

# 20W-COMSCIM146 midterm

SENYANG JIANG

TOTAL POINTS

**79 / 80**

QUESTION 1

true/false 18 pts

1.1 1 2 / 2

- 2 pts Incorrect

✓ - 0 pts Correct

1.2 2 2 / 2

✓ - 0 pts Correct

- 2 pts Incorrect

1.3 3 2 / 2

✓ - 0 pts Correct

- 2 pts Incorrect

1.4 4 2 / 2

✓ - 0 pts Correct

- 0 pts Incorrect but all are given points

1.5 5 2 / 2

✓ - 0 pts Correct

- 2 pts Incorrect

1.6 6 2 / 2

✓ - 0 pts Correct

- 2 pts Incorrect

1.7 7 2 / 2

✓ - 0 pts Correct

- 2 pts Incorrect

1.8 8 2 / 2

✓ - 0 pts Correct

- 2 pts Incorrect

1.9 9 2 / 2

✓ - 0 pts Correct

- 2 pts Incorrect

QUESTION 2

multiple choice 28 pts

2.1 10 4 / 4

✓ - 0 pts Correct

- 1 pts (a) is not selected

- 1 pts (c) is not selected

- 1 pts (d) is not selected

- 1 pts (b) is selected

2.2 11 4 / 4

✓ - 0 pts Correct

- 1 pts a is selected

- 1 pts b is NOT selected

- 1 pts c is selected

- 1 pts d is selected

2.3 12 3 / 4

- 0 pts Correct

✓ - 1 pts a) is not selected

- 1 pts b) is not selected

- 1 pts c) is selected

- 1 pts d) is not selected

2.4 13 4 / 4

✓ - 0 pts Correct

- 1 pts a) is true

- 1 pts b) is true

- 1 pts c) is false

- 1 pts d) is true

2.5 14 4 / 4

✓ - 0 pts Correct

- 1 pts a) is false
- 1 pts b) is true
- 1 pts c) is false
- 1 pts d) is true

## 2.6 15 4 / 4

- ✓ - 0 pts Correct
- 1 pts a) is not selected
- 1 pts b) is not selected
- 1 pts c) is selected
- 1 pts d) is selected

## 2.7 16 4 / 4

- ✓ - 0 pts Correct
- 1 pts a) is true
- 1 pts b) is false
- 1 pts c) is false
- 1 pts d) is false

## QUESTION 3

### decision tree 13 pts

#### 3.1 a 2 / 2

- ✓ - 0 pts Correct
- 0.5 pts correct expression. should use the provided identities for final answer.
- 2 pts wrong
- 1 pts partially correct

#### 3.2 b 5 / 5

- ✓ - 0 pts Correct
- 2 pts incorrect  $H(Y|X=1)$
- 2 pts incorrect  $H(Y|X=0)$
- 1 pts correct (form of)  $H(Y|X=0)$ ,  $H(Y|X=1)$ , wrong (form of)  $H(Y|X)$
- 1 pts wrong/missing IG(information gain)

#### 3.3 C 5 / 5

- ✓ - 0 pts Correct
- 1 pts missing IG
- 2 pts wrong  $H[y|x_2=0]$
- 2 pts wrong  $H[y|x_2=1]$

- 1 pts correct  $H[y|x_2=1]$ , correct  $H[y|x_2=0]$ , wrong  $H[y|x_2]$
- 1 pts wrong form of  $H[y|x]$

#### 3.4 d 1 / 1

- ✓ - 0 pts Correct
- 1 pts wrong

## QUESTION 4

### MLE 8 pts

#### 4.1 a 3 / 3

- ✓ - 0 pts Correct
- 2 pts Click here to replace this description.
- 1 pts Click here to replace this description.
- 3 pts Click here to replace this description.

#### 4.2 b 3 / 3

- ✓ - 0 pts Correct
- 1 pts Click here to replace this description.
- 2 pts Click here to replace this description.
- 3 pts Click here to replace this description.

#### 4.3 C 2 / 2

- ✓ - 0 pts Correct
- 1 pts Click here to replace this description.
- 2 pts Click here to replace this description.

## QUESTION 5

### logistic regression 12 pts

#### 5.1 a 2 / 2

- ✓ - 0 pts Correct
- 2 pts The answer is not correct.
- 1 pts Your expression is correct, but your result of C is not correct.

