Midterm

● Graded

Student

HARRIS DOAN

Total Points

45.75 / 50 pts

Question 1

Copying Instances when Training Decision Tree

2 / 2 pts

✓ **– 0 pts Correct**: False

– 2 pts Incorrect: True

Question 2

Perceptron Maximum Iteration

2 / 2 pts

✓ **– 0 pts Correct**: False

– 2 pts Incorrect: True

Question 3

Pruned Decision Tree

2 / 2 pts

✓ **– 0 pts Correct**: True

– 2 pts Incorrect: False

Question 4

1-Nearest Neighbors

2 / 2 pts

✓ **– 0 pts Correct**: False

– 2 pts Incorrect: True

Question 5

Gradient Descent on Function

0 / 2 pts

– 0 pts Correct: False

✓ **– 2 pts Incorrect**: True

– 1 pt Incorrect: Correct work shown

Question 6

1-NN and Linear Separability

2 / 2 pts

✓ **– 0 pts Correct**: True

– 2 pts Incorrect: False

– 0 pts Correct: Valid explanation

Question 7

Training Error of Perceptron

2 / 2 pts

✓ **– 0 pts Correct**: False

– 2 pts Incorrect: True

Question 8

Perceptron vs. Logistic Regression

2 / 2 pts

✓ **– 0 pts Correct**: False

– 2 pts Incorrect: True

– 1 pt Incorrect: Has comparison of cost functions

Question 9

Permuting Instances

2 / 2 pts

✓ **– 0 pts Correct**: True

– 2 pts Incorrect: False

– 1 pt Incorrect: References Convergence Theorem/properties

Question 10

Log Likelihood

2 / 2 pts

✓ **– 0 pts Correct**: True

– 2 pts Incorrect: False

– 1 pt Incorrect: Comparison of derivatives

Question 11

Nearest Neighbors

3 / 3 pts

✓ **+ 0.75 pts Correct**: Option A ***not*** selected

✓ **+ 0.75 pts Correct**: Option B selected

✓ **+ 0.75 pts Correct**: Option C selected

✓ **+ 0.75 pts Correct**: Option D ***not*** selected

Question 12

Comparing Accuracies

1.5 / 3 pts

✓ **+ 0.75 pts Correct**: Option A ***not*** selected

+ 0.75 pts Correct: Option B selected

+ 0.75 pts Correct: Option C ***not*** selected

✓ **+ 0.75 pts Correct**: Option D ***not*** selected

Question 13

Reducing Overfitting in Decision Trees

3 / 3 pts

+ 0.75 pts Correct: Option A selected

+ 0.75 pts Correct: Option B ***not*** selected

+ 0.75 pts Correct: Option C selected

+ 0.75 pts Correct: Option D selected

✓ **+ 3 pts All Correct**: A,C,D selected, B not selected

+ 0 pts incorrect

Question 14

XOR Training Error

3 / 3 pts

✓ **+ 0.75 pts Correct**: Option A selected

✓ **+ 0.75 pts Correct**: Option B ***not*** selected

✓ **+ 0.75 pts Correct**: Option C ***not*** selected

✓ **+ 0.75 pts Correct**: Option D selected

Question 15

Logistic Regression for Yelp Reviews

2.25 / 3 pts

✓ **+ 0.75 pts Correct**: Option A selected

✓ **+ 0.75 pts Correct**: Option B selected

✓ **+ 0.75 pts Correct**: Option C ***not*** selected

+ 0.75 pts Correct: Option D ***not*** selected

💬 $\sigma(-0.5 + 0 + 0) < 0.5$; hence, a negative review.

