EE 5184 Machine Learning, Fall 2023
Final Exam

Lecturer: Pei-Yuan Wu
TA: Chun-Lin Huang, Yuan-Chia Chang, and Jie-Xiang Fan

December 22, 2023

This exam contains 7 questions and 120 pts in total. In this exam,

- $[\![m,n]\!]$ denotes the set of integers from m to n.
- $\mathbf{I}$ denotes the identity matrix. We may use notation $\mathbf{I}_m$ to emphasize it as an $m \times m$ identity matrix.
- 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%) Linear Regression with Weighted Sample**
   Consider linear regression model $f(\mathbf{x})=\mathbf{w}^\mathsf{T}\mathbf{x}$. Given data points $\mathbf{x}_1,...,\mathbf{x}_N\in\mathbb{R}^m$, find the optimal $\mathbf{w}$ that minimizes the following L^2-regularized weighed squared error loss function

$$L(\mathbf{w})=\frac{1}{2}\sum_{i=1}^{N}\kappa_i(y_i-\mathbf{w}^\mathsf{T}\mathbf{x}_i)^2+\frac{\lambda}{2}\|\mathbf{w}\|_2^2.$$

   where λ and each κ_i are fixed positive numbers.

2. **(10%) Gradient boosting of regression models**
   Let $\mathcal{X}$ be the input space, H be a collection of regression models that map from $\mathcal{X}$ to $\mathbb{R}$. Let $((x_i,y_i))_{i=1}^m$ be the training data set, where $x_i\in\mathcal{X}$ and $y_i\in\mathbb{R}$. Given $T\in\mathbb{N}$, suppose we want to build regression model of the form

$$g_{T+1}(x)=\sum_{t=1}^{T}\alpha_t h_t(x),$$

   where $h_t\in H$ and $\alpha_t\in\mathbb{R}$ for all $t\in[\![1,T]\!]$. Please show how the functions h_t and coefficients α_t are chosen by gradient boosting with an aim to minimize the following loss function:

$$L(g_{T+1})=\frac{1}{2}\sum_{i=1}^{N}(g_{T+1}(x_i)-y_i)^2$$

3. **(10%) Laplacian eigenmap**
Consider a tiny social network as in Figure 1.

   (a) **(2 pt)** Write down the adjacency matrix $\mathbf{W}$, where the (i,j)-th element $w_{i,j}$ indicates whether or not an edge exists between x_i and x_j.

   (b) **(2 pt)** Write down the graph Laplacian $\mathbf{L}=\mathbf{D}-\mathbf{W}$, where $\mathbf{D}=\text{diag}(d_1,...,d_7)$ with d_i being the number of edges incident on x_i.

   (c) **(6 pt)** To embed each node x_i to some $\mathbf{z}_i\in\mathbb{R}^3$, we formulate the following optimization problem:

$$\begin{array}{ll} \text{minimize} & \frac{1}{2}\sum_{1\le i,j\le 7} w_{i,j}\|\mathbf{z}_i-\mathbf{z}_j\|^2 \\ \text{subject to} & \sum_{i=1}^{7}(1-2^{-d_i})\mathbf{z}_i\mathbf{z}_i^{\mathsf{T}}=\mathbf{I}_3 \\ \text{variables} & \mathbf{z}_1,...,\mathbf{z}_7\in\mathbb{R}^3 \end{array} \tag{1}$$

   Suppose you have tool at hand that allows you to compute eigenvalues and eigenvectors of a symmetric matrix, describe how to solve (1).

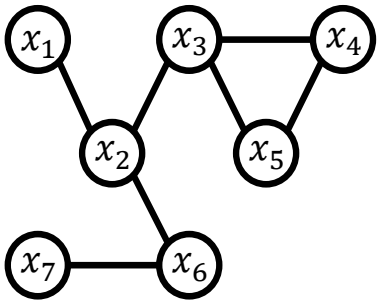

Figure 1: A tiny social network.

