UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN

CS440/ECE448 Artificial Intelligence

Exam 1

Spring 2022

Exam 1 will be February 21, 2022

Your Name: _____

Your NetID: _____

Your Section: _____

Instructions

- Please write your name and NetID on the top of every page.
- This will be a CLOSED BOOK exam. No additional written materials (books, cheat-sheets, etc.) or electronic devices (phones, tablets, calculators, computers etc.) are allowed.
- No calculators are permitted. You need not simplify explicit numerical expressions.

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

Let A and B be independent binary random variables with $P(A=1)=0.1$, $P(B=1)=0.4$. Let $C = A \vee B$, and let $D = A \oplus B$ (A XOR B).

(a) What is $P(C=1)$?

(b) What is $P(D=1)$?

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

(d) Are A and D independent? Why?

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

Use the axioms of probability to prove that, for a binary random variable A, $P(A = 0) = 1–P(A = 1)$.

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

20% of students at U of I live north of University Ave. 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 does not live north of University Ave?

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

Consider the following joint probability distribution, for binary random variables A and B:

$$P(A = 1, B = 1) = 0.12$$
$$P(A = 1, B = 0) = 0.18$$
$$P(A = 0, B = 1) = 0.28$$
$$P(A = 0, B = 0) = 0.42$$

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

Question 5 ***(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"?

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

Y is a random variable denoting the class of a newspaper title: $Y = 0$ means the article is about sports, $Y = 1$ means the article is about science. The title is only three words long; its three words are the random variables W_1, W_2, and W_3. Depending on whether the article is about sports or science, the title may contain any word from the following vocabulary: $\{\text{Illini}, \text{win}, \text{discover}, \text{everything}\}$. The prior probability of an article about science is $P(Y = 1) = 0.4$. Assume a naïve Bayes model, with word likelihoods of

$$\begin{aligned} P(W_i = \text{Illini} | Y = 0) &= 0.3 \\ P(W_i = \text{Illini} | Y = 1) &= 0.3 \\ P(W_i = \text{win} | Y = 0) &= 0.3 \\ P(W_i = \text{win} | Y = 1) &= 0.1 \\ P(W_i = \text{discover} | Y = 0) &= 0.1 \\ P(W_i = \text{discover} | Y = 1) &= 0.4 \end{aligned}$$

Now you download the article, and discover that its title is $X = \text{Illini discover everything}$. What is $P(Y = 1 | W_1 = \text{Illini}, W_2 = \text{discover}, W_3 = \text{everything})$? Leave your answer in the form of an expression composed of numbers; do not simplify.

Question 7 ***(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?

(b) What is the Bayes error rate?

(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?

Question 8 ***(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 $\vec{x} = [x_1, x_2, x_3]^T$ and $\vec{w} = [w_1, w_2, w_3]^T$, let $\vec{x}^T\vec{w}$ denote the dot product of vectors $\vec{x}$ and $\vec{w}$, 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 $y = u(w^T x + b)$.

Question 9 ***(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 implemented by a linear classifier? If so, construct a linear classifier that does it; if not, say why not.

Question 10 ***(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, $f(\vec{x})$, for a naive Bayes classifier, where $\vec{x}=[x_1,\ldots,x_{100}]^T$ is the set of observed values of the evidence variables. You might find it useful to define $N(\vec{x})=$ the number of nonzero elements of the binary vector $\vec{x}$; note that $0\leq N(\vec{x})\leq 100$.

(b) The naive Bayes classifier can be written as

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

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

Question 11 ***(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 two features: $x_1 =$ score given by reviewer A, and $x_2 =$ 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 $f(\vec{x}) = 1$ if the movie returns a profit, $f(\vec{x}) = -1$ otherwise. The initial weights are $b = -1, w_1 = 0, w_2 = 0$. 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?

(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 $f(\vec{x}) = +1$ when the total of the two reviewer scores is more than 8, and $f(\vec{x}) = -1$ otherwise. Is this possible? If so, what are the weights b, w_1, and w_2 that will make this possible?

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

Question 12 ***(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 $\vec{w}_k$, 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?

Question 13 ***(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.

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

The softmax function is defined as

$$f_k = \frac{\exp(\xi_k)}{\sum_j \exp(\xi_j)}$$

Find $df_5/d\xi_3$ in terms of f_3, f_5, ξ_3 and/or ξ_5.

Question 15 ***(0 points)***

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

Table 1: Variables used in Problem 15.

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

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

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

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

You have a two-layer neural network trained as an animal classifier. The input feature vector is $\vec{x} = [x_1, x_2, x_3]^T$, where x_1, x_2, and x_3 are some features. There are two hidden nodes $\vec{h}^{(1)} = [h_1^{(1)}, h_2^{(1)}]^T$, and three output nodes, $\vec{f} = [f_1, f_2, f_3]^T$, corresponding to the three output classes f_1 =Pr(Y=dog|X=x), f_2 =Pr(Y=cat|X=x), and f_3 =Pr(Y=skunk|X=x). The hidden layer uses a sigmoid nonlinearity, the output layer uses a softmax. The weight matrices have elements $w_{j,k}^{(l)}$, and the biases are $b_j^{(l)}$.

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

(b) Let $w_{i,j}^{(2)}$ be the weight connecting the i^{th} output node to the j^{th} hidden node. What is $df_2/dw_{2,1}^{(2)}$? Write your answer in terms of $h_i^{(2)}$, $w_{i,j}^{(2)}$, and/or the hidden node activations $h_j^{(1)}$, for any appropriate values of i and/or j.

(c) Suppose that you are presented with an all-zero feature vector $\vec{x} = [0, 0, 0]^T$. Suppose that the first-layer weight matrix is also all zero, $w_{j,k}^{(1)} = 0$, but the bias is nonzero, specifically, it has the value $\vec{b}^{(1)} = [12, 13]^T$. Suppose that, for this particular training token, $df_2/dh_1^{(1)} = 15$. What is $df_2/db_1^{(1)}$? Write your answer as a product of fractions involving exponentials of integers; there should be only constants in your answer, no variables, but you need not simplify.

Question 17 ***(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.

Question 18 ***(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.

Question 19 ***(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?

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

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