Question 16

Decision Tree: Entropy of Y

3 / 3 pts

✓ **− 0 pts Correct**: Option C

− 1 pt Partial: Most work correct, but selected wrong option

− 2 pts Partial: Work shows general understanding, but with a major flaw

− 3 pts Incorrect: Work shows little to no understanding; potentially missing

Question 17

Decision Tree: Conditional Entropy rel. X1

3 / 3 pts

✓ **− 0 pts Correct**: Option C

− 1 pt Partial: Most work correct, but selected wrong option

− 2 pts Partial: Work shows general understanding, but with a major flaw

− 3 pts Incorrect: Work shows little to no understanding; potentially missing

Question 18

Decision Tree: Conditional Entropy rel. X2

3 / 3 pts

✓ **− 0 pts Correct**: Option D

− 1 pt Partial: Most work correct, but selected wrong option

− 2 pts Partial: Work shows general understanding, but with a major flaw

− 3 pts Incorrect: Work shows little to no understanding; potentially missing

Question 19

Maximum Likelihood: Expression

3 / 3 pts

✓ **– 0 pts Correct**: Option A

– 1 pt Partial: Most work correct, but selected wrong option

– 2 pts Partial: Work shows general understanding, but with a major flaw

– 3 pts Incorrect: Work shows little to no understanding; potentially missing

Question 20

Maximum Likelihood: Sample Mean

3 / 3 pts

✓ **– 0 pts Correct**: Option A

– 1 pt Partial: Most work correct, but selected wrong option

– 2 pts Partial: Work shows general understanding, but with a major flaw

– 3 pts Incorrect: Work shows little to no understanding; potentially missing

CM146: Introduction to Machine Learning **Winter 2024**

Midterm exam

Feb 13, 2024

- Please do not open the exam unless you are instructed to do so.
- This is an open book and open notes exam.
- Everything you need in order to solve the problems is supplied in the body of this exam OR in a cheatsheet at the end of the exam.
- Mark your answers ON THE EXAM ITSELF. If you make a mess, clearly indicate your final answer (box it).
- For true/false questions, CIRCLE True OR False. Justification for your choice is not needed but could be provided for partial credit.
- Unless otherwise instructed, for multiple-choice questions, CIRCLE ALL CORRECT CHOICES (in some cases, there may be more than one).
- You may use scratch paper if needed (provided at the end of the exam).
- You have 1 hour 45 minutes.

Good Luck! Legibly write your name and UID in the space provided below.

Name: Harris Doan

UID: 605317270

True/False		/20
Multiple choice		/30
Total		/50

True/False (20 pts)

1. (2 pts) You are given a training dataset with attributes $A_1, \ldots, A_m$ and instances $x^{(1)}, \ldots, x^{(n)}$ and you use the ID3 algorithm to build a decision tree D_1. You then take one of the instances, add a copy of it to the training set (so your new training set will have $n+1$ instances), and rerun the decision tree learning algorithm (with the same random seed) to create D_2. D_1 and D_2 are necessarily identical decision trees.

   True [✓ False]

   ID3 based into gain of the feature has a large P of being changed. So D_1, D_2 not guaranteed to be

2. (2 pts) You run the PerceptronTrain algorithm with $maxIter = 100$. The algorithm terminates at the end of 100 iterations with a classifier that attains a training error of 1%. This means that the training data is not linearly separable.

   True [✓ False]

   It is possible to change only after 100 iteration and reduce TR from 1% to 0%. So it IS linearly sep.

3. (2 pts) You learn a decision tree with the *MaxDepth* parameter set to infinity and then prune the resulting decision tree. The resulting pruned decision tree is less likely to overfit compared to the original decision tree.

   [✓ True] False

   Since pruning is a way to avoid overfitting, if both have ∞ depth, but 1 is pruned, it'll be less overf

4. (2 pts) We want to use 1-Nearest Neighbors (1-NN) to classify houses into one of two classes (cheap vs expensive) given a single feature that measures the area of the house. The predictions made by the 1-NN classifier data can change if the area of the house is measured in square metres instead of square feet. (You can neglect the effect of ties *i.e.*, two training instances that are both nearest neighbors to a test instance.)

   True [✓ False]

   For K-NN always normalize data. So if area changes, the model changes. Thus, this can change the classifier.

5. (2 pts) You run gradient descent to minimize the function $f(x) = (2x-3)^3$. Assume the step size has been chosen appropriately and you run gradient descent till convergence. Then gradient descent will return the global minimum of f.

   ↳ always reach a local minima. In this since it's convex, global minima.

   [✓ True] False

   $f'(x) = 2(2x-3) \times 2 = 8x - 12$

   ↳ $f''(x) = 8 > 0$ → $f(x)$ is convex

6. (2 pts) On a dataset that is <u>not</u> linearly separable, the 1-nearest neighbors classifier obtains zero training error.

[✓ True] False

1-NN always obtain zero error.

7. (2 pts) The training error of the perceptron never increases with each iteration of the perceptron algorithm.

