Name:

Roll No.: **Dept.**:

IIT Kanpur
CS771A (IML)
Quiz-1
Date: August 23, 2023

Instructions: *Total:* **30 marks**

1. Please write your name, roll number, department on **all pages** of this question paper.
2. Write your answers clearly in the provided box. Keep your answer precise and concise.

Section 1 (15 very short answer questions: 15 x 2 = 30 marks). .

1. For the Learning with Prototypes (LwP) model with Euclidean distances, suppose the means of the positive and negative class are denoted by $\boldsymbol{\mu}_+$ and $\boldsymbol{\mu}_-$, respectively. A new test input $\boldsymbol{x}_*$ is classified as positive if (write down the expression):

$$||\mu_- - \boldsymbol{x}_n||^2 - ||\mu_+ - \boldsymbol{x}_n||^2 > 0$$

2. Write down the expressions for the loss functions for least squares linear regression and ridge regression. Assume training data to be denoted as $\{\boldsymbol{x}_n, y_n\}_{n=1}^N$ and the weight vector to be denoted as $\boldsymbol{w}$.

$$\text{Least Squares:} \quad L(\boldsymbol{w}) = \sum_{n=1}^{N}(y_n - \boldsymbol{w}^\top \boldsymbol{x}_n)^2$$

$$\text{Ridge:} \quad L(\boldsymbol{w}) = \sum_{n=1}^{N}(y_n - \boldsymbol{w}^\top \boldsymbol{x}_n)^2 + \lambda \boldsymbol{w}^\top \boldsymbol{w}$$

3. Briefly explain (in at most 1-2 sentences) overfitting in terms of the training and test error of a machine learning model.

   A machine learning model is likely to have overfit if its traning error is very small but the test/validation error is very large, i.e., the gap between training and test errors is quite large.

4. Consider a linear regression model $y_n = wx_n + b$ with scalar inputs, scalar weight w, and assume the bias term b as well. Also assume that the inputs x_n have already been centered (i.e., by subtracting off their mean from the original inputs) so their mean is 0. Assuming you are given N input-response pairs $\{x_n, y_n\}_{n=1}^N$, write down the expression for the bias term b (please do show the basic steps).

   Summing both side over all training examples, we have $\sum_{n=1}^N y_n = \sum_{n=1}^N (wx_n + b)$. Since the inputs are already centered, their mean will be zero, so $\sum_{n=1}^N x_n = 0$, and thus $\sum_{n=1}^N y_n = w \times 0 + Nb$. Therefore $b = \frac{1}{N}\sum_{n=1}^N y_n$, i.e., the bias in this case is simply the mean of the responses.

5. Consider two simple decision stumps D_1 and D_2. The first leaf of D_1 has 200 positive and 0 negative examples, and its second leaf has 0 positive and 200 negatives. The first leaf of D_2 has 100 positives and 100 negatives, and its second leaf also has 100 positives and 100 negatives. Which is these two decision stumps has a higher information gain and why? Explain only using words in at most 1-2 sentences.

   For D_1, both its leaf nodes are completely pure, whereas for D_2, the leaf nodes are not pure (in fact, both classes are present in equal measure for both the leaves). Clearly, D_1 split was much better and, even without computing, we can say it has a higher information gain.

6. For a multi-class classification model (e.g., softmax classification), given the learned weight vectors $\boldsymbol{w}_1, \boldsymbol{w}_2, \ldots,$ and a test input $\boldsymbol{x}_*$, how would you predict the class $\boldsymbol{x}_*$ belongs to? How would you predict the probability of $\boldsymbol{x}_*$ belonging to any class $i \in \{1, 2, \ldots, K\}$?

   The predicted class will be the one whose weight vector results in the largest inner product with $\boldsymbol{x}_*$, i.e., $\hat{y}_* = \arg\max_{i \in \{1,2,\ldots,K\}} \boldsymbol{w}_i^\top \boldsymbol{x}_*$. To calculate the probability, we can convert the logits using the softmax function, e.g., $p(y_* = i|\mathbf{W}, \boldsymbol{x}_*) = \frac{\exp(\boldsymbol{w}_i^\top \boldsymbol{x}_*)}{\sum_{j=1}^K \exp(\boldsymbol{w}_j^\top \boldsymbol{x}_*)}$.

