CS 182: Introduction to Machine Learning, Fall 2022
Homework 2

(Due on Wednesday, Oct. 26 at 11:59pm (CST))

Notice:

- Please submit your assignments via Gradescope. The entry code is G2V63D.
- Please make sure you select your answer to the corresponding question when submitting your assignments.
- Each person has a total of five days to be late without penalty for all the assignments. Each late delivery less than one day will be counted as one day.

1. [20 points] [*Bayesian Decision Theory*]

   (a) Suppose that in an experiment 10,000 in total Salmon and Sea bass were observed and recorded their features, one of which is the lightness of skin. Recorders evaluated the lightness by 10-point criterion. If the lightness value is larger or equal to 5 points, it is labelled as "light"; otherwise "dark". Here gives the table recording the amount of fish in different classes and lightness.

	C_1: Salmon	C_2: Sea bass
$x =$ light	2,125	1,000
$x =$ dark	6,375	500

Table 1: The number of fish in different classes and lightness

   Specify decision rules via likelihood and posterior respectively, and give examples with $x =$ light to use the rules. [10 points]

   (b) For a K-class classification problem, the loss matrix $[\lambda_{ik}]_{i,k=1}^{K}$ records loss values for misclassification. Specifically, λ_{ik} is the loss for classifying a datapoint belonging the class C_k into C_i. And the loss incurred for selecting the reject option with threshold θ is λ. For an input $\mathbf{x}$, if $P(C_k|\mathbf{x}) \leq \theta$ (with $0 \leq \theta \leq 1$), the action of classifying $\mathbf{x}$ into C_k is rejected, even if $\mathbf{x}$ may actually belong to C_k.

   (i) Find the optimal decision rule that will give the minimum expected loss. [5 points]

   (ii) If the loss matrix is given by $[\lambda_{ik}]_{i,k=1}^{K} = \mathbf{1} - \mathbf{I}$, where $\mathbf{1}$ and $\mathbf{I}$ denote the all-one matrix and the identity respectively, and we set $0 \leq \lambda \leq 1$. State the relationship between λ and the rejection threshold θ. [5 points]

2. [20 points] [*Parameter Estimation*] Given a a training set $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^n$, we assume that feature $\mathbf{x}_i \in \mathbb{R}^p$ and label $y_i \in \mathbb{R}$ are related via the equation

$$y_i = \mathbf{w}^T \mathbf{x}_i + e_i, \qquad i = 1, \ldots, n$$

where $\mathbf{w}$ is the parameter to be learned and e_i is an error term.

(a) Let $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_n]^T$ and $\mathbf{y} = [y_1, \ldots, y_n]^T$. Assume that $e_i \sim \mathcal{N}(0, \sigma^2)$ and $\mathbf{X}^T\mathbf{X}$ is a full-rank matrix. Derive the maximum likelihood estimate $\mathbf{w}_{ML} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$. [10 points]

(b) Assume $\mathbf{w} \sim \mathcal{N}(\mathbf{0}, \nu^2\mathbf{I})$. Derive the maximum a posteriori estimate $\mathbf{w}_{MAP} = (\mathbf{X}^T\mathbf{X} + \frac{\sigma^2}{\nu^2}\mathbf{I})^{-1}\mathbf{X}^T\mathbf{y}$. [10 points]

3. [20 points] [*Parameter Estimation*] Consider a multi-class classification problem with L training samples $(\mathbf{x}_1, \mathbf{y}_1), \ldots, (\mathbf{x}_L, \mathbf{y}_L)$, where the input $\mathbf{x}_l \in \mathbb{R}^N$ contains N features and output $\mathbf{y}_l \in \mathbb{R}^M$ is a zero vector with one entry equals one to indicate the class of sample l, $l = 1, \ldots, L$. Suppose we use a two-layer neural network following the batch learning scheme, then the loss function can be defined as

$$\ell\left(\mathbf{W}_1, \mathbf{W}_2, \mathbf{b}_1, \mathbf{b}_2\right) = -\sum_{l=1}^{L} \mathbf{y}_l^T \log\left(\text{softmax}\left(\mathbf{W}_2 \text{sigmoid}\left(\mathbf{W}_1 \mathbf{x}_l + \mathbf{b}_1\right) + \mathbf{b}_2\right)\right),$$

where $\mathbf{W}_1$, $\mathbf{W}_2$, $\mathbf{b}_1$, and $\mathbf{b}_2$ are the training parameters. Derive the update rule for all parameters using gradient descent.

4. [20 points] [*Bayesian Decision Theory, Linear Discrimination*]

(a) Consider a binary classification problem, with the class conditional density

$$p(\mathbf{x} \mid C_i) = \frac{1}{(2\pi)^{\frac{d}{2}} \det(\mathbf{\Sigma}_i)^{\frac{1}{2}}} \exp\left[-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_i)^T \mathbf{\Sigma}_i^{-1}(\mathbf{x}-\boldsymbol{\mu}_i)\right], \ i = 1, 2.$$

Assume that $\mathbf{\Sigma}_i = \sigma^2 \mathbf{I}$ for all i, and that the priors for the two classes are not equal, i.e., $P(C_1) \neq P(C_2)$. If our target is to minimize the expected loss, a.k.a. conditional risk, (as in the lecture notes, we denote by λ_{ij} the loss incurred for assigning an input $\mathbf{x}$ to class C_i when the actual state is C_j), derive the decision boundary, and explain how geometrically it differs from that when one minimizes the misclassification error, a.k.a. probability of error. [10 points]

(b) Assume that $\mathbf{x}$ is 2-dimensional, and that the two dimensions are independent following the Laplace distribution:

$$p(x_j \mid C_i) = \frac{1}{2\sigma} \exp(-\frac{|x_j - \mu_{ij}|}{\sigma}), \ j = 1, 2.$$

By minimizing the misclassification error, obtain and draw the decision boundary when $\mu_{11} = 1$, $\mu_{12} = 1$, $\mu_{21} = 3$, $\mu_{22} = 5$, $\sigma = 1$, and $P(C_1) = P(C_2)$. [10 points]

5. [20 points] [*Coding: Logistic Regression*] Complete "HW2-Coding.ipynb". After completion, you should convert your notebook to PDF, and concatenate the writing part and the coding part into one PDF which is the file to submit.