	0.1
True,True	0.9

(a) What is $P(C=T)$? Write your answer in numerical form, but you don't need to simplify.

Solution:

$$\begin{aligned}P(C=T) &= \Pr(\neg A,\neg B,C)+\Pr(\neg A,B,C)+\Pr(A,\neg B,C)+\Pr(A,B,C)\\ &= (0.6)(0.9)(0.7)+(0.6)(0.1)(0.7)+(0.4)(0.9)(0.1)+(0.4)(0.1)(0.9)\end{aligned}$$

(b) What is $P(A=T|B=T,C=T)$? Write your answer in numerical form, but you don't need to simplify.

Solution:

$$\begin{aligned}P(A=T|B=T,C=T) &= \frac{\Pr(A,B,C)}{\Pr(A,B,C)+\Pr(\neg A,B,C)}\\ &= \frac{(0.4)(0.1)(0.9)}{(0.4)(0.1)(0.9)+(0.6)(0.1)(0.7)}\end{aligned}$$

(c) You've been asked to re-estimate the parameters of the network based on the following observations:

Observation	A	B	C
1	True	False	False
2	False	False	True
3	True	True	False
4	False	False	False

Given the data in the table, what are the maximum likelihood estimates of the model parameters? If there is a model parameter that cannot be estimated from these data, mark it "UNKNOWN."

Solution: $P(A=T)=2/4$, $P(B=T)=1/4$, and

a,b	$P(C=T\|A=a,B=b)$
F, F	$1/2$
F, T	UNKNOWN
T, F	$0/1$
T, T	$0/1$

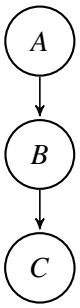
(d) Use the table of data given in part (c), but this time, estimate the data using Laplace smoothing, with a smoothing parameter of $k = 1$.

Solution: $P(A = T) = 3/6$, $P(B = T) = 1/3$, and

a, b	$P(C = T \mid a, b)$
F, F	$2/4$
F, T	$1/2$
T, F	$1/3$
T, T	$1/3$

Question 11 ***(0 points)***

Consider the following Bayes network (all variables are binary):



$$P(A=T)=0.8$$

a	$P(B=T\|A=a)$
False	0.7
True	0.3

b	$P(C=T\|B=b)$
False	0.5
True	0.7

(a) What is $P(C=T)$? Write your answer in numerical form, but you don't need to simplify.

Solution:

$$\begin{aligned}P(C=T) &= \Pr(\neg A,\neg B,C)+\Pr(\neg A,B,C)+\Pr(A,\neg B,C)+\Pr(A,B,C)\\ &= (0.2)(0.3)(0.5)+(0.2)(0.7)(0.7)+(0.8)(0.7)(0.5)+(0.8)(0.3)(0.7)\end{aligned}$$

(b) What is $P(A=T|B=T,C=T)$? Write your answer in numerical form, but you don't need to simplify.

Solution:

$$\begin{aligned}P(A=T|B=T,C=T) &= \frac{\Pr(A,B,C)}{\Pr(A,B,C)+\Pr(\neg A,B,C)}\\ &= \frac{(0.8)(0.3)(0.7)}{(0.8)(0.3)(0.7)+(0.2)(0.7)(0.7)}\end{aligned}$$

(c) You've been asked to re-estimate the parameters of the network based on the following observations:

Observation	a	b	c
1	True	False	False
2	False	False	True
3	True	True	False
4	False	False	False

Given the data in the table, what are the maximum likelihood estimates of the model parameters? If there is a model parameter that cannot be estimated from these data, mark it "UNKNOWN."

Solution:

$$P(A=T)=2/4$$

a	$P(B=T\|A=a)$
False	0/2
True	1/2

b	$P(C=T\|B=b)$
False	1/3
True	0/1

(d) Use the table of data given in part (c), but this time, estimate the data using Laplace smoothing, with a smoothing parameter of $k=1$.

Solution: $P(A=T)=3/6$, $P(B=T)=1/3$, and

a	$P(B=T\|A=a)$
False	1/4
True	2/4

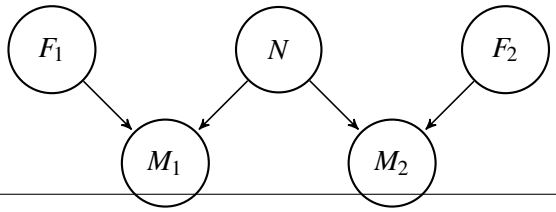
b	$P(C=T\|B=b)$
False	2/5
True	1/3

Question 12 ***(0 points)***

Two astronomers in different parts of the world make measurements M_1 and M_2 of the number of stars N in some small region of the sky, using their telescopes. Under normal circumstances, this experiment has three possible outcomes: either the measurement is correct (with probability $1-2e-f$), or the measurement overcounts the stars by one (one star too high) with probability e, or the measurement undercounts the stars by one (one star too low) with probability e. There is also the possibility, however, of a large measurement error in either telescope (events F_1 and F_2, respectively, each with probability f), in which case the measured number will be **at least** three stars too low (regardless of whether the scientist makes a small error or not), or, if N is less than 3, fail to detect any stars at all.

(a) Draw a Bayesian network for this problem.

Solution: A solution must include the variables N, M_1, M_2 with the dependencies shown below. The variables F_1, F_2 are optional:



(b) Write out a conditional distribution for $P(M_1|N)$ for the case where $N \in \{1,2,3\}$ and $M_1 \in \{0,1,2,3,4\}$. Each entry in the conditional distribution table should be expressed as a function of the parameters e and/or f.

Solution:

	M_1				
N	0	1	2	3	4
1	$e+f$	$1-2e-f$	e	0	0
2	f	e	$1-2e-f$	e	0
3	f	0	e	$1-2e-f$	e

(c) Suppose $M_1 = 1$ and $M_2 = 3$. What are the possible numbers of stars if you assume no prior constraint on the values of N?

Solution: $N = 2$ is possible, if both made small mistakes. $N = 4$ is possible, if M_2 made a small and M_1 a big mistake. $N \geq 6$ is possible, if both M_1 and M_2 made big mistakes.

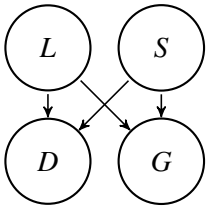
(d) What is the most likely number of stars, given the observations $M_1 = 1, M_2 = 3$? Explain how to compute this, or if it is not possible to compute, explain what additional information is needed and how it would affect the result.

Solution: We need to find the value of N that maximizes $P(N, M_1 = 1, M_2 = 3)$. We have that $P(N = 2, M_1 = 1, M_2 = 3) = P(N = 2)e^2$. We know that $P(N = 4, M_1 = 1, M_2 = 3) \leq P(N = 4)fe$; we don't know exactly how much it is, because we don't know $P(M_1 = 1|N = 4)$, but we know that $P(M_1 = 1|N = 4) \leq f$. So if $P(N = 2)e > P(N = 4)f$, $N = 2$ is the most probable value. If $P(N = 2)e \leq P(N = 4)f$, then it depends on the way in which big errors are distributed among the various values that are "at least three stars" too small.

Question 13 ***(0 points)***

Maria likes ducks and geese. She notices that when she leaves the heat lamp on (in her back yard), she is likely to see ducks and geese. When the heat lamp is off, she sees ducks and geese in the summer, but not in the winter.

(a) The following Bayes net summarizes Maria's model, where the binary variables D,G,L, and S denote the presence of ducks, geese, heat lamp, and summer, respectively:



On eight randomly selected days throughout the year, Maria makes the observations shown in the table below:

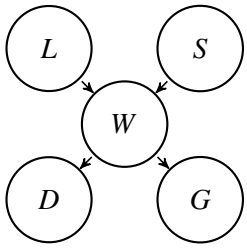
day	D	G	L	S	day	D	G	L	S
1	0	1	1	0	5	1	0	0	1
2	1	0	1	0	6	1	0	1	1
3	0	0	0	0	7	0	1	1	1
4	0	0	0	0	8	0	1	0	1

Write the maximum-likelihood conditional probability tables for D, G, L and S.

Solution: We have that $P(S=T)=0.5$, $P(L=T)=0.5$, and

s	l	$P(D=T\|S=s,L=l)$	$P(G=T\|S=s,L=l)$
0	0	0	0
0	1	0.5	0.5
1	0	0.5	0.5
1	1	0.5	0.5

(b) Maria speculates that ducks and geese don't really care whether the lamp is lit or not, they only care whether or not the temperature in her yard is warm. She defines a binary random variable, W, which is 1 when her back yard is warm, and she proposes the following revised Bayes net:



She forgot to measure the temperature in her back yard, so W is a hidden variable. Her initial guess is that $\Pr(D|W) = \frac{2}{3}$, $\Pr(D|\neg W) = \frac{1}{3}$, $\Pr(G|W) = \frac{2}{3}$, $\Pr(G|\neg W) = \frac{1}{3}$, $\Pr(W|L \wedge S) = \frac{2}{3}$, $\Pr(W|\neg(L \wedge S)) = \frac{1}{3}$. Find the posterior probability $\Pr(W|\text{day})$ for each of the 8 days, day $\in \{1,\ldots,8\}$, whose observations are shown in the Table in part (a).

Solution:

day	1	2	3	4	5	6	7	8
$P(W\|\text{day})$	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{9}$	$\frac{1}{9}$	$\frac{1}{3}$	$\frac{2}{3}$	$\frac{2}{3}$	$\frac{1}{3}$

Question 14 ***(0 points)***

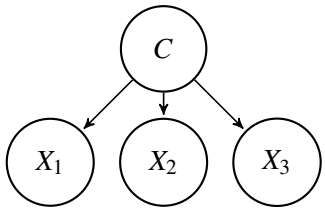
Consider the data points in Table 1, representing a set of seven patients with up to three different symptoms. We want to use the Naïve Bayes assumption to diagnose whether a person has the flu based on the symptoms.

