Instructions:

- You have until Tuesday May 5th 2:59pm PST to finish and upload your answers to CCLE. For the complete regulations, see the announcement on CCLE.
- On the first page of your submission, write down your full name, University ID, and the following academic integrity statement, then sign and date.

  > I agree to the UCLA Student Code of Conduct for academic integrity. I agree to use NO resource other than lecture notes and the textbook for this exam. I agree to communicate with NO ONE regarding this exam. Violations of this code will result in immediate FAILURE of this course.

- Use blank sheets of paper to write down answers with requested supporting work.

Problem 1

Given a data set

$$\mathcal{D} = \{(\mathbf{x}^{(i)}, t^{(i)})\}_{i=1}^{N}, \quad \mathbf{x}^{(i)} \in \mathbb{R}^D, t^{(i)} \in \mathbb{R}.$$

Fixing an integer $M \in \mathbb{N}$ and a basis function $\boldsymbol{\phi} : \mathbb{R}^D \to \mathbb{R}^M$. The regression problem is about finding the best parameter $\mathbf{w} \in \mathbb{R}^M$ so

$$t^{(i)} = \mathbf{w}^T \boldsymbol{\phi}(\mathbf{x}^{(i)}) + \boldsymbol{\epsilon}^{(i)}$$

where $\epsilon^{(i)} \sim \mathcal{N}(0, \beta^{-1})$ are independent identical (unbiased) Gaussian noise.

(a) Write down a formula for the likelihood function $p(\mathcal{D}|\mathbf{w}, \beta)$. (10 points)

(b) Show that the maximum likelihood solution

$$\mathbf{w}_\beta^* = \arg\max_{\mathbf{w}} p(\mathcal{D}|\mathbf{w}, \beta)$$

for any value of $\beta > 0$ is the same as the least square solution

$$\overline{\mathbf{w}} = \arg\min_{\mathbf{w}} \frac{1}{2} \left| t_i - y(x^{(i)}, \mathbf{w}) \right|^2. \qquad \text{(10 points)}$$

(c) Fixing the model complexity $M \in \mathbb{N}$, give three examples of the basis function $\boldsymbol{\phi}(\mathbf{x})$. (10 points)

Problem 2

Suppose a data set $\mathcal{D} = \{(\mathbf{x}^{(i)}, t^{(i)})\}_{i=1}^{N}$ is given. $\mathbf{x}^{(i)} \in \mathbb{R}^D, t^{(i)} \in \mathbb{R}$ for $i = 1, \cdots, N$.

(a) Show that the optimal solution $\mathbf{w}^* = \arg\min J(\mathbf{w})$ for a regularized sum-of-squares error function

$$J(\mathbf{w}) = \frac{1}{2}\sum_{i=1}^{N}\left(t^{(i)} - \mathbf{w}^T\boldsymbol{\phi}(\mathbf{x}^{(i)})\right)^2 + \frac{\lambda}{2}\|\mathbf{w}\|^2,$$

where $\lambda > 0$, is a linear combination of the vectors $\{\boldsymbol{\phi}(\mathbf{x}^{(i)})\}_{i=1}^{N}$. In other words, show that

$$\mathbf{w}^* = \sum_{i=1}^{N} a^{(i)}\boldsymbol{\phi}(\mathbf{x}^{(i)})$$

for some scalars $a^{(i)} \in \mathbb{R}, i = 1, \cdots, N$. (10 points)

(b) We define the Gram matrix

$$\mathbf{K} = [\ K_{ij}\ ] = [\ \boldsymbol{\phi}(\mathbf{x}^{(i)})^T\boldsymbol{\phi}(\mathbf{x}^{(j)})\ ] \in \mathbb{R}^{N\times N}.$$

Show that $\mathbf{K}$ is symmetric semi-positive definite. (A matrix $\mathbf{A} \in \mathbb{R}^{N\times N}$ is symmetric semi-positive definite if and only if $\mathbf{x}^T\mathbf{A}\mathbf{x} \geq 0$ for all vector $\mathbf{x} \in \mathbb{R}^N$.) (10 points)