True [✓ False]

Points classified correctly could be wrong upon the next iteration.

8. (2 pts) On a linearly separable dataset, the perceptron and logistic regression learn the same separating hyperplane.

True [✓ False]

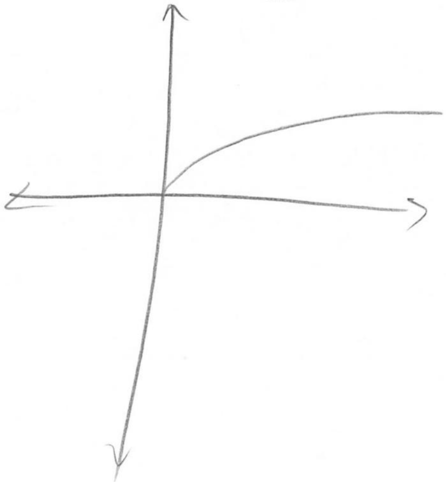
Not true, same features different labels.

9. (2 pts) Permuting the order of instances in the training data can affect the number of iterations for convergence of the perceptron algorithm (assuming the data is linearly separable).

[✓ True] False

10. (2 pts) The value of x at which $f(x)$ attains its maximum is the same as the value of x at which $log(f(x))$ attains its maximum (assume that $f(x) > 0$ for all x).

[✓ True] False



Multiple choice (30 pts)

CIRCLE ALL CORRECT CHOICES (in some cases, there may be more than one)

11. (3 pts) In k-nearest neighbor classification, which of the following statements are true? (circle all that are correct)

    (a) The decision boundary is smoother with smaller values of k.
    [✓ (b)] k-NN does not require any parameters to be learned in the training step (for a fixed value of k and a fixed distance function).
    [✓ (c)] If we set k equal to the number of instances in the training data, k-NN will predict the same class for any input.
    (d) For larger values of k, it is more likely that the classifier will overfit than underfit.

12. (3 pts) Assume we are given a set of one-dimensional inputs and their corresponding output (that is, a set of $\{(x_i, y_i)\}, x_i \in \mathbb{R}, y_i \in \mathbb{R}$). We would like to compare the following two models where $\theta \in \mathbb{R}$:

$$\text{A} : y = \theta^2 x$$
$$\text{B} : y = \theta x$$

    For each model, we split our data into training and testing data to evaluate the generalization accuracy of the learned model (assume that the number of instances in the training and the test data are large). Which of the following is correct?

    (a) There are datasets for which A would be more *accurate* than B.
    (b) There are datasets for which B would be more *accurate* than A.
    [✓ (c)] Both (a) and (b) are correct.
    (d) They would perform equally well on all datasets.

13. (3 pts) What strategy can help reduce over-fitting in decision trees.

    [✓ (a)] Pruning
    (b) Make sure it achieves zero training error
    [✓ (c)] Adding more training data
    [✓ (d)] Enforce a maximum depth for the tree

14. (3 pts) Which of the following algorithms can achieve zero training error on the XOR problem?

[✓ (a)] Decision tree

(b) Logistic regression

(c) Perceptron

[✓ (d)] 1-Nearest Neighbors

XOR

X	Y	XOR
0	0	0
0	1	1
1	0	1
1	1	0

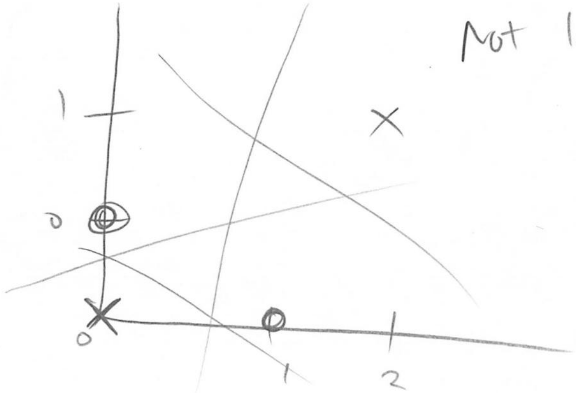
15. (3 pts) Consider a logistic regression model to predict if a yelp review is positive or not ($y = 1$ means the review is positive) based on two features: x_1 and x_2. x_1 is the number of times the word "great" appears and x_2 is the number of times the word "not" appears. The logistic regression model $P(y = 1|x; \boldsymbol{\theta}) = \sigma(\theta_0 + \theta_1 x_1 + \theta_2 x_2)$ with $\boldsymbol{\theta} = (\theta_0, \theta_1, \theta_2) = (-0.5, 1, -2)$. Which of the following is true ?