Name:

Roll No.: **Dept.**:

IIT Kanpur
CS771A (IML)
Quiz-1
Date: August 23, 2023

7. What is the advantage of using gradient descent to learn the weight vector of a linear/ridge regression model instead of learning it using the closed form solution?

   When using gradient descent for linear/ridge regression, we do not need to perform any expensive matrix inversion. In contrast, when using the closed form solution for these problems, we need to invert a $D \times D$ matrix where D is the number of features.

8. The cross entropy loss on an example $(\boldsymbol{x}_n, y_n)$ where $y_n \in \{0, 1\}$ is defined as $\ell(y_n, \boldsymbol{x}_n, \boldsymbol{w}) = -[y_n \log \mu_n + (1 - y_n) \log(1 - \mu_n)]$ where $\mu_n = \sigma(\boldsymbol{w}^\top \boldsymbol{x}_n)$ is the model's predicted probability of the label being 1. Briefly explain why this loss makes sense for a binary classification problem?

   The loss makes sense because when y_n is 1 and μ_n is too small (which means a bad prediction), the loss will be very large. Likewise, when y_n is 0 and μ_n is very large, e.g., close to 1 (which means a bad prediction), the loss will again be very large.

9. Consider learning a decision tree, given some training data, where each input has D binary-valued features. Let's assume that we will not test any feature that has been tested at one of the previous levels (but we can possibly test a feature at multiple nodes at the same level). How many information gain calculations would be needed to construct the full decision tree (i.e., assuming no pruning)? Just give the basic expression; no need to try simplifying it too much to get a more compact expression.

   D information gain (IG) calculations will be needed at level-1, i.e.. the root node (one IG calculation for each feature). At level-2, we have 2 nodes (since the features are binary, no matter which features gets selected at the root node, we will have two outgoing branches and thus 2 (internal) nodes at level-2. For each of these 2 nodes, we have to do $D-1$ IG calculations for (since we have $D-1$ features left to be tested at level-2), resulting in $2 \times (D-1)$ IG calculations at level-2. Proceeding in a similar manner, at level-3, we will have 4 nodes and $D-2$ IG calculations for each (since we have $D-2$ features left to be tested at level-3), resulting in $2^2 \times (D-2)$ IG calculations at level-3. The same idea applies at other levels as well. The total number of IG calculations will therefore be $D + 2 \times (D-1) + 2^2 \times (D-2) + 2^3 \times (D-3) + \ldots$.

10. In what situation, an ϵ-ball nearest neighbors method may fail to make a prediction at test time?

    It might happen that, for the chosen value of ϵ, there is no neighbor within the ϵ-ball.

11. Briefly explain the difference between multi-class classification and multi-label classification. You may use an example to explain.

    In multi-class classification, given an input (say an image), the task is to predict a single exclusive label from a set of K labels. In multi-label classification, given an input, the task is to predict all relevant labels from a set of K labels.

12. Rank (fastest to slowest) the following methods in terms of their prediction speed (i.e., how long they take to predict the label of a new test input) for a binary classification problem: LwP, K-nearest neighbors, decision tree, logistic regression. If some of these take roughly equal time, you may say so.

    (1) Decision Tree, (2 and 3) LwP and logistic regression (both take roughly equal time), (4) K-nearest neighbors

13. Can we use a linear regression model to learn a nonlinear regression function? Briefly explain your answer.

    Yes, if we transform the original features $\boldsymbol{x}$ through some fixed/pre-defined transformation (e.g., a polynomial transformation or kernel methods) or learned transformation (e.g., deep learning) then it is possible. If we denote the inputs after such a transformation as $\phi(\boldsymbol{x})$, then a linear model $\boldsymbol{w}^\top \phi(\boldsymbol{x})$ on these transformed features can also learn a nonlinear mapping between $\boldsymbol{x}$ and y

Name:

Roll No.: **Dept.**:

IIT Kanpur
CS771A (IML)
Quiz-1
Date: August 23, 2023

14. **(MCQ)** Which of these classifiers learns a linear separator (select all options that you think are correct)? (1) LwP when using Euclidean distance, (2) K-nearest neighbors classifier for K=1 and any distance function, (3) K-nearest neighbors classifier for any value of K and Euclidean distance, (4) logistic regression.

    Correct options: 1 and 4

15. **(MCQ)** Which of these can be used for regression (select all options that you think are correct)? (1) Decision Tree, (2) K-nearest neighbor, (3) Learning with Prototypes, (4) Logistic Regression.

    Correct options: 1 and 2