EE 5184 Machine Learning, Fall 2022
Final Exam - Solution

Lecturer: Pei-Yuan Wu

December 19, 2022

This exam contains 8 questions and 115 pts in total. In this exam,

- $[\![m, n]\!]$ denotes the set of integers from m to n.
- For any set A, the indicator function $1_A(x)$ is defined as

$$1_A(x) = \begin{cases} 1 & \text{, if } x \in A \\ 0 & \text{, if } x \notin A \end{cases}$$

- The sigmoid function is defined as $\sigma(z) = \frac{1}{1+e^{-z}}$.
- The p-norm of a vector $\mathbf{x} = (x_1, ..., x_n)$ is denoted as

$$\|\mathbf{x}\|_p = (x_1^p + \cdots + x_n^p)^{1/p}.$$

1. **(10%) Weighted ridge regression**
   Consider the linear regression model $f_{\mathbf{w}} : \mathbf{w} \in \mathbb{R}^M \mapsto \mathbf{w} \cdot \mathbf{x}$, where $\mathbf{w} \in \mathbb{R}^M$ is a vector of weights for each feature. The weighted ridge regression solves the (column) weight vector $\mathbf{w} \in \mathbb{R}^M$ through minimizing the following loss function:

$$L(\mathbf{w}) = \sum_{i=1}^{N} \omega_i (y_i - X_i \mathbf{w})^2 + \lambda \|\mathbf{w}\|_2^2 \quad (1)$$

   where $((X_i, y_i))_{i=1}^N$ is the training data of N input-output pairs, with each $X_i \in \mathbb{R}^{1 \times M}$ being a row vector, and $\omega_i > 0$ denotes the "importance" of the i'th observation, and $\lambda > 0$ denotes the regularization coefficient. We may rewrite (1) in a more compact form

$$L(\mathbf{w}) = (\mathbf{y} - \mathbf{X}\mathbf{w})^T \Omega (\mathbf{y} - \mathbf{X}\mathbf{w}) + \lambda \|\mathbf{w}\|_2^2$$

   where $\mathbf{y} = [y_1 \ \cdots \ y_N]^T$ is a column vector of observed outputs, $\mathbf{X} \in \mathbb{R}^{N \times M}$ is a matrix with X_i being its i'th row, and $\Omega = \mathrm{diag}(\omega_1, ..., \omega_N)$. Suppose $L(\mathbf{w})$ is minimized when $\mathbf{w} = \mathbf{w}_0$.

   (a) **(5%)** Find $\mathbf{w}_0$ in explicit form of $\mathbf{X}$, Ω, λ, and $\mathbf{y}$.

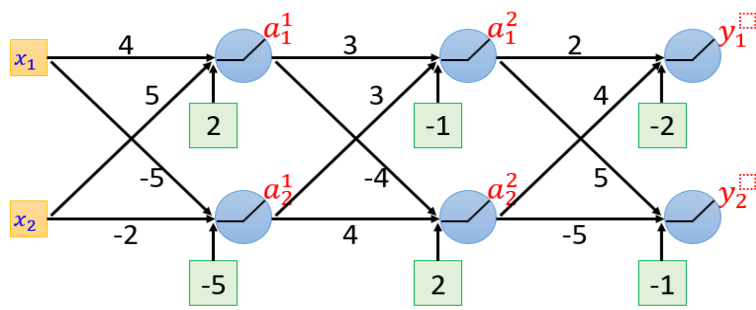

Figure 1: Fully connected neural newtork with ReLU activation function.

(b) **(5%)** Find $L(\mathbf{w}_0)$ in explicit form of $\mathbf{X}$, Ω, λ, and $\mathbf{y}$.

(Hint: Rewrite the loss function in quadratic form $L(\mathbf{w}) = (\mathbf{w}-\mathbf{b})^T\mathbf{A}(\mathbf{w}-\mathbf{b}) + \mathbf{C}$.)

2. **(10%) Principle component analysis classics**
Let $(\mathbf{x}_i)_{i=1}^N$ be N data points, where $\mathbf{x}_i = (x_{i,1}, x_{i,2}) \in \mathbb{R}^2$ for each $i \in [\![1, N]\!]$. Suppose your calculator tells you that

$$\sum_{i=1}^N x_{i,1} = \sum_{i=1}^N x_{i,2} = 0$$

$$\sum_{i=1}^N x_{i,1}^2 = 363, \quad \sum_{i=1}^N x_{i,1}x_{i,2} = -60, \quad \sum_{i=1}^N x_{i,2}^2 = 482,$$

(a) **(5%)** Find the first principle axis after performing PCA on this data set.

(b) **(5%)** Denote $\hat{\mathbf{x}}_i \in \mathbb{R}^2$ as the projection of $\mathbf{x}_i$ to the first principle axis. Find reconstruction error $\sum_{i=1}^N \|\mathbf{x}_i - \hat{\mathbf{x}}\|_2^2$.

