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

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

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

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

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

Solution: Higher recall= $\frac{TP}{TP+FN}$ means catching more spam emails (higher true positive TP). This result 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

9 folds for training

$3 \times 9 = 27$

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

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

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

$\exp(-w_0) = 9$

$w_0 = -\ln 9$

6. (5 points) For $C = 4$ classes and $D = 5$ features, what is the predicted class for input $\mathbf{x} = [1\,0\,0\,1\,0]$, when using a multiclass regression model with the following weights: $\mathbf{W} = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 1 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ 0 & 1 & 1 & 1 \end{bmatrix}$

$y = xW$

$=$

$[1\,0\,0\,1\,0]$ 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:

- $||y - \hat{y}||_2^2$
- $\sum_{c=1}^{C} -y_c \log \hat{y}_c$
- $-y \log \hat{y} - (1 - y) \log(1 - \hat{y})$
- $\sum_{c=1}^{C} \mathbb{I}[\hat{y}_c \neq y_c]$

- log Bernoulli
- log Binomial
- log Categorical
- 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

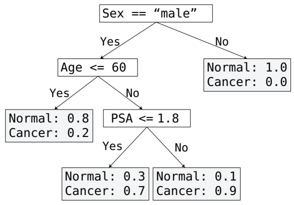
   C. Overfitting can be detected using a validation set ← detect overfitting by comparing training and validation error
   D. Overfitting occurs when the learning rate is too large
   E. Overfitting can occur when the K value for the KNN is too large
   F. Overfitting can occur when the tree depth is too 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.

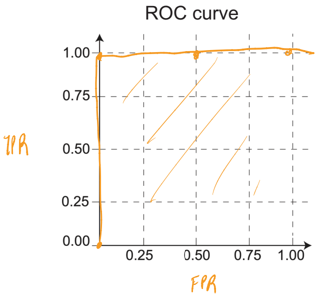


Patient ID	Age	PSA	Sex	Status
1	55	1.8	male	Normal
2	76	1.7	male	Cancer
3	56	0	female	Normal

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

	1	0.6	0.1	-0.1
TPR	0	1	1	1
FPR	0	0	0.5	1



$TPR = \frac{TP}{TP+FN} =$

$FPR = \frac{FP}{FP+TN}$

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
AN	2	0	2	0	1	1	0	2
AP	1	0	0	1	0	1	0	1

	1	0.6	0.1	-0.1
TPR	0	1	1	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
  ↳ consider 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 justificaitons)

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., $\frac{\partial L}{\partial x_d}$):

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

→ 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 int $(\hat{y} - y)x_d$, 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:

$$\mathbf{x}^* = \begin{bmatrix} 0 & 0 & 1 \end{bmatrix}, \quad y^* = \begin{bmatrix} 0 \end{bmatrix}$$

Hamming distance → counts nb of differing features

[0, 0, 0] → 1 diff ✓ ⇒ 0 K=3 nearest neighbors

[0, 1, 0] → 2 diff ✓ ⇒ 1

[1, 1, 0] → 3 diff

[0, 1, 1] → 1 diff ✓ ⇒ 0

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

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

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 ↗

$\frac{dJ}{dw} = \frac{d}{dw}\left(\frac{1}{2} \sum_n (y^{(n)} - wx^{(n)})^2 \right)$

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

$= \frac{1}{2} \sum_n \cdot 2(y^{(n)} - wx^{(n)}) \cdot \frac{d}{dw} (y^{(n)} - wx^{(n)})$

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

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

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

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

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

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

COMP 551 Applied Machine Learning Midterm Exam

Feb 19, 2024, 2:35 pm - 3:35 pm

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 12 pages.
- 6 multiple-choice questions: circle only ONE correct answer per question.
- 4 multiple-select questions, circle ALL correct answers per question. Scoring: Right minus Wrong.
- 6 short-answer questions: write your answer directly below each question.
- Advice: Try not to spend too much time searching answers through your notes as it will slow you down and you will not have enough time to complete this exam in 1 hour. Good luck!

