EE 5184 Machine Learning, Fall 2025
Final Exam

Lecturer: Pei-Yuan Wu
TAs: Lu-Fang Chiang, Yi-Chen Lee, Yi-Lun Lee
TAs: Yu-Cheng Lin, Che-Wei Tsai

December 19, 2025

Students are required to read, write out, and sign the following Honor Code Declaration on the answer sheet before beginning the examination. Answer sheets that do not satisfy this requirement will not be graded.

Honor Code Declaration

I pledge my honour that this examination paper represents my own work and has been completed independently in full accordance with University regulations. I confirm that the only permitted reference material used during this examination is one handwritten or printed double-sided A4 sheet, with no additional attachments, inserts, or adhesives, and that no unauthorized assistance has been received or provided.

Signature: _____ Date: _____

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.
- For a $\mathbb{R}^m$-valued random column vector $\mathbf{x}$ with mean $\boldsymbol{\mu} = \mathbb{E}[\mathbf{x}]$, the covariance matrix is
$$\mathbb{V}[\mathbf{x}] = \mathbb{E}[(\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})^\mathsf{T}].$$
- $\mathbf{I}$ denotes the identity matrix. We may use notation $\mathbf{I}_m$ to emphasize it as an $m \times m$ identity matrix.

1. **(15 pts) Logistic Regression**
Consider a binary classification problem with training data $((\mathbf{x}_i,y_i))_{i=1}^n$ where $\mathbf{x}_i\in\mathbb{R}^m$ and $y_i\in\{0,1\}$. Assume the logistic regression model parameterized by $\theta=(\mathbf{w},b)\in\mathbb{R}^m\times\mathbb{R}$,

$$P_\theta(Y=1|X=\mathbf{x})=\sigma(\mathbf{w}^\mathsf{T}\mathbf{x}+b), \quad \text{where } \sigma(z)=\frac{1}{1+e^{-z}}.$$

   (a) (3 pts) Show that maximizing the likelihood of the data is equivalent to minimizing the empirical cross-entropy loss

$$\mathcal{L}(\mathbf{w},b)=\frac{1}{n}\sum_{i=1}^{n}[-y_i\log\sigma(z_i)-(1-y_i)\log(1-\sigma(z_i))], \quad z_i=\mathbf{w}^\mathsf{T}\mathbf{x}_i+b.$$

   (b) (3 pts) Explain why squared error loss is generally inappropriate for logistic regression in classification problems.

   (c) (3 pts) Derive the gradient of $\mathcal{L}(\mathbf{w},b)$ with respect to $\mathbf{w}$ and b.

   (d) (3 pts) Explain why gradient descent is guaranteed to converge to a *global* minimizer (assuming a sufficiently small step size).

   (e) (3 pts) Show that logistic regression induces a linear decision boundary in feature space.

2. **(25 pts) Hat Matrix in Linear Regression**
Consider a linear model where the outputs are contaminated by noise:

$$\mathbf{y}=\mathbf{X}\mathbf{w}+\boldsymbol{\epsilon}, \;\; \boldsymbol{\epsilon}\sim\mathcal{N}(0,\sigma^2\mathbf{I}_n),$$

where $\mathbf{w}\in\mathbb{R}^p$, and the feature matrix $\mathbf{X}\in\mathbb{R}^{n\times p}$ has full rank $p\leq n$. In class, we obtain the unbiased estimator $\hat{\mathbf{w}}$ for $\mathbf{w}$ which can be written as

$$\hat{\mathbf{w}}=f(\mathbf{X},\mathbf{y}):=(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}\mathbf{y}.$$

The vector of fitted values of the dependent variable is given by:

$$\hat{\mathbf{y}}=\mathbf{X}\hat{\mathbf{w}}=\mathbf{H}\mathbf{y},$$

where $\mathbf{H}=g(\mathbf{X}):=\mathbf{X}(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}\in\mathbb{R}^{n\times n}$ is the *hat matrix* for $\mathbf{X}$, and the vector of residuals is given by:

$$\hat{\boldsymbol{\epsilon}}=(\mathbf{I}_n-\mathbf{H})\mathbf{y}.$$

   (a) (2 pts) Prove that $\mathbf{H}$ is an orthogonal projection matrix. i.e. $\mathbf{H}^\mathsf{T}=\mathbf{H}$ and $\mathbf{H}^2=\mathbf{H}$.

