CS440/ECE448 Lecture 13: Exam 1 Review

CC0, Mark Hasegawa-Johnson

2/2023

UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN

CS440/ECE448 Artificial Intelligence

Exam 1

Spring 2023

February 20, 2023

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 $x = 0.524979$.

Outline

- How to take the exam
- What the exam will cover
- Sample problems

How to take the exam

- Be in Lincoln Hall Theater by 12:58pm on Monday 2/20
- Bring pencil and eraser
- Bring your ID: we will check it when we collect your exam

What's allowed

- You can bring one page of notes, both sides, hand-written or in a font size comparable to handwriting
- No calculators, computers, or cellphones

Outline

- How to take the exam
- What the exam will cover
- Sample problems

What the exam will cover

- Random variables
- Decision theory
- Naïve Bayes
- Fairness
- Learning
- Linear regression
- Linear classifier
- Multilayer networks
- Optimization
- Privacy

Outline

- How to take the exam
- What the exam will cover
- Sample problems

Sample problem: Random variables

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?

Sample problem: Decision theory

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:

$$P(X = \text{noise}|Y = \text{cpu}) = 0.8;\ P(X = \text{noise}|Y = \text{powersupply}) = 0.4$$

(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?

Sample problem: Naïve Bayes

You've been asked to create a naive 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, 24 are blueberry. In terms of the Laplace smoothing parameter k, estimate the probability of each of the three flavors. Note that it's possible that other flavors besides strawberry, raspberry, and blueberry may exist.

Sample problem: Fairness

Suppose that Y is a reference label, $\hat{Y}$ is the output of your classifier, and A is some attribute that should not influence the relationship between Y and $\hat{Y}$. $P(Y = 1|A = 1) = 2/3$. Given $A = 1$, the relationship between Y and $\hat{Y}$ is described by the following confusion matrix. What is $P(Y = 1|\hat{Y} = 1, A = 1)$?

	$\boldsymbol{P(\hat{Y} = 0 \mid Y, A = 1)}$	$\boldsymbol{P(\hat{Y} = 1 \mid Y, A = 1)}$
$Y = \mathbf{0}$	0.8	0.2
$Y = \mathbf{1}$	0.4	0.6

Sample problem: Learning

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

Sample problem: Linear regression

In stochastic gradient descent, we train using one training token at a time. Suppose $\mathcal{L} = (w@x - y)^2$, $w = [w_1, w_2, b]$, $x = [x_1, x_2, 1]$. In terms of x_1, x_2, w_1, w_2, b and/or y, what is $\frac{d\mathcal{L}}{dw_2}$?

Sample problem: Linear classifier

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.

Sample problem: Multilayer networks

Suppose that

$$f = w_{2,1,1}h_1 + w_{2,1,2}h_2 + b_2$$
$$h_1 = \mathrm{ReLU}\left(w_{1,1,1}x_1 + w_{1,1,2}x_2 + b_{1,1}\right)$$
$$h_2 = \mathrm{ReLU}\left(w_{1,2,1}x_1 + w_{1,2,2}x_2 + b_{1,2}\right)$$

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

Sample problem: Optimization

Suppose you have an m-dimensional weight vector, $w = [w_0, \cdots, w_{m-1}]$. Suppose that each weight can be discretized to one of just n possible values. You are considering two methods of optimization: (a) an exhaustive search (a.k.a. a grid search), or (b) a series of p different coordinate descent searches, each with a different random starting weight vector, and each continued for up to q iterations. How many random restarts would you need (how large would p be) before the coordinate descent algorithm becomes as computationally expensive as the grid search? Express your answer in terms of $m, n,$ and q.

Sample problem: Privacy

In order to estimate traffic flow without compromising user privacy, the cell phone company asks users to install a navigation app that uses differential privacy. Every time the app is opened, it generates a random variable R with distribution $P(R = 1) = 1 - P(R = 0) = a$. If $R = 1$, then the app reports the user location accurately. If $R = 0$, then the app lies: it chooses one of the m neighborhoods in the city uniformly at random, and reports that neighborhood as the user's location. After studying the data, the cell phone company finds that the fraction of users whose apps report their location as "campustown" on a Saturday night at 6pm is $P(X = \text{campustown}) = b$. In terms of a, b, and m, what is the actual fraction of users who are in campustown at 6pm on a Saturday night?