4. **(15%) Axis-aligned cuboids**
Let $\mathcal{X}=\mathbb{R}^3$ be the input space and consider the set of concepts of the form $c=\{(x,y,z):a_1\le x\le b_1,\ a_2\le y\le b_2,\ a_3\le z\le b_3\}$ for some real numbers a_1,b_1,a_2,b_2,a_3,b_3. Show that this class can be (ϵ,δ)-PAC-learned from training data of size $m\ge\frac{6}{\epsilon}\log\frac{6}{\delta}$.

5. **(20%) EM for Mixture of Linear Models**
Consider the generative model parameterized by $\theta=((\pi_k,\mathbf{w}_k,\rho_k))_{k=1}^K$, where $\mathbf{w}_k\in\mathbb{R}^m$, $\rho_k>0$, $\pi_k>0$ for each k, and that $\sum_{k=1}^K\pi_k=1$, so that the probabilistic density function (PDF) of generating a real-valued number y given $\mathbf{x}\in\mathbb{R}^m$ is

$$p(y|\mathbf{x};\theta)=\sum_{k=1}^K\pi_k\mathcal{N}(y;f_k(\mathbf{x};\theta),\rho_k)$$

where $f_k(\mathbf{x};\theta)=\mathbf{w}_k^\mathsf{T}\mathbf{x}$ and

$$\mathcal{N}(y;\mu,\rho)=\frac{1}{\sqrt{2\pi\rho}}\exp\left(-\frac{(y-\mu)^2}{2\rho}\right)$$

denotes the PDF of Gaussian distribution with mean μ and variance ρ. That is, $p(\cdot|\mathbf{x};\theta)$ describes how the output is generated based on a mixture of linear models $f_k(\cdot;\theta)$ with mixing coefficient π_k and uncertainty ρ_k.

Suppose we observe N data inputs $\mathbf{x}_1,...,\mathbf{x}_N\in\mathbb{R}^m$ and their corresponding outputs $y_1,...,y_N$. The maximum likelihood estimation is given by

$$\theta^{opt}\in\operatorname*{argmin}_\theta\sum_{i=1}^N\log p(y_i|\mathbf{x}_i;\theta).$$

Derive the E-step and M-step equations of the EM algorithm for optimizing the linear model coefficients $\mathbf{w}_k$, the mixing coefficients π_k, as well as the variances ρ_k.

6. **(35%) Hinge loss with L^1 regularization**
Given data points $\mathbf{x}_1,...,\mathbf{x}_N \in \mathbb{R}^m$ as well as their labels $y_1,...,y_N \in \{\pm 1\}$ and panelty coefficients $C_1,...,C_N > 0$, where each $\mathbf{x}_i = [x_{i,1},...,x_{i,m}]^\mathsf{T}$ is a column vector, consider the following optimization problem:

$$\begin{array}{ll} \text{minimize} & \|\mathbf{w}\|_1 + \sum_{i=1}^N C_i \xi_i \\ \text{subject to} & \begin{array}{l} y_i(\mathbf{w}^\mathsf{T}\mathbf{x}_i + b) \geq 1 - \xi_i \\ \xi_i \geq 0 \end{array} \quad , \ i=1,...,N \\ \text{variables} & \mathbf{w} \in \mathbb{R}^m, b \in \mathbb{R}, \boldsymbol{\xi} \in \mathbb{R}^N \end{array} \tag{2}$$

Note that in this formulation, we replace the L^2-regularization term $\frac{1}{2}\|\mathbf{w}\|^2$ by the L^1-regularization term $\|\mathbf{w}\|_1 = \sum_{j=1}^m |w_j|$.