   (b) (4 pts) Prove that $\mathbb{V}[\hat{\mathbf{y}}]=\sigma^2\mathbf{H}$ and $\mathbb{V}[\hat{\boldsymbol{\epsilon}}]=\sigma^2(\mathbf{I}_n-\mathbf{H})$.

   (c) (2 pts) Prove that $\mathrm{Tr}(\mathbf{H})=p$.

   (d) (2 pts) Prove that $g(\mathbf{X})=g(\mathbf{X}\mathbf{G})$ for any invertible matrix $\mathbf{G}\in\mathbb{R}^{p\times p}$.

(e) (4 pts) Consider the case where $\mathbf{X}$ is $n\times(p+1)$ with a constant term and p independent variables for each row:

$$\mathbf{X}=\begin{bmatrix} 1 & x_{11} & \cdots & x_{1p} \\ 1 & x_{21} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{np} \end{bmatrix} \quad \mathbf{G}=\begin{bmatrix} 1 & -\bar{x}_1 & -\bar{x}_2 & \cdots & -\bar{x}_p \\ 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ 0 & \vdots & \vdots & \ddots & 0 \\ 0 & 0 & 0 & \cdots & 1 \end{bmatrix}$$

where $\bar{x}_j=\dfrac{\sum_{i=1}^{n} x_{ij}}{n}$.

(i) (2 pts) Under the transformation $\mathbf{X}'=\mathbf{XG}$, derive the unbiased estimator $\hat{\mathbf{w}}'=f(\mathbf{X}',\mathbf{y})$ in terms of $\hat{\mathbf{w}}$ and $\mathbf{G}$.

(ii) (2 pts) Write

$$\hat{\mathbf{w}}'=\begin{bmatrix} \hat{w}'_0 \\ \hat{w}'_1 \\ \vdots \\ \hat{w}'_p \end{bmatrix}, \quad \hat{\mathbf{w}}=\begin{bmatrix} \hat{w}_0 \\ \hat{w}_1 \\ \vdots \\ \hat{w}_p \end{bmatrix}.$$

Express each $\hat{w}'_i$ in terms of $\hat{w}_0$,...,$\hat{w}_p$ and $\bar{x}_1$,...,$\bar{x}_p$.

(f) (5 pts) Show that $(\mathbf{I}_n+\mathbf{uv}^\mathsf{T})^{-1}$, should it exists, must take the form $\mathbf{I}_n-\mathbf{ua}^\mathsf{T}$ for some column vector $\mathbf{a}$. Give an expression of $\mathbf{a}$ in terms of $\mathbf{u}$ and $\mathbf{v}$. For what values of $\mathbf{v}^\mathsf{T}\mathbf{u}$ will $\mathbf{I}_n+\mathbf{uv}^\mathsf{T}$ be invertible? (Hint: Let $\mathbf{A}=(\mathbf{I}_n+\mathbf{uv}^\mathsf{T})^{-1}$ and consider $(\mathbf{I}_n+\mathbf{uv}^\mathsf{T})\mathbf{A}=\mathbf{I}_n$.)

(g) (2 pts) If the i-th row (whose transpose is denoted as the column vector $\mathbf{x}_i$) is removed from the training data matrix $\mathbf{X}$, show that

$$\mathbf{X}_{-i}^\mathsf{T}\mathbf{X}_{-i}=\mathbf{X}^\mathsf{T}\mathbf{X}-\mathbf{x}_i\mathbf{x}_i^\top, \quad \mathbf{X}_{-i}^\mathsf{T}\mathbf{y}_{-i}=\mathbf{X}^\mathsf{T}\mathbf{y}-\mathbf{x}_i y_i$$

where $\mathbf{X}_{-i},\mathbf{y}_{-i}$ denote the matrix/vector obtained by deleting the i-th row of $\mathbf{X},\mathbf{y}$, respectively.

(h) (4 pts) Let $\hat{\mathbf{w}}_{-i}=f(\mathbf{X}_{-i},\mathbf{y}_{-i})$. Show that

$$\hat{\mathbf{w}}-\hat{\mathbf{w}}_{-i}=\frac{(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{x}_i\hat{\epsilon}_i}{1-h_{ii}},$$

where $\hat{\epsilon}_i=y_i-\mathbf{x}_i^\mathsf{T}\hat{\mathbf{w}}$, and $h_{ii}=\mathbf{x}_i^\mathsf{T}(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{x}_i$.
(Hint: You may verify that $(\mathbf{A}+\mathbf{uv}^\mathsf{T})^{-1}=\mathbf{A}^{-1}-\frac{\mathbf{A}^{-1}\mathbf{uv}^\mathsf{T}\mathbf{A}^{-1}}{1+\mathbf{v}^\mathsf{T}\mathbf{A}^{-1}\mathbf{u}}$ whenever $\mathbf{A}$ is invertible and $1+\mathbf{v}^\mathsf{T}\mathbf{A}^{-1}\mathbf{u}\neq 0$.)

