Name:

Roll No.: **Dept.**:

IIT Kanpur
CS771A (IML)
Mid-sem Exam
Date: September 23, 2023

Instructions: *Total:* **60 marks**

1. Total duration: **2 hours**. Please write your name, roll number, department on **all pages**.
2. This booklet has 6 pages (5 pages + 1 page for rough work). No part of your answers should be on pages designated for rough work. Additional rough sheets may be provided if needed.
3. Write/mark your answers clearly in the provided space. Please keep your answers precise and concise.
4. Avoid showing very detailed derivations. You may do those on rought sheet and only show key steps.
5. **Answer each question using information provided in the question (no clarifications during the exam). If you want to make any assumptions, please state them in your answer.**

Section 1 (2 Multiple Choice Questions: Total 4 marks). (Tick/circle all options that you think are true)

1. Which of the following is true about MLE? (1) It is equivalent to finding the best parmaters by minimizing the loss function on the training data, (2) It is guranteed to give the global optimal solution for the parameters, (3) It yields the same answer as maximum-a-posteriori (MAP) estimation if MAP is done using zero-mean Gaussian prior on the parameters, (4) It has the risk of overfitting.

2. Which of the following is true about probabilistic linear regression with Gaussian likelihood and Gaussian prior on the weight vector and assuming all other hyperparameters as fixed? (1) MLE solution for the weight vector is the same as the solution of least squares regression, (2) Posterior distribution of the weight vector is also Gaussian, (3) The mode of the posterior distribution is the same as the MAP solution (4) The mean of the posterior distribution is the same as the MLE solution.

Section 2 (8 Descriptive Answer Questions: Total 56 marks). .

1. You wish to learn a linear regression model with the loss function $L(\boldsymbol{w}) = \sum_{i=1}^{N} |y_i - \boldsymbol{w}^\top \boldsymbol{x}_i|$ subject to constraints on $\boldsymbol{w} \in \mathbb{R}^D$ that its entries can only be non-negative. Briefly describe an optimization algorithm to learn $\boldsymbol{w}$ and write all the necessary mathematical equations/updates required by the algorithm. Your approach must not involve introducing any additional variables to solve this problem. (**6 marks**).

Name:

Roll No.: **Dept.**:

IIT Kanpur
CS771A (IML)
Mid-sem Exam
Date: September 23, 2023

2. Suppose we are given N training examples $\{(x_i, y_i)\}_{i=1}^N$ with each input $x_n \in \mathbf{R}$ being a scalar value and each output $y_n \in \{1, 2, \ldots, K\}$. Assume the inputs from class c are generated i.i.d. by univariate Gaussian $\mathcal{N}(x|\mu_c, \sigma_c^2)$ and, for simplicity, assume that only μ_c is unknown and σ_c^2 is known.

   - Derive (only show key steps) and write down the expression for the MLE solution of μ_c. (**3 marks**)

   - Given the MLE solution of $\{\mu_c\}_{c=1}^K$, to predict the label y_* for a new test input x_*, suppose this model uses the following rule at test time: predict y_* to be the class under which the test input x_* has the largest probability density. Compare and contrast this model with the Learning with the Prototypes (LwP) model. What would be the advantage of this model as compared to LwP? (**5 marks**)

3. Given a set P with N inputs $\{\boldsymbol{x}_i\}_{i=1}^N$ and another set Q with M inputs $\{\boldsymbol{z}_j\}_{j=1}^M$, one notion of distance between these two sets is the squared Euclidean distance $||\phi_P - \phi_Q||^2$ between their means ϕ_P and ϕ_Q in a feature space ϕ defined by a kernel function k. Derive the expression of this distance. Your final answer must be written only in terms of k and not in terms of the feature mapping ϕ. (**6 marks**)

Name:

Roll No.: **Dept.**:

IIT Kanpur
CS771A (IML)
Mid-sem Exam
Date: September 23, 2023

4. Suppose we have rolled a K faced die a total of N times and the number of times each face showed is given by $N_1, N_2, \ldots, N_K$. Assume each outcome $\boldsymbol{x}_n$ (a one-hot vector of length K) of the die roll to be a random variable modeled by a multinoulli $p(\boldsymbol{x}_n|\boldsymbol{\pi})$ where $\boldsymbol{\pi} = \{\pi_1, \pi_2, \ldots, \pi_K\}$ is the parameter of the multinoulli distribution. Note that the multinomial distribution also has the constraint $\sum_{i=1}^{K} \pi_i = 1$.

   Using Lagrange multiplier based constrained optimization, derive the MLE solution of $\{\pi_i\}_{i=1}^{K}$. (**8 marks**)

5. Define an ϵ-insensitive loss for linear regression as $\ell_\epsilon(y_n, \hat{y}_n) = 0$ if $|y_n - \hat{y}_n| \leq \epsilon$, otherwise $\ell_\epsilon(y_n, \hat{y}_n) = (|y_n - \hat{y}_n| - \epsilon)^2$. Here $\hat{y}_n = \boldsymbol{w}^\top \boldsymbol{x}_n$ denotes the model's prediction and y_n is the true label. Write down the expression for likelihood $p(y_n|\boldsymbol{x}_n, \boldsymbol{w})$ that corresponds to this loss function, and plot its PDF. (**6 marks**)

Name:

Roll No.: **Dept.**:

IIT Kanpur
CS771A (IML)
Mid-sem Exam
Date: September 23, 2023

6. The ridge regression problem has the solution $\boldsymbol{w} = (\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I}_D)^{-1}\mathbf{X}^\top \boldsymbol{y}$, where $\mathbf{X}$ is $N \times D$ feature matrix (row i contains the input $\boldsymbol{x}_i$) and $\boldsymbol{y}$ is the $N \times 1$ response vector, and $\lambda > 0$ is the regularization hyperparameter. Using the result $(\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I}_D)^{-1}\mathbf{X}^\top = \mathbf{X}^\top(\mathbf{X}\mathbf{X}^\top + \lambda \mathbf{I}_N)^{-1}$ (due to Woodbury matrix identity), show that the ridge regression model can be kernelized. More specifically, show that, given a test input $\boldsymbol{x}*$, we can compute a kernelized prediction which is given by $y_* = \sum_{i=1}^N \alpha_i k(\boldsymbol{x}_i, \boldsymbol{x}_*)$. (**6 marks**)

7. Consider a logistic regression model $p(y_n|\boldsymbol{x}_n, \boldsymbol{w}) = \frac{1}{1+\exp(-y_n \boldsymbol{w}^\top \boldsymbol{x}_n)}$, with a zero-mean Gaussian prior $p(\boldsymbol{w}) = \mathcal{N}(\mathbf{0}, \lambda^{-1}\mathbf{I})$. Note that this loss function for logistic regression assumes $y_n \in \{-1, +1\}$ instead of $\{0, 1\}$. Show that the MAP estimate for $\boldsymbol{w}$ can be written as $\boldsymbol{w} = \sum_{n=1}^N \alpha_n y_n \boldsymbol{x}_n$ where each α_n itself is a function of $\boldsymbol{w}$. Based on the expression of α_n, you would see that it has a precise meaning. Briefly explain what α_n means, and also briefly explain why the result $\boldsymbol{w} = \sum_{n=1}^N \alpha_n y_n \boldsymbol{x}_n$ makes sense for this model. (**8 marks**)

Name:

Roll No.: **Dept.:**

IIT Kanpur
CS771A (IML)
Mid-sem Exam
Date: September 23, 2023

---

8. Consider the dual problem for soft-margin linear SVM: $\arg\max_{0 \leq \boldsymbol{\alpha} \leq C} f(\boldsymbol{\alpha})$, where $f(\boldsymbol{\alpha}) = \boldsymbol{\alpha}^\top \mathbf{1} - \frac{1}{2}\boldsymbol{\alpha}^\top \mathbf{G}\boldsymbol{\alpha}$, $\mathbf{G}$ is an $N \times N$ matrix such that $G_{nm} = y_n y_m \boldsymbol{x}_n^\top \boldsymbol{x}_m$ (assume labels as -1/+1), and $\boldsymbol{\alpha} = [\alpha_1, \alpha_2, \ldots, \alpha_N]$ are the Lagrange multipliers. Given the optimal $\boldsymbol{\alpha}$, the SVM weight vector is $\boldsymbol{w} = \sum_{n=1}^N \alpha_n y_n \boldsymbol{x}_n$.

   Your goal is to derive a **co-ordinate ascent** procedure for the vector $\boldsymbol{\alpha}$, such that each iteration updates a uniformly randomly chosen entry α_n of the vector $\boldsymbol{\alpha}$. However, instead of updating $\boldsymbol{\alpha}$ via standard co-ordinate descent as $\alpha_n = \alpha_n + \eta g_n$ where g_n denotes the n-th entry of the gradient vector $\nabla_{\boldsymbol{\alpha}} f(\boldsymbol{\alpha})$, we will update it as $\alpha_n = \alpha_n + \delta_*$ where $\delta_* = \arg\max_\delta f(\boldsymbol{\alpha} + \delta \mathbf{e}_n)$ and $\mathbf{e}_n$ denotes a vector of all zeros except a 1 at entry n. Essentially, this will give the new α_n that guarantees the maximum increase in f, with all other α_n's fixed at their current value. Derive the expression for δ_*.

   Note that your expression for δ_* should be such that the constraint $0 \leq \alpha_n \leq C$ is maintained. (**8 marks**)

Some distributions and their properties:

- For $x \in \mathbb{R}$, the PDF of univariate Gaussian: $\mathcal{N}(x|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\{-\frac{(x-\mu)^2}{2\sigma^2}\}$. If using precision $\beta = 1/\sigma^2$, the PDF is $\mathcal{N}(x|\mu, \beta^{-1}) = \sqrt{\frac{\beta}{2\pi}} \exp\{-\frac{\beta}{2}(x-\mu)^2\}$.
- Given $\boldsymbol{\pi} = [\pi_1, \ldots, \pi_K]$, s.t. $\sum_{k=1}^K \pi_k = 1$, the multinoulli is defined as $\text{multinoulli}(\boldsymbol{x}|\boldsymbol{\pi}) = \prod_{i=1}^K \pi_i^{x_i}$, where the random variable $\boldsymbol{x} = [x_1, x_2, \ldots, x_K]$ is a one-hot vector.

Some other useful results:

- $\frac{\partial}{\partial \boldsymbol{x}} \boldsymbol{a}^\top \boldsymbol{x} = \boldsymbol{a}$, $\frac{\partial}{\partial \boldsymbol{x}} \boldsymbol{x}^\top \mathbf{A} \boldsymbol{x} = 2\mathbf{A}\boldsymbol{x}$. $|x| = \max\{x, -x\}$. If a function $h(x) = \max\{f(x), g(x)\}$ then, at any point x_*, (1) if $f(x_*) > g(x_*)$ then the subgradient $\partial h(x_*) = \partial f(x_*)$, (2) if $g(x_*) > f(x_*)$ then $\partial h(x_*) = \partial g(x_*)$, and (3) if $f(x_*) = g(x_*)$ then $\partial h(x_*) = \{\alpha a + (1-\alpha) b\}$, where $a \in \partial f(x_*)$, $b \in \partial g(x_*)$, and $\alpha \in [0, 1]$.

Name:

Roll No.: **Dept.**:

IIT Kanpur
CS771A (IML)
Mid-sem Exam
Date: September 23, 2023