AI503: Mathematics for Artificial Intelligence

Prof. Jinwoo Shin

Midterm Exam

(i) Exam is held from 9:00am to 11:45am.
(ii) You should solve the exam questions on a paper (prepared by yourself).
(iii) You should log in to the course zoom link and turn on the camera during the exam to prevent cheating.
(iv) The exam is closed-book and closed-note.
(v) Until 11:45 am, you should scan (or take a picture of) your answers and send the file to jihoontack@kaist.ac.kr. **Late submission is not allowed.**

Problems	Score
Problem 1, (15)	
Problem 2, (15)	
Problem 3, (15)	
Problem 4, (15)	
Problem 5, (20)	
Problem 6, (20)	
Total (100)	

Problem 1 - (15pt)

Prove or disprove the following statements.

(a) For any $n \times n$ matrix A, the largest singular value σ_1 is equal to or larger than all eigenvalues.
(b) For any $n \times n$ matrix A, there exists $n \times n$ positive semidefinite matrix P and $n \times n$ orthogonal matrix Q such that $A = PQ$.
(c) For any $m \times n$ matrix A, matrix $A^T A$ is symmetric and positive definite.

Problem 2 - (15pt)

Consider the following real-world investigation problem known as Markowitz mean-variance portfolio optimization, which solves to minimize the variance by adjusting the ratio between two bi-variate Gaussian distribution investment model. This optimization problem is formulated as:

$$\begin{aligned} \min \quad & f(w_1, w_2) = w_1^2\sigma_1^2 + w_2^2\sigma_2^2 + 2w_1w_2\sigma_{12} \\ \text{subject to} \quad & w_1 + w_2 = 1 \\ & w_1 \geq 0 \\ & w_1 \leq 1 \end{aligned}$$

Solve the above optimization problem with Lagrangian method where $\sigma_1^2 = 0.3, \sigma_2^2 = 0.2, \sigma_{12} = 0.1$.

Problem 3 - (15pt)

Let $\boldsymbol{\theta} = [\theta_0, \theta_1, ..., \theta_D], \mathbf{x}_t \in \mathbb{R}^D, y_t \in \mathbb{R}$ for all $t \in \{1, ..., N\}$. Consider a linear model of the form

$$f(\mathbf{x}, \boldsymbol{\theta}) = \theta_0 + \sum_{i=1}^{D} \theta_i x_i$$

together with a sum-of-squares error function of the form

$$\text{error}_D(\boldsymbol{\theta}) = \frac{1}{2} \sum_{t=1}^{N} \{f(\mathbf{x}_t, \boldsymbol{\theta}) - y_t\}^2$$

Now suppose that Gaussian noise ϵ_i with zero mean and variance σ^2 is added independently to each of the input variables x_i. Show that minimizing error_D averaged over the noise distribution is equivalent to minimizing the sum-of-squares error for noise-free input variables with the addition of some regularization term.

Problem 4 - (15pt)

Suppose we have a sample of N pairs x_i, y_i drawn i.i.d. from the distribution characterized as follows:

$$\begin{aligned} x_i &\sim D(x), && \text{the data distribution with } x_i \in \mathbb{R} \\ y_i &= f(x_i) + \epsilon_i, && \text{the regression function with } y_i \in \mathbb{R} \\ \epsilon_i &\sim \mathcal{N}(0, \sigma^2), && \text{the noise distribution with } \epsilon_i \in \mathbb{R} \end{aligned}$$

We construct an estimator for f linear in the y_i for test data $x_0 \in \mathbb{R}$,

$$\hat{f}(x_0) = \sum_{i=1}^{N} \{w_i(x_0; \mathcal{X}) y_i\},$$

where the weights $w_i(x_0; \mathcal{X})$ do not depend on the y_i, but can depend on the training sequence of x_i, denoted here by $\mathcal{X}$. Show that the least squares estimator is a member of this class of estimators. Describe explicitly the weight $w_i(x_0; \mathcal{X})$ in this case.

Problem 5 - (20pt)

(a) Let

$$J(\mathbf{v}_2, \mathbf{z}_2) = \frac{1}{n} \sum_{i=1}^{n} (\mathbf{x}_i - z_{i1}\mathbf{v}_1 - z_{i2}\mathbf{v}_2)^T (\mathbf{x}_i - z_{i1}\mathbf{v}_1 - z_{i2}\mathbf{v}_2),$$

where $x_i \in \mathbb{R}^D$, $v_i \in \mathbb{R}^D$, $\mathbf{z}_i = [z_{1i}, z_{2i}, \ldots, z_{ni}]^T \in \mathbb{R}^n$ and v_1 is first principal component (eigenvector with largest eigenvalue). Show that $\frac{\partial J}{\partial z_2} = 0$ yields $z_{i2} = \mathbf{v}_2^T \mathbf{x}_i$.

(b) Show that the value of $\mathbf{v}_2$ that minimizes

$$\tilde{J}(\mathbf{v}_2, \lambda_2, \lambda_{12}) = -\mathbf{v}_2^T \mathbf{C} \mathbf{v}_2 + \lambda_2(\mathbf{v}_2^T \mathbf{v}_2 - 1) + \lambda_{12}(\mathbf{v}_2^T \mathbf{v}_1 - 0) \quad (\text{lagrangian multiplier } \lambda_2, \lambda_{12} \geq 0)$$

is given by the eigenvector of $\mathbf{C}$ ($= \frac{1}{n} \sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i^T$) with the second largest eigenvalue.
Hint: orthonormal eigenvector $\mathbf{v}_1$ satisfy $\mathbf{C}\mathbf{v}_1 = \lambda_1 \mathbf{v}_1$ and $\frac{\partial \mathbf{x}^T \mathbf{A} \mathbf{x}}{\partial \mathbf{x}} = (\mathbf{A} + \mathbf{A}^T)\mathbf{x}$.

Problem 6 - (20pt)

Consider a mixture of K Gaussians, $p(\mathbf{x}|\theta) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$,
and define the log-likelihood as

$$l(\theta) = \sum_{n=1}^{N} \log p(\mathbf{x}_n|\theta)$$

Define the posterior responsibility that cluster k has for data point n as follows:

$$r_{nk} \triangleq p(z_n = k|\mathbf{x}_n, \boldsymbol{\theta}) = \frac{\pi_k \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{k'=1}^{K} \pi_{k'} \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_{k'}, \boldsymbol{\Sigma}_{k'})}$$

(a) Show that the gradient of the log-likelihood w.r.t. $\boldsymbol{\mu}_k is$

$$\frac{d}{d\boldsymbol{\mu}_k} l(\boldsymbol{\theta}) = \sum_n r_{nk} \Sigma_k^{-1} (\mathbf{x}_n - \boldsymbol{\mu}_k)$$

(b) One way to handle the constraint that $\sum_{k=1}^{K} \pi_k = 1$ is to reparameterize using the softmax function, i.e., $\pi_k \triangleq \frac{e^{w_k}}{\sum_{k'=1}^{K} e^{w_{k'}}}$. Show that the gradient of the log-likelihood w.r.t w_k is

$$\frac{d}{dw_k} l(\theta) = \sum_n r_{nk} - \pi_k$$

Hint: use the chain rule and the fact that

$$\frac{d\pi_j}{dw_k} = \begin{cases} \pi_j(1 - \pi_j), & \text{if} \quad j = k \\ -\pi_j \pi_k, & \text{if} \quad j \neq k \end{cases}$$