

COMP 551 Applied Machine Learning Midterm Exam

October 18, 2022, 8:40 am - 9:40 am

Name:

Student ID number:

- You have 60 minutes to write the exam.
- Hard-copies notes, books, and printed slides are allowed but electronic devices are NOT allowed.
- This exam contains 16 questions on 10 pages.
- For each of the 6 multiple-choice questions, circle only ONE correct answer.
- For each of 4 multiple-select questions, circle ALL correct answers.
- For each of 6 short-answer questions, write your answer directly below the question.

1 Multiple-choice questions (30 points)

1. (5 points) After *correctly* training a logistic regression model on the training data, if we evaluate the trained model on N test data points using $\hat{y}^{(n)} = \frac{\exp(-\mathbf{x}^{(n)}\mathbf{w})}{1+\exp(-\mathbf{x}^{(n)}\mathbf{w})}$ as $p(y^{(n)} = 1|\mathbf{x}^{(n)})$ for each data point and obtain the area under the receiver operating characteristic curve (AUROC) equal to 0.2. What would be the AUROC if $\hat{y}^{(n)} = \frac{1}{1+\exp(-\mathbf{x}^{(n)}\mathbf{w})}$ were used as the prediction for $p(y^{(n)} = 1|\mathbf{x}^{(n)})$ on the same test data instead? Choose the correct answer

- A. 0.0
B. 0.2
C. 0.8
D. 1.0

$$\hat{y}^{(n)} = 1 - \frac{1}{1 + \exp(-\mathbf{x}^{(n)}\mathbf{w})}$$

2. (5 points) Given the following confusion table, what are the precision and recall rates:

	Predicted negative	Predicted positive
Actual negative	1	2
Actual positive	1	3

A. precision: 0.60; recall: 0.75

B. precision: 0.75; recall: 0.66

C. precision: 0.60; recall: 0.33

D. precision: 0.75; recall: 0.25

$$\text{Precision} = \frac{TP}{TP+FP} = \frac{3}{3+2} = 0.60$$

$$\text{Recall} = \frac{TP}{TP+FN} = \frac{3}{3+1} = 0.75$$

Solution: TP=3; PP=5; FN=1. Precision=TP/PP=3/5=0.60; Recall=TP/(TP+FN)=3/(3+1)=0.75

3. (5 points) Suppose you have a spam detection model (spam = 1, not spam = 0) to filter the messages sent to you. You can control a parameter in the model to increase or decrease the recall. If you set the parameter to have higher recall, do you expect to see more or less spam in your inbox? Choose the correct answer

A. less

B. more

C. no change

$$\text{Recall} = \frac{TP}{TP+FN} \quad \uparrow \Rightarrow TP \uparrow \Rightarrow \text{more spams caught}$$

Solution: Higher recall = $\frac{TP}{TP+FN}$ means catching more spam emails (higher true positive TP). This results in less spams in the inbox (FN gets lower) but also more non-spam emails being marked as spam (FP gets higher).

4. (5 points) Suppose you are conducting 10-fold cross-validation to choose the best tree depth from a set {1, 5, 10} for the Decision Tree. How many times will the same data point be used for training after the entire 10-fold CV experiment? Choose the correct answer.

A. 1

B. 10

C. 27

D. 30

tree depth: 3

$$3 \times 9 = 27$$

9 folds for training

5. (5 points) Suppose you would like to simulate data for binary classification using a logistic regression model:

$$\hat{y} = \frac{1}{1 + \exp(-\mathbf{x}\mathbf{w} - w_0)}$$

What value of the bias term w_0 will you set so that the expected fraction of the positive label is 0.1 when the input features $\mathbf{x} = \mathbf{0}$? Choose the correct answer.