(a) **(5%)** Show that $(\bar{\mathbf{w}}, \bar{b}, \bar{\boldsymbol{\xi}})$ is an optimal solution of (2) if and only if $\bar{\mathbf{w}} = \bar{\mathbf{u}} - \bar{\mathbf{v}}$ where $(\bar{\mathbf{u}}, \bar{\mathbf{v}}, \bar{b}, \bar{\boldsymbol{\xi}})$ is an optimal solution of the following problem:

$$\begin{array}{ll} \text{minimize} & f(\mathbf{u},\mathbf{v},b,\boldsymbol{\xi}) = \sum_{j=1}^m (u_j + v_j) + \sum_{i=1}^N C_i \xi_i \\ \text{subject to} & \begin{array}{ll} y_i((\mathbf{u}-\mathbf{v})^\mathsf{T}\mathbf{x}_i + b) \geq 1 - \xi_i & , \ i=1,...,N \\ \xi_i \geq 0 & \\ u_j \geq 0, \ v_j \geq 0 & , \ j=1,...,m \end{array} \\ \text{variables} & \mathbf{u} \in \mathbb{R}^m, \mathbf{v} \in \mathbb{R}^m, b \in \mathbb{R}, \boldsymbol{\xi} \in \mathbb{R}^N \end{array} \tag{3}$$

Following (a), we can now rewrite (3) as the following primal problem:

$$\begin{array}{ll} \text{minimize} & f(\mathbf{u},\mathbf{v},b,\boldsymbol{\xi}) = \sum_{j=1}^m (u_j + v_j) + \sum_{i=1}^N C_i \xi_i \\ \text{subject to} & \begin{array}{ll} g_i^1(\mathbf{u},\mathbf{v},b,\boldsymbol{\xi}) = 1 - \xi_i - y_i((\mathbf{u}-\mathbf{v})^\mathsf{T}\mathbf{x}_i + b) \leq 0 & , \ i=1,...,N \\ g_i^2(\mathbf{u},\mathbf{v},b,\boldsymbol{\xi}) = -\xi_i \leq 0 & \\ g_j^3(\mathbf{u},\mathbf{v},b,\boldsymbol{\xi}) = -u_j \leq 0 & , \ j=1,...,m \\ g_j^4(\mathbf{u},\mathbf{v},b,\boldsymbol{\xi}) = -v_j \leq 0 & \end{array} \\ \text{variables} & \mathbf{u} \in \mathbb{R}^m, \mathbf{v} \in \mathbb{R}^m, b \in \mathbb{R}, \boldsymbol{\xi} \in \mathbb{R}^N \end{array} \tag{4}$$

as well as its Lagrangian dual problem:

$$\begin{array}{ll} \text{maximize} & \theta(\alpha,\beta,\boldsymbol{\mu},\boldsymbol{\nu}) = \inf\{L(\mathbf{u},\mathbf{v},b,\boldsymbol{\xi},\alpha,\beta,\boldsymbol{\mu},\boldsymbol{\nu}) : \mathbf{u} \in \mathbb{R}^m, \mathbf{v} \in \mathbb{R}^m, b \in \mathbb{R}, \boldsymbol{\xi} \in \mathbb{R}^N\} \\ \text{subject to} & \begin{array}{ll} \alpha_i \geq 0, \ \beta_i \geq 0 & , \ i=1,...,N \\ \mu_j \geq 0, \ \nu_j \geq 0 & , \ j=1,...,m \end{array} \\ \text{variables} & \alpha \in \mathbb{R}^N, \beta \in \mathbb{R}^N, \boldsymbol{\mu} \in \mathbb{R}^m, \boldsymbol{\nu} \in \mathbb{R}^m \end{array} \tag{5}$$

where L denotes the Lagranian function.

(b) **(2%)** Associate dual variables $\alpha_i, \beta_i, \mu_j, \nu_j$ to constraints $g_i^1, g_i^2, g_j^3, g_j^4$, respectively. Show that L can be written in the following explicit form:

$$\begin{aligned} & L(\mathbf{u},\mathbf{v},b,\boldsymbol{\xi},\alpha,\beta,\boldsymbol{\mu},\boldsymbol{\nu}) \\ &= \mathbf{1}^\mathsf{T}(\mathbf{u}+\mathbf{v}) + \sum_{i=1}^N C_i \xi_i + \sum_{i=1}^N \alpha_i \Big(1 - \xi_i - y_i((\mathbf{u}-\mathbf{v})^\mathsf{T}\mathbf{x}_i + b)\Big) + \sum_{i=1}^N \beta_i(-\xi_i) - \boldsymbol{\mu}^\mathsf{T}\mathbf{u} - \boldsymbol{\nu}^\mathsf{T}\mathbf{v} \end{aligned}$$

where $\mathbf{1}$ denotes the all one vector.