Sore Throat	Stomachache	Fever	Flu
No	No	No	No
No	No	Yes	Yes
No	Yes	No	No
Yes	No	No	No
Yes	No	Yes	Yes
Yes	Yes	No	Yes
Yes	Yes	Yes	No

Table 1: Symptoms of seven patients, three of whom had the flu.

(a) Define random variables, and show the structure of the Bayes network representing a Naïve Bayes classifier for the flu, using the variables shown in Table 1.

Solution: The binary variables could be called F, T, S, and E, representing the presence of flu, sore throat, stomach ache, and fever, respectively. The Bayes net is then



(b) Calculate the maximum likelihood conditional probability tables.

Solution:

f	$P(F=f)$	$\Pr(T\|F=f)$	$\Pr(S\|F=f)$	$\Pr(E\|F=f)$
0	$4/7$	$1/2$	$1/2$	$1/4$
1	$3/7$	$2/3$	$1/3$	$2/3$

(c) If a person has stomachache and fever, but no sore throat, what is the probability of him or her having the flu (according to the conditional probability tables you calculated in part (b))?

Solution:

$$\begin{aligned}
\Pr(F|\neg T,S,E) &= \frac{\Pr(\neg T,S,E,F)}{\Pr(\neg T,S,E)} \\
&= \frac{\Pr(F,\neg T,S,E)}{\Pr(F,\neg T,S,E)+\Pr(\neg F,\neg T,S,E)} \\
&= \frac{(3/7)(1/3)(1/3)(2/3)}{(3/7)(1/3)(1/3)(2/3)+(4/7)(1/2)(1/2)(1/4)} \\
&= \frac{8}{17}
\end{aligned}$$

Question 15 ***(7 points)***

There is a lion in a cage in the dungeons under Castle Rock.

- The zookeeper goes on vacation with a probability of P.
- If the zookeeper is on vacation, the lion doesn't get fed. If not, the lion gets fed with probability Q, and goes hungry with probability $1-Q$.
- If the lion has not been fed, and you try to pet it, then it will bite your hand with probability R. If it has been fed, it will only bite you with probability S.

(a) (2 points) Draw a Bayes network with three random variables: $Z = 1$ if the zookeeper is on vacation, $F = 1$ if the lion gets fed today, $B = 1$ if it will bite the hand of the next person who tries to pet it. Draw edges to show the dependencies specified by the problem statement above.

Solution:

$$Z \longrightarrow F \longrightarrow B$$

(b) (3 points) Circe pets the lion, and it bites her hand. In terms of the unknown parameters P, Q, R, and S, what is the probability that the zookeper is on vacation?

Solution:

$$\Pr(Z|B) = \frac{\Pr(Z,F,B)+\Pr(Z,\neg F,B)}{\Pr(Z,F,B)+\Pr(Z,\neg F,B)+\Pr(\neg Z,F,B)+\Pr(\neg Z,\neg F,B)}$$

$$= \frac{PR}{PR+(1-P)QS+(1-P)(1-Q)R}$$

(c) (2 points) Lord Lucky, the Lord of Castle Rock, hires a troupe of circus performers to pet the lion, once per day, in an attempt to learn the parameters P, Q, R, and S. Over the course of seven days, he collects the following observations. Based on these observations, find maximum-likelihood estimates of P, Q, R, and S.

Day	Z	F	B
1	1	0	1
2	0	1	1
3	0	1	0
4	1	0	0
5	0	0	1
6	0	1	0
7	0	1	0

Solution: The probability estimates are

$$P = \Pr(Z) = \frac{2}{7}$$

$$Q = \Pr(F|\neg Z) = \frac{4}{5}$$

$$R = \Pr(B|\neg F) = \frac{2}{3}$$

$$S = \Pr(B|F) = \frac{1}{4}$$

Question 16 ***(0 points)***

You're creating sentiment analysis. You have a training corpus with four movie reviews:

Review #	Sentiment	Review
1	+	what a great movie
2	+	I love this film
3	-	what a horrible movie
4	-	I hate this film

Let $Y = 1$ for positive sentiment, $Y = 0$ for negative sentiment.

(a) What's the maximum likelihood estimate of $P(Y = 1)$?

Solution: Maximum likelihood estimate is

$$P(Y = 1) = \frac{\text{\# times } Y = 1}{\text{\# training tokens}} = \frac{2}{4} = \frac{1}{2}$$

(b) Find maximum likelihood estimates $P(W|Y = 1)$ and $P(W|Y = 0)$ for the following ten words:

$$w \in \{\text{what,a,movie,I,this,film,great,love,horrible,hate}\}$$

Solution: There are three cases. For the words $w \in \{\text{what,a,movie,I,this,film}\}$, $P(W = w|Y = 0) = P(W = w|Y = 1) = 1/8$. For the words $w \in \{\text{great,love}\}$, $P(W = w|Y = 0) = 0$, and $P(W = w|Y = 1) = 1/8$. For the words $w \in \{\text{horrible,hate}\}$, $P(W = w|Y = 1) = 0$, and $P(W = w|Y = 0) = 1/8$.

(c) Use Laplace smoothing, with a smoothing parameter of $k = 1$, to estimate $P(W|Y = 1)$ and $P(W|Y = 0)$ for the ten words $w \in \{\text{what,a,movie,I,this,film,great,love,horrible,hate}\}$.

Solution: There are three cases. For the words $w \in \{\text{what,a,movie,I,this,film}\}$, $P(W = w|Y = 0) = P(W = w|Y = 1) = 2/18$. For the words $w \in \{\text{great,love}\}$, $P(W = w|Y = 0) = 1/18$, and $P(W = w|Y = 1) = 2/18$. For the words $w \in \{\text{horrible,hate}\}$, $P(W = w|Y = 1) = 1/18$, and $P(W = w|Y = 0) = 2/18$.

(d) Using some other method (unknown to you), your professor has estimated the following conditional probability table:

y	$P(\text{great}\|Y = y)$	$P(\text{love}\|Y = y)$	$P(\text{horrible}\|Y = y)$	$P(\text{hate}\|Y = y)$
1	0.01	0.01	0.005	0.005
0	0.005	0.005	0.01	0.01

and $P(Y = 1) = 0.5$. All other words (except great, love, horrible, and hate) can be considered out-of-vocabulary, and you can assume that $P(W|Y = y) = 1$ for all out-of-vocabulary words. Under these assumptions, what is the probability $P(Y = 1|R)$ that the following 14-word review is a positive review?

$$R = \{\text{"I'm horrible fond of this movie, and I hate anyone who insults it."}\}$$

Solution:

$$P(Y=1|R) = \frac{P(Y=1,R)}{P(Y=1,R)+P(Y=0,R)} = \frac{(0.5)(0.005)(0.005)}{(0.5)(0.005)(0.005)+(0.5)(0.01)(0.01)} = \frac{1}{5}$$

Question 17 ***(0 points)***

Laplace invented "Laplace smoothing" in order to estimate the probability that the sun will rise tomorrow. Suppose he had historical records indicating that the sun had been observed to rise on 1,826,200 consecutive days (and the event "the sun did not rise today" has never been observed). What probability would Laplace smoothing estimate for the event "The sun wil rise tomorrow"?

Solution:

$$\Pr(R) = \frac{1826200+1}{1826200+2}$$

Question 18 ***(0 points)***

You're on a phone call with your friend, trying to help figure out why their computer won't start. There are only two possibilities, $Y = \text{CPU}$, or $Y = \text{PowerSupply}$, with prior probability $P(Y = \text{CPU}) = 0.3$.

You ask your friend whether the computer makes noise when they try to turn it on. There are two possibilities, $X = \text{quiet}$, and $X = \text{loud}$. You know that a power supply problem often leaves a quiet computer, but that the relationship is stochastic, as shown:

$$P(X = \text{noise}|Y = \text{CPU}) = 0.8, \quad P(X = \text{noise}|Y = \text{PowerSupply}) = 0.4$$

(a) What is the MAP classifier function $f(X)$, as a function of X?

Solution: The joint probabilities of evidence and label are:

$$P(\text{noise}, \text{CPU}) = 0.24, \quad P(\text{noise}, \text{PowerSupply}) = 0.28$$
$$P(\text{quiet}, \text{CPU}) = 0.06, \quad P(\text{quiet}, \text{PowerSupply}) = 0.42$$

Choosing the maximum *a posteriori* label given each observation gives

$$f(\text{noise}) = \text{PowerSupply}, \quad f(\text{quiet}) = \text{PowerSupply}$$

In other words, regardless of whether the computer is noisy or quiet, the power supply is always the most probable source of the problem.

(b) What is the Bayes error rate?

Solution: The Bayes error rate is the probability of error of the optimal classifier, which is $P(\text{noise}, \text{CPU}) + P(\text{quiet}, \text{CPU}) = 0.3$.

(c) CPU damage is more expensive than power supply damage, so let's define a false alarm to be the case where your classifier says $f(X) = \text{CPU}$, but the actual problem is $Y = \text{PowerSupply}$. Under this definition, what are the false-alarm rate and missed-detection rate of the MAP classifier?

Solution: The MAP classifier always guesses "Power Supply," so the false alarm and missed detection rates are

$$P(f(X) = \text{CPU}|Y = \text{PowerSupply}) = 0.0$$
$$P(f(X) = \text{PowerSupply}|Y = \text{CPU}) = 1.0$$

Question 19 ***(0 points)***

You're trying to determine whether a particular newspaper article is of class $Y = 0$ or $Y = 1$. The prior probability of class $Y = 1$ is $P(Y = 1) = 0.4$. The newspaper is written in a language that only has four words, so that the i^{th} word in the article must be $W_i \in \{0, 1, 2, 3\}$, with probabilities given by:

$$\begin{array}{ll} P(W_i = 0|Y = 0) = 0.3 & P(W_i = 0|Y = 1) = 0.1 \\ P(W_i = 1|Y = 0) = 0.1 & P(W_i = 1|Y = 1) = 0.1 \\ P(W_i = 2|Y = 0) = 0.1 & P(W_i = 2|Y = 1) = 0.3 \end{array}$$

The article is only three words long; it contains the words

$$A = (W_1 = 3, W_2 = 2, W_3 = 0)$$

What is $P(Y = 1, A)$?

