EE 5184 Machine Learning, Fall 2024
Final Exam

Lecturer: Pei-Yuan Wu
TA: Po-Yang Hsieh, Le-Rong Hsu, and Chao-Chi Lan

December 19, 2024

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

- $\mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu}, \Sigma)$ indicates that $\mathbf{x}$ is a random vector following Gaussian distribution with mean vector $\boldsymbol{\mu}$ and covariance matrix Σ.
- $[\![m, n]\!]$ denotes the set of integers from m to n.
- $\mathbf{I}$ denotes the identity matrix.

1. **(10%) Principal component analysis**
   Given N samples $\mathbf{x}_1,...,\mathbf{x}_N \in \mathbb{R}^2$. Suppose

$$\frac{1}{N}\sum_{i=1}^{N}\mathbf{x}_i=\mathbf{0},\quad \frac{1}{N}\sum_{i=1}^{N}\mathbf{x}_i\mathbf{x}_i^\mathsf{T}=\begin{bmatrix} 801 & -300 \\ -300 & 1396 \end{bmatrix}.$$

   (a) (3%) Find $\frac{1}{N}\sum_{i=1}^{N}\|\mathbf{x}_i\|^2$.
   (b) (4%) Find the first principal axis after performing PCA on this dataset.
   (c) (3%) Denote $\mathbf{u}_i$ as the projection of $\mathbf{x}_i$ to the first principal axis. Find $\frac{1}{N}\sum_{i=1}^{N}\|\mathbf{u}_i\|^2$.

2. **(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 $\mathbf{w}=[w_1 \ \cdots \ w_m]^\mathsf{T}$ that minimizes the following L^2-regularized weighed squared error loss function

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

   where κ_i and λ_j are fixed positive numbers.

3. **(20%) Revisit of binary classification with generative models**
Consider the setting in binary classification where the positive and negative samples are assumed to be independently generated from two Gaussian distributions with identical convariance matrices. More elaborately, consider generative models p_θ parameterized by $\theta=(\boldsymbol{\mu}_1,\boldsymbol{\mu}_2,\Sigma)$ for which the positive sample $\mathbf{x}^+$ and negative sample $\mathbf{x}^-$ are generated by

$$\mathbf{x}^+\sim\mathcal{N}(\boldsymbol{\mu}_1,\Sigma),\ \ \mathbf{x}^-\sim\mathcal{N}(\boldsymbol{\mu}_2,\Sigma).$$

(a) (8%) Suppose Professor Wu first flips a coin, which with $p_1=3/4$ probability will turn heads and $p_2=1/4$ probability will turn tails. Depending on the outcome of the coin:

- If heads, Professor Wu then writes "positive" on a piece of paper and draws a positive sample $\mathbf{x}$ from generative model p_θ.
- If tails, Professor Wu then writes "negative" on a piece of paper and draws a negative sample $\mathbf{x}$ from generative model p_θ.

Professor Wu now tells you $\mathbf{x}$ and asks you to guess what he wrote on that piece of paper. Suppose you will gain r_1=\$20 if you correctly guess the sample to be positive, and r_2=\$4 if you correctly guess it to be negative. How would you make a guess so as to maximize your expected gain? Please write down the formulas based on which you make your guess (the more explicit the better), and explain why.

Suppose we are now given training data $\mathcal{X}$ which consists of N_1 positive samples $\mathbf{x}_1^+,...\mathbf{x}_{N_1}^+$ and N_2 negative samples $\mathbf{x}_1^-,...\mathbf{x}_{N_2}^-$ (here we assume N_1 and N_2 are both positive).