(c) **(3%)** Show that (4) satisfies the Slater's condition.

(d) **(6%)** Show that

(i) **(3%)** $\theta(\alpha,\beta,\boldsymbol{\mu},\boldsymbol{\nu})=-\infty$ unless the following conditions hold:

$$\sum_{i=1}^{N}\alpha_i y_i=0,\ \ \boldsymbol{\mu}=\mathbf{1}-\sum_{i=1}^{N}\alpha_i y_i\mathbf{x}_i,\ \ \boldsymbol{\nu}=\mathbf{1}+\sum_{i=1}^{N}\alpha_i y_i\mathbf{x}_i \tag{6a}$$

$$\alpha_i+\beta_i=C_i\ \ \forall i=1,...,N \tag{6b}$$

at which case $\theta(\alpha,\beta,\boldsymbol{\mu},\boldsymbol{\nu})=\sum_{i=1}^{N}\alpha_i$.

(ii) **(3%)** The stationary condition holds if and only if (6) is satisfied.

(e) **(2%)** Show that the dual problem (5) can be simplified as

$$\begin{array}{ll}\text{maximize} & \sum_{i=1}^{N}\alpha_i \\ \text{subject to} & \sum_{i=1}^{N}\alpha_i y_i=0 \\ & -\mathbf{1}\le\sum_{i=1}^{N}\alpha_i y_i\mathbf{x}_i\le\mathbf{1} \\ \text{variables} & 0\le\alpha_i\le C_i,\ \ i=1,...,N\end{array} \tag{7}$$

(f) **(4%)** Write down the KKT conditions for primal/dual problems (4)(5).

(g) **(10%)** Suppose $(\bar{\mathbf{u}},\bar{\mathbf{v}},\bar{b},\bar{\boldsymbol{\xi}})$ and $(\bar{\alpha},\bar{\beta},\bar{\boldsymbol{\mu}},\bar{\boldsymbol{\nu}})$ are optimal solutions to (4) and (5) respectively. Denote $\bar{\mathbf{w}}=\bar{\mathbf{u}}-\bar{\mathbf{v}}$. Show that

(i) **(3%)** $\bar{u}_j=\max(\bar{w}_j,0)$ and $\bar{v}_j=\max(-\bar{w}_j,0)$. (Hint: Consider the cases $\sum_{i=1}^{N}\bar{\alpha}_i y_i x_{i,j}<1$ and $\sum_{i=1}^{N}\bar{\alpha}_i y_i x_{i,j}>-1$.)

(ii) **(3%)** $\bar{w}_j=0$ unless $\sum_{i=1}^{N}\bar{\alpha}_i y_i x_{i,j}\in\{\pm 1\}$.

(iii) **(4%)**

$$\begin{cases}\bar{\alpha}_i=C_i, & \bar{\xi}_i=1-y_i(\bar{\mathbf{w}}^{\mathsf{T}}\mathbf{x}_i+\bar{b}), & \text{if } y_i(\bar{\mathbf{w}}^{\mathsf{T}}\mathbf{x}_i+\bar{b})<1\\ \bar{\alpha}_i=0, & \bar{\xi}_i=0, & \text{if } y_i(\bar{\mathbf{w}}^{\mathsf{T}}\mathbf{x}_i+\bar{b})>1\\ 0\le\bar{\alpha}_i\le C_i, & \bar{\xi}_i=0, & \text{if } y_i(\bar{\mathbf{w}}^{\mathsf{T}}\mathbf{x}_i+\bar{b})=1\end{cases}$$

(h) **(3%)** Is it true that $\bar{\mathbf{w}}$ must be a linear combination of $\mathbf{x}_1,...,\mathbf{x}_N$? Justify your answers.