(c) Show that the coefficients from part (a) satisfy

$$(\mathbf{K} + \lambda\mathbf{I}_N)\begin{bmatrix} a^{(1)} \\ a^{(2)} \\ \vdots \\ a^{(N)} \end{bmatrix} = \begin{bmatrix} t^{(1)} \\ t^{(2)} \\ \vdots \\ t^{(N)} \end{bmatrix}. \qquad \text{(10 points)}$$

Problem 3

Consider the two-class classification problem. Denote the data set $\mathcal{D} = \{(\mathbf{x}^{(i)}, t^{(i)})\}_{i=1}^N$ where

$$t^{(i)} = \begin{cases} 1, & i \in C_{[1]} \\ 0, & i \in C_{[2]} \end{cases}$$

is the target variable encoding the class membership.

(a) Suppose $p(\mathbf{x}|C_{[1]}) \sim \mathcal{N}(\boldsymbol{\mu}_{[1]}, \boldsymbol{\Sigma})$ and $p(\mathbf{x}|C_{[2]}) \sim \mathcal{N}(\boldsymbol{\mu}_{[2]}, \boldsymbol{\Sigma})$, that is, data from two classes scatter around different class-specific mean but share the same covariance matrix. Denote $p(C_{[1]}) = \pi$, hence $p(C_{[2]}) = 1 - \pi$. The likelihood function is given by

$$p(\mathcal{D}|\pi, \boldsymbol{\mu}_{[1]}, \boldsymbol{\mu}_{[2]}, \boldsymbol{\Sigma}) = \left(\prod_{i \in C_{[1]}} \pi \mathcal{N}(\mathbf{x}^{(i)}|\boldsymbol{\mu}_{[1]}, \boldsymbol{\Sigma}) \right) \cdot \left(\prod_{i \in C_{[2]}} (1-\pi) \mathcal{N}(\mathbf{x}^{(i)}|\boldsymbol{\mu}_{[2]}, \boldsymbol{\Sigma}) \right)$$

Show that the maximum likelihood estimate of the class probability π is given by the fraction of data points in $C_{[1]}$, i.e.

$$\arg\max_{\pi} p(\mathcal{D}|\pi, \boldsymbol{\mu}_{[1]}, \boldsymbol{\mu}_{[2]}, \boldsymbol{\Sigma}) = \frac{\#\{i : i \in C_{[1]}\}}{N}. \qquad \text{(10 points)}$$

(b) The *logistic sigmoid* function is defined by

$$\sigma(b) = \frac{1}{1 + e^{-b}}.$$

Show that (i) $\sigma(-b) = 1 - \sigma(b)$, (ii) σ is a monotonically increasing function, and (iii) σ maps all of $\mathbb{R}$ onto the interval $(0, 1)$. (15 points)

(c) In an approach different from part (a), we suppose that $p(C_{[1]}|\mathbf{x}) = \sigma(\mathbf{w}^T \boldsymbol{\phi}(\mathbf{x}))$ where $\mathbf{w} \in \mathbb{R}^M$ is a coefficient vector to be trained. According to part (b), $p(C_{[2]}|\mathbf{x}) = \sigma(-\mathbf{w}^T \boldsymbol{\phi}(\mathbf{x}))$. The likelihood function is then given by this different formula,

$$p(\mathcal{D}|\mathbf{w}) = \left(\prod_{i \in C_{[1]}} \sigma(\mathbf{w}^T \boldsymbol{\phi}(\mathbf{x}^{(i)})) \right) \cdot \left(\prod_{i \in C_{[2]}} \sigma(-\mathbf{w}^T \boldsymbol{\phi}(\mathbf{x}^{(i)})) \right).$$

Define the error function $E(\mathbf{w}) = -\log p(\mathcal{D}|\mathbf{w})$ to be the negative logarithm of the likelihood. Compute $\nabla E(\mathbf{w})$ and explain why the maximum likelihood estimate $\nabla E(\mathbf{w}^*) = 0$ doesn't have an analytical solution. (15 points)

(d) **(Bonus)** Provide a strategy to compute the optimal solution $\mathbf{w}^*$ numerically. (up to 10 points)