Introduction to ML (CS771), 2024-2025-Sem-I Midsem exam. September 15, 2024		Total Marks	60
		Duration	2 hours
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 may not be able to provide clarifications during the exam. If any aspect of some question appears ambiguous/unclear to you, please state your assumption(s) and answer accordingly.
4.	The last page (page 6) of this booklet can be used for rough work.

Section 1 (Short Answer Questions): Answer the following questions concisely in the space provided below the question.

1.1	Assume a scalar-valued function $f(\boldsymbol{w})$, with $\boldsymbol{w} = [w_1, w_2, w_3, w_4]^\top \in \mathbb{R}^4$, defined as $f(\boldsymbol{w}) = (w_1 - w_2)^2 + (w_2 - w_3)^2 + (w_3 - w_4)^2 + (w_4 - w_1)^2$. Write down the final expression (don't show the steps) of gradient of $f(\boldsymbol{w})$ w.r.t. $\boldsymbol{w}$. Without using the gradient, can you answer what is the minima of this function? If so, answer what it is. If not, state why you can't. Also, is $f(\boldsymbol{w})$ convex? Briefly justify the answer. **(2+2+2 =6 marks)**

1.2	Which of these are examples of convex sets? (1) the set of D-dimensional vectors with all non-negative entries, (2) set of D-dimensional vectors that have at most $K < D$ nonzero entries. Briefly justify your answers. **(2+2 =4 marks)**

1.3	Given the K real-valued logits $z_1, z_2, \dots, z_K$, the softmax function produces a vector of probabilities with each probability of the form $\mu_i = \frac{\exp(z_i)}{\sum_{j=1}^{K} \exp(z_j)}$. Write down the expressions of derivative of μ_i w.r.t. z_i and also the derivative of μ_i w.r.t. $z_j, j \neq i$. **(3+3 =6 marks)**

1.4	Assume two scalar inputs x and z. The RBF kernel based similarity between these two inputs is defined as $k(x, z) = \exp(-\gamma(x - z)^2)$. Assume $\gamma = 1$ for simplicity. Show that $k(x, z) = \phi(x)^\top \phi(z)$ and write down the expression of the feature mapping ϕ. You may use the series expansion result $\exp(x) = \sum_{n=0}^{\infty} \frac{x^n}{n!}$ **(6 marks)**

1.5	Suppose we want to train a Perceptron algorithm for binary classification but in a manner that the hyperplane has a margin, and not just equal margin γ on both sides but uneven margins (γ_+ towards the positive side and γ_- on the negative side; assume $\gamma_+ > \gamma_-$). Briefly describe using the necessary equation(s) how you would accomplish this. In what cases, having an uneven margin may be desirable? **(6 marks)**

1.6	Write down the expression for the gradient descent (GD) update for the ridge regression model which is the same as least squares (LS) linear regression with an additional ℓ_2-squared regularization on the weight vector $\boldsymbol{w}$. How is this GD update of the weight vector different from GD update for the standard LS linear regression, and what effect does it have on the weight vector for ridge regression. **(6 marks)**

Section 2 (Long Answer Questions): Answer the following questions concisely in the space provided below the question.

2.1	The binary cross-entropy loss used in logistic regression with weight vector $\boldsymbol{w} \in \mathbb{R}^D$ is given by $L(\boldsymbol{w}) = \sum_{n=1}^{N} -[y_n \log \sigma(w^\top \boldsymbol{x}_n) + (1 - y_n)\log(1 - \sigma(\boldsymbol{w}^\top \boldsymbol{x}_n))]$, where $\sigma(.)$ denotes the sigmoid function. Derive the Newton's method updates for $\boldsymbol{w}$ for optimizing this loss. **Note:** If you find it more convenient with matrix-vector notation then note that the above loss can also be written as $L(\boldsymbol{w}) = -\boldsymbol{y}^\top \log \sigma(\boldsymbol{X}\boldsymbol{w}) - (1 - \boldsymbol{y})^\top \log(1 - \sigma(\boldsymbol{X}\boldsymbol{w}))$ where $\boldsymbol{X}$ denotes the $N \times D$ feature matrix and $\boldsymbol{y}$ is the $N \times 1$ label vector, and log and sigmoid functions are applied element-wise on their vector arguments **(12 marks)**

2.2	Given training data $(\boldsymbol{X}, \boldsymbol{y}) = \{(\boldsymbol{x}_n, y_n)\}_{n=1}^{N}$ with $y_n \in \{1,2,\dots,K\}$, show that class prototype vectors $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \dots, \boldsymbol{\mu}_K$ in the learning with prototypes (LwP) problem are the solution to the minimizing the following objective: $\underset{\boldsymbol{\mu}_1,\boldsymbol{\mu}_2,\dots,\boldsymbol{\mu}_K}{\arg\min} \sum_{n=1}^{N} \left\Vert \boldsymbol{x}_n - \boldsymbol{\mu}_{y_n} \right\Vert^2$. Now consider a variant of the LwP problem where the labels $\boldsymbol{y} = \{y_n\}_{n=1}^{N}$ are not known. In this case, we must estimate not just the class prototype vectors $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \dots, \boldsymbol{\mu}_K$ but also the labels $\boldsymbol{y} = \{y_n\}_{n=1}^{N}$. For this "unsupervised" learning problem, give an ALT-OPT algorithm that alternates between estimating both sets of unknowns. Clearly write the update equations for the prototypes and labels in this ALT-OPT algorithm. **(14 marks)**

Only for Rough Work