1 Multiple-choice questions (30 points)

1. (5 points) Suppose we know that *apriori* majority of the input features are irrelevant to the target label. Which method will likely perform the worst when using all features for prediction?

   A. KNN ⟶ treats all features equally

   B. Decision tree

   C. Logistic regression

Solution: KNN will perform the worst as the distance function takes into account all features. DT and LR will perform relatively well as they have internal feature selection.

2. (5 points) After training a binary classifier that can produce probability for the positive class, what threshold guarantees to produce 100% *Recall Rate* on the test data? Note we set the predicted class to 1 if the model predicted probability is *greater* than the threshold.

   A. -1

   B. 0.01

   C. 0.5

   D. 1

   $TPR = \frac{TP}{TP+FN}$

   at -1, everything positive

E. 2

Solution: For Recall or TPR=TP/(TP+FN), we can have TPR equal to 1 when the threshold is -1. That is, every data point is predicted to be positive and have zero false negative.

3. (5 points) Suppose you have a hate-speech detection model to detect hate-speech in online comments (hate-speech = 1, normal = 0). Training was successful and you have a pretty good model which performs much better than random but is less than perfect. You can control the threshold parameter α such that if you label a comment as hate-speech if $p(\text{hate-speech}|\text{comment}) > \alpha$. If you increase α, what happens to the model's precision and recall? Select one for each of precision and recall?

   A. recall decreases; precision decreases

   B. recall decreases; precision increases

   C. recall increases; precision decreases

   D. recall increases; precision increases

Solution: Increasing α increases the number of negatives, both TN and FN, and decreases the number of positives, both TP and FP, if the model does better than chance FP should be reduces by a higher proportion than TP. Recall$= \frac{TP}{TP+FN}$ will thus decrease. And precision $= \frac{TP}{TP+FP}$ will increase. 2.5 pts for each.

4. (5 points) What loss does the estimate $\hat{\mathbf{w}} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}$ minimize?

   A. $J(\mathbf{w}) = \sqrt{\sum_n (y^{(n)} - \hat{y}^{(n)})^2}$, where $\hat{y}^{(n)} = \mathbf{x}^{(n)}\mathbf{w}$

   B. $J(\mathbf{w}) = \sum_n |y^{(n)} - \hat{y}^{(n)}|$, where $\hat{y}^{(n)} = \mathbf{x}^{(n)}\mathbf{w}$

   C. $J(\mathbf{w}) = \sum_n (y^{(n)} - \hat{y}^{(n)})^2$, where $\hat{y}^{(n)} = \mathbf{x}^{(n)}\mathbf{w}$

   D. $J(\mathbf{w}) = \sum_n -y^{(n)} \log \hat{y}^{(n)} - (1 - y^{(n)}) \log(1 - \hat{y}^{(n)})$, where $\hat{y}^{(n)} = \frac{1}{1+\exp(-\mathbf{x}^{(n)}\mathbf{w})}$

5. (5 points) Which of the following methods can only be trained using gradient decent?

   A. Decision tree

   B. Linear regression

   C. Linear regression with basis transformed features

   D. Logistic regression

Solution: DT are not trained by GD. Linear regression and linear regression with basis-transformed feature can be fit by analytical solution. Only Logistic regression can only be trained with GD.

6. (5 points) For $C = 3$ classes and $D = 2$ features, what is the class probabilities for input $\mathbf{x} = [1\ 0]$, when using a multiclass regression with the following weights (assuming natural log ln):

$$\mathbf{W} = \begin{bmatrix} 0 & \ln 3 & 0 \\ \ln 3 & 0 & \ln 4 \end{bmatrix}$$