A. 0

B. 0.1

C. -ln9

D. -ln10

$$0.1 = \frac{1}{1 + \exp(-w_0)}$$

$$\exp(-w_0) = 9$$

$$w_0 = -\ln 9$$

$$1 + \exp(-w_0) = \frac{1}{0.1}$$

6. (5 points) For $C = 4$ classes and $D = 5$ features, what is the predicted class for input $\mathbf{x} = [1 \ 0 \ 0 \ 1 \ 0]$,

$$\mathbf{W} = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 1 & 1 \\ 1 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 & 1 \\ 0 & 1 & 1 & 1 & 1 \end{bmatrix}$$

when using a multiclass regression model with the following weights: $\mathbf{W} =$

$$y = \mathbf{x}\mathbf{w}$$

=

$$[1 \ 0 \ 0 \ 1 \ 0] \cdot [0 \ 2 \ 0 \ 1]$$

↑

A. 1

B. 2

C. 3

D. 2 or 4

E. 4

2 Multiple-select questions (20 points)

7. (5 points) Assuming you are helping doctors to predict cancer stages I, II, III, IV of lung cancer patients. Which of the following method(s) would you experiment? Circle ALL that are appropriate.

A. K nearest neighbours

B. Decision tree

C. Linear regression

D. Logistic regression

E. Multiclass regression

Solution: KNN, DT, Multiclass are capable of predicting multiclass labels but not linear regression and logistic regression

8. (5 points) Suppose \hat{y} is our predicted value and y is the true value of a target variable. For multiclass prediction, \hat{y}_c denotes the probabilities of class c , and y_c is the binary indicator for whether the true label is class c . Connect each cost listed on the left to the equivalent log likelihood on the right by drawing a line between them. Leave the ones which do not have matching log likelihood:

- | | |
|-------------------------------------------------|------------------------------------|
| • $\ y - \hat{y}\ _2^2$ | • log Bernoulli |
| • $\sum_{c=1}^C -y_c \log \hat{y}_c$ | • log Binomial |
| • $-y \log \hat{y} - (1 - y) \log(1 - \hat{y})$ | • log Categorical |
| • $\sum_{c=1}^C \mathbb{I}[\hat{y}_c \neq y_c]$ | • log Gaussian with $\sigma^2 = 1$ |

Solution: ($\|y - \hat{y}\|_2^2$, log Gaussian) ($\sum_{c=1}^C -y_c \log \hat{y}_c$, Categorical) ($-y \log \hat{y} - (1 - y) \log(1 - \hat{y})$, log Bernoulli)

Give them full mark if they link CE to log Binomial and full mark if they link CE to log Bernoulli (1.25 pt) and 0 if they link CE to anything else on the right side.

9. (5 points) Choose ALL ML methods that can be trained by gradient descent.

A. K nearest neighbours

B. Decision tree

C. Linear regression

D. Logistic regression

E. Multiclass regression

all regressions

10. (5 points) Choose ALL correct statement(s) below

A. Overfitting occurs when the training error starts to increase

B. Test error will always be lower than validation error

C. Overfitting can be detected using a validation set

← detect overfitting by comparing training and validation error

D. Overfitting occurs when the learning rate is too large

E. Overfitting can occur when the K value for the KNN is too large

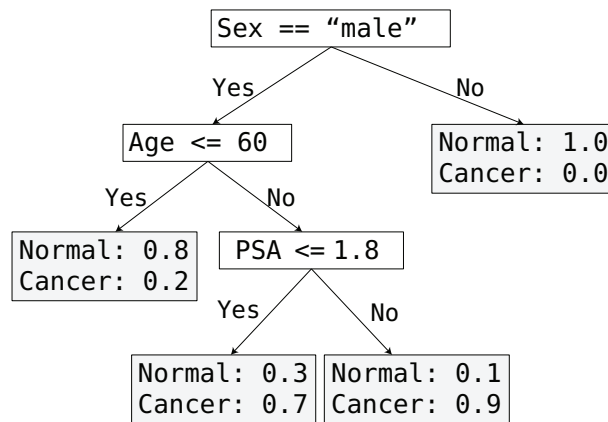
F. Overfitting can occur when the tree depth is too large

Solution: Rubrics:

For multiple select question, students get points for making choices correctly (including TP and TN). For example, if someone selects A and B out of A, B, and D correct choices in a 5-point M-S Question of 5 choices (A-E). He/she will earn $4 - 1 = 3$ points. He/she earns points for making 2 correct selection A and B and He/she gets 2 points for not choosing C and E. Then He/she loses 1 point for not choosing D. Therefore, the marking scheme is Right Minus Wrong