(b) (4%) Show that the log likelihood function is given by (denote $N=N_1+N_2$)

$$L(\theta)=N\log\frac{1}{\sqrt{(2\pi)^d|\Sigma|}}-\frac{1}{2}\sum_{i=1}^{N_1}(\mathbf{x}_i^+-\boldsymbol{\mu}_1)^\intercal\Sigma^{-1}(\mathbf{x}_i^+-\boldsymbol{\mu}_1)-\frac{1}{2}\sum_{i=1}^{N_2}(\mathbf{x}_i^--\boldsymbol{\mu}_2)^\intercal\Sigma^{-1}(\mathbf{x}_i^--\boldsymbol{\mu}_2). \tag{1}$$

For simplicity, from now on we only consider generative models with isotropic Gaussian distributions, namely $\Sigma=\rho\mathbf{I}$ where $\rho>0$. In such cases, $\theta=(\boldsymbol{\mu}_1,\boldsymbol{\mu}_2,\rho)$ and that (1) can be simplified as (assuming each data point is d-dimensional)

$$L(\theta)=N\log\frac{1}{(2\pi\rho)^{d/2}}-\frac{1}{2\rho}\left(\sum_{i=1}^{N_1}\|\mathbf{x}_i^+-\boldsymbol{\mu}_1\|^2+\sum_{i=1}^{N_2}\|\mathbf{x}_i^--\boldsymbol{\mu}_2\|^2\right). \tag{2}$$

(c) (8%) The maximum likelihood estimator is given by

$$\hat{\theta}=(\hat{\boldsymbol{\mu}}_1,\hat{\boldsymbol{\mu}}_2,\hat{\rho})\in\operatorname*{argmax}_\theta L(\theta).$$

Write down the explicit forms of $\hat{\boldsymbol{\mu}}_1,\hat{\boldsymbol{\mu}}_2,\hat{\rho}$.

4. **(20%) Gradient boosting of rank-based classifier**
Let $\mathcal{X}$ be the input space, H be a collection of binary classifiers that map from $\mathcal{X}$ to $\{\pm 1\}$. Let $((x_i,y_i))_{i=1}^m$ be the training data set, where $x_i \in \mathcal{X}$ and $y_i \in \{\pm 1\}$. 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:

$$L(g) = \sum_{i \in \mathcal{I}_+} \sum_{j \in \mathcal{I}_-} e^{g(x_j) - g(x_i)}.$$

where $\mathcal{I}_+ = \{i : y_i = 1\}$ and $\mathcal{I}_- = \{i : y_i = -1\}$ denote the positive and negative classes, respectively.
Note: In other words, by minimizing the loss function, we aim to find a regression model which in general assigns greater values to positive samples than to negative samples. Such regression model can be readily turned into a classifier by setting a threshold.

5. **(20%) A coin-flipping experiment**
Professor Wu asks his assistant Yang-Yang to perform a coin-flipping experiment, in which Yang-Yang is given a pair of coins A and B of unknown biases θ_A and θ_B, respectively (that is, on any given flip, coin A will land on heads with probability θ_A and tails with probability $1-\theta_A$ and similarly for coin B). The goal is to estimate $\theta = (\theta_A, \theta_B)$ by repeating the following procedure N times: randomly choose one of the two coins (with equal probability), and perform M independent coin tosses with the selected coin. Thus, the entire procedure involves a total of MN coin tosses.

During the experiment, Yang-Yang keeps track of two vectors $x = (x_1,...,x_N)$ and $z = (z_1,...,z_N)$, where $x_i \in [\![0,M]\!]$ is the number of heads observed during the ith set of tosses, and $z_i \in \{A,B\}$ is the identity of the coin used during the ith set of tosses.

After the experiment is done, Yang-Yang got so carried away that he accidentally spilled ink, smudging the experiment records of z. As a kind-hearted classmate of Yang-Yang who has learned Expectation Maximization in class, could you please help Yang-Yang and Professor Wu to estimate θ_A and θ_B only with the record x? Please derive the E-step and M-step explicitly.

6. **(25%) Rank-based SVM**

Given $C>0$ and data points $\mathbf{x}_1,...,\mathbf{x}_N \in \mathbb{R}^M$ and their labels $y_1,...,y_N \in \{\pm 1\}$, where each $\mathbf{x}_i=[x_{i,1},...,x_{i,M}]^\mathsf{T}$ is a column vector. Denote $\mathcal{I}_+=\{i:y_i=1\}$ and $\mathcal{I}_-=\{i:y_i=-1\}$ as the positive and negative classes, respectively.