   A. [1/3, 1/3, 1/3]

B. [0.3, 0.3, 0.4]
C. [0.6, 0.2, 0.2]
D. [0.2, 0.6, 0.2]
E. [0, 1, 0]

Solution:

$$\mathbf{a} = \mathbf{x}\mathbf{W} = \begin{bmatrix}1 & 0\end{bmatrix}\begin{bmatrix}0 & \ln 3 & 0\\ \ln 3 & 0 & \ln 4\end{bmatrix} = \begin{bmatrix}0 & \ln 3 & 0\end{bmatrix}$$

$$\hat{\mathbf{y}} = \begin{bmatrix}\frac{\exp(0)}{\exp(0)+\exp(\ln 3)+\exp(0)} & \frac{\exp(\ln 3)}{\exp(0)+\exp(\ln 3)+\exp(0)} & \frac{\exp(0)}{\exp(0)+\exp(\ln 3)+\exp(0)}\end{bmatrix} = \begin{bmatrix}0.2 & 0.6 & 0.2\end{bmatrix}$$

2 Multiple-select questions (20 points)

7. (5 points) In binary classification, what method(s) below is or are equivalent to simply taking the positive fraction in the training data as the predicted value for all test data points?

A. Decision tree with only the root node
B. K-nearest neighbours with K set to be 1
C. K-nearest neighbours with K set to be the number of training examples
D. Maximum likelihood estimate of the Bernoulli rate over the N binary labels from the training data.

Solution: DT and KNN with K=N predicts based on the average of target labels. In the case of binary classification, it is $\hat{y} = \frac{1}{N}\sum_n y^{(n)}$ (i.e., positive fraction). The MLE of Bernoulli rate is $\pi = \frac{N_1}{N}$, where N_1 is the number of positive examples.

8. (5 points) What model(s) are suitable to predict the monthly grocery cost of average Canadian household using economic factors as input features?

A. K nearest neighbours
B. Decision tree
C. Linear regression
D. Logistic regression
E. Multiclass regression

|| ∅ continuous

Solution: KNN, DT, and linear regression can predict real value response but not logistic or multiclass regression.

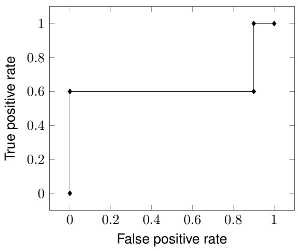
9. (5 points) Choose ALL action(s) below that may cause overfitting

A. Increasing K in KNN
B. Increasing tree depth in decision tree
C. Training logistic regression on more input features
D. Training logistic regression on more training examples

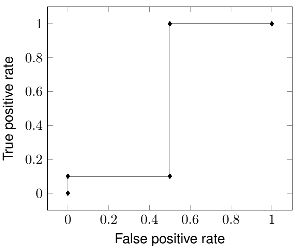
Solution: Since KNN takes the average over K neighbours, Increasing K will not introduce overfitting but rather decrease K will. Increasing tree depth in DT is prone to overfitting as it tries to separate the fine-grained training data points. So as increasing features.

10. (5 points) You evaluated two ML methods on a test dataset with balanced positive and negative examples and obtained the following ROC curves. Reminder: TPR = recall = TP/(TP+FN); FPR = FP/(TN+FP); precision = TP/(TP+FP)

Model 1



Model 2



Circle ALL correct statement(s) below.

A. At the 2nd point of Model 1, we have a perfect precision.
B. Model 1 does not have a perfect recall at any threshold.
C. At the 4th point of Model 2 we have perfect recall.
D. Model 1 has higher AUROC than Model 2.

Solution:

1. Choosing the 4th point of Model 2 with Recall=TPR=1 and FPR=0.5 , we have perfect recall with the lowest possible FPR of 0.5.
2. Model 1 reaches perfect recall at FPR = 0.9.
3. Choosing the 2nd point of Model 1 we have FPR=0 meaning that FP=0, and so precision = TP/(TP+FP)=1 so we have a perfect precision with the highest possible TPR=0.6.
4. AUROC$_1$ = 0.64, AUROC$_2$ = 0.45 so Model 1 has better AUROC.

Solution: Rubrics: The points are calculated based on how many correct and incorrect answers you have chosen, e.g., $points = \frac{5}{correct\_answer\_in\_total} * (correct\_answer\_you\_selected - incorrect\_answer\_you\_selected)$.