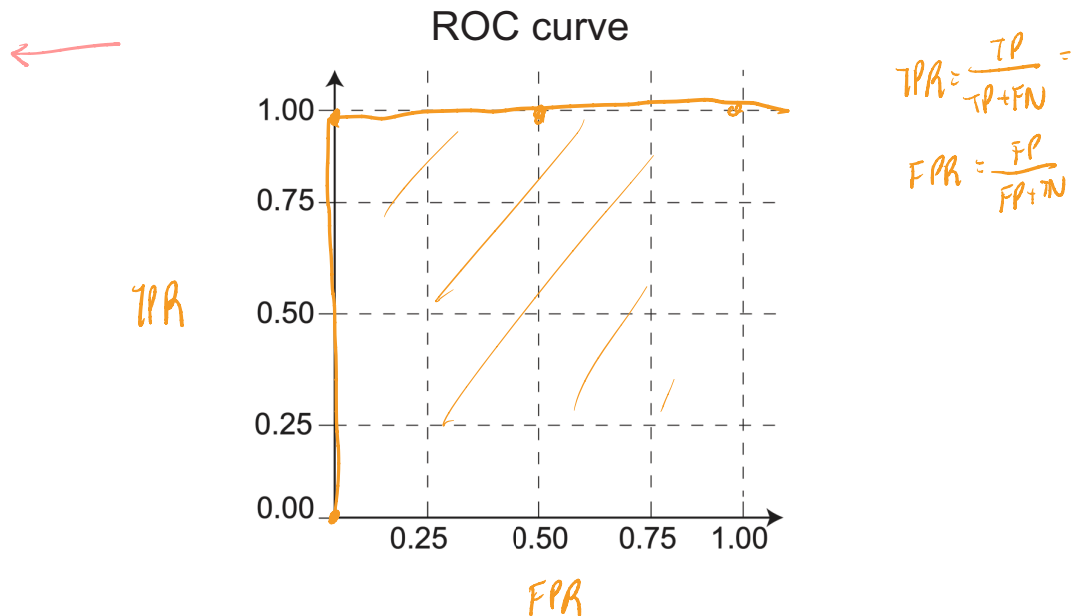
3 Short answer questions (50 points)

11. (15 points) The following decision tree is used to diagnose prostate cancer based on age, PSA, and sex. The decision tree is used to predict the probabilities of cancer for 3 patients in the table beside. Based on the thresholds 1, 0.6, 0.1, and -1, draw the ROC curve treating cancer as positive label, i.e., $y = 1$ and normal as negative label, i.e., $y = 0$. You may use the tables below to help you compute the ROC curve and get partial marks. But you still get full mark for this question if you draw the correct ROC curve without filling out the tables.



Patient ID	Age	PSA	Sex	Status	patient index	$p(y = 1 x)$	predicted label	true label
1	55	1.8	male	Normal	1	0.2	0	0
2	76	1.7	male	Cancer	2	0.7	1	1
3	56	0	female	Normal	3	0	0	0

		1		0.6		0.1		-0.1			
		PN	PP	PN	PP	PN	PP	PN	PP		
AN		2	0	2	0	1	1	0	2	TPR	1 0.6 0.1 -0.1
AP		1	0	0	1	0	1	0	1	FPR	0 1 1 1
											0 0 0.5 1



Patient ID	Age	PSA	Sex	Status
1	55	1.8	male	Normal
2	76	1.7	male	Cancer
3	56	0	female	Normal

Solution:

patient index	$p(y = 1 x)$	predicted label	true label
1	0.2	0	0
2	0.7	1	1
3	0	0	0

	PN	PP	PN	PP	PN	PP	PN	PP		1	0.6	0.1	-0.1
AN	2	0	2	0	1	1	0	2	TPR	0	1	1	1
AP	1	0	0	1	0	1	0	1	FPR	0	0	0.5	1

ROC FPR and TPR coordinates (0,0), (0,1), (0.5,1), (1,1)

Grading Rubrics:

If the ROC graph is correct, full mark;

If not:

Correct "patient index table": 3 points

Correct "Confusion matrix": 4 points

Correct "ROC points": 3 points

Correct ROC graph: 5 points

We will remove marks proportionally. Eg, the student calculates one thing wrong in the patient index table, and it causes cascade mistakes, so one point will be removed for all the following tables. (In this case, 4 points) (If half of the table is wrong, we remove half of the corresponding point.)