3. **(15 pts) 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 $y_i \in \mathbb{R}$. 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 $t \in [\![1,T]\!]$, so as to minimize the following loss function:

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

Recall that with gradient boosting, we may initialize $g_1 = 0$, and for each iteration t we update $g_{t+1} = g_t + \alpha_t h_t$ by

$$h_t \in \operatorname*{argmin}_{h \in H} \frac{\partial}{\partial \alpha} L(g_t + \alpha h)\Big|_{\alpha \downarrow 0}, \quad \alpha_t \in \operatorname*{argmin}_{\alpha \geq 0} L(g_t + \alpha h_t).$$

(a) (3 pts) Let $f_{x,y}(\alpha) = |x + \alpha y|$ where $x, y \in \mathbb{R}$, compute $\lim_{\alpha \downarrow 0} \frac{f_{x,y}(\alpha) - f_{x,y}(0)}{\alpha}$.

(b) (6 pts) Show that

$$h_t \in \operatorname*{argmin}_{h \in H} \left(\sum_{i: g_t(x_i) > y_i} h(x_i) - \sum_{i: g_t(x_i) < y_i} h(x_i) + \sum_{i: g_t(x_i) = y_i} |h(x_i)| \right)$$

$$\alpha_t \in \operatorname*{argmin}_{\alpha \geq 0} \sum_{i \in \mathcal{I}_t} |h_t(x_i)| |\alpha - \theta_t^i|$$

where $\mathcal{I}_t = \{i \in [\![1,N]\!] : h_t(x_i) \neq 0\}$, $\theta_t^i = \frac{y_i - g_t(x_i)}{h_t(x_i)}$.

(c) (6 pts) It turns out that

$$\operatorname*{argmin}_{\alpha \geq 0} \sum_{i \in \mathcal{I}_t} |h_t(x_i)| |\alpha - \theta_t^i| = [a_t, b_t].$$

Give the closed form expression for a_t and b_t.

4. **(15 pts) EM for Mixture of Poisson Models**
Consider the generative model parameterized by $\theta = ((\pi_k, \lambda_k))_{k=1}^K$, where $\lambda_k > 0$, $\pi_k > 0$ for each k, and $\sum_{k=1}^K \pi_k = 1$. The probability mass function (PMF) of generating a count $x \in \{0,1,2,...\}$ is

$$p_\theta(x) = \sum_{k=1}^{K} \pi_k \text{Pois}(x; \lambda_k),$$

where

$$\text{Pois}(x; \lambda) = \frac{\lambda^x e^{-\lambda}}{x!}$$

is the Poisson PMF with rate λ. Suppose we observe N i.i.d. count samples $x_1,...,x_N$, the maximum likelihood estimator is

$$\theta^{\text{opt}}=\underset{\theta}{\text{argmax}}\prod_{i=1}^{N}p_\theta(x_i).$$

Derive the E-step and M-step equations of the EM algorithm for estimating the priors π_k and rates λ_k.

5. **(15 pts) Optimality of Principal Component Analysis (PCA)**
In class we have introduced Eckart-Young-Mirsky theorem as stated below:

Theorem 1 (Eckart-Young-Mirsky)**.** *Let $\mathbf{A}\in\mathbb{R}^{m\times n}$ with singular value decomposition $\mathbf{A}=\mathbf{U}\Sigma\mathbf{V}^\mathsf{T}$, where $\mathbf{U}\in\mathbb{R}^{m\times m}$ and $\mathbf{V}\in\mathbb{R}^{n\times n}$ are orthogonal matrices, $\Sigma\in\mathbb{R}^{m\times n}$ is a diagonal matrix with non-increasing diagonal entries $\sigma_1\geq\sigma_2\geq\cdots\geq\sigma_{m\wedge n}\geq 0$. Let $k\leq m\wedge n$, then both low rank approximation problems*

$$\begin{array}{lll}\underset{\mathbf{A}_k\in\mathbb{R}^{m\times n}}{\text{minimize}} & \|\mathbf{A}-\mathbf{A}_k\|_2 & \text{subject to} \quad \text{rank}(\mathbf{A}_k)\leq k \\ \underset{\mathbf{A}_k\in\mathbb{R}^{m\times n}}{\text{minimize}} & \|\mathbf{A}-\mathbf{A}_k\|_F & \text{subject to} \quad \text{rank}(\mathbf{A}_k)\leq k\end{array}$$

have optimal solution $\mathbf{A}_k=\sum_{i=1}^{k}\sigma_i\mathbf{u}_i\mathbf{v}_i^\mathsf{T}$. Here $\mathbf{u}_i$ and $\mathbf{v}_i$ denote the i'th column in matrices $\mathbf{U}$ and $\mathbf{V}$, respectively.