[✓ (a)] The decision boundary is given by the line $x_1 - 2x_2 - 0.5 = 0$

[✓ (b)] If the word "great" appears more often (assuming everything else about the review is the same), probability that the review is classified as positive becomes closer to 1.

(c) If the word "not" appears more often (assuming everything else about the review is the same), probability that the review is classified as positive becomes closer to 1.

[✓ (d)] If the review contains neither the word "great" nor the word "not", it will be classified as positive.



Not linearly separable

Decision Tree learning

Suppose you want to build a decision tree for a problem. In the dataset, there are two classes (*i.e.*, Y can take one of two possible values), with 60 examples in the $+$ class and 30 examples in the $-$ class. Recall that the information gain for target label Y and feature X is defined as $Gain = H[Y] - H[Y|X]$, where $H[Y] = -E[\log_2 P(Y)]$ is the entropy. See cheatsheet at the end of this exam for entropy values.

16. (3 pts) What is the entropy of the response variable Y? $+ = \frac{60}{90} = \frac{2}{3}$

(a) 0.73 $H[Y] = B(\frac{2}{3}) = 0.92$
(b) 0.81
[✓ (c)] 0.92 $- = \frac{1}{3}$
(d) 0.97

17. (3 pts) For this data, we are interested in computing the information gain of a binary feature X_1. In the $+$ class, the number of instances that have $X_1 = 0$ and $X_1 = 1$ respectively: $(30, 30)$. In the $-$ class, these numbers are: $(0, 30)$. What is the conditional entropy of Y relative to X_1?

(a) 0 $X_1 = 1 \Rightarrow (+) : 30,\ (-) : 30 \ /\ X_1 = 0 : (+)\ 30$
(b) 0.33
[✓ (c)] 0.67 $H[Y|X_1 = 1] = B(\frac{1}{2}) = 1,\ H[Y|X_1 = 0] = 0$
(d) 0.92

$H[Y|X_1] = \frac{2}{3} \cdot 1 + \frac{1}{3} \cdot 0 \Rightarrow \frac{2}{3} + 0 = \frac{2}{3} \approx 0.6$

18. (3 pts) We are interested in computing the information gain of a binary feature X_2. In the $+$ class, the number of instances that have $X_2 = 0$ and $X_2 = 1$ respectively are: $(40, 20)$. In the $-$ class, these numbers are: $(20, 10)$. What is the conditional entropy of Y relative to X_2?

$X_2 = 1 \Rightarrow (+) : 20,\ (-) : 10 \Rightarrow \frac{2\cancel{0}}{3\cancel{0}} =$

(a) 0
(b) 0.33 $X_2 = 0 \Rightarrow (+) : 40,\ (-) : 20 = \frac{4\cancel{0}}{6\cancel{0}} =$
(c) 0.67
[✓ (d)] 0.92

$H[Y|X_2 = 1] = B(\frac{2}{3}) = 0.92$

$H[Y|X_2 = 0] = B(\frac{2}{3}) = 0.92$

$H[Y|X_2] = \frac{3\cancel{0}}{9\cancel{0}} \frac{1}{3} (0.92) + \frac{6\cancel{0}}{9\cancel{0}} \frac{2}{3} (0.92)$

$H[Y|X_2] \approx 0.92$

$$\sum_{n=1}^{n} \underbrace{\log(e^{-\lambda})}_{-\lambda} + \sum_{n=1}^{n} \log\left(\frac{\lambda^{x_n}}{x_n!}\right)$$ is constant

$$N \Rightarrow -\lambda N + \log(\lambda) \sum_{n=1}^{n} x_n$$

MLE

We observe a data set consisting of N samples: $x_1, \ldots, x_N$. $x_1, \ldots, x_N$ are i.i.d. random variables where each random variable is distribute as $Poisson(\lambda)$. The probability mass function for $X \sim Poisson(\lambda)$ is:

$$p(x; \lambda) = \frac{e^{-\lambda}\lambda^x}{x!}$$

19. (3 pts) What is the expression for the log-likelihood $l(\lambda)$ (all terms that do not depend on λ are refered to as *const*)?