Solution:

$P(Y = 1, A) = P(Y = 1)P(W_1 = 3|Y = 1)P(W_2 = 2|Y = 1)P(W_3 = 0|Y = 1) = (0.4)(0.5)(0.3)(0.1)$

Question 20 ***(0 points)***

The University of Illinois Vaccavolatology Department has four professors, named Aya, Bob, Cho, and Dale. The building has only one key, so we take special care to protect it. Every day Aya goes to the gym, and on the days she has the key, 60% of the time she forgets it next to the bench press. When that happens one of the other three TAs, equally likely, always finds it since they work out right after. Bob likes to hang out at Einstein Bagels and 50% of the time he is there with the key, he forgets the key at the shop. Luckily Cho always shows up there and finds the key whenever Bob forgets it. Cho has a hole in her pocket and ends up losing the key 80% of the time somewhere on Goodwin street. However, Dale takes the same path to campus and always finds the key. Dale has a 10% chance to lose the key somewhere in the Vaccavolatology classroom, but then Cho picks it up. The professors lose the key at most once per day, around noon (after losing it they become extra careful for the rest of the day), and they always find it the same day in the early afternoon.

(a) Let X_t = the first letter of the name of the person who has the key ($X_t \in \{A, B, C, D\}$). Find the maximum likelihood estimates of the Markov transition probabilities $P(X_t|X_{t-1})$.

Solution:

	X_t			
X_{t-1}	A	B	C	D
A	0.4	0.2	0.2	0.2
B	0	0.5	0.5	0
C	0	0	0.2	0.8
D	0	0	0.1	0.9

(b) Sunday night Bob had the key (the initial state distribution assigns probability 1 to $X_0 = B$ and probability 0 to all other states). The first lecture of the week is Tuesday at 4:30pm, so one of the professors needs to open the building at that time. What is the probability for each professor to have the key at that time? Let X_0, X_{Mon} and X_{Tue} be random variables corresponding to who has the key Sunday, Monday, and Tuesday evenings, respectively. Fill in the probabilities in the table below.

Professor	$P(X_0)$	$P(X_{Mon})$	$P(X_{Tue})$
A	0		
B	1		
C	0		
D	0		

Solution:

Professor	$P(X_0)$	$P(X_{Mon})$	$P(X_{Tue})$
A	0	0	0
B	1	0.5	0.25
C	0	0.5	0.35
D	0	0	0.4

Question 21 ***(0 points)***

Consider a hidden Markov model (HMM) whose hidden variable denotes part of speech (POS), $X_t \in \{N, V\}$ where N =noun, V =verb, the initial state probability is $P(X_1 = N) = 0.8$, and the transition probabilities are $P(X_t = N|X_{t-1} = N) = 0.1$ and $P(X_t = V|X_{t-1} = V) = 0.1$. Suppose we have the observation probability matrix given in Table 2. You are given the sentence "bill rose." You want to

E_t	rose	bill	likes
$P(E_t\|X_t = N)$	0.4	0.4	0.2
$P(E_t\|X_t = V)$	0.2	0.2	0.6

Table 2: Observation probabilities for a simple POS HMM.

figure out whether each of these two words, "bill" and "rose", is being used as a noun or a verb.

(a) List the four possible combinations of (X_1, X_2). For each possible combination, give $P(X_1, E_1, X_2, E_2)$.

Solution:

$P(X_1, E_1, X_2, E_2)$	$X_2 = N$	$X_2 = V$
$X_1 = N$	$(0.8)(0.4)(0.1)(0.4)$	$(0.8)(0.4)(0.9)(0.2)$
$X_1 = V$	$(0.2)(0.2)(0.9)(0.4)$	$(0.2)(0.2)(0.1)(0.2)$

(b) Find $P(X_2 = V|E_1 = \text{bill}, E_2 = \text{rose})$.

Solution:

$$\begin{aligned} P(E, X_2 = V) &= P(X_1 = N, E_1, X_2 = V, E_2) + P(X_1 = V, E_1, X_2 = V, E_2) \\ &= (0.8)(0.4)(0.9)(0.2) + (0.2)(0.2)(0.1)(0.2) \\ P(E, X_2 = N) &= P(X_1 = N, E_1, X_2 = N, E_2) + P(X_1 = V, E_1, X_2 = N, E_2) \\ &= (0.8)(0.4)(0.1)(0.4) + (0.2)(0.2)(0.9)(0.4) \end{aligned}$$

Dividing the first row by the sum of the two rows, we get

$$P(X_2 = V|E) = \frac{(0.8)(0.4)(0.9)(0.2) + (0.2)(0.2)(0.1)(0.2)}{(0.8)(0.4)(0.9)(0.2) + (0.2)(0.2)(0.1)(0.2) + (0.8)(0.4)(0.1)(0.4) + (0.2)(0.2)(0.9)(0.4)}$$

(c) Use the Viterbi algorithm to find the most likely state sequence for this sentence.

Solution:

- To find the backpointer from $X_2 = N$, we find the maximum among the two possibilities $P(X_1 = N, E_1, X_2 = N, E_2)$ and $P(X_1 = V, E_1, X_2 = N, E_2)$. The larger of the two is $P(X_1 = V, E_1, X_2 = N, E_2) = (0.2)(0.2)(0.9)(0.4)$, so the backpointer from $X_2 = N$ points to $X_1 = V$.
- To find the backpointer from $X_2 = V$, we find the maximum among the two possibilities $P(X_1 = N, E_1, X_2 = V, E_2)$ and $P(X_1 = V, E_1, X_2 = V, E_2)$. The larger of the two is $P(X_1 =$

$N, E_1, X_2 = V, E_2) = (0.8)(0.4)(0.9)(0.2)$, so the backpointer from $X_2 = V$ points to $X_1 = N$.

- To find the best terminal state, then, we find the maximum among the two possibilities $P(X_1 = V, E_1, X_2 = N, E_2)$ and $P(X_1 = N, E_1, X_2 = V, E_2)$. The larger of the two is $P(X_1 = N, E_1, X_2 = V, E_2) = (0.8)(0.4)(0.9)(0.2)$, so the maximum likelihood state sequence is $(X_1, X_2) = (N, V)$.

Question 22 ***(0 points)***

A particular hidden Markov model (HMM) has state variable X_t, and observation variables E_t, where t denotes time. Suppose that this HMM has two states, $X_t \in \{0,1\}$, and three possible observations, $E_t \in \{0,1,2\}$. The initial state probability is $P(X_1 = 1) = 0.3$. The transition and observation probability matrices are

X_{t-1}	$P(X_t = 1 \mid X_{t-1})$	X_t	$P(E_t = 0 \mid X_t)$	$P(E_t = 1 \mid X_t)$
0	0.6	0	0.4	0.1
1	0.4	1	0.1	0.6

Suppose that, in a particular test of the HMM, the observation sequence is

$$\{E_1, E_2\} = \{2, 1\}$$

(a) What is the joint probability $P(X_1 = 1, E_1 = 2, X_2 = 0)$?

Solution:

$$\begin{aligned} P(X_1 = 1, E_1 = 2, X_2 = 0) &= P(X_1 = 1)P(E_1 = 2 \mid X_1 = 1)P(X_2 = 0 \mid X_1 = 1) \\ &= (0.3)(0.3)(0.6) \end{aligned}$$

(b) What is the probability of the most likely state sequence ending in $X_2 = 0$? In other words, what is $\max_{X_1} P(X_1, E_1 = 2, X_2 = 0, E_2 = 1)$?

Solution:

$$\begin{aligned} \max_{X_1} P(X_1, E_1 = 2, X_2 = 0, E_2 = 1) &= \max_{X_1} P(X_1)P(E_1 = 2 \mid X_1)P(X_2 = 0 \mid X_1)P(E_2 = 1 \mid X_2 = 0) \\ &= \max\left((0.7)(0.5)(0.4)(0.1), (0.3)(0.3)(0.6)(0.1)\right) \\ &= (0.7)(0.5)(0.4)(0.1) \end{aligned}$$

Question 23 ***(0 points)***

Consider the following binary logic function:

$$y = \neg\left((x_1 \wedge \neg x_2 \wedge \neg x_3) \vee (x_1 \wedge x_2 \wedge \neg x_3)\right)$$

Convert truth values to numbers in the obvious way: let $x_i = 1$ be a synonoym for $x_i =$ **True**, and let $x_i = 0$ by a synonym for $x_i =$ **False**. Let $\mathbf{x} = [x_1, x_2, x_3]^T$ and $\mathbf{w} = [w_1, w_2, w_3]^T$, let $\mathbf{x}^T\mathbf{w}$ denote the dot product of vectors $\mathbf{x}$ and $\mathbf{w}$, and let $u(\cdot)$ denote the unit step function. Find a set of parameters w_1, w_2, w_3 and b such that the logic function shown above can be computed as $y = u(w^T x + b)$.

Solution: Drawing up a truth table, we get

x_1	x_2	x_3	y
0	0	0	1
0	0	1	1
0	1	0	1
0	1	1	1
1	0	0	0
1	0	1	1
1	1	0	0
1	1	1	1

If we plot these eight points in 3D space, we find that the dots for which $Y = 0$ and the dots for which $Y = 1$ are separated, for example, by the plane $-x_1 + x_3 + 0.5 = 0$, with the $Y = 1$ dots on the side of this plane where $-x_1 + x_3 + 0.5 > 0$. The linear classifier can therefore use $\mathbf{w} = [-1, 0, 1]^T$, $b = 0.5$.

