UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN

CS440/ECE448 Artificial Intelligence

Exam 1

Spring 2026

February 16, 2026

Your Name: ______________________________

Your NetID: ______________________________

Instructions

- Please write your name on the top of every page.
- Have your ID ready; you will need to show it when you turn in your exam.
- This will be a CLOSED BOOK, CLOSED NOTES exam. You are permitted to bring and use only one 8.5x11 page of notes, front and back, handwritten or typed in a font size comparable to handwriting.
- No electronic devices (phones, tablets, calculators, computers etc.) are allowed.
- Make sure that your answer includes only the variables that it should include, but DO NOT simplify explicit numerical expressions. For example, the answer $x = \frac{1}{1+\exp(-0.1)}$ is MUCH preferred (much easier for us to grade) than the answer $x = 0.524979$.

Possibly Useful Formulas

$$P(X=x|Y=y)P(Y=y) = P(Y=y|X=x)P(X=x)$$
$$P(X=x) = \sum_y P(X=x, Y=y)$$
$$E\left[f(X,Y)\right] = \sum_{x,y} f(x,y)P(X=x,Y=y)$$

Precision,Recall $= \dfrac{TP}{TP+FP}, \dfrac{TP}{TP+FN}$

MPE=MAP: $f(x) = \arg\max\left(\log P(Y=y) + \log P(X=x|Y=y)\right)$

Naive Bayes: $P(X=x|Y=y) \approx \prod_{i=1}^{n} P(W=w_i|Y=y)$

Laplace Smoothing w/OOV: $P(W=w_i) = \dfrac{k + \text{Count}(W=w_i)}{k + \sum_v \left(k + \text{Count}(W=v)\right)}$

Fairness: $P(Y|A) = \dfrac{P(Y|f(X),A)P(f(X)|A)}{P(f(X)|Y,A)}$

Score: $v_{t+1}(j) = \max_i v_t(i) + \log a_{ij} + \log b_j(x_{t+1})$

Backpointer: $\psi_{t+1}(j) = \arg\max_i v_t(i) + \log a_{ij} + \log b_j(x_{t+1})$

Linear Regression: $\mathscr{L} = \dfrac{1}{n}\sum_{i=1}^{n}\left(\mathbf{w}^T\mathbf{x}_i - y_i\right)^2$

Question 1 ***(26 points)***

Valentine's baskets and Easter baskets differ in the candies that they contain. Let $P(E=\top)$ be the probability that you have received an Easter basket, and let $P(E=\bot)=1-P(E=\top)$ be the probability of receiving a Valentine's basket. Let $P(M=m|E=e)P(C=c|E=e)$ be the probability that you receive m marshmallow candies and c caramel candies given $E=e$.

(a) (13 points) A Bayes classifier observes m and c, then chooses the value of E with the highest *a posteriori* probability. In terms of the distributions $P(E)$, $P(M|E)$, and $P(C|E)$, what is the error rate of the Bayes classifier? Your answer may contain explicit summations, minimizations, and/or maximizations; please specify the variable(s) over which you are summing, maximizing, and/or minimizing.

Solution:

$$\begin{aligned} P(\text{Error}) &= \sum_c \sum_m \min_e P(E=e)P(M=m|E=e)P(C=c|E=e) \\ &= 1-\sum_c \sum_m \max_e P(E=e)P(M=m|E=e)P(C=c|E=e) \end{aligned}$$

(b) (13 points) You are currently five years old; you have received five Easter baskets and five Valentine's baskets. The contents of the Easter baskets, written as (c,m) tuples, were

$$\mathscr{D}_{\text{Easter}} = \{(3,3),(4,4),(5,5),(3,5),(4,5)\}$$

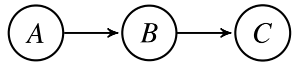
In terms of a Laplacian smoothing parameter k, estimate $P(C=c|E=\top)$ and $P(M=m|E=\top)$ for $c,m\in\{3,4,5,\text{other}\}$.

Solution:

$$P(C=c|E=\top)=\begin{cases} \frac{2+k}{5+4k} & c\in\{3,4\} \\ \frac{1+k}{5+4k} & c=5 \\ \frac{k}{5+4k} & c=\text{other} \end{cases},\quad P(M=m|E=\top)=\begin{cases} \frac{3+k}{5+4k} & m=5 \\ \frac{1+k}{5+4k} & m\in\{3,4\} \\ \frac{k}{5+4k} & m=\text{other} \end{cases},$$

Question 2 ***(30 points)***

Consider three binary events A, B, and C, related by the Bayes network shown here:



The parameters of this Bayes network are given by the unknown constants v through z, as follows:

$$P(A) = v$$
$$P(B|A) = w, \quad P(B|\neg A) = x$$
$$P(C|B) = y, \quad P(C|\neg B) = z$$

(a) (10 points) Are events A and B independent, conditionally independent given knowledge of C, both, or neither? Explain your answer.

Solution: They are neither independent nor conditionally independent. **Acceptable explanation #1:** there is an arrow connecting A to B, so B always depends on A regardless of the rest of the graph. **Acceptable explanation #2:** The are not independent because they have an unknown common ancestory (A). They are not conditionally independent given C because then they would have a known common descendant (C).

(b) (10 points) Suppose C is unknown. Find the four probabilities $P(A,B)$, $P(A,\neg B)$, $P(\neg A,B)$, and $P(\neg A,\neg B)$.