3. **(15%) Forward and backward propagation**
Consider the fully connected neural network in Figure.1 where each neuron adopts ReLU activation: The network can be represented as a function f_θ, namely

$$\begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = f_\theta\left(\begin{bmatrix} x_1 \\ x_2 \end{bmatrix}\right),$$

where the parameter θ records all the weights and biases in the neural network. (You may omit the derivation, however, partial credits may be granted if you provide derivation though answers being incorrect)

(a) **(3%)** If $(x_1, x_2) = (2, 3)$, please compute a_1^1, a_2^1, a_1^2, a_2^1, y_1, y_2.

(b) **(12%)** Following (a), if the groundtruth is $(\hat{y}_1, \hat{y}_2) = (22, 12)$, show how each weight and bias in the neural network is updated by gradient descent with an aim to minimize the loss function

$$L(\theta) = \left\| \begin{bmatrix} \hat{y}_1 \\ \hat{y}_2 \end{bmatrix} - f_\theta \left(\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \right) \right\|_2^2$$

where we assume the learning rate is $\eta = 0.01$.

4. **(10%) Logistic regression with miss labels**
Consider a binary classification problem in which each data point $\mathbf{x}_i \in \mathbb{R}^M$ is known to belong to one of two classes, as specified by class label $\xi_i \in \{\pm 1\}$, and suppose that the procedure for collecting training data is imperfect, so that training data are sometimes mislabelled. More elaborately, for every data point x_i, instead of observing its class label ξ_i, we instead observe a perturbed class label $y_i = (-1)^{z_i} \xi_i$ where z_i is $\{0, 1\}$-valued and follows Bernoulli distribution

$$1 - \mathbb{P}[z_i = 0] = \mathbb{P}[z_i = 1] = \pi_i$$

where $\pi_i \in [0, 1]$ is the probability of mis-labeling the i'th data point.

Given training data of N input-output pairs $\mathcal{D} = ((\mathbf{x}_i, y_i))_{i=1}^N$, where $\mathbf{x}_i \in \mathbb{R}^M$ and $y_i \in \{\pm 1\}$. Consider the generative model

$$p_{\mathbf{w},b}(\xi = 1|\mathbf{x}) = 1 - p_{\mathbf{w},b}(\xi = -1|\mathbf{x}) = \sigma(\mathbf{w} \cdot \mathbf{x} + b),$$

where σ denotes the sigmoid function.

(a) **(5%)** Write down the error function $L(\mathbf{w}, b)$ to be minimized so as to maximize the log-likelihood function of generative model $p_{\mathbf{w},b}$, assuming each of the N data points (x_i, y_i) is generated (and labelled/mislabelled) independently.

(b) **(5%)** Describe how to perform maximum likelihood estimation of $\mathbf{w}$ and b by minimizing $L(\mathbf{w}, b)$ with gradient descent algorithm. Please write down the update equation for gradient descent.

5. **(10%) Boosting**

(a) **(7%)** Consider training a boosting classifier using decision stumps on the data set illustrated in Figure.2:

i. **(2%)** Which examples will have their weights increased at the end of the first iteration? Circle them.

ii. **(5%)** How many iterations will it take to achieve zero training error? Justify your answers.

(b) **(3%)** Suppose AdaBoost is run on N training examples, and suppose on each round that the weighted training error ϵ_t of the t'th weak hypothesis is at most $1/2 - \gamma$, for some number $0 < \gamma < 1/2$. After

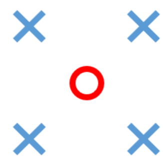

Figure 2: AdaBoost Data set

how many iterations, T, will the combined hypothesis be consistent with the N training examples, i.e., achieves zero training error? Your answer should only be expressed in terms of N and γ. (Hint: Recall that exponential loss is an upper bound for 0-1 loss. What is the training error when 1 example is misclassified?)

6. **(10%) Manhattan k-means**
In this problem, we design K-means algorithm with an aim to minimize the in-cluster Manhattan distance. Given a set of N data points $\mathbf{x}_i = (x_{i,1}, ..., x_{i,M}) \in \mathbb{R}^M$ $(i \in [\![1, N]\!])$, we aim to minimize the loss function:

$$L(\mathcal{C}, \Xi) = \sum_{i=1}^{N} \|\mathbf{x}_i - \boldsymbol{\xi}_{\mathcal{C}(i)}\|_1 = \sum_{i=1}^{N} \sum_{m=1}^{M} |x_{i,m} - \xi_{\mathcal{C}(i),m}| \tag{2}$$

where $\Xi = (\boldsymbol{\xi}_k)_{k=1}^K$ for which $\boldsymbol{\xi}_k = (\xi_{k,1}, ..., \xi_{k,M}) \in \mathbb{R}^M$ is the centroid of the k'th cluster, and $\mathcal{C} : [\![1, N]\!] \to [\![1, K]\!]$ is the cluster assignment function where the i'th data point is assigned to cluster $\mathcal{C}(i)$.