Question 24 ***(0 points)***

We want to implement a classifier that takes two input values, where each value is either 0, 1 or 2, and outputs a 1 if at least one of the two inputs has value 2; otherwise it outputs a 0. Can this function be learned by a Perceptron? If so, construct a Perceptron that does it; if not, why not.

Solution: In this case the input space of all possible examples with their target outputs is:

	0	1	2
2	1	1	1
1	0	0	1
0	0	0	1

Since there is clearly no line that can separate the two classes, this function is not linearly separable and so it cannot be learned by a Perceptron.

Question 25 ***(0 points)***

Consider a problem with a binary label variable, Y, whose prior is $P(Y=1)=0.4$. Suppose that there are 100 binary evidence variables, $X=[X_1,\ldots,X_{100}]$, each with likelihoods given by $P(X_i=1|Y=0)=0.3$ and $P(X_i=1|Y=1)=0.8$ for $1\leq i\leq 100$.

(a) Specify the classifier function, $f(\mathbf{x})$, for a naive Bayes classifier, where $\mathbf{x}=[x_1,\ldots,x_{100}]^T$ is the set of observed values of the evidence variables. You might find it useful to define $N(\mathbf{x})=$ the number of nonzero elements of the binary vector $\mathbf{x}$; note that $0\leq N(\mathbf{x})\leq 100$.

Solution:

$$f(\mathbf{x})=\begin{cases} 1 & P(Y=1)\prod_{i:x_i=1}P(X_i=1|Y=1)\prod_{i:x_i=0}P(X_i=0|Y=1) > \\ & (1-P(Y=1))\prod_{i:x_i=1}P(X_i=1|Y=0)\prod_{i:x_i=0}P(X_i=0|Y=0) \\ 0 & \text{otherwise} \end{cases}$$

Plugging in the parameter values, we get:

$$f(\mathbf{x})=\begin{cases} 1 & 0.4(0.8)^{N(\mathbf{x})}(0.2)^{100-N(\mathbf{x})} > 0.6(0.3)^{N(\mathbf{x})}(0.7)^{100-N(\mathbf{x})} \\ 0 & \text{otherwise} \end{cases}$$

Another way to write this would be:

$$f(\mathbf{x})=\begin{cases} 1 & 0.4\prod_{i=1}^{100}(0.8)^{x_i}(0.2)^{1-x_i} > 0.6\prod_{i=1}^{100}(0.3)^{x_i}(0.7)^{1-x_i} \\ 0 & \text{otherwise} \end{cases}$$

(b) The naive Bayes classifier can be written as

$$f(\mathbf{x})=\begin{cases} 1 & \mathbf{w}^T\mathbf{x}+b>0 \\ 0 & \text{otherwise} \end{cases},$$

where $\mathbf{w}^T\mathbf{x}$ is the dot product between the vectors $\mathbf{w}$ and $\mathbf{x}$. Find $\mathbf{w}$ and b (write them as expressions in terms of constants; don't simplify).

Solution:

$$f(\mathbf{x})=\begin{cases} 1 & \ln(0.4)+\sum_{i=1}^{100}x_i\ln(0.8)+(1-x_i)\ln(0.2) > \ln(0.6)+\sum_{i=1}^{100}x_i\ln(0.3)+(1-x_i)\ln(0.7) \\ 0 & \text{otherwise} \end{cases}$$

so the parameters are

$$\begin{aligned} b &= \ln(0.4)-\ln(0.6)+100\ln(0.2)-100\ln(0.7) \\ w_i &= \ln(0.8)-\ln(0.2)-\ln(0.3)+\ln(0.7), \quad 1\leq i\leq 100 \end{aligned}$$

Question 26 ***(10 points)***

You are a Hollywood producer. You have a script in your hand and you want to make a movie. Before starting, however, you want to predict if the film you want to make will rake in huge profits, or utterly fail at the box office. You hire two critics A and B to read the script and rate it on a scale of 1 to 5 (assume only integer scores). Each critic reads it independently and announces their verdict. Of course, the critics might be biased and/or not perfect, therefore you may not be able to simply average their scores. Instead, you decide to use a perceptron to classify your data. There are three features: a constant bias, and the two reviewer scores. Thus $f_0 = 1$ (a constant bias), $f_1 =$ score given by reviewer A, and $f_2 =$ score given by reviewer B.

Movie Name	A	B	Profit
Pellet Power	1	1	No
Ghosts!	3	2	Yes
Pac is bac	4	5	No
Not a Pizza	3	4	Yes
Endless Maze	2	3	Yes

(a) (5 points) Train the perceptron to generate $\hat{Y} = 1$ if the movie returns a profit, $\hat{Y} = -1$ otherwise. The initial weights are $w_0 = -1, w_1 = 0, w_2 = 0$. Present each row of the table as a training token, and update the perceptron weights before moving on to the next row. Use a learning rate of $\alpha = 1$. After each of the training examples has been presented once (one epoch), what are the weights?

Solution: The first row of the table is correctly classified, therefore the weights are not changed. The second row is incorrectly classified, therefore the weights are updated as $W = W + YF = [0,3,2]$. Using these weights results in misclassification of the third row, therefore the weights are updated again to $[-1,-1,-3]$. Using these weights results in misclassification of the fourth row, therefore the weights are updated again to $[0,2,1]$. These weights correctly classify the fifth row.

(b) (3 points) Suppose that, instead of learning whether or not the movie is profitable, you want to learn a perceptron that will always output $\hat{Y} = +1$ when the total of the two reviewer scores is more than 8, and $\hat{Y} = -1$ otherwise. Is this possible? If so, what are the weights w_0, w_1, and w_2 that will make this possible?

Solution: Yes, a perceptron can learn this function. Any weights such that $w_1 = w_2$ and $w_0 = -8w_1$ are correct; for example, the weights $[-8,1,1]$.

(c) (2 points) Instead of either part (a) or part (b), suppose you want to learn a perceptron that will always output $\hat{Y} = +1$ when the two reviewers agree (when their scores are exactly the same), and will output $\hat{Y} = -1$ otherwise. Is this possible? If so, what are the weights w_0, w_1 and w_2 that will make this possible?

Solution: This problem is the arithmetic complement of the XOR problem, therefore it is not linearly separable, and cannot be learned by a perceptron.

Question 27 ***(0 points)***

An image classification algorithm is being trained using the multiclass perceptron learning rule. There are 10 classes, each parameterized by a weight vector $\mathbf{w}_k$, for $0 \leq k \leq 9$. During the last round of training, all of the training tokens were correctly classified. Which of the weight vectors were updated, and why?

Solution: None. The perceptron learning rule updates the weight vectors only if the classifier makes a mistake.

Question 28 ***(0 points)***

Logistic regression is trained using gradient descent, with the goal of achieving the Bayes error rate (the lowest possible error rate) on testing data. There are many reasons why gradient descent might not successfully minimize the number of test-corpus errors. List at least three.

Solution: Here are a few:

1. **Wrong criterion:** The number of errors is not a differentiable criterion, so gradient descent has to minimize a differentiable approximation. Minimizing the differentiable approximation might not actually minimize the number of errors.
2. **Generalization error:** Minimizing error on the training corpus might not minimize error on the test corpus.
3. **Approximation error:** The Bayes error rate might not be achievable by a linear classifier. Since logistic regression learns a linear classifier, it might not be possible to achieve the Bayes error rate.
4. **Local optimum:** Gradient descent converges to a local minimum of the training criterion, not a global minimum.
5. **Computational limitations:** The amount of computation available for training might not be enough for gradient descent to fully converge.

Question 29 ***(0 points)***

The softmax function is defined as

$$a_k = \frac{\exp(z_k)}{\sum_j \exp(z_j)}$$

Find da_5/dz_3 in terms of z_3, z_5, a_3 and/or a_5.

Solution:

$$\frac{da_5}{dz_3} = -\frac{\exp(z_5)}{(\sum_j \exp(z_j))^2}\frac{d\sum_j \exp(z_j)}{dz_3} = -\frac{\exp(z_5)\exp(z_3)}{(\sum_j \exp(z_j))^2} = -a_5 a_3$$

Question 30 ***(0 points)***

A particular two-layer neural net has input vector $\mathbf{x} = [x_1, x_2]^T$, hidden layer activations $\mathbf{h}^{(1)} = [h_1^{(1)}, h_2^{(1)}]^T$, and a scalar output f. Its weights and biases are stored in the first-layer weight matrix $\mathbf{W}^{(1)}$ and bias vector $\mathbf{b}^{(1)}$, and the second-layer bias vector $\mathbf{w}^{(2)}$ and bias $b^{(2)}$, respectively. The weights and biases are given to you; their values are also provided in Table 3. The hidden layer nonlineary is ReLU; the output nonlinearity is a logistic sigmoid.

Table 3: Variables used in Problem 30.

$$\mathbf{W}^{(1)} = \begin{bmatrix} 3 & 4 \\ 0 & 9 \end{bmatrix}$$
$$\mathbf{b}^{(1)} = [-3, 3]^T$$
$$\mathbf{w}^{(2)} = [5, 4]^T$$
$$b^{(2)} = -7$$

(a) Suppose the input is $\mathbf{x} = [9, -6]^T$. What is $\mathbf{h}^{(1)}$? Write your answer as a vector of ReLUs of sums of products; do not simplify.

Solution:

$$\mathbf{h}^{(1)} = [\text{ReLU}((3)(9) + (4)(-6) + -3), \text{ReLU}((0)(9) + (9)(-6) + 3)]^T$$

(b) Suppose the hidden layer is $\mathbf{h}^{(1)} = [4, 5]^T$. What is f? Write your answer as a ratio of terms involving the exponential of a sum of products; do not simplify.