Solution:

$$P(A,B) = vw$$
$$P(A,\neg B) = v(1-w)$$
$$P(\neg A,B) = (1-v)x$$
$$P(\neg A,\neg B) = (1-v)(1-x)$$

(c) (10 points) Suppose C is known to be true. Find $P(C)$, and in terms of $P(C)$ find $P(A,B|C)$, $P(A,\neg B|C)$, $P(\neg A,B|C)$, and $P(\neg A,\neg B|C)$.

Solution:

$$P(C) = vwy + v(1-w)z + (1-v)xy + (1-v)(1-x)z$$

$$P(A,B|C) = \frac{vwy}{P(C)}$$

$$P(A,\neg B|C) = \frac{v(1-w)z}{P(C)}$$

$$P(\neg A,B|C) = \frac{(1-v)xy}{P(C)}$$

$$P(\neg A,\neg B|C) = \frac{(1-v)(1-x)z}{P(C)}$$

Question 3 ***(20 points)***

$Y_t = 1$ means that your office is being robbed at time t, otherwise $Y_t = 0$. Suppose you know that $Y_0 = 0$. $X_t = 1$ means that the security system sends a notification to your phone at time t, otherwise $X_t = 0$. Define $a_{ij} = P(Y_{t+1} = j|Y_t = i)$, and $b_{ij} = P(X_t = j|Y_y = i)$. Suppose there is some normalizer, z, and base, c, such that $\log_c(a_{ij}/z) = \log_c(b_{ij}/z) = 0$ for $i = j$, but for $i \neq j$, $\log_c(a_{ij}/z) = -3$ and $\log_c(b_{ij}/z) = -2$. Suppose that $\{X_1,\ldots,X_9\} = \{1,0,1,1,1,1,0,0,1\}$. Draw a trellis specifying the numerical values of $v_t(i) = \max_{Y_\tau,\tau<t} \log_c P(Y_1,\ldots,Y_t = i|X_1,\ldots,X_t)/z$ for $i \in \{0,1\}$ and $1 \leq t \leq 9$, and specify a sequence of state variables $\{Y_1,\ldots,Y_9\}$ that maximizes the score. If there are more than one state sequences with the same maximum score, you only need to provide one of them.

Solution: A correct answer needs to show the following numerical values. The arrows need not be shown, but may be helpful to you in solving the second part of the question:

t	1	2	3	4	5	6	7	8	9
X_t	1	0	1	1	1	1	0	0	1
$v_t(0)$	$\leftarrow -2$	$\leftarrow -2$	$\leftarrow -4$	$\leftarrow -6$	$\leftarrow -8$	-10	$\swarrow -8$	$\leftarrow -8$	$\leftarrow -10$
$v_t(1)$	$\nwarrow -3$	-5	-5	$\leftarrow -5$	$\leftarrow -5$	$\leftarrow -5$	$\leftarrow -7$	$\leftarrow -9$	$\leftarrow -9$

There are three maximum-likelihood state sequences, any of which is a correct answer: $\{0,0,1,1,1,1,1,1,1\}$, $\{0,1,1,1,1,1,1,1,1\}$, or $\{1,1,1,1,1,1,1,1,1\}$.

Question 4 ***(24 points)***

Consider a linear regression model $f(\mathbf{x}_i) = \mathbf{w}^T \mathbf{x}_i$ where $\mathbf{x}_i$ is an m-dimensional vector drawn from a size-n training corpus, $1 \le i \le n$. Suppose you want to choose the vector $\mathbf{w}$ to minimize $\mathscr{L}_{\text{train}}$, defined as

$$\mathscr{L}_{\text{train}} = \frac{1}{n} \sum_{i=1}^{n} |f(\mathrm{x}_i) - y_i|^3,$$

where y_i is a real-valued scalar for $1 \le i \le n$.

(a) (12 points) Find the gradient $\frac{\partial \mathscr{L}_{\text{train}}}{\partial \mathbf{w}}$.

Solution:

$$\frac{\partial \mathscr{L}_{\text{train}}}{\partial \mathbf{w}} = \frac{1}{n} \sum_{i=1}^{n} 3\mathbf{x}_i |f(\mathbf{x}_i) - y_i|^2 \operatorname{sign}(f(\mathbf{x}_i) - y_i)$$

(b) (12 points) Suppose you also have a development corpus, with tokens numbered $n+1 \le i \le n+j$, and a test corpus, with tokens numbered $n+j+1 \le i \le n+j+k$. The losses in these corpora are

$$\mathscr{L}_{\text{dev}} = \frac{1}{j} \sum_{i=n+1}^{n+j} |f(\mathrm{x}_i) - y_i|^3,$$

$$\mathscr{L}_{\text{test}} = \frac{1}{k} \sum_{i=n+j+1}^{n+j+k} |f(\mathrm{x}_i) - y_i|^3.$$

Your client, a hardware manufacturer, is very clever about finding things to measure. They propose the following experiment: (1) increase m by finding something else to measure about each product, (2) re-train the new $\mathbf{w}$, finding the value that minimizes $\mathscr{L}_{\text{train}}$, (3) measure $\mathscr{L}_{\text{dev}}$ and $\mathscr{L}_{\text{test}}$, (4) repeat.

Do you expect $\mathscr{L}_{\text{dev}}$ to be smallest when $m = 1$, when $m = \frac{n}{2}$, or when $m = \frac{3n}{2}$? Why?

Solution: $\mathscr{L}_{\text{dev}}$ is the sum of optimization error plus generalization error. Optimization error gets smaller as m increases, but generalization error gets larger as m increases, therefore I would expect $m = \frac{n}{2}$ to have the smallest $\mathscr{L}_{\text{dev}}$.

This page is scratch paper