#### 5.2 b 5 / 5

- ✓ - 0 pts Correct
- 5 pts Wrong answer

#### 5.3 C 2 / 2

✓ - 0 pts Correct

- 2 pts wrong answer

- 1 pts Your answer is partially correct (you know the answer should be the sum of the answer you get for (b)).

- 0.5 pts C varies for different n, so you cannot get  $N \log C$  or  $C^N$

5.4 d 3 / 3

✓ - 0 pts Correct

- 3 pts Wrong answer

- 2 pts You did not show the equivalence.

- 1 pts Your result of  $l(\theta)$  is not correct.

- 1 pts your answers are partially correct.

QUESTION 6

6 name 1 / 1

✓ - 0 pts Correct

## Midterm

Feb. 10<sup>th</sup>, 2020

- Please do not open the exam unless you are instructed to do so.
- This is a closed book and closed 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 and provide a brief justification for full credit.
- Unless otherwise instructed, for multiple-choice questions, CIRCLE ALL CORRECT CHOICES (in some cases, there may be more than one) and provide a brief justification if the question asks for one.
- If you think something about a question is open to interpretation, feel free to ask the instructor or make a note on the exam.
- If you run out of room for your answer in the space provided, please use the blank pages at the end of the exam and indicate clearly that you've done so.
- Do NOT put answers on the back of any page of the exam.
- You may use scratch paper if needed (provided at the end of the exam).
- You have 1 hour 45 minutes.
- Besides having the correct answer, being concise and clear is very important. For full credit, you must show your work and explain your answers.

Good Luck!

Legibly write your name and UID in the space provided below to earn 2 points.

Name: Jiang Senyang

UID: 505111806

Name and UID		/1
True/False		/18
Multiple choice		/28
Decision tree		/13
Maximum likelihood		/8
Logistic regression		/12
<b>Total</b>		<b>/80</b>

## True/False (18 pts)

1. (2 pts) We are using linear regression to predict height from weight and age. The optimal value of the residual sum of squares (RSS) will change if we measure weight in pounds instead of kilograms for each instance.

True

False

The optimal value will not change because the parameter  $w$  for weight will scale by a constant, making the prediction unchanged.

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

True

False

1-nearest neighbors classifier always obtains zero training error.

3. (2 pts) As the number of training examples grows toward infinity, the probability that logistic regression will over-fit the training data goes to zero.

True

False

If we have infinite training examples, the model will have the true data distribution and it's impossible to overfit.

4. (2 pts) The K-nearest neighbor algorithm often uses Euclidean distance as the default distance metric:  $d(\mathbf{x}_i, \mathbf{x}_j) = \|\mathbf{x}_i - \mathbf{x}_j\|_2$ . Suppose we instead use a new distance metric:  $d(\mathbf{x}_i, \mathbf{x}_j) = \|\mathbf{x}_i - \mathbf{x}_j\|_1$ . The classification results will change as a result of this distance metric.

True

False

The nearest neighbor will change if we change from  $\|\mathbf{x}_i - \mathbf{x}_j\|_2$  to  $\|\mathbf{x}_i - \mathbf{x}_j\|_1$ .

5. (2 pts) The training error of the perceptron never increases with each iteration of the perceptron algorithm.

True

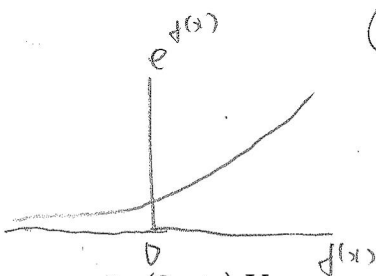
False

Previous points classified correctly could be incorrect after the next iteration

6. (2 pts) The value of  $x$  at which  $f(x)$  attains its maximum is the same as the value of  $x$  at which  $e^{f(x)}$  attains its maximum.

True

False



$e^{f(x)}$  is an increasing function

7. (2 pts) You compute  $x^* = \operatorname{argmin}_x f(x)$  and find that  $x^* = -\infty$ .  $f(x)$  is not convex.

True

False

counter example:  $\operatorname{argmin}_x e^x = -\infty$ , but  $e^x$  is convex

$$(e^x)'' = e^x \geq 0$$

8. (2 pts) A decision tree learned with *MaxDepth* parameter set to  $\infty$  always attains zero training error.

True

False

two instances may have same feature values but different labels

9. (2 pts) Stochastic Gradient Descent is faster per iteration than Batch Gradient Descent.