7. **(20%) Conjunction of Boolean literals**
Consider learning the concept class $C_{n,k}$ of conjunctions of at most k literals from n Boolean literals. A Boolean literal is either a variable x_i ($i \in [\![1,n]\!]$), or its negation $\bar{x}_i$. For $n=5$ and $k=4$, an example is the conjunction $x_1 \wedge \bar{x}_2 \wedge x_4$ (which consists of $3 \leq 4$ literals), where $\bar{x}_2$ denotes the negation of the Boolean literal x_2. (1,0,0,1,1) is a positive example for this concept while (1,0,0,0,0) is a negative example.

Observe that for $n=5$, a positive example (1,0,1,0,0) implies that the target concept cannot contain the literals $\bar{x}_1$ and $\bar{x}_3$ and that it cannot contain the literals x_2, x_4 and x_5. In contrast, a negative example is not as informative since it is not known which of its n bits are incorrect. A simple algorithm $\mathbb{A}$ for finding a consistent hypothesis is thus based on positive examples and consists of the following: for each positive example $(b_1,...,b_n)$ and $i \in [\![1,n]\!]$, if $b_i=1$ then $\bar{x}_i$ is ruled out as a possible literal, while if $b_i=0$ then x_i is ruled out. The conjunction of all the literals not ruled out is thus a hypothesis consistent with the target. Figure 2 shows an example training sample as well as the consistent hypothesis returned by $\mathbb{A}$ for the case $n=6$.

0	1	1	0	1	1	+
0	1	1	1	1	1	+
0	0	1	1	0	1	-
0	1	1	1	1	1	+
1	0	0	1	1	0	-
0	1	0	0	1	1	+
0	1	?	?	1	1	

Figure 2: Each of the first six rows of the table represents a training example with its label, + or −, indicated in the last column. The last row contains 0 (respectively 1) in column $i \in [\![1,6]\!]$ if the i-th entry is 0 (respectively 1) for all the positive examples. It contains "?" if both 0 and 1 appear as an i-th entry for some positive example. Thus, for this training sample, the hypothesis returned by $\mathbb{A}$ is $\bar{x}_1 \wedge x_2 \wedge x_5 \wedge x_6$.

(a) **(5%)** For $n=8$ and $k=4$, consider the training sample as given by Table 1.

(i) **(2%)** What is the hypothesis conjunction returned by $\mathbb{A}$?

(ii) **(3%)** Are you certain that the hypothesis returned by $\mathbb{A}$ in (i) is exactly the target concept? If so, please justify your answer; otherwise, please describe how it might go wrong.

Input	Output
(1,1,1,0,1,1,0,0)	+
(1,1,1,0,1,1,0,1)	+
(1,1,1,0,1,1,1,1)	+
(1,1,1,0,0,1,0,0)	+
(1,0,1,0,1,0,1,0)	-
(0,1,0,1,1,0,1,0)	-
(0,0,1,0,0,0,0,1)	-
(1,0,1,0,1,0,0,1)	-
(1,1,1,0,0,1,1,0)	+
(1,1,0,1,1,1,0,1)	+
(1,1,1,1,0,1,0,1)	+
(1,1,1,0,0,1,1,1)	+
(0,0,1,1,1,0,1,1)	-
(1,0,0,0,1,0,1,1)	-
(1,0,0,1,0,1,0,0)	-
(0,1,0,0,1,0,1,0)	-
(0,1,0,1,1,0,1,1)	-
(1,1,1,0,0,1,0,1)	+
(0,1,1,0,0,0,0,0)	-
(1,1,0,0,1,1,0,1)	+

Table 1: A training sample for learning Boolean conjunction.

(b) **(15%)** Suppose you can actively collect training samples through the following process: you toss a fair coin to decide each of the Boolean literal, that is, for each example $(b_1,...,b_n)$ the literals $b_1,...,b_n$ are independently generated from Bernoulli distribution. Suppose you randomly draw a training sample S of m examples through the aforementioned process, and apply $\mathbb{A}$ to learn a hypothesis conjunction h_S. Show that the probability of h_S to be exactly the same as the target concept $c \in C_{n,k}$ is at least $1-2n\cdot\exp\left(-\frac{m}{2^{k+1}}\right)$.