If one table is correct, we can give all marks above(if the student only gives the confusion matrix table, and it is correct, (and there is something wrong later), we give 3+4=7 points)

If one table is incorrect without precedent, we do not give the points for the tables above(if the student only gives the confusion matrix table but it is incorrect, and there is an empty patient index table, we minus 3 points for that.)

Note: if the x axis of ROC curve is TPR instead of FPR (and the y axis of ROC curve is FPR instead of TPR), 2 points will be removed as a warning.

12. (5 points) Suppose you are using K -nearest neighbour model to predict test data based on $N = 50$ training data points. What is the value for K (if any) that your prediction is constant for any test data point regardless of the input features? Provide your answer below.

- Model treats all pts identically if $K=50$

↳ considers all data points

↳ in classification : most frequent class label

↳ in regression : average

Solution: When $K=50$, we are basically predict every test data point by the most frequent class label in classification or the average of the response values in regression computed using the entire training data.

Grading Rubrics:

$K=50$, 2 points

some fair and related descriptions, 3 points. (But must have some justifications)

13. (5 points) What are the two machine learning methods that you know of that can produce zero training error in any classification or regression task? Is there any assumption about the data that is required to guarantee 0 training error by such ML methods? Provide your answer and a brief explanation for each method and the data assumption below.

Solution: KNN and Decision Tree (DT). For KNN, when $K=1$, we are using each training data point itself to make prediction; For DT, when tree depth is large enough such that each tree node contains only one training example, the training error is zero.

To have 0 training error, data points of different target values or labels must not have identical feature values. This ensures that the 1-NN model always pick up the training data point itself and that DT can always split the data points down to homogeneous leaf nodes either including one

single data point per leaf or all data points having the same target label or value.

Grading Rubrics:

Some fair explanation on KNN K=1: 2 points.

Some fair explanation on Decision Tree: 2 points

Assumption of "different target values must not have identical feature values": 1 points. (Very few student got this.)

If the student mentions something else: delete 1 point (unless it is already 0 points)

If students mentioned large number of basis features in a linear regression as we covered in Module 4.1 give them **2.5 points** for that method. But they still another to receive full mark.

14. (5 points) Given $\hat{y} = \frac{1}{1 + \exp(-\sum_d x_d w_d)}$, compute the partial derivative of the following function with respect to the input x_d (i.e., $\frac{\partial L}{\partial x_d}$):

$$L = y \log \hat{y} + (1 - y) \log(1 - \hat{y}) \quad \rightarrow \text{check logistic regression gradient calculation}$$

Solution: For $L = y \log \hat{y} + (1 - y) \log(1 - \hat{y})$:

$$(y - \hat{y})w_d$$

For $L = -y \log \hat{y} - (1 - y) \log(1 - \hat{y})$:

$$(\hat{y} - y)w_d$$

Grading Rubrics:

Use your judgement...

Basically, if the answer is correct, full points.

If the answer is the minus of that answer, remove 1 points.

If they put in $(\hat{y} - y)x_d$, which is directly from the lecture notes. It shows that they didn't understand partial derivative but know where to look. Give them **3 points**.

If they have correct partial derivative on chain rule

$$\frac{\partial J(\mathbf{x})}{\partial x_d} = \frac{\partial J(\mathbf{w})}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial a} \frac{\partial a}{\partial x_d}$$

Give them 1.5 point for the correct derivative but partial marks if they got some of the partial correctly.

15. (10 points) Gini Index (GI) is computed as $GI = 1 - \sum_c \pi_c^2$, where π_c is the probability for class c . Compute the GI score on a test data point for K-nearest neighbour using Hamming distance $\sum_{d=1}^D \mathbb{I}(x_d^{(i)} \neq x_d^{(j)})$ and $K = 3$ using the following training data ($N = 4, D = 3$). Show your work.
Training data:

$$\mathbf{X} = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 1 & 0 \\ 1 & 1 & 0 \\ 0 & 1 & 1 \end{bmatrix}, \quad \mathbf{y} = \begin{bmatrix} 0 \\ 1 \\ 0 \\ 0 \end{bmatrix}$$