Our goal here is to find a function $h(\mathbf{x})=\mathbf{w}^\mathsf{T}\mathbf{x}$, which in general assigns higher values to positive samples than to negative samples, with $\mathbf{w}$ determined by solving the following optimization problem:

$$\begin{array}{ll} \text{minimize} & \frac{1}{2}\|\mathbf{w}\|_2^2+\sum_{i\in\mathcal{I}_+}\sum_{j\in\mathcal{I}_-}C_{i,j}\xi_{i,j} \\ \text{subject to} & \mathbf{w}^\mathsf{T}(\mathbf{x}_i-\mathbf{x}_j)+\xi_{i,j}\geq 1 \quad (i\in\mathcal{I}_+,j\in\mathcal{I}_-) \\ & \xi_{i,j}\geq 0 \\ \text{variables} & \mathbf{w}\in\mathbb{R}^M,\boldsymbol{\xi}=[\xi_{i,j}]_{i\in\mathcal{I}_+,j\in\mathcal{I}_-}\in\mathbb{R}^{\mathcal{I}_+\times\mathcal{I}_-} \end{array} , \tag{3}$$

where $C_{i,j}>0$ are panelty terms. We may rewrite (3) in primal form:

$$\begin{array}{ll} \text{minimize} & f(\mathbf{w},\boldsymbol{\xi})=\frac{1}{2}\|\mathbf{w}\|_2^2+\sum_{i\in\mathcal{I}_+}\sum_{j\in\mathcal{I}_-}C_{i,j}\xi_{i,j} \\ \text{subject to} & g_{1,i,j}(\mathbf{w},\boldsymbol{\xi})=1-\mathbf{w}^\mathsf{T}(\mathbf{x}_i-\mathbf{x}_j)-\xi_{i,j}\leq 0 \quad (i\in\mathcal{I}_+,j\in\mathcal{I}_-) \\ & g_{2,i,j}(\mathbf{w},\boldsymbol{\xi})=-\xi_{i,j}\leq 0 \\ \text{variables} & \mathbf{w}\in\mathbb{R}^M,\boldsymbol{\xi}=[\xi_{i,j}]_{i\in\mathcal{I}_+,j\in\mathcal{I}_-}\in\mathbb{R}^{\mathcal{I}_+\times\mathcal{I}_-} \end{array} . \tag{4}$$

as well as its dual problem as

$$\begin{array}{ll} \text{maximize} & \theta(\alpha,\beta)=\inf_{\mathbf{w}\in\mathbb{R}^M,\boldsymbol{\xi}\in\mathbb{R}^{\mathcal{I}_+\times\mathcal{I}_-}} L(\mathbf{w},\boldsymbol{\xi},\alpha,\beta) \\ \text{subject to} & \alpha_{i,j}\geq 0,\ \beta_{i,j}\geq 0 \quad (i\in\mathcal{I}_+,j\in\mathcal{I}_-) \\ \text{variables} & \alpha_{i,j}\in\mathbb{R},\ \beta_{i,j}\in\mathbb{R} \quad (i\in\mathcal{I}_+,j\in\mathcal{I}_-) \end{array} . \tag{5}$$

where L denotes the Lagranian.

(a) (3%) Write down $L(\mathbf{w},\boldsymbol{\xi},\alpha,\beta)$ in closed form.

(b) (4%) Write down $\theta(\alpha,\beta)$ in closed form.

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

$$\begin{array}{ll} \text{maximize} & \sum_{i\in\mathcal{I}_+}\sum_{j\in\mathcal{I}_-}\alpha_{i,j}-\frac{1}{2}\left\|\sum_{i\in\mathcal{I}_+}\sum_{j\in\mathcal{I}_-}\alpha_{i,j}(\mathbf{x}_i-\mathbf{x}_j)\right\|^2 \\ \text{variables} & 0\leq\alpha_{i,j}\leq C_{i,j} \quad (i\in\mathcal{I}_+,j\in\mathcal{I}_-) \end{array} . \tag{6}$$

(d) (3%) Show that (4) satisfies Slator's conditions.