True

False

SAD  $O(D)$  per iteration  
BAD  $O(MD)$  per iteration

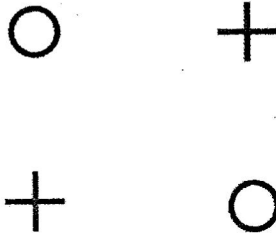
## Multiple choice (28 pts)

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

10. (4 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

11. (4 pts) Consider the following data set: Circle the classifiers that will achieve zero training error on this data set.



- (a) Logistic regression
- (b) Depth-2 ID3 decision trees
- (c) Perceptron
- (d) 3NN classifier

12. (4 pts) Which of the following are true about linear regression?

- (a) Its cost function is always convex
- (b) Its cost function is always convex after adding  $l_2$  regularization.
- (c) The cost function always has a unique minimum.
- (d) The cost function always has a unique minimum after adding  $l_2$  regularization.



13. (4 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  $P(y = 1|x; \theta) = \sigma(\theta_0 + \theta_1 x_1 + \theta_2 x_2)$  with  $\theta = (\theta_0, \theta_1, \theta_2) = (2, 1, -2)$ . Which of the following is true?
- (a) The decision boundary is given by the line  $x_1 - 2x_2 + 2 = 0$
- (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.
14. (4 pts) Which of the following is true of the Perceptron learning algorithm?
- (a) If the algorithm does not converge within 100 iterations, the data is not linearly separable.
- (b) If the algorithm converges within 100 iterations, the data is linearly separable.
- (c) If the data is linearly separable, it will converge within 100 iterations.
- (d) If the data is linearly separable, it may not converge in 100 iterations but may converge in 200 iterations.
15. The entropy of a distribution over a set of 3 items with probability mass function  $p$  is defined as  $-\sum_{k=1}^3 p(k) \log_2 p(k)$ . Which of the following distributions has the lowest entropy?
- (a) (1, 0, 0)
- (b) (0, 1, 0)
- (c) ( $\frac{1}{3}, \frac{1}{3}, \frac{1}{3}$ )
- (d) (0.1, 0.3, 0.6)

16. (4 pts) Given the same training data consisting of  $N$  instances  $\{(x_n, y_n)\}_{n=1}^N$  where  $x_n \in \mathbb{R}, y_n \in \mathbb{R}$ , we fit two models:  $h_1(x) = \theta_0 + \theta_1 x$  and  $h_2(x) = \theta_0 + \theta_1 x + \theta_2 x^2$ . For each model, we estimate its parameters by minimizing the residual sum of squares (RSS). Let  $RSS_m$  denote the minimum value of the residual sum of squares cost function evaluated on the training dataset for model  $m$  ( $m \in \{1, 2\}$ ). Which of the following is always true?

(a)  $RSS_1 \geq RSS_2$

(b)  $RSS_1 \leq RSS_2$

(c)  $RSS_1 = RSS_2$

(d)  $RSS_1$  can sometimes be greater and sometimes lesser than  $RSS_2$ .

## Decision Tree (13 pts)

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.

- (a) (2 pts) What is the entropy of the class variable  $Y$ ?

$$P(Y=+) = \frac{2}{3} \quad P(Y=-) = \frac{1}{3}$$

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

- (b) (5 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). Write down conditional entropy and information gain of  $X_1$  relative to  $Y$ ?

$$X_1 = 1 : 30 + , 30 -$$

$$X_1 = 0 : 30 +$$

$$H[Y | X_1 = 1] = B\left(\frac{1}{2}\right) = 1$$

$$H[Y | X_1 = 0] = B(1) = 0$$

$$\begin{aligned} H[Y | X_1] &= \frac{60}{90} \cdot 1 + \frac{30}{90} \cdot 0 \\ &= \frac{2}{3} \approx 0.67 \end{aligned}$$

$$\begin{aligned} \text{information gain} &= 0.92 - 0.67 \\ &= 0.25 \end{aligned}$$

- (c) (5 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). Write down conditional entropy and information gain of  $X_2$  relative to  $Y$ ?

$$X_2 = 1 : 20+, 10-$$

$$X_2 = 0 : 40+, 20-$$

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

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

$$\begin{aligned} H[Y|X_2] &= \frac{30}{90} \cdot 0.92 + \frac{60}{90} \cdot 0.92 \\ &= 0.92 \end{aligned}$$

$$\text{Information gain} = 0.92 - 0.92 = 0$$

(d) (1 pts) Using information gain, which attribute will the ID3 decision tree learning algorithm choose as the root?