(a) **(1%)** Find $\theta \in \mathbb{R}$ that minimizes $\sum_{i=1}^{5} |\theta - i^2|$.

Randomly initialize $\mathcal{C}^{(0)}$ be a cluster assignment.

(b) **(3%)** Given cluster assignment $\mathcal{C}^{(t-1)}$, what are the cluster centroids $\Xi = (\boldsymbol{\xi}_k)_{k=1}^K$ that minimizes $L(\mathcal{C}^{(t-1)}, \Xi)$? Denote such optimal cluster centroids as $\Xi^{(t)}$.

(c) **(3%)** Given cluster centroids $\Xi^{(t)} = (\boldsymbol{\xi}_k^{(t)})_{k=1}^K$, what is the cluster assignment $\mathcal{C}$ that minimizes $L(\mathcal{C}, \Xi^{(t)})$? Denote such optimal cluster assignment as $\mathcal{C}^{(t)}$.

(d) **(3%)** Alternatively optimize the cluster centroids and assignments by iterating (b) and (c) through $t = 0, 1, 2, \cdots$. Does there exist $T < \infty$ such that $L(\mathcal{C}^{(T-1)}, \Xi^{(T-1)}) = L(\mathcal{C}^{(t)}, \Xi^{(t)})$? Justify your answers.

7. **(20%) EM algorithm for mixture of uniform model**
Consider the generative model parameterized by $\theta = (\pi_k, b_k)_{k=1}^K$, where $b_k > 0$ and $\pi_k > 0$ for each k, and that $\sum_{k=1}^K \pi_k = 1$, so that the probability density function of generating a $[0, \infty)$-valued number x is

$$p(x; \theta) = \sum_{k=1}^{K} \frac{\pi_k}{b_k} 1_{[0,b_k]}(x).$$

That is, $p(\cdot;\theta)$ is a mixture of uniform distributions. Suppose we observe training data of N numbers $x_1, ..., x_N \in [0, \infty)$, derive the E-step and M-step equations of the EM algorithm for optimizing the mixing coefficients π_k and the scalars b_k by maximum likelihood. You may assume the initial guess of the parameters $\theta^{(0)} = (\pi_k^{(0)}, b_k^{(0)})_{k=1}^K$ satisfy

$$\max_{1\le k\le K} b_k^{(0)} \ge \max_{1\le i\le N} x_i, \quad \min_{1\le k\le K} b_k^{(0)} \ge \min_{1\le i\le N} x_i$$

8. **(30%) Spherical one class SVM**

Suppose we aim to fit a hypersphere which encompasses a majority of data points $\mathbf{x}_1, ..., \mathbf{x}_N \in \mathbb{R}^M$ by considering the following optimization problem: (here $\boldsymbol{\mu}$ and each $\mathbf{x}_i$ are considered as column vectors)

$$\begin{array}{ll} \text{minimize} & R^2 + \frac{1}{\nu}\sum_{i=1}^N C_i\xi_i \\ \text{subject to} & \left.\begin{array}{l} \|\mathbf{x}_i - \boldsymbol{\mu}\|^2 \le R^2 + \xi_i \\ \xi_i \ge 0 \end{array}\right\} \forall i \in [\![1, N]\!] \\ & R \ge 0 \\ \text{variables} & R \in \mathbb{R}, \boldsymbol{\mu} \in \mathbb{R}^M, \boldsymbol{\xi} = (\xi_1, ..., \xi_N) \in \mathbb{R}^N \end{array} \tag{3}$$

where $C_i > 0$ for each $i \in [\![1, N]\!]$, and $0 < \nu < \sum_{i=1}^N C_i$. Let $\rho = R^2$ and rewrite (3) in the form of primal problem:

$$\begin{array}{ll} \text{minimize} & f(\rho, \boldsymbol{\mu}, \boldsymbol{\xi}) = \rho + \frac{1}{\nu}\sum_{i=1}^N C_i\xi_i \\ \text{subject to} & \left.\begin{array}{l} g_{1,i}(\rho, \boldsymbol{\mu}, \boldsymbol{\xi}) = \|\mathbf{x}_i - \boldsymbol{\mu}\|^2 - \rho - \xi_i \le 0 \\ g_{2,i}(\rho, \boldsymbol{\mu}, \boldsymbol{\xi}) = -\xi_i \le 0 \end{array}\right\} \forall i \in [\![1, N]\!] \\ & g_3(\rho, \boldsymbol{\mu}, \boldsymbol{\xi}) = -\rho \le 0 \\ \text{variables} & \rho \in \mathbb{R}, \boldsymbol{\mu} \in \mathbb{R}^M, \boldsymbol{\xi} \in \mathbb{R}^N \end{array} \tag{4}$$