(e) (4%) Write down the KKT conditions corresponding to (4) and (5).

(f) (3%) Let $(\bar{\mathbf{w}},\bar{\boldsymbol{\xi}})$ and $(\bar{\alpha},\bar{\beta})$ be optimal primal and dual solutions to (4) and (5), respectively. Show that

$$\bar{\xi}_{i,j}=\max(1-\mathbf{w}^\mathsf{T}(\mathbf{x}_i-\mathbf{x}_j),0),\ \ \bar{\alpha}_{i,j}=\begin{cases} 0 & \text{, if } \bar{\mathbf{w}}^\mathsf{T}\mathbf{x}_i>\bar{\mathbf{w}}^\mathsf{T}\mathbf{x}_j+1 \\ C_{i,j} & \text{, if } \bar{\mathbf{w}}^\mathsf{T}\mathbf{x}_i<\bar{\mathbf{w}}^\mathsf{T}\mathbf{x}_j+1 \end{cases} .$$

(g) (5%) Suppose each sample (both positive and negative) $\mathbf{x}_i$ is the image of some $\mathbf{z}_i\in\mathbb{R}^m$ through some mapping $\Phi:\mathbb{R}^m\to\mathbb{R}^M$. Suppose $M\gg N\gg m$, while the kernel function $K(\mathbf{z},\mathbf{z}')=\Phi(\mathbf{z})^\mathsf{T}\Phi(\mathbf{z}')$ can be easily computed. Show how the kernel trick can be applied to compute $h(\Phi(\mathbf{z}))$ efficiently. (Don't forget to also show how the "weight" of each training sample should be effectively solved)

7. **(15%) Markov Decision Process (MDP)**
Consider a discounted MDP $M=(\mathcal{S},\mathcal{A},P,r,\gamma,\mu)$, where $\mathcal{S}$ is the state space, $\mathcal{A}$ is the action space, $P:\mathcal{S}\times\mathcal{A}\to\Delta(S)$ is the transition probability, $r:\mathcal{S}\times\mathcal{A}\to[0,1]$ is the reward function, $\gamma\in[0,1)$ is the discount factor, $\mu\in\Delta(S)$ is the initial state distribution, for which $\Delta(S)$ denotes the space of probability distributions over $\mathcal{S}$. For each policy π, denote the value function V^π and action-value function Q^π as follows:

$$V^\pi(s)=\mathbb{E}\left[\sum_{t=0}^{\infty}\gamma^t r(s_t,a_t)\middle|\pi,s_0=s\right]$$
$$Q^\pi(s,a)=\mathbb{E}\left[\sum_{t=0}^{\infty}\gamma^t r(s_t,a_t)\middle|\pi,s_0=s,a_0=a\right].$$

Denote V^* and Q^* as the optimal value function and optimal action-value function, respectively. You may assume both $\mathcal{S}$ and $\mathcal{A}$ are finite.

(a) (2%) Show that both V^π and Q^π are bounded by $\frac{1}{1-\gamma}$.
(b) (2%) Write down the Bellman consistency equations.
(c) (2%) Write down the Bellman optimality equations.
(d) (3%) Denote $\mathbb{T}$ as the Bellman optimality operator, and suppose $Q:\mathcal{S}\times\mathcal{A}\to[0,1]$. Given $\epsilon>0$, find $N(\epsilon)\in\mathbb{N}$ such that

$$|(\mathbb{T}^k Q)(s,a)-Q^*(s,a)|<\epsilon \quad \forall(s,a)\in\mathcal{S}\times\mathcal{A},\ k\geq N(\epsilon).$$

Please find $N(\epsilon)$ as small as possible, and elaborate your derivations.
(e) (6%) Let π be a policy satisfying $\pi=\pi_{Q^\pi}$, where π_{Q^π} denotes the greedy policy w.r.t. Q^π. Show that $Q^\pi=Q^*$.