Testing data:

$$\mathbf{x}^* = [0 \ 0 \ 1], \quad y^* = [0]$$

Hamming distance \rightarrow counts nb of differing features

$[0, 0, 0] \rightarrow 1 \text{ diff} \quad \checkmark \Rightarrow 0$
 $[0, 1, 0] \rightarrow 2 \text{ diff} \quad \checkmark \Rightarrow 1$
 $[1, 1, 0] \rightarrow 3 \text{ diff}$
 $[0, 1, 1] \rightarrow 1 \text{ diff} \quad \checkmark \Rightarrow 0$
 $K=3$ nearest neighbors

Probability class 0 = $\frac{2}{3}$
 " " 1 = $\frac{1}{3}$

$$GI = 1 - \sum_c \pi_c^2$$

$$= 1 - \left(\left(\frac{2}{3}\right)^2 + \left(\frac{1}{3}\right)^2\right)$$

$$= 1 - \frac{5}{9} = \frac{4}{9}$$

Solution: First compute the Hamming distances between the training data point and the test data point. We have 1, 2, 3, 1. For $K=3$, we will choose data point 1, 2, and 4 with label 0, 1, 0, respectively. The class fractions for $y=0$ and $y=1$ are $2/3$ and $1/3$, respectively. $GI = 1 - 4/9 - 1/9 = 4/9$

Grading Rubrics:

Calculate distance: 3 points;

Choose the correct points: 3 points;

Have the correct class probabilities: 2 points.

Calculate the correct GI: 2 points.

16. (10 points) What's the maximum likelihood estimate for the parameter w using the following cost function and training data $\mathcal{D} = \{(0, 1), (3, 2), (1, 3)\}$, where each tuple contains input and response (i.e., $(x^{(n)}, y^{(n)})$), respectively. Write down your derivations and provide your estimated value.

$$J(w) = \frac{1}{2} \sum_n (y^{(n)} - wx^{(n)})^2$$

Solution: $w^* = \operatorname{argmin}_w \frac{1}{2} \sum_n (y^{(n)} - wx^{(n)})^2$. We solve this by setting the derivative to zero:

$$\frac{\partial J}{\partial w} = \sum_n -x^{(n)}(y^{(n)} - wx^{(n)}) = 0 \rightarrow \sum_n x^{(n)}y^{(n)} = \sum_n wx^{(n)}x^{(n)} \rightarrow w = \frac{\sum_n y^{(n)}x^{(n)}}{\sum_n x^{(n)}x^{(n)}} = \frac{9}{10} = 0.9$$

Grading Rubrics:

Derivatives: ($\frac{\partial J}{\partial w} = \sum_n -x^{(n)}(y^{(n)} - wx^{(n)}) = 0$): 5 points

Correct formula for w (and some fair derivations): ($\frac{\sum_n y^{(n)}x^{(n)}}{\sum_n x^{(n)}x^{(n)}}$), 4 points.

Correct answer: 1 point.

(One number itself does not worth 10 points; it only worth 1 point.)

Note: please also give fair points to some other works.

$$J(w) = \frac{1}{2} \sum_n (y^{(n)} - wx^{(n)})^2$$

gradient \rightarrow

$$\begin{aligned} \frac{dJ}{dw} &= \frac{d}{dw} \left(\frac{1}{2} \sum_n (y^{(n)} - wx^{(n)})^2 \right) \\ &= \frac{1}{2} \sum_n \frac{d}{dw} (y^{(n)} - wx^{(n)})^2 \\ &= \frac{1}{2} \sum_n 2(y^{(n)} - wx^{(n)}) \cdot \frac{d}{dw} (y^{(n)} - wx^{(n)}) \\ &= \sum_n (y^{(n)} - wx^{(n)}) \cdot -x^{(n)} \end{aligned}$$

*Set to $\frac{dJ}{dw} = 0$ for optimal w

$$\Rightarrow \sum_n (y^{(n)} - wx^{(n)}) \cdot -x^{(n)} = 0$$

$$\sum_n -x^{(n)}y^{(n)} + \sum_n wx^{(n)}x^{(n)} = 0$$

$$w = \frac{\sum_n x^{(n)}y^{(n)}}{\sum_n x^{(n)}x^{(n)}} = \frac{(0 \cdot 1) + (3 \cdot 2) + (1 \cdot 3)}{0^2 + 3^2 + 1^2} = \frac{9}{10} = 0.9$$

$$D = \{(0, 1), (3, 2), (1, 3)\}$$