3 Short answer questions (50 points)

11. (10 points) Given the simple logistic regression:

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

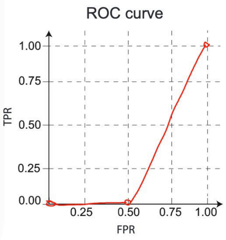
where the fitted value $\hat{w} = -\log 4$. Given 3 test data points with input features $\mathbf{x} = \begin{bmatrix}1\\0\\1\end{bmatrix}$ and true label $\mathbf{y} = \begin{bmatrix}0\\0\\1\end{bmatrix}$. Using thresholds $t \in \{0, 0.25, 1\}$, draw the ROC curve. You may use the tables below to help 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.

	$\hat{y}^{(n)}$	$\hat{y} > t$	y
1			
2			
3			

	PN	PP	PN	PP	PN	PP
AN						
AP						

t	0	0.25	1
TPR			
FPR			

ROC curve



Solution:

	$\hat{y}^{(n)}$	$\hat{y} > t$			y
1	0.2	1	0	0	0
2	0.5	1	1	0	0
3	0.2	1	0	0	1

	PN	PP	PN	PP	PN	PP
AN	0	2	1	1	2	0
AP	0	1	1	0	1	0

t	0	0.25	1
TPR	1	0	0
FPR	1	0.5	0

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

Grading Rubrics:

If the ROC graph is correct, full mark;
If not:
Correct "Prediction table": 2 points

Correct "Confusion matrix": 4 points
Correct "ROC points": 2 points
Correct ROC graph: 2 points

Some students used log with base other than natural log for $\hat{w} = \log 4$. This is not what i intended as it requires a calculator. But if they somehow computed the correct ROC using that log, we should still give them full mark.

12. (5 points) For a small dataset say $N = 50$, how would you obtain a stable estimate of the performance of your ML method? Describe and justify the key technique that you will use.

leave - one - out cross - validation

Solution: We will use leave-one-out cross-validation (LOOCV) with each fold containing one data point. This way the model will train on $N-1$ data points and validated on the held-out data points for N times to obtain a stable estimate of the performance. 4 points when students only mention cross-validation. No points on other solutions.

13. (10 points) The following dataset contains 4 examples with 1 feature x and a continuous target variable y.

id	x	y
1	2	3
2	-1	-2
3	-3	-2
4	0	0

You decide to use 4-fold cross-validation (CV) with 1 data point for the validation set (i.e., leave-one-out CV) and 3 for training to determine whether you should use K=1 or K=2 in K-Nearest Neighbours. What is the cross-validation sum of squared error (SSE) loss for K=1? What is the cross-validation SSE for K=2? Should you use K=1 or K=2? Show your work.

Solution:

- 1's NN: 4, 2
- 2's NN: 4, 3
- 3's NN: 2, 4
- 4's NN: 2, 1

1 is validation set: $K = 1, y = 0 \Rightarrow e = 3^2$; $K = 2, y = -1 \Rightarrow e = 4^2$

2 is validation set: $K = 1, y = 0 \Rightarrow e = 2^2$; $K = 2, y = -1 \Rightarrow e = 1^2$

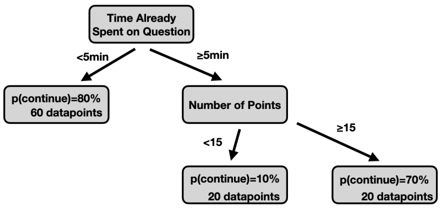
3 is validation set: $K = 1, y = -2 \Rightarrow e = 0^2$; $K = 2, y = -1 \Rightarrow e = 1^2$

4 is validation set: $K = 1, y = -2 \Rightarrow e = 2^2$; $K = 2, y = 0.5 \Rightarrow e = 0.5^2$

So the average of cross-validation loss for K=1 is $(9+4+0+4)/4 = 17/4$ and for K=2: $(16+1+1+0.25)/4 = 18.25/4$ So we should choose K=1.

