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 $x = 0.524979$.

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 $= \frac{TP}{TP+FP}, \frac{TP}{TP+FN}$

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: $P(W=w_i)=\frac{k+\text{Count}(W=w_i)}{k+\sum_v\left(k+\text{Count}(W=v)\right)}$

Fairness: $P(Y|A)=\frac{P(Y|\hat{Y},A)P(\hat{Y}|A)}{P(\hat{Y}|Y,A)}$

Linear Regression: $\varepsilon_i=f(x_i)-y_i=b+w@x_i-y_i$

Mean Squared Error: $\text{MSE}=\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i^2$

Linear Classifier: $f(x)=\arg\max_k w_k@x+b$

Cross-Entropy: $\mathscr{L}=-\frac{1}{n}\sum_{i=1}^{n}\log f_{y_i}(x_i)$

Softmax: $\underset{c}{\text{softmax}}(w@x+b)=\frac{\exp(w_c@x+b_c)}{\sum_{k=0}^{V-1}\exp(w_k@x+b_k)}$

Softmax Error: $\varepsilon_{i,c}=\begin{cases} f_c(x_i)-1 & c=y_i \\ f_c(x_i)-0 & \text{otherwise}\end{cases}$

Gradient Descent: $w\leftarrow w-\eta\nabla_w\mathscr{L}$

Neural Net: $h=\text{ReLU}\left(b_0+w_0@x\right),\quad f=\text{softmax}\left(b_1+w_1@h\right)$

Back-Propagation: $\frac{\partial\mathscr{L}}{\partial h_j}=\sum_k\frac{\partial\mathscr{L}}{\partial f_k}\times\frac{\partial f_k}{\partial h_j},\quad\frac{\partial\mathscr{L}}{\partial w_{0,k,j}}=\frac{\partial\mathscr{L}}{\partial h_k}\times\frac{\partial h_k}{\partial w_{0,k,j}}$

Question 1 ***(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 2 ***(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 3 ***(12 points)***

You are trying to make a classifier that can distinguish between crows and ravens. You have a training set with 100 crows and 100 ravens, and a development test set with 20 crows and 20 ravens.

(a) (6 points) The 1-nearest-neighbors algorithm gets 100% accuracy on the training set, but only 60% accuracy on the development test set. The 3-nearest neighbors algorithm gets only 90% accuracy on the training set, but it gets 70% accuracy on the development test set. If you want your algorithm to work well for birds you've never seen before, should you choose $k = 1$ or $k = 3$? Why?

Solution: Choose $k = 3$. The KNN is trained using the training set, therefore its performance on the development test set is a better estimate of how well it will perform on previously unseen test data.

(b) (6 points) Suppose that, instead of 100 crows and 100 ravens, your training set has only 3 crows and 3 ravens. The crows are named Larry, Moe, and Curly, and they are 12, 18, and 13 inches long, respectively. The ravens are named Ingrid, Bette, and Marlene, and they are 24, 16, and 22 inches long, respectively. Bird X is 19 inches long, and is either a crow or a raven. Specify two different values of k for which the k-nearest neighbors algorithm gives different estimates of Bird X's species, and list the k nearest neighbors for each of these two values of k.

Solution: There are many possible answers. One possible answer: for $k = 1$, the nearest neighbor is Moe, so KNN classifies bird X as a crow. For $k = 3$, the nearest neighbors are Moe, Bette, and Marlene, so KNN classifies bird X as a raven.

Question 4 ***(13 points)***

You have a machine learning problem in which the input is a 3-dimensional vector, 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 w and a bias term b, and a softmax linear classifier algorithm that uses weight vectors w_0 and 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:

$$\textbf{Linear Regression:} w \leftarrow w - \eta \varepsilon_i x_i,$$
$$\textbf{Linear Classifier:} w_c \leftarrow w_c - \eta \varepsilon_{i,c} x_i,$$

where $x_i = [x_{i,0}, x_{i,1}, x_{i,2}]$ 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

$$f(x_i) = w@x_i + b$$

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

Solution:

$$\begin{aligned} \varepsilon_i &= w@x_i - y_i \\ &= -\rho + \theta + \gamma - 1 \end{aligned}$$

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

$$f_c(x_i) = \underset{c}{\text{softmax}}(w@x_i + b)$$

Suppose that $x_i = [-1, 0, 1]$ and $y_i = 1$. Suppose w is initialized to $w_0 = [0, 0, 0]$, $w_1 = [\rho, \phi, \theta]$, $b_0 = 0$, and $b_1 = \gamma$. In terms of ρ, ϕ, θ, and γ, what is $\varepsilon_{i,1}$?

Solution: The target output is: $f_0(x_i)$ should be 0, $f_1(x_i)$ should be 1, therefore

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

This page is scratch paper