Solution:

$$f = 1/(1 + \exp(-(5)(4) - (4)(5) + 7))$$

Question 31 ***(5 points)***

You have a two-layer neural network trained as an animal classifier. The input feature vector is $\mathbf{x} = [x_1, x_2, x_3, 1]^T$, where x_1, x_2, and x_3 are some features, and 1 is multiplied by the bias. There are two hidden nodes, and three output nodes, $\mathbf{y}^* = [y_1^*, y_2^*, y_3^*]^T$, corresponding to the three output classes $y_1^* = \Pr(\text{dog}|\mathbf{x})$, $y_2^* = \Pr(\text{cat}|\mathbf{x})$, $y_3^* = \Pr(\text{skunk}|\mathbf{x})$. The hidden layer uses a sigmoid nonlinearity, the output layer uses a softmax.

(a) (2 points) A Maltese puppy has the feature vector $\mathbf{x} = [2, 20, -1, 1]^T$. Suppose all weights and biases are initialized to zero. What is $\mathbf{y}^*$?

Solution: If all weights and biases are zero, then the excitation of each hidden node is $0 \times 2 + 0 \times 20 + 0 \times (-1) + 0 \times 1 = 0$. With zero input, the sigmoid $1/(1+\exp(-f)) = 0.5$, but weights in the last layer are also all zero, so the excitations at the last layer are all zero. With a softmax nonlinearity, every output node is computing $\exp(0)/\sum_{i=1}^{3}\exp(0) = 1/3$. So

$$\mathbf{y}^* = \left[\frac{1}{3}, \frac{1}{3}, \frac{1}{3}\right]$$

(b) (3 points) Let w_{ij} be the weight connecting the i^{th} output node to the j^{th} hidden node. What is dy_2^*/dw_{21}? Write your answer in terms of y_i^*, w_{ij}, and/or the hidden node activations h_j, for any appropriate values of i and/or j.

Solution: Let's use the notation f_i as the excitation of the i^{th} output node. That allows us to write the softmax as:

$$y_2^* = \frac{\exp(f_2)}{\sum_{j=1}^{3}\exp(f_j)}, \quad f_j = \sum_i w_{ji} h_i$$

Then:

$$\begin{aligned}
\frac{dy_2^*}{dw_{21}} &= \frac{1}{\sum_{i=1}^{3}\exp(f_i)}\frac{d\exp(f_2)}{dw_{21}} + \exp(f_2)\frac{d(1/\sum_i \exp(f_i)))}{dw_{21}} \\
&= \frac{1}{\sum_{i=1}^{3}\exp(f_i)}\exp(f_2)\frac{df_2}{dw_{21}} + \exp(f_2)\left(-\frac{1}{(\sum_{i=1}^{3}\exp(f_i))^2}\right)\frac{d(\sum\exp(f_i)))}{dw_{21}} \\
&= \frac{\exp(f_2)}{\sum_{i=1}^{3}\exp(f_i)}h_1 - \frac{\exp(f_2)}{(\sum_{i=1}^{3}\exp(f_i))^2}\frac{d\exp(f_2)}{dw_{21}} \\
&= \frac{\exp(f_2)}{\sum_{i=1}^{3}\exp(f_i)}h_1 - \frac{\exp(f_2)}{(\sum_{i=1}^{3}\exp(f_i))^2}\exp(f_2)\frac{df_2}{dw_{21}} \\
&= \frac{\exp(f_2)}{\sum_{i=1}^{3}\exp(f_i)}h_1 - \frac{\exp(f_2)^2}{(\sum_{i=1}^{3}\exp(f_i))^2}h_1 \\
&= y_2^* h_1 - (y_2^*)^2 h_1
\end{aligned}$$

Question 32 ***(0 points)***

In a pinhole camera, a light source at (x,y,z) is projected onto a pixel at $(x',y',-f)$ through a pinhole at $(0,0,0)$. Write $\sqrt{(x')^2+(y')^2}$ in terms of x, y, z, and f.

Solution: The pinhole camera equations are

$$x' = \frac{-fx}{z}, \quad y' = \frac{-fy}{z}$$

from which we derive

$$\sqrt{(x')^2+(y')^2} = \frac{f}{z}\sqrt{x^2+y^2}$$

Question 33 ***(0 points)***

The real world contains two parallel infinite-length lines, whose equations, in terms of the coordinates (x,y,z), are parameterized as $ax+by+cz=d$ and $ax+by+cz=e$; in addition, both of these lines are on the ground plane, $y=g$, for some constants (a,b,c,d,e,g). Show that the images of these two lines, as imaged by a pinhole camera, converge to a vanishing point, and give the coordinates (x',y') of the vanishing point.

Solution: The pinhole camera equations are

$$x' = \frac{-fx}{z}, \quad y' = \frac{-fy}{z}$$

From which we derive

$$x = \frac{-zx'}{f}, \quad y = \frac{-zy'}{f}$$

So the equations of the two lines are

$$-\frac{ax'}{f} - \frac{by'}{f} + c = \frac{d}{z}$$
$$-\frac{ax'}{f} - \frac{by'}{f} + c = \frac{e}{z}$$

As $z \to \infty$, the right-hand-sides of these two equations both go to zero, and the equations of both lines converge to

$$ax' + by' = cf$$

In addition, we have $y=g$, so $y' = -fg/z \to 0$, and therefore $x' = cf/a$. The coordinates are $(x',y') = (cf/a, 0)$.

Question 34 ***(0 points)***

Consider the convolution equation

$$z(x',y') = \sum_m \sum_n h(m,n) y(x'-m, y'-n)$$

Where $y(x',y')$ is the original image, $z(x',y')$ is the filtered image, and the filter $h(m,n)$ is given by

$$h(m,n) = \begin{cases} \frac{1}{21} & 1 \le m \le 3, \quad -3 \le n \le 3 \\ -\frac{1}{21} & -3 \le m \le -1, \quad -3 \le n \le 3 \end{cases}$$

Would this filter be more useful for smoothing, or for edge detection? Why?

Solution: The sum of $h(m,n)$, over all m and n, is 0. So if it is filtering a constant-color region, the output would always be zero, regardless of the input color. So it's not very useful for smoothing.

Any given pixel of $Z(x',y')$ is the difference between the pixels $Y(x',y')$ to its left, minus those to its right. Since it's computing a difference, it would be useful for edge detection.

Question 35 ***(0 points)***

Under what circumstances is a difference-of-Gaussians filter more useful for edge detection than a simple pixel difference?

Solution: A difference-of-Gaussians filter first smooths the input image (using a Gaussian smoother), then computes a pixel difference. The smoothing step can reduce random noise. Therefore, this procedure is more useful if the input image has some random noise in it.

Question 36 ***(0 points)***

The pinhole camera equations are

$$x' = \frac{-fx}{z}, \quad y' = \frac{-fy}{z}$$

Explain in words how these equations can be used to show that the image of any object gets smaller as the object gets farther from the camera.

Solution: Two points, on opposite sides of the object, project images onto the film at positions that are inversely proportional to the distance (z) from the object to the camera. Since the positions of these two points on the image are inversely proportional to z, the distance between them is also inversely proportional to z, therefore as the object gets farther from the camera, the distance between the opposite sides of the object (in the image) decreases.

Question 37 ***(0 points)***

Consider two binary random variables, X and Y. Suppose that

$$P(X=1) = a$$
$$P(Y=1) = b$$
$$P(X=1, Y=0) = c$$

In terms of a, b, and/or c, what is $P(Y=1|X=1)$?

Solution:

$$P(Y=1|X=1) = \frac{a-c}{a}$$

Question 38 ***(0 points)***

You've been asked to create a naïve Bayes model of the candy produced by the Santa Claus Candy Company. As your training dataset, you've been given a box containing 80 pieces of candy, of which 8 are strawberry, 48 are raspberry, and 24 are blueberry. In terms of the Laplace smoothing parameter k, estimate the following probabilities:

Solution:

$$P(\text{flavor} = \text{strawberry}|\text{Santa Claus Candy Company}) = \frac{8+k}{80+4k}$$
$$P(\text{flavor} = \text{raspberry}|\text{Santa Claus Candy Company}) = \frac{48+k}{80+4k}$$
$$P(\text{flavor} = \text{blueberry}|\text{Santa Claus Candy Company}) = \frac{24+k}{80+4k}$$
$$P(\text{flavor} = \text{other}|\text{Santa Claus Candy Company}) = \frac{k}{80+4k}$$

Question 39 ***(0 points)***

Describe, in one sentence each, the purpose of (1) a training set, (2) a development test set, (3) an evaluation test set.

Solution: A training set is used to train the model parameters. A development test set is used to compare many different fully-trained models; we choose the one with the best performance on the development test set. An evaluation test set is used to estimate how well the chosen model will perform in the real world.

Question 40 ***(0 points)***

You're trying to create a multi-class perceptron that will classify animals as being either fish, birds, or reptiles. Your feature vector is $\vec{x} = [x_1, x_2, x_3, 1]^T$, where

$$x_1 = \text{fraction of time the animal spends under water}$$
$$x_2 = \text{fraction of time the animal spends on land}$$
$$x_3 = \text{fraction of time the animal spends flying}$$

- Based on your extensive prior knowledge of zoology, you initialize your perceptron with the following weight vectors: $\vec{w}_{\text{fish}} = [1,0,0,0]^T$, $\vec{w}_{\text{reptile}} = [0,1,0,0]^T$, and $\vec{w}_{\text{bird}} = [0,0,1,0]^T$.
- Your first training token is a crocodile, for which $y =$reptile, and $\vec{x} = [0.7, 0.3, 0, 1]^T$.

After training with this training token, what are the numerical values of $\vec{w}_{\text{fish}}$, $\vec{w}_{\text{reptile}}$, and $\vec{w}_{\text{bird}}$? Assume a learning rate of $\eta = 1$.