Explain how the Eckart-Young-Mirsky theorem justifies PCA as an optimal low-dimensional representation of the data.

6. **(10 pts) Markov Decision Process (MDP)**
Consider a discounted MDP $M=(\mathcal{S},\mathcal{A},P,r,\gamma,\mu)$ defined as follows:

States. $\mathcal{S}=\{A,B,C\}$, where C is the terminal state.

Actions.

- In state A, the available actions are $\mathcal{A}(A)=\{\alpha_1,\alpha_2\}$
- In state B, the available actions are $\mathcal{A}(B)=\{\beta_1,\beta_2\}$
- State C has no available actions

Transitions and Rewards.

- From state A:
  - Action α_1: $P(B|A,\alpha_1)=0.75$, $P(C|A,\alpha_1)=0.25$, $r(A,\alpha_1)=2$
  - Action α_2: $P(C|A,\alpha_2)=1$, $r(A,\alpha_2)=4$
- From state B:
  - Action β_1: $P(A|B,\beta_1)=0.625$, $P(C|B,\beta_1)=0.375$, $r(B,\beta_1)=1$
  - Action β_2: $P(C|B,\beta_2)=1$, $r(B,\beta_2)=3$

The discount factor is $\gamma=0.8$. State C is terminal state. Upon reaching C, the process terminates and no further rewards are obtained.

For any policy π, the value function $V^\pi(s)$ is the expected discounted return starting from state s:

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

The optimal value function is defined by $V^*(s)=\max_\pi V^\pi(s)$.

(a) (4 pts) Given deterministic stationary policy π where $\pi(A)=\alpha_1$, $\pi(B)=\beta_1$:
   (i) (2 pts) Write the Bellman consistency equations for $V^\pi(A)$ and $V^\pi(B)$.
   (ii) (2 pts) Solve for $V^\pi(A)$ and $V^\pi(B)$.

(b) (4 pts) Using the Bellman optimality equations, express $V^*(A)$ and $V^*(B)$ as a system of equations, and solve this system to obtain $V^*(A)$ and $V^*(B)$.

(c) (2 pts) Give an optimal policy, and justify why such policy is optimal.

7. **(25 pts) Weight Bounded SVM**

Given data points $\mathbf{x}_1,...,\mathbf{x}_N\in\mathbb{R}^m$ as well as their labels $y_1,...,y_N\in\{\pm1\}$ and penalty 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 weight bounded SVM problem:

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

where $L_j<U_j$ are given bounds on each weight component. This formulation is useful when we have prior knowledge about the magnitude of feature importance or want to prevent any single feature from dominating.

The original problem (1) can be equivalently written as:

$$\begin{array}{ll}\text{minimize} & f(\mathbf{w},b,\boldsymbol{\xi})=\frac{1}{2}\|\mathbf{w}\|^2+\sum_{i=1}^N C_i\xi_i \\ \text{subject to} & g_i^1(\mathbf{w},b,\boldsymbol{\xi})=1-\xi_i-y_i(\mathbf{w}^\mathsf{T}\mathbf{x}_i+b)\leq 0,\quad i\in[\![1,N]\!] \\ & g_i^2(\mathbf{w},b,\boldsymbol{\xi})=-\xi_i\leq 0,\quad i\in[\![1,N]\!] \\ & g_j^3(\mathbf{w},b,\boldsymbol{\xi})=w_j-U_j\leq 0,\quad j\in[\![1,m]\!] \\ & g_j^4(\mathbf{w},b,\boldsymbol{\xi})=L_j-w_j\leq 0,\quad j\in[\![1,m]\!] \\ \text{variables} & \mathbf{w}\in\mathbb{R}^m, b\in\mathbb{R}, \boldsymbol{\xi}\in\mathbb{R}^N\end{array} \tag{2}$$

as well as its Lagrangian dual problem:

