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

11 & 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

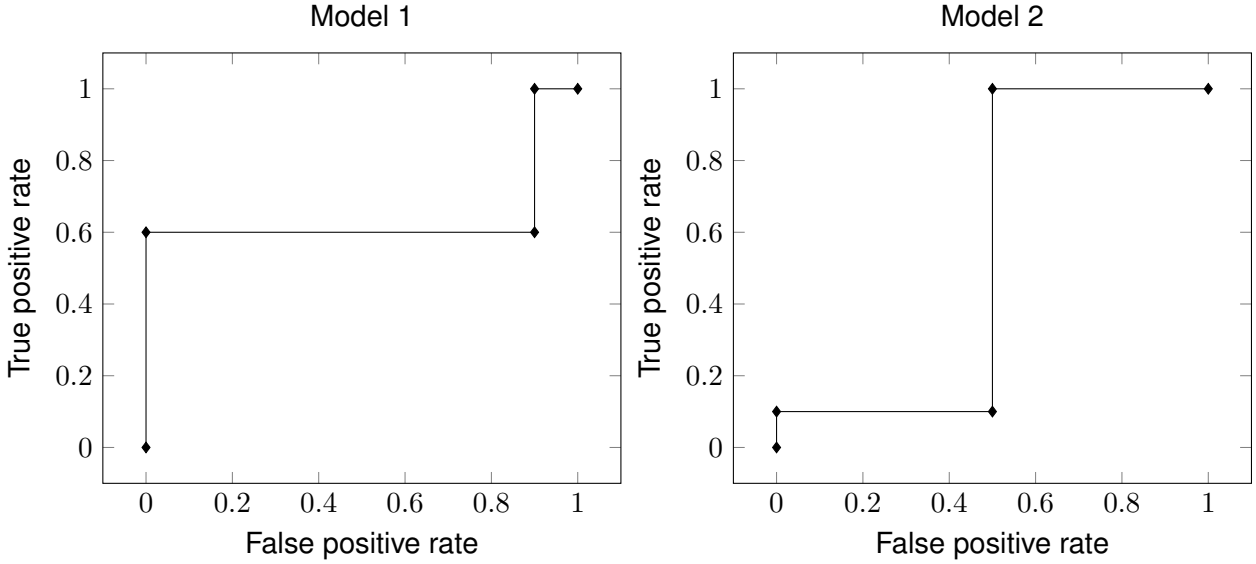
B. Increasing tree depth in decision tree

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)



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:

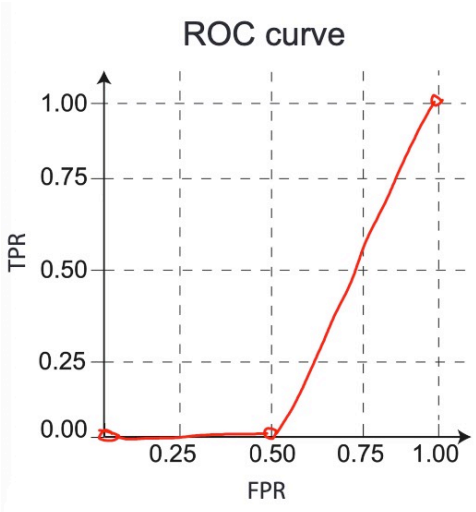
$$\hat{y}^{(n)} = \frac{1}{1 + \exp(-x^{(n)}\hat{w})}$$

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			



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

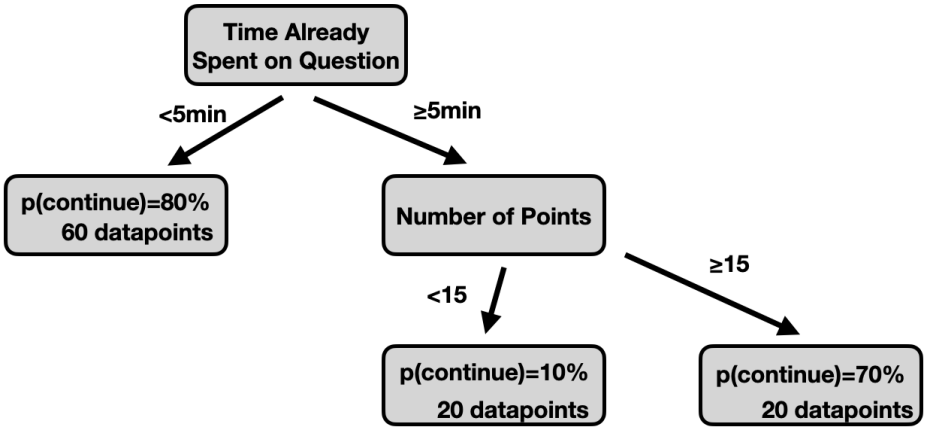
3 is validation set: $K = 1, y = -2 \Rightarrow e = 0^2$; $K = 2, y = -1 \Rightarrow e = 1^2$

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:



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:

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

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.