Solution:

$$\vec{w}_{\text{fish}} = [0.3, -0.3, 0, -1]^T$$
$$\vec{w}_{\text{reptile}} = [0.7, 1.3, 0, 1]^T$$
$$\vec{w}_{\text{bird}} = [0, 0, 1, 0]^T$$

Question 41 ***(0 points)***

In stochastic gradient descent, we train using one training token at a time. Suppose

$$\mathscr{L} = (\mathbf{w}^T\mathbf{x} - y)^2$$

$$\mathbf{w} = \begin{bmatrix} w_1 \\ w_2 \\ b \end{bmatrix}, \quad \mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ 1 \end{bmatrix}$$

In terms of $\mathbf{x}$, $\mathbf{w}$, w_1, w_2, b, x_1, x_2, and/or y, what is $\frac{d\mathscr{L}}{dw_2}$?

Solution:

$$\frac{d\mathscr{L}}{dw_2} = 2(\mathbf{w}^T\mathbf{x} - y)x_2$$

Question 42 ***(0 points)***

Suppose that

$$f = w_{1,1}^{(2)}h_1 + w_{1,2}^{(2)}h_2 + b^{(2)}$$
$$h_1 = \text{ReLU}\left(w_{1,1}^{(1)}x_1 + w_{1,2}^{(1)}x_2 + b_1^{(1)}\right)$$
$$h_2 = \text{ReLU}\left(w_{2,1}^{(1)}x_1 + w_{2,2}^{(1)}x_2 + b_2^{(1)}\right)$$

Assume, for a particular training token, that $h_1 > 0$ and $h_2 > 0$. For that particular training token, what is $\frac{\partial f}{\partial w_{1,1}^{(1)}}$? Express your answer in terms of x_j, h_j, $w_{j,k}^{(l)}$, and/or $b_k^{(l)}$ for any values of j, k, and/or l that may be useful to you.

Solution:

$$\frac{\partial f}{\partial w_{1,1}^{(1)}} = \frac{\partial f}{\partial h_1}\frac{\partial h_1}{\partial w_{1,1}^{(1)}}$$
$$= w_{1,1}^{(2)}x_1$$

Question 43 ***(0 points)***

In 1963, the US Air Force photographed a flying saucer, using a camera fixed on the roof of the Rantoul Air Force Base. The flying saucer appears in the photographic plate at position $(x', y') = (7\text{mm}, -2\text{mm})$. The camera had a focal length of $f = 6\text{mm}$. From this information, what can you say about the real-world location of the flying saucer? List or describe all of the real-world coordinates (x, y, z) that are consistent with these measurements.

Solution:

$$-\frac{x'}{f} = \frac{x}{z}$$
$$-\frac{y'}{f} = \frac{y}{z}$$

So the flying saucer could be at any of the points in the following set:

$$\left\{(x, y, z) : x = -\frac{7}{6}z,\ y = \frac{1}{3}z\right\}$$

Question 44 ***(0 points)***

You are standing on a downward-sloping hillside, with your camera pointed straight ahead of you. Parallel to your line of sight, on your left-hand side (at position $x = -2$ meters), there is a low fence (height 1 meter). The fence descends the hill in front of you, vanishing into a point far in the distance. Let (x', y') denote the position of the fence's vanishing point on your photograph, where x' is horizontal position, y' is vertical position, and $(0, 0)$ is the point directly corresponding to your line of sight.

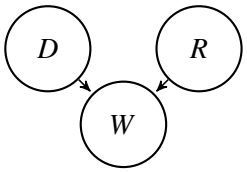
- Is $x' < 0$, $x' = 0$, or $x' > 0$? Explain.
- Is $y' < 0$, $y' = 0$, or $y' > 0$? Explain.

Solution:

- $x' = 0$. The fence is parallel to your line of sight, thus $x = -b$ for some offset b. If we divide by z, and let z go to infinity, we find that $x' = 0$.
- $y' > 0$. The fence is descending, so $y = az + c$ for some negative value of a. Dividing by z, substituting $y/z = -y'/f$, and letting z go to infinity, we find that $y' = -af$, which is positive.

Question 45 ***(0 points)***

Consider the following Bayesian network:



Suppose that the model parameters are as follows:

$$P(D=d) = \frac{1}{2} \text{ for } d \in \{1,2\}$$
$$P(R=r) = \frac{1}{2} \text{ for } r \in \{1,2\}$$
$$P(W=T|D=d,R=r) = \begin{cases} \frac{2}{3} & d \geq r \\ \frac{1}{3} & d < r \end{cases}$$

What is $P(D=2|W=T)$?

Solution:

$$P(D=2|W=T) = \frac{\sum_{r=1}^{2} P(D=2,R=r,W=T)}{\sum_{d=1}^{2}\sum_{r=1}^{2} P(D=d,R=r,W=T)}$$
$$= \frac{\left(\frac{1}{2}\right)\left(\frac{1}{2}\right)\left(\frac{2}{3}\right) + \left(\frac{1}{2}\right)\left(\frac{1}{2}\right)\left(\frac{2}{3}\right)}{\left(\frac{1}{2}\right)\left(\frac{1}{2}\right)\left(\frac{2}{3}\right) + \left(\frac{1}{2}\right)\left(\frac{1}{2}\right)\left(\frac{2}{3}\right) + \left(\frac{1}{2}\right)\left(\frac{1}{2}\right)\left(\frac{2}{3}\right) + \left(\frac{1}{2}\right)\left(\frac{1}{2}\right)\left(\frac{1}{3}\right)}$$

If you wish, you can simplify the last formula to $4/7$, but it's not required.

Question 46 ***(0 points)***

Suppose that you have a hidden Markov model in which the state variables (Y_t) and observation variables (X_t) are binary. The initial-state probability is $P(Y_1 = T) = 0.6$, and the other model parameters are as follows:

Y_{t-1}	$P(Y_t = T \mid Y_{t-1})$
F	0.2
T	0.3

Y_t	$P(X_t = T \mid Y_t)$
F	0.2
T	0.9

What is $P(Y_1 = T \mid X_1 = T, X_2 = T)$?

Solution:

$$
\begin{aligned}
&P(Y_1 = T \mid X_1 = T, X_2 = T) \\
&= \frac{\Pr(Y_1, X_1, Y_2, X_2) + \Pr(Y_1, X_1, \neg Y_2, X_2)}{\Pr(Y_1, X_1, Y_2, X_2) + \Pr(Y_1, X_1, \neg Y_2, X_2) + \Pr(\neg Y_1, X_1, Y_2, X_2) + \Pr(\neg Y_1, X_1, \neg Y_2, X_2)} \\
&= \frac{(0.6)(0.9)(0.3)(0.9) + (0.6)(0.9)(0.7)(0.2)}{(0.6)(0.9)(0.3)(0.9) + (0.6)(0.9)(0.7)(0.2) + (0.4)(0.2)(0.2)(0.9) + (0.4)(0.2)(0.8)(0.2)}
\end{aligned}
$$

Question 47 ***(0 points)***

Suppose that you have a hidden Markov model in which the state variables (Y_t) and observation variables (X_t) are binary. The initial-state probability is $P(Y_1 = T) = 0.6$, and the other model parameters are as follows:

Y_{t-1}	$P(Y_t = F \mid Y_{t-1})$
F	0.0
T	1.0

Y_t	$P(X_t = T \mid Y_t)$
F	0.9
T	0.3

Suppose that the observations are $X_1 = F, X_2 = T$. What is the most likely sequence of values Y_1, Y_2?

Solution: In this problem, the sequences (F,F) and (T,T) are impossible. The other two sequences have the probabilities:

$$
\begin{aligned}
P(Y_1 = T, X_1 = F, Y_2 = F, X_2 = T) &= (0.6)(0.7)(1)(0.9) \\
P(Y_1 = F, X_1 = F, Y_2 = T, X_2 = T) &= (0.4)(0.1)(1)(0.3)
\end{aligned}
$$

The sequence (T,F) has higher probability.

Question 48 ***(0 points)***

Your apartment is haunted by a ghost. Like most ghosts, your ghost tends to sleep for several days at a time. Let $Y_t = T$ if the ghost is awake on day t. You can't see the ghost, but if the ghost is awake, your cat tends to hide under the bed; let $X_t = T$ if your cat is hiding under the bed on day t. Suppose that these probabilities are given by the following distribution, where a, b, c and d are arbitrary constants:

Y_{t-1}	$P(Y_t = T \mid Y_{t-1})$
F	a
T	b

Y_t	$P(X_t = T \mid Y_t)$
F	c
T	d

Suppose you know that the ghost was asleep on day 0 ($Y_0 = F$). You don't know whether or not it was awake on day 1, but you know that your cat hid under the bed ($X_1 = T$). In terms of a, b, c and/or d, find $P(Y_1 = T \mid Y_0 = F, X_1 = T)$.

Solution:

$$P(Y_1 = T \mid Y_0 = F, X_1 = T) = \frac{P(Y_1 = T, X_1 = T \mid Y_0 = F)}{P(X_1 = T \mid Y_0 = F)}$$
$$= \frac{ad}{ad + (1-a)c}$$

Question 49 ***(0 points)***

Consider an HMM with state variables $Y_1, \ldots, Y_T$ and observations $X_1, \ldots, X_T$. Suppose that the model has the following parameters, where a, b, c and d are some arbitrary constants:

Y_{t-1}	$P(Y_t = 2 \mid Y_{t-1})$
1	a
2	b

Y_t	$P(X_t = 2 \mid Y_t)$
1	c
2	d

For a particular observation sequence $X_1 = x_1, \ldots, X_T = x_T$, define the Viterbi vertex probability to be

$$v_{j,t} = \max_{y_1, \ldots, y_{t-1}} P(Y_1 = y_1, X_1 = x_1, \ldots, Y_{t-1} = y_{t-1}, X_{t-1} = x_{t-1}, Y_t = j, X_t = x_t)$$

Suppose that $v_{j,t}$ has been calculated, and has been found to have the following values, where e and f are some arbitrary constants:

$$v_{1,t} = e$$
$$v_{2,t} = f$$

Furthermore, suppose that $x_{t+1} = 2$. In terms of the constants a, b, c, d, e, and/or f, what are $v_{1,t+1}$ and $v_{2,t+1}$?