$$\begin{array}{ll}\text{maximize} & \theta(\alpha,\beta,\lambda,\boldsymbol{\rho})=\inf_{\mathbf{w},b,\boldsymbol{\xi}} L(\mathbf{w},b,\boldsymbol{\xi},\alpha,\beta,\lambda,\boldsymbol{\rho}) \\ \text{subject to} & \alpha_i\geq 0,\beta_i\geq 0,\quad i\in[\![1,N]\!] \\ & \lambda_j\geq 0,\rho_j\geq 0,\quad j\in[\![1,m]\!] \\ \text{variables} & \alpha\in\mathbb{R}^N,\beta\in\mathbb{R}^N,\lambda\in\mathbb{R}^m,\boldsymbol{\rho}\in\mathbb{R}^m\end{array} \tag{3}$$

where L denotes the Lagrangian function, with dual variables α_i,β_i,λ_i,ρ_i associate to constraints g_i^1,g_i^2,g_i^3,g_i^4 respectively.

(a) (2 pts) Write down $L(\mathbf{w},b,\boldsymbol{\xi},\alpha,\beta,\lambda,\boldsymbol{\rho})$ in closed form.

(b) (3 pts) Show that (2) satisfies Slater's condition.

(c) (4 pts) Show that:

(i) (2 pts) $\theta(\alpha,\beta,\lambda,\boldsymbol{\rho})=-\infty$ unless the following conditions hold:

$$\sum_{i=1}^{N}\alpha_i y_i=0 \tag{4a}$$

$$\alpha_i+\beta_i=C_i \quad \forall i\in[\![1,N]\!]. \tag{4b}$$

(ii) (2 pts) The stationary condition holds if and only if (4) and $\mathbf{w}=\sum_{i=1}^{N}\alpha_i y_i\mathbf{x}_i-\lambda+\boldsymbol{\rho}$ are satisfied.

(d) (3 pts) Show that the dual problem (3) can be simplified as:

$$\begin{array}{ll}\text{maximize} & \sum_{i=1}^{N}\alpha_i-\frac{1}{2}\left\|\sum_{i=1}^{N}\alpha_i y_i\mathbf{x}_i-\lambda+\boldsymbol{\rho}\right\|^2-\sum_{j=1}^{m}(U_j\lambda_j-L_j\rho_j)\\ \text{subject to} & \sum_{i=1}^{N}\alpha_i y_i=0\\ & 0\le\alpha_i\le C_i,\quad i\in[\![1,N]\!]\\ & \lambda_j\ge 0,\rho_j\ge 0,\quad j\in[\![1,m]\!]\\ \text{variables} & \alpha\in\mathbb{R}^N,\lambda\in\mathbb{R}^m,\boldsymbol{\rho}\in\mathbb{R}^m.\end{array} \tag{5}$$

(e) (4 pts) Write down the complete KKT conditions for the primal (2) and dual problems (3).

(f) (9 pts) Suppose $(\bar{\mathbf{w}},\bar{b},\bar{\boldsymbol{\xi}})$ and $(\bar{\alpha},\bar{\beta},\bar{\lambda},\bar{\boldsymbol{\rho}})$ are optimal primal and dual solutions respectively.

(i) (4.5 pts) For the margin conditions, show that:

- If $y_i(\bar{\mathbf{w}}^\mathsf{T}\mathbf{x}_i+\bar{b})>1$, then $\bar{\xi}_i=0$, $\bar{\alpha}_i=0$.
- If $y_i(\bar{\mathbf{w}}^\mathsf{T}\mathbf{x}_i+\bar{b})=1$, then $\bar{\xi}_i=0$, $0\le\bar{\alpha}_i\le C_i$.
- If $y_i(\bar{\mathbf{w}}^\mathsf{T}\mathbf{x}_i+\bar{b})<1$, then $\bar{\xi}_i=1-y_i(\bar{\mathbf{w}}^\mathsf{T}\mathbf{x}_i+\bar{b})$, $\bar{\alpha}_i=C_i$.

(ii) (4.5 pts) For the weight bounds, show that:

- If $\sum_{i=1}^{N}\bar{\alpha}_i y_i x_{i,j}<L_j$, then $\bar{w}_j=L_j$.
- If $\sum_{i=1}^{N}\bar{\alpha}_i y_i x_{i,j}>U_j$, then $\bar{w}_j=U_j$.
- If $\sum_{i=1}^{N}\bar{\alpha}_i y_i x_{i,j}\in[L_j,U_j]$, then $\bar{w}_j=\sum_{i=1}^{N}\bar{\alpha}_i y_i x_{i,j}$.