use  $X_1$  as the root

## Maximum Likelihood (8 pts)

Let  $X_1, \dots, X_N$  be i.i.d. random variables where  $X_n \sim \text{Exp}(\lambda), n \in \{1, \dots, N\}$ . The probability density function for  $X \sim \text{Exp}(\lambda)$  is:

$$p(x; \lambda) = \begin{cases} \lambda e^{-\lambda x}, & \text{if } x \geq 0 \\ 0, & \text{if } x < 0 \end{cases}$$

- (a) (3 pts) Give an expression for the log likelihood  $l(\lambda)$  as a function of  $\lambda$  given this specific dataset. You may assume all values,  $X_1, \dots, X_N \geq 0$ .

$$\begin{aligned} l(\lambda) &= \sum_{n=1}^N \log P(X_n = x_n; \lambda) \\ &= \sum_{n=1}^N -\lambda x_n + \log \lambda \\ &= -\lambda \sum_{n=1}^N x_n + N \log \lambda \end{aligned}$$

(b) (3 pts) Compute the derivative of the log likelihood for this specific dataset.

$$\frac{dl(\lambda)}{d\lambda} = -\sum_{n=1}^N x_n + \frac{N}{\lambda}$$

(c) (2 pts) What is the maximum likelihood estimate  $\hat{\lambda}$  of  $\lambda$ ?

$$-\sum_{n=1}^N x_n + \frac{N}{\lambda} = 0$$
$$\hat{\lambda} = \frac{N}{\sum_{n=1}^N x_n}$$



## Alternative to logistic regression (12 pts)

In class, we discussed the logistic regression model for binary classification problem. Here, we consider an alternative model. We have a training set  $\{(\mathbf{x}_n, y_n)\}_{n=1}^N$  where  $\mathbf{x}_n \in \mathbb{R}^{D+1}$  and  $y_n \in \{0, 1\}$ . Like in logistic regression, we will construct a probabilistic model for the probability that  $y_n$  belongs to class 0 or 1, given  $\mathbf{x}_n$  and the model parameters,  $\boldsymbol{\theta}_0$  and  $\boldsymbol{\theta}_1$  ( $\boldsymbol{\theta}_0, \boldsymbol{\theta}_1 \in \mathbb{R}^{D+1}$ ). More specifically, we model the target  $y_n$  as:

$$\begin{aligned} p(y_n = 0 | \mathbf{x}_n; \boldsymbol{\theta}_0, \boldsymbol{\theta}_1) &= C e^{\boldsymbol{\theta}_0^T \mathbf{x}_n} \\ p(y_n = 1 | \mathbf{x}_n; \boldsymbol{\theta}_0, \boldsymbol{\theta}_1) &= C e^{\boldsymbol{\theta}_1^T \mathbf{x}_n} \end{aligned} \quad (1)$$

- (a) (2 pts) Find the value of  $C$  that makes Equation 1 a valid probability distribution for  $y_n$ .

$$C e^{\boldsymbol{\theta}_0^T \mathbf{x}_n} + C e^{\boldsymbol{\theta}_1^T \mathbf{x}_n} = 1$$

$$C (e^{\boldsymbol{\theta}_0^T \mathbf{x}_n} + e^{\boldsymbol{\theta}_1^T \mathbf{x}_n}) = 1$$

$$C = \frac{1}{e^{\boldsymbol{\theta}_0^T \mathbf{x}_n} + e^{\boldsymbol{\theta}_1^T \mathbf{x}_n}}$$

- (b) (5 pts) Write the log likelihood of the parameters for a single instance  $(x_n, y_n)$ :  $l(\theta_0, \theta_1) = \log p(y_n | x_n; \theta_0, \theta_1)$ . Express your answer in terms of  $y_n, x_n, \theta_0, \theta_1$ . (Hint: use the notation in class where for a Bernoulli random variable  $Y_n$  that takes values 1 with probability  $\theta$  and 0 with probability  $1 - \theta$ , we can write its probability mass function:  $p(y_n) = \theta^{y_n} (1 - \theta)^{1 - y_n}$ ).