Solution:

$$v_{1,t+1} = \max\left(e(1-a)c, f(1-b)c\right)$$
$$v_{2,t+1} = \max\left(ead, fbd\right)$$

Question 50 ***(0 points)***

Consider two binary random variables, X and Y. Suppose that

$$P(Y=0) = b$$
$$P(X=1, Y=0) = c$$

In terms of b and/or c, what is the largest possible value of $P(X=1)$?

Solution:

$$\begin{aligned} P(X=1) &= P(X=1,Y=0) + P(X=1,Y=1) \\ &\leq P(X=1,Y=0) + P(Y=1) \\ &= P(X=1,Y=0) + (1 - P(Y=0)) \\ &= 1 - b + c \end{aligned}$$

Question 51 ***(0 points)***

Suppose you are training a naïve Bayes model. There are two classes, $Y=0$ and $Y=1$, with the following observations:

- Training text for class $Y=0$: "apple apple apple apple apple apple".
- Training text for class $Y=1$: "banana banana banana banana banana apple".

Use this example to discuss, in a few sentences, the importance of Laplace smoothing.

Solution: Without Laplace smoothing, using these training data, the probability of the word "banana" given class $Y=0$ would be zero. A reasonable person might suppose that the sentence "apple apple apple banana" is from class $Y=0$, but the model would assign it zero probability, because it contains the word "banana." Laplace smoothing is important because it gives a small nonzero probability to words that were never seen during training, so it would be possible to label the sentence "apple apple apple banana" as being from class $Y=0$.

Question 52 ***(0 points)***

Imagine training a perceptron with a training dataset that contains only two training tokens: $\vec{x}_1 = [1,1]^T, y_1 = 1$ and $\vec{x}_2 = [-1,-1]^T, y_2 = -1$. Suppose you begin with the weight vector $\vec{w} = [0,0]^T$ and bias $b = -1$, then present the data in alternating order $\{(\vec{x}_1, y_1), (\vec{x}_2, y_2), (\vec{x}_1, y_1), (\vec{x}_2, y_2), \ldots\}$, with a learning rate of $\eta = 1$, until $\vec{w}$ and b converge. What are the final converged values of $\vec{w}$ and b?

Solution:

$$\vec{w} = [1,1]^T$$
$$b = 0$$

Question 53 ***(0 points)***

In stochastic gradient descent, we train using one training token at a time. Suppose x is a scalar input, and suppose we have

$$\mathscr{L} = -\ln f_2(x)$$

where

$$f_k(\vec{x}) = \frac{e^{w_k x + b_k}}{e^{w_1 x + b_1} + e^{w_2 x + b_2}} \text{ for } k \in \{1,2\}$$

In terms of x, w_1, w_2, b_1, b_2, $f_1(x)$ and/or $f_2(x)$, what is $\frac{d\mathscr{L}}{db_1}$?

Solution:

$$\frac{d\mathscr{L}}{db_1} = f_1(x) = \frac{e^{w_1 x + b_1}}{e^{w_1 x + b_1} + e^{w_2 x + b_2}}$$

Question 54 ***(0 points)***

Consider a two-layer neural network with a scalar input, x. Assume that all of the weights and biases are nonzero, and that the output $f(x)$ is computed as:

$$f(x) = w_{1,1}^{(2)} h_1 + w_{1,2}^{(2)} h_2 + b^{(2)}$$
$$h_1 = \text{ReLU}\left(w_{1,1}^{(1)} x + b_1^{(1)}\right)$$
$$h_2 = \text{ReLU}\left(w_{2,1}^{(1)} x + b_2^{(1)}\right)$$

For what values of x is $\frac{\partial f}{\partial w_{1,1}^{(1)}} \neq 0$? Express your answer in terms of h_j, $w_{j,k}^{(l)}$, and/or $b_k^{(l)}$ for any values of j, k, and/or l that may be useful to you.

Solution:

$$\frac{\partial f}{\partial w_{1,1}^{(1)}} = \frac{\partial f}{\partial h_1} \frac{\partial h_1}{\partial w_{1,1}^{(1)}}$$

The first derivative is $\frac{\partial f}{\partial h_1} = w_{1,1}^{(1)}$, which the problem statement declares to be nonzero. The second derivative is nonzero only if $w_{1,1}^{(1)} x + b_1^{(1)} > 0$, i.e.

$$x > -\frac{b_1^{(1)}}{w_{1,1}^{(1)}}$$

Question 55 ***(13 points)***

Cryptids have variable numbers of arms. Let X be the number of arms a cryptid has; then $P(X=2)=a$, $P(X=3)=b$, and $P(X=4)=c$, where $a+b+c=1$.

(a) (6 points) The number of skills that a cryptid can learn is equal to the number of distinct pairs of arms it has: a 2-arm cryptid can learn 1 skill, a 3-arm cryptid can learn 3 skills, and a 4-arm cryptid can learn 6 skills. In terms of the parameters a, b and c, what is the expected number of skills a cryptid can learn?

Solution:

$$E[\#\text{ skills}] = a+3b+6c$$

(b) (7 points) Only cryptids with political skill are allowed to run for Congress, so there is no reason for voters to prefer 4-arm cryptids, yet they do. Let $Y=1$ if a cryptid is elected to Congress, and $Y=0$ otherwise; cryptid voter bias is measured by the fact that $P(Y=1|X=4)=\frac{3}{5}$, but $P(Y=1|X<4)=\frac{2}{5}$. You have developed an algorithm that generates candidate endorsements, $\hat{Y}\in\{0,1\}$, with perfect demographic parity ($P(\hat{Y}=1|X=4)=P(\hat{Y}=1|X<4)=\frac{1}{2}$) and with perfect predictive parity ($P(Y=1|\hat{Y}=1,X=4)=P(Y=1|\hat{Y}=1,X<4)=p$). In terms of p, what is $P(\hat{Y}=1|Y=1,X=4)$?

Solution:

$$\begin{aligned}P(\hat{Y}=1|Y=1,X=4) &= \frac{P(Y=1|\hat{Y}=1,X=4)P(\hat{Y}=1|X=4)}{P(Y=1|X=4)}\\ &= \frac{p\left(\frac{1}{2}\right)}{\left(\frac{3}{5}\right)}\end{aligned}$$

Question 56 ***(12 points)***

Every Easter, the Chicago Cubs hide 6000 Easter eggs at Wrigley Field. After an hour of searching, you've found 4 blue eggs, 5 orange eggs, and 2 green eggs.

(a) (6 points) Use Laplace smoothing to estimate the fraction of all eggs at Wrigley Field that are blue. Note that colors other than orange, blue and green may exist. Your answer should be a function of the Laplace smoothing hyperparameter, k.

Solution:

$$P(\text{blue}|\text{Wrigley}) = \frac{4+k}{4+5+2+4k}$$

(b) (6 points) Your significant other has been collecting Easter eggs at Soldier Field, where the Bears have hidden 10,000 eggs (note: this means that the probability any given egg was at Soldier Field on Easter is larger than the probability that it was at Wrigley Field). Based on your observations, you deduce that the distribution of colors is different at Soldier Field versus Wrigley Field: $P(X = \text{blue}|Y = \text{wrigley}) = p$, but $P(X = \text{blue}|Y = \text{soldier}) = q$. Your friend Al brings you a blue egg, that he found at either Soldier Field or Wrigley Field. Under what condition should you believe that he found it at Soldier Field? Your answer should be an inequality in terms of p and q.

Solution: Estimated $P(\text{Wrigley}) = \frac{6000}{16000}$ ($\frac{6000+k}{16000+2k}$ is also an acceptable answer). You should decide that the egg is from Soldier Field if

$$P(Y = \text{wrigley}|X = \text{blue}) < P(Y = \text{soldier}|X = \text{blue}),$$

which is true if

$$P(X = \text{blue}|Y = \text{wrigley})P(Y = \text{wrigley}) < P(X = \text{blue}|Y = \text{soldier})P(Y = \text{soldier})$$

which happens if

$$\frac{6000}{16000}p < \frac{10000}{16000}q$$

Question 57 ***(13 points)***

You have a machine learning problem in which the input is a 3-dimensional vector, $\mathbf{x}$, and the output is binary, $y \in \{0,1\}$. You are considering two possible solutions: a linear regression algorithm that uses a weight vector $\mathbf{w}$ and a bias term b, and a softmax linear classifier algorithm that uses weight vectors $\mathbf{w}_0$ and $\mathbf{w}_1$ and bias coefficients b_0 and b_1. As you know, the stochastic gradient descent algorithm has a similar form in both cases:

$$\textbf{Linear Regression:} \mathbf{w} \leftarrow \mathbf{w} - \eta \varepsilon_i \mathbf{x}_i,$$
$$\textbf{Linear Classifier:} \mathbf{w}_c \leftarrow \mathbf{w}_c - \eta \varepsilon_{i,c} \mathbf{x}_i,$$

where $\mathbf{x}_i = [x_{i,0}, x_{i,1}, x_{i,2}]^T$ and y_i are the stochastically sampled training token, ε_i is the linear regression error term, and $\varepsilon_{i,0}, \varepsilon_{i,1}$ are the linear classifier errors.

(a) (6 points) Consider a linear regression algorithm, whose output is

$$f(\mathbf{x}_i) = \mathbf{w}^T \mathbf{x}_i + b$$

Suppose that $\mathbf{x}_i = [-1,0,1]^T$ and $y_i = 1$. Suppose $\mathbf{w}$ is initialized to $\mathbf{w} = [\rho, \phi, \theta]^T$, and b is initialized as $b = \gamma$. In terms of ρ, ϕ, θ, and γ, what is ε_i?