[✓ (a)] $l(\lambda) = -N\lambda + \log(\lambda)(\sum_n x_n) + const$

(b) $l(\lambda) = \lambda^N e^{-\lambda \sum_n x_n} + const$

(c) $l(\lambda) = -N\log(\lambda) + \log(\lambda)\sum_n x_n + const$

(d) $l(\lambda) = \sum_n \lambda e^{-\lambda x_n} + const$

$$\mathcal{LL}(\lambda) = \log(p(x_1 \ldots x_n) ; \lambda)$$

$$= \log(x_1 ; \lambda) \log(x_2, \lambda) \ldots \log($$

$$\downarrow$$

$$\log(p(x_1 ; \lambda)) + \log(p(x_2 ; \lambda)) \ldots$$

$$\downarrow$$

$$\sum_{n=1} \log(p(x_n ; \lambda)) \to *$$

$$* \sum_{n=1}^{n} \log\left(\frac{e^{-\lambda}\lambda^{x_n}}{x!}\right) \nearrow 0$$

$\bar{x}N$

20. (3 pts) Let $\bar{x} = \frac{\sum_n x_n}{N}$ denote the sample mean of $(x_1, \ldots, x_N)$. What is the MLE, $\hat{\lambda}$, of λ?

[✓ (a)] $\hat{\lambda} = \bar{x}$

(b) $\hat{\lambda} = \frac{1}{\bar{x}}$

(c) $\hat{\lambda} = e^{\bar{x}}$

(d) $\hat{\lambda} = \sum_n x_n$

$$\frac{d\,\mathcal{LL}(\lambda)}{d\lambda} = d\left(-N\lambda + \log(\lambda)\sum_n x_n\right)$$

$$\frac{d\,\mathcal{LL}(\lambda)}{d\lambda}(-N\lambda) + \frac{d\,\mathcal{LL}(\lambda)}{d(\lambda)} \overset{a}{\log(\lambda)} \overset{b}{\sum_n x_n} + \cancel{\frac{d\,\mathcal{LL}(\lambda)}{d\lambda}}\ \text{con}$$

$$\downarrow$$

$$-N +$$

product rule

$$a'b + ab'$$

$$\frac{1}{\lambda}\sum_n x_n + \cancel{\log(\lambda)}\left(\cancel{\frac{d\mathcal{LL}}{d(\lambda)}} \cancel{\sum_n} x_n\right) \nearrow \overset{0}{\bar{x}N}$$

$$-N + \frac{1}{\lambda}\sum_n x_n \qquad \frac{\sum_n x_n}{\lambda} = \frac{\bar{x}N}{\lambda} - N = 0$$

$$\bar{\lambda} = \bar{x} \qquad \frac{\bar{x}\cancel{N}}{\cancel{\ }} = \lambda \qquad N$$

Identities

Probability density/mass functions for some distributions

$$
\begin{aligned}
\text{Normal} &: P(x;\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \\
\text{Multinomial} &: P(\boldsymbol{x};\boldsymbol{\pi}) = \prod_{k=1}^{K} {\pi_k}^{x_k} \\
&\quad \boldsymbol{x} \text{ is a length } K \text{ vector with exactly one entry equal to 1} \\
&\quad \text{and all other entries equal to 0} \\
\text{Poisson} &: P(x;\lambda) = \frac{\lambda^x \exp(-\lambda)}{x!}
\end{aligned}
$$

Matrix calculus

Here $\boldsymbol{x} \in \mathbb{R}^n, \boldsymbol{b} \in \mathbb{R}^n, \boldsymbol{A} \in \mathbb{R}^{n\times n}$. $\boldsymbol{A}$ is symmetric.

$$\nabla \boldsymbol{x}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{x} = 2\boldsymbol{A}\boldsymbol{x}, \qquad \nabla \boldsymbol{b}^{\mathrm{T}}\boldsymbol{x} = \boldsymbol{b}$$

Entropy

The entropy $H(X)$ of a Bernoulli random variable $X \sim \mathit{Bernoulli}(p)$ for different values of p:

p	$H(X)$
$\frac{1}{2}$	1
$\frac{1}{3}$	0.92
$\frac{1}{4}$	0.81
$\frac{1}{5}$	0.73
$\frac{2}{5}$	0.97