$$p(y_n | x_n; \theta_0, \theta_1) = (C e^{\theta_1^T x_n})^{y_n} (C e^{\theta_0^T x_n})^{1 - y_n}$$

$$\begin{aligned} L(\theta_0, \theta_1) &= \log p(y_n | x_n; \theta_0, \theta_1) \\ &= y_n \log C e^{\theta_1^T x_n} + (1 - y_n) \log C e^{\theta_0^T x_n} \\ &= y_n (\log C + \theta_1^T x_n) + (1 - y_n) (\log C + \theta_0^T x_n) \\ &= y_n \theta_1^T x_n + (1 - y_n) \theta_0^T x_n + \log C \\ &= y_n \theta_1^T x_n + (1 - y_n) \theta_0^T x_n - \log [e^{\theta_0^T x_n} + e^{\theta_1^T x_n}] \\ &= y_n (\theta_1 - \theta_0)^T x_n + \theta_0^T x_n - \log [e^{\theta_0^T x_n} (1 + e^{(\theta_1 - \theta_0)^T x_n})] \\ &= y_n (\theta_1 - \theta_0)^T x_n - \log (1 + e^{(\theta_1 - \theta_0)^T x_n}) \end{aligned}$$

- (c) (2 pts) Write the log likelihood of the parameters  $\mathcal{L}(\theta_0, \theta_1)$  for the full training data.

$$\begin{aligned}\mathcal{L}(\theta_0, \theta_1) &= \sum_{n=1}^N \ell(\theta_0, \theta_1) \\ &= \sum_{n=1}^N y_n (\theta_1 - \theta_0)^T x_n - \sum_{n=1}^N \log(1 + e^{(\theta_1 - \theta_0)^T x_n})\end{aligned}$$

(d) (3 pts) Recall that in logistic regression, we model the target  $y_n$  as :

$$\begin{aligned} p(y_n = 0 | x_n; \theta) &= 1 - \sigma(\theta^T x_n) \\ p(y_n = 1 | x_n; \theta) &= \sigma(\theta^T x_n) \end{aligned}$$

Here  $\sigma(x) = \frac{1}{1+e^{-x}}$  is the sigmoid function. Show that our new model (Equation 1) is exactly the same as the logistic regression model with  $\theta = \theta_1 - \theta_0$ .

from equation 1:  $p(y_n = 0 | x_n; \theta_0, \theta_1) = \frac{e^{\theta_0^T x_n}}{e^{\theta_0^T x_n} + e^{\theta_1^T x_n}}$

$$\begin{aligned} &= \frac{e^{\theta_0^T x_n}}{e^{\theta_0^T x_n} + e^{\theta_1^T x_n}} \\ &= 1 - \frac{e^{\theta_1^T x_n}}{e^{\theta_0^T x_n} + e^{\theta_1^T x_n}} \\ &= 1 - \frac{1}{1 + e^{(\theta_0 - \theta_1)^T x_n}} \\ &= 1 - \frac{1}{1 + e^{-\theta^T x_n}} = 1 - \sigma(\theta^T x_n) \end{aligned}$$

$$\begin{aligned} p(y_n = 1 | x_n; \theta_0, \theta_1) &= \frac{e^{\theta_1^T x_n}}{e^{\theta_0^T x_n} + e^{\theta_1^T x_n}} \\ &= \frac{1}{e^{(\theta_0 - \theta_1)^T x_n} + 1} \\ &= \frac{1}{1 + e^{-\theta^T x_n}} \\ &= \sigma(\theta^T x_n) \end{aligned}$$

So these two models are the same.

## Identities

### Probability density/mass functions for some distributions

$$\text{Normal : } P(x; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x - \mu)^2}{2\sigma^2}\right)$$

$$\text{Multinomial : } P(\mathbf{x}; \boldsymbol{\pi}) = \prod_{k=1}^K \pi_k^{x_k}$$

$\mathbf{x}$  is a length  $K$  vector with exactly one entry equal to 1  
and all other entries equal to 0

$$\text{Poisson : } P(x; \lambda) = \frac{\lambda^x \exp(-\lambda)}{x!}$$

## Matrix calculus

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

$$\nabla_{\mathbf{x}}^T \mathbf{A} \mathbf{x} = 2\mathbf{A} \mathbf{x}, \quad \nabla_{\mathbf{b}}^T \mathbf{x} = \mathbf{b}$$

## Entropy

The entropy  $H(X)$  of a Bernoulli random variable  $X \sim \text{Bernoulli}(p)$  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

You may use this page for scratch space.

You may use this page for scratch space.