University of Toronto at Scarborough
Department of Computer and Mathematical Sciences

CSCC11: Introduction to Machine Learning

Practice Midterm Test
Winter 2024

Duration: 60 minutes
No aids allowed

There are 8 pages total (including this page)

Last name: _____

First name: _____

Student number: _____

Signature: _____

Circle your tutorial: Wed 2-3 Wed 3-4 Thurs 4-5 Fri 11-12

Question	Marks	
1	_____	/ 16 marks
2	_____	/ 16 marks
3	_____	/ 14 marks
4	_____	/ 9 marks
Total	_____	/ 55 marks

1. **Multiple Choice and Short Answer Questions [16 marks]**

(a) *[2 marks]* For LS regression, circle those cases below that you expect to occur as your training set size increases and briefly explain why:

(a) training error increases and test error increases

(b) training error increases and test error decreases

(c) training error decreases and test error increases

(d) training error decreases and test error decreases

(b) *[2 marks]* Of the mathematical expressions below, circle (only) those that are true (assume that all vectors are d-dimensional column vectors and all matrices have size $d \times d$).

$\mathbf{y}^T A \mathbf{x}^T$ is symmetric $\quad \frac{\partial}{\partial \mathbf{w}} \sum_i ||\mathbf{x}_i^T \mathbf{w}||^2 = \sum_i \mathbf{x}_i \mathbf{x}_i^T \mathbf{w} \quad \mathbf{a}^T \mathbf{b} = \mathbf{b}^T \mathbf{a} \quad \frac{\partial Q\, Q^T \mathbf{x}}{\partial \mathbf{x}} = Q\, Q^T$

(c) *[3 marks]* Suppose you fit linear, cubic, quartic, and quintic polynomials to some data. Circle the one most likely to overfit the data (assuming one or more do overfit)?

linear quadratic cubic quintic

Circle the one most likely to underfit the data (assuming one or more do underfit)?

linear quadratic cubic quintic

Circle the one you expect to have the smallest squared error on the training data.

linear quadratic cubic quintic

(c) *[2 marks]* The decision boundaries for Gaussian class conditional models with two classes have what mathematical form (what degree of polynomial)?

(d) *[2 marks]* Briefly explain what is the difference between Maximum Likelihood (ML) and Maximum a Posteriori (MAP) parameter estimation.

(e) *[3 marks]* How do you know the critical point of the least squares objective (i.e., where all elements of the gradient of the objective function are zero) corresponds to a minima and not a maxima?

(f) *[2 marks]* True or False: The risk of over-fitting is often be mitigated by obtaining more training data, and explain why.

Circle one of **T** or **F** and then explain why below.

2. **Decision Trees [16 marks]**

Suppose you want to predict whether a mushroom is edible. You are given training data with measurements of height (*tall* or *short*), colour (*white* or *brown*), and weight (*light* or *heavy*). Use the approach to decision tree learning outlined in class, adapted for the fact that here our measurements are categorical. In this case, all splits will test for equality with one of the features (e.g., colour = white), sending data to the left when true, and otherwise to the right.

Height	Colour	Weight	Edible
short	brown	light	**N**
short	brown	light	**N**
tall	white	light	**N**
tall	brown	light	**N**
short	white	heavy	**N**
short	white	heavy	**N**
tall	white	heavy	**Y**
tall	brown	heavy	**Y**
short	brown	heavy	**Y**
short	brown	heavy	**Y**

(a) *[2 marks]* Entropy is used to help select split functions. Define entropy for a random variable having K possible values (or outcomes), with probabilities P_j, for $j = 1...K$.

(b) *[3 marks]* Let $\mathcal{D}_j$ be the data set that arrives at node j. Let $t_j(\mathbf{x})$ be the split function at node j, where $\mathbf{x}$ is a feature vector. The split function returns -1 or 1. Define information gain (and any other notation that you might need) for $t_j(\mathbf{x})$.

(c) *[3 marks]* To begin building the tree, find the entropy of the target label, ie Edible. (You don't need to find the numerical value without a calculator, but rather give the expression to compute the value.)

(d) *[3 marks]* Which attribute would the decision tree building algorithm choose to use for the root and why (i.e., what would be the first split function)? You can determine this by eye-balling the table, without computing information gain for all possible splits.

(e) *[2 marks]* What would be the information gain be for the split function you chose for the root (in the last question)? Give a mathematical expression rather than the numerical value.

(f) *[3 marks]* Draw the full decision tree that would be learned for this data. (Hint: The tree should not have more than three or four splits. You should be able to choose good splits without having to compute the information gain for each possible split.)

3. **Linear Regression [14 marks]** Suppose you have a linear regression problem. The training dataset has N data points $\{(x_i, y_i)\}_{i=1}^N$, where both x and y are real-valued. Let $\mathbf{w}^* = (w_0^*, w_1^*)^T$ be the least-squares solution (assumed to be unique). In other words, it minimizes

$$E(\mathbf{w}) = \frac{1}{N}\sum_{i=1}^{N}(y_i - (w_0 + w_1 x_i))^2$$

Circle all statements below that must be true if $\mathbf{w}^*$ is indeed the LS solution

$$\frac{1}{N}\sum_i (y_i - w_0^* - w_1^* x_i) y_i = 0$$

$$\frac{1}{N}\sum_i (y_i - w_0^* - w_1^* x_i)(y_i - \bar{y}) = 0$$

$$\frac{1}{N}\sum_i (y_i - w_0^* - w_1^* x_i)(x_i - \bar{x}) = 0$$

$$\frac{1}{N}\sum_i (y_i - w_0^* - w_1^* x_i)(w_0^* + w_1^* x_i) = 0$$

where $\bar{x}$ and $\bar{y}$ are the sample means from the same dataset. **Explain** your answer below by deriving the solution to the optimization problem and showing how the necessary conditions you circled follow from your solution.

4. **Probability [9 marks]** Rob and Alice are alternately and independently flipping a coin. The first player to get a head wins. Alice flips the coin first.

   (a) *[2 marks]* If $P(\text{head}) = 1$ what is the probability that Alice wins, and why?

   (b) *[7 marks]* If $P(\text{head}) = 0.5$ what is the probability that Alice wins? (hint: Try to enumerate the different settings under which Alice can win. Note that for $0 \leq a \leq 1$, it is given that $\sum_{i=0}^{\infty} a^i = \frac{1}{1-a}$.)

This page for rough work.