Solution:

$$\begin{aligned} \varepsilon_i &= \mathbf{w}^T \mathbf{x}_i - y_i \\ &= -\rho + \theta + \gamma - 1 \end{aligned}$$

(b) (7 points) Consider a softmax classifier,

$$f_c(\mathbf{x}_i) = \underset{c}{\text{softmax}}(\mathbf{w}^T \mathbf{x}_i + b)$$

Suppose that $\mathbf{x}_i = [-1,0,1]^T$ and $y_i = 1$. Suppose $\mathbf{w}$ is initialized to $\mathbf{w}_0 = [0,0,0]^T$, $\mathbf{w}_1 = [\rho, \phi, \theta]$, $b_0 = 0$, and $b_1 = \gamma$. In terms of ρ, ϕ, θ, and γ, what is $\varepsilon_{i,1}$?

Solution: The target output is: $f_0(\mathbf{x}_i)$ should be 0, $f_1(\mathbf{x}_i)$ should be 1, therefore

$$\begin{aligned} \varepsilon_{i,1} &= f_1(\mathbf{x}_i) - 1 \\ &= \frac{\exp(-\rho + \theta + \gamma)}{1 + \exp(-\rho + \theta + \gamma)} - 1 \end{aligned}$$

Question 58 ***(13 points)***

Ctulhoids are small slug-like animals with many eyes (they are very cute). 40% of all ctulhoids have 3 blue eyes, while 60% have 4 blue eyes. The number of orange eyes a ctulhoid has is either one less than the number of its blue eyes (with probability $1-a$) or one more than the number of its blue eyes (with probability a).

(a) (6 points) What is a ctulhoid's expected total number of eyes, including both blue eyes and orange eyes?

Solution:

$$E[\#\text{ eyes}] = 0.4(1-a)\times 5 + 0.4a\times 7 + 0.6(1-a)\times 7 + 0.6a\times 9$$

(b) (7 points) Let $A=1$ if a ctulhoid has more orange eyes than blue eyes, and let $A=0$ otherwise. Let $Y=1$ if somebody adopts the ctulhoid as a pet, and let $Y=0$ otherwise. People like orange eyes: $P(Y=1|A=1)=\frac{2}{3}$, but $P(Y=1|A=0)=\frac{1}{3}$. You have decided that this bias in favor of orange-eyed ctulhoids is unfair, so you have created an algorithm that makes pet recommendations ($\hat{Y}$) with perfect demographic parity ($P(\hat{Y}=1|A=1)=P(\hat{Y}=1|A=0)=\frac{1}{2}$), and with perfect equal opportunity ($P(\hat{Y}=1|Y=1,A=1)=P(\hat{Y}=1|Y=1,A=0)=p$). In terms of p, what is $P(Y=1|\hat{Y}=1,A=1)$?

Solution:

$$P(Y=1|\hat{Y}=1,A=1) = \frac{P(\hat{Y}=1|Y=1,A=1)P(Y=1|A=1)}{P(\hat{Y}=1|A=1)}$$
$$= \frac{p\left(\frac{2}{3}\right)}{\left(\frac{1}{2}\right)}$$

Question 59 ***(12 points)***

You have been comparing emoji usage among messages on the Telegram and WhatsApp messaging systems. After extensive research, your Telegram database contains m examples of the rofl emoji, n examples of the halo emoji, and no examples of any other emoji.

(a) (6 points) Use Laplace smoothing to estimate the fraction of all emojis on Telegram that are halo emojis. Note that emojis other than rofl and halo may exist, even though there are none in your training dataset. Your answer should be a function of m, n, and the Laplace smoothing hyperparameter, k.

Solution:

$$P(\text{halo}|\text{telegram}) = \frac{n+k}{m+n+3k}$$

(b) (6 points) Based on extensive research, you conclude that 87% of all text messages are sent via WhatsApp, and 13% are sent via Telegram. The likelihood of rofl on each of these two platforms is $P(X = \text{rofl}|Y = \text{whatsapp}) = p$, and $P(X = \text{rofl}|Y = \text{telegram}) = q$. A journalist shows you a text message containing a rofl emoji (and no other emojis), and asks you to guess whether it came from WhatsApp or Telegram. Under what condition should you say that it came from Telegram? Your answer should be an inequality in terms of p and q.

Solution: You should decide that the emoji is from Telegram if

$$P(Y = \text{telegram}|X = \text{rofl}) > P(Y = \text{whatsapp}|X = \text{rofl}),$$

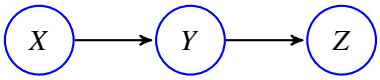
which is true if

$$P(X = \text{rofl}|Y = \text{telegram})P(Y = \text{telegram}) > P(X = \text{rofl}|Y = \text{whatsapp})P(Y = \text{whatsapp})$$

which happens if

$$0.13q > 0.87p$$

Question 60 ***(7 points)***



X, Y, and Z are random variables, each of which can take the values -1, 0, or 1. Their causal dependencies are shown in the Bayes network above. The parameters of this model are:

	x		
	-1	0	1
$P(X=x)$	a	b	c
$P(Y=-1\|X=x)$	d	e	f
$P(Y=0\|X=x)$	g	h	i
$P(Y=1\|X=x)$	j	k	l

	y		
	-1	0	1
$P(Z=-1\|Y=y)$	m	n	o
$P(Z=0\|Y=y)$	p	q	r
$P(Z=1\|Y=y)$	s	t	u

In terms of the parameters a through u, what is $P(X=-1|Z=0)$?

Solution:

$$\begin{aligned} P(X=-1|Z=0) &= \frac{P(X=-1,Z=0)}{P(Z=0)} \\ &= \frac{\sum_y P(X=-1,Y=y,Z=0)}{\sum_y \sum_x P(X=x,Y=y,Z=0)} \\ &= \frac{a(dp+gq+jr)}{a(dp+gq+jr)+b(ep+hq+kr)+c(fp+iq+lr)} \end{aligned}$$

Question 61 ***(7 points)***

Slarti is in Paris, attempting to walk home from a pizza restaurant. He is on a sidewalk whose west edge is a steep drop into the river; he wants to make sure he does not fall into the river. Let Y_t be the true distance between Slarti and the cliff edge at time t, measured in meters (m). Suppose you know that $Y_0 = 3$m, for sure. Since Slarti is too full to walk straight, he wobbles as he walks. Since it's foggy, he does not always see clearly: X_t is how far away the cliff edge looks, at time t, which may or may not be equal to the true distance Y_t. The transition probabilities and observation probabilities are

$$P(Y_t = k|Y_{t-1} = j) = \begin{cases} \frac{1}{4} & k = j-1 \\ \frac{1}{2} & k = j \\ \frac{1}{4} & k = j+1 \end{cases}, \quad P(X_t = k|Y_t = j) = \begin{cases} \frac{1}{3} & k = j-1 \\ \frac{1}{3} & k = j \\ \frac{1}{3} & k = j+1 \end{cases}$$

What is $P(Y_2 = 2, X_2 = 2)$?

Solution:

$$\begin{aligned} P(Y_2 = 1, X_2 = 2) &= \sum_{y_1} P(Y_0 = 3, Y_1 = y_1, Y_2 = 2, X_2 = 2) \\ &= P(Y_0 = 3, Y_1 = 2, Y_2 = 2, X_2 = 2) + P(Y_0 = 3, Y_1 = 3, Y_2 = 2, X_2 = 2) \\ &= \left(\frac{1}{2}\right)\left(\frac{1}{4}\right)\left(\frac{1}{3}\right) + \left(\frac{1}{4}\right)\left(\frac{1}{2}\right)\left(\frac{1}{3}\right) \end{aligned}$$

Question 62 ***(7 points)***

Your friend is sending you a message by flashing red and green lights in a coded pattern. The first light is red or green with equal probability. After that, each time, the lights stay the same with probability a:

$$P(Y_t = \text{green}|Y_{t-1} = \text{green}) = P(Y_t = \text{red}|Y_{t-1} = \text{red}) = a,$$

otherwise they change color. Because of the fog, you only see the correct color with probability b:

$$P(X_t = \text{green}|Y_t = \text{green}) = P(X_t = \text{red}|Y_t = \text{red}) = b,$$

otherwise you mistakenly see the wrong color. Suppose you see the sequence $X_0 = R\text{red}, X_1 = \text{green}$. In terms of the parameters a and b, what is the joint probability $P(Y_0 = \text{red}, Y_1 = \text{green}, X_0 = \text{red}, X_1 = \text{green})$?

Solution:

$$\begin{aligned}
&P(Y_0 = \text{red}, X_0 = \text{red}, Y_1 = \text{green}, X_1 = \text{green}) \\
&= P(Y_0 = \text{red})P(X_0 = \text{red}|Y_0 = \text{red})P(Y_1 = \text{green}|Y_0 = \text{red})P(X_1 = \text{green}|Y_1 = \text{green}) \\
&= \frac{1}{2}(1-a)b^2
\end{aligned}$$

Question 63 ***(10 points)***

A particular CNN has a grayscale image input, $x[n_1, n_2]$, and a one-channel output:

$$\xi[n_1, n_2] = w[n_1, n_2] * x[n_1, n_2],$$

where $*$ denotes convolution. The output is then max-pooled over the entire image:

$$\hat{y} = \max_{0 \le n_1 < N_1} \max_{0 \le n_2 < N_2} \xi[n_1, n_2]$$

Suppose the weights and the input image are given by

$$w[n_1, n_2] = \begin{cases} e^{-(n_1^2 + n_2^2)} & -3 \le n_1 \le 3, \ -3 \le n_2 \le 3 \\ 0 & \text{otherwise} \end{cases}$$

$$x[n_1, n_2] = \begin{cases} e^{-((n_1-15)^2 + (n_2-12)^2)} & 0 \le n_1 \le 63, \ 0 \le n_2 \le 63 \\ 0 & \text{otherwise} \end{cases}$$

What is $\frac{d\hat{y}}{dw[2,1]}$? Your answer should be an explicit function of numerical constants; there should not be any variables in your answer.

Solution:

$$\frac{d\hat{y}}{dw[2,1]} = e^{-5}$$