as well its Lagrangian dual problem:

$$\begin{array}{ll} \text{maximize} & \theta(\alpha, \beta, \gamma) = \inf_{\rho\in\mathbb{R}, \boldsymbol{\mu}\in\mathbb{R}^M, \boldsymbol{\xi}\in\mathbb{R}^N} L(\rho, \boldsymbol{\mu}, \boldsymbol{\xi}, \alpha, \beta, \gamma) \\ \text{subject to} & \alpha_i \ge 0, \beta_i \ge 0 \ \forall i \in [\![1, N]\!] \\ & \gamma \ge 0 \\ \text{variables} & \alpha = (\alpha_1, ..., \alpha_N) \in \mathbb{R}^N, \beta = (\beta_1, ..., \beta_N) \in \mathbb{R}^N, \gamma \in \mathbb{R} \end{array} \tag{5}$$

(a) **(3%)** Write down the Lagrangian function $L(\rho, \boldsymbol{\mu}, \boldsymbol{\xi}, \alpha, \beta, \gamma)$ in explicit form of $\rho, \boldsymbol{\mu}, \boldsymbol{\xi}, \alpha, \beta, \gamma$.

(b) **(4%)** Show that the duality gap between (4) and (5) is zero.

(c) **(4%)** Derive $\theta(\alpha, \beta, \gamma)$ in explicit form of dual variables α, β, γ.

(d) **(3%)** Show that the dual problem can be simplified as

$$\begin{array}{ll} \text{maximize} & \|\alpha\|_1 \left(\sum_{i=1}^N \hat{\alpha}_i\|\mathbf{x}_i\|^2 - \sum_{1\le i,j\le N} \hat{\alpha}_i\hat{\alpha}_j\mathbf{x}_i^T\mathbf{x}_j\right) \\ \text{subject to} & \sum_{i=1}^N \alpha_i \le 1 \\ \text{variables} & 0 \le \alpha_i \le \frac{C_i}{\nu}, i \in [\![1, N]\!] \end{array} \tag{6}$$

where $\|\alpha\|_1 = \sum_{i=1}^N \alpha_i$ and $\alpha_i = \|\alpha\|_1\hat{\alpha}_i$.

(e) **(14%)** Suppose $(\bar{\rho}, \bar{\boldsymbol{\mu}}, \bar{\boldsymbol{\xi}})$ and $(\bar{\alpha}, \bar{\beta}, \bar{\gamma})$ are the optimal solutions to problems (4) and (5), respectively.

i. **(2%)** Show that $\|\bar{\alpha}\|_1 \bar{\boldsymbol{\mu}} = \sum_{i=1}^{N} \bar{\alpha}_i \mathbf{x}_i$.

ii. **(3%)** Show that

$$\bar{\rho} \in \arg\min_{\rho \geq 0} \left(\rho + \frac{1}{\nu} \sum_{i=1}^{N} C_i \max(\|\mathbf{x}_i - \bar{\boldsymbol{\mu}}\|^2 - \rho, 0) \right),$$

iii. **(3%)** Show that

$$\min \left\{ \rho \geq 0 : \sum_{i:\|\mathbf{x}_i - \bar{\boldsymbol{\mu}}\|^2 \leq \rho}^{N} C_i \geq \nu \right\} \leq \bar{\rho} \leq \min \left\{ \rho \geq 0 : \sum_{i:\|\mathbf{x}_i - \bar{\boldsymbol{\mu}}\|^2 \leq \rho}^{N} C_i > \nu \right\}. \tag{7}$$

iv. **(3%)** Prove that $\bar{\xi}_i = \max\left(\|\mathbf{x}_i - \bar{\boldsymbol{\mu}}\|^2 - \bar{\rho}, 0\right)$ for each $i \in [\![1, N]\!]$.

v. **(3%)** Prove that

$$\begin{cases} \bar{\alpha}_i = C_i/\nu & , \text{if } \|\mathbf{x}_i - \bar{\boldsymbol{\mu}}\|^2 > \bar{\rho} \\ \bar{\alpha}_i = 0 & , \text{if } \|\mathbf{x}_i - \bar{\boldsymbol{\mu}}\|^2 < \bar{\rho} \\ 0 \leq \bar{\alpha}_i \leq C_i/\nu & , \text{if } \|\mathbf{x}_i - \bar{\boldsymbol{\mu}}\|^2 = \bar{\rho} \end{cases}.$$

(f) **(2%)** Suppose $C_i = 1/n$ for each $i \in [\![1, n]\!]$. What is the physical meaning of ν?