Introduction to ML (CS771), 2024-2025-Sem-I		Total Marks	25
Quiz 2. September 7, 2024		Duration	45 minutes
Name		Roll No.	

Instructions:

1.	Clearly write your name (in block letters) and roll number in the provided boxes above.
2.	Write your final answers concisely in the provided space. You may use blue/black pen.
3.	We won't be able to provide clarifications during the quiz. If any aspect of some question appears ambiguous/unclear to you, please state your assumption(s) and answer accordingly.

Question 1: Write **T** or **F** for True/False in the box next to each question given below, with a brief (1-2 sentences at most) explanation in the provided space in the box below the question. Marks will be awarded only when the answer (T/F) and explanation both are correct. (**3 x 2 = 6 marks**)

1.1	In any iteration $t = 1,2,\dots,T$ of gradient descent (GD) for linear regression, the gradient expression is more highly influenced by those training examples $(\boldsymbol{x}_n, y_n)$ on which the current $\boldsymbol{w}^{(t)}$ has a small error (i.e., difference between y_n and ${\boldsymbol{w}^{(t)}}^\top \boldsymbol{x}_n$).	

1.2	The absolute value loss function $\lvert y_n - \boldsymbol{w}^\top \boldsymbol{x}_n \rvert$ for linear regression cannot be optimized using first-order optimality to get a closed form solution for the weight vector $\boldsymbol{w}$	

1.3	The Perceptron loss function defined as $\max\{0, -y_n \boldsymbol{w}^\top \boldsymbol{x}_n\}$ is not differentiable but the Hinge loss function defined as $\max\{0, 1 - y_n \boldsymbol{w}^\top \boldsymbol{x}_n\}$ is differentiable.	

Question 2: Answer the following questions concisely in the space provided below the question.

2.1	Mention two advantages of Newton's method for optimization as compared to gradient descent, and also one disadvantage of the former. **(4 marks)**

2.2	Given the confusion matrix for the test data in a multi-class classification problem, can you compute the accuracy? If yes, how? If not, why not? **(3 marks)**

2.3	The soft-margin SVM problem for binary classification minimizes the following loss function: $L(\boldsymbol{w}, b) = \frac{\lVert \boldsymbol{w} \rVert^2}{2} + C \sum_{n=1}^{N} \xi_n$ where $\xi_n > 0$ denotes the slack on the n^{th} training example. What would be the effect of using a very-very large value of C? Would the model tend to overfit or underfit? Also, what about the margin of the classifier? Will we get a large margin or small margin? Briefly justify your answer. **(4 marks)**

2.4	Assuming multi-class classification given N training examples and a total of C classes, write down the expression of the multi-class cross-entropy loss function, clearly and succinctly defining the terms/notation involved in the expression, and briefly explain why this is a suitable loss function for multi-class classification problems. **(4 marks)**

2.5	Given a linear regression problem with non-negativity constraints on each entry of the weight vector $\boldsymbol{w}$, which of these two approaches would you prefer and why: (1) Projected Gradient Descent, and (2) Lagrangian based Optimization? **(4 marks)**