2 points for validation set predictions being ok and. 2 points for calculating validation losses. 2 points for splitting into 4 validation sets. 2 points for averaging (or summing) validation losses. 2 points for choosing the lowest avg loss for K. -1pt for every minor calculation mistake.

14. (10 points) You are curious about what strategy your fellow students use during midterm exams. Specifically, how do they decide whether to continue working on a hard question or skip it for the moment and pass to the next question. You collected 100 data points from a survey. After training a Decision Tree (DT) on this data, you get the following model:


Time Already Spent on Question
<5min
≥5min
p(continue)=80% 60 datapoints
Number of Points
<15
≥15
p(continue)=10% 20 datapoints
p(continue)=70% 20 datapoints

Calculate the expected Gini Index (GI) of the above DT. Recall $GI = 1 - \sum_c \pi_c^2$ where π_c is the probability for class $c \in \{\text{continue}, \text{skip}\}$. You do not need to obtain the final value. For example, you can leave the your answer as $0.5 \times (1 - 0.9^2)$ (which is obviously not the solution for this question).

Solution:

$$GI = 0.6 * (1 - 0.8^2 - 0.2^2) + 0.2 * (1 - 0.9^2 - 0.1^2) + 0.2 * (1 - 0.7^2 - 0.3^2) = 0.312$$

5 points for taking the expectation properly. 5 points for applying the appropriate expression correctly inside the expectation. Or 2-3 points if students only try to use $GI = 1 - \sum_c \pi_c^2$ but not use it correctly.

15. (5 points) You have fit a multiple regression and obtained the regression coefficients $\hat{\mathbf{W}}$ on D features and N training examples, which took up a lot of compute time. However, someone gives you a new feature $\mathbf{x}_{D+1}$ recorded over the same N training examples. Derive the analytical ordinary least square (OLS) solution for only coefficient w_{D+1} *conditioned on* $\hat{\mathbf{y}} = \mathbf{X}\hat{\mathbf{W}}$ without refitting the previous D coefficients. Computing w_{D+1} should only takes $O(N)$ time.

Solution:

$$\begin{aligned} \Delta\mathbf{y} &= \mathbf{y} - \hat{\mathbf{y}} \\ L &= ||\Delta\mathbf{y} - \mathbf{x}_{D+1} w_{D+1}||_2^2 \\ \frac{\partial L}{\partial w_{D+1}} &\overset{set}{=} 0 \quad \implies \quad \hat{w}_{D+1} = (\mathbf{x}_{D+1}^\top \mathbf{x}_{D+1})^{-1} \mathbf{x}_{D+1}^\top \Delta y \end{aligned}$$

Grading Rubrics:
3 points for setting up the problem correctly (finding the correct loss function for w_{D+1}. 2 extra points for finding correct solution (i.e. minimizing the loss function correctly).

16. (10 points) Suppose we use a basis function $z = \frac{1}{1+\exp(-x\beta)}$ to transform the linear feature $x \in \mathbb{R}$ onto $z \in [0, 1]$ before performing linear regression. Compute the partial derivative of the squared loss $L = (y - zw)^2$ for one data point with respect to β.

Solution:
Let $a = x\beta$ and $z = \frac{1}{1+\exp(-a)}$

$$\begin{aligned} \frac{\partial L}{\partial \beta} &= \frac{\partial L}{\partial z}\frac{\partial z}{\partial a}\frac{\partial a}{\partial \beta} \\ \frac{\partial L}{\partial z} &= -2(y - zw)w \\ \frac{\partial z}{\partial a} &= z(1 - z) \\ \frac{\partial a}{\partial \beta} &= x \end{aligned}$$

Therefore,

$$\frac{\partial L}{\partial \beta} = -2(y - zw)wz(1 - z)x$$

Grading Rubrics:
2 points on each step, but if students write in other ways that also make sense or get the derivative directly and correctly, they will get full marks.