CS182 - Introduction to Machine Learning, 2022-23 Fall
Delayed Final Exam

10:30AM - 12:30PM, Saturday, Feb. 11th, 2023

4 pages, 6 problems, 100 points in total

Guidelines:

1) The exam is closed-book and closed-notes. You need to finish it all on your own.
2) Please write down your answers on a separate paper.
3) Please submit your answers to Gradescope **no later than 15 minutes** after the exam is finished. The entry code is **G2V63D**. No late submission is accepted.

Problem 1. (17 points) *[Logistic Regression]*

Consider the following binary classification problem. A dataset consists of R datapoints, denoted as $\left\{(\mathbf{x}^{(r)}, y^{(r)})\right\}_{r=1}^{R}$, where $\mathbf{x}^{(r)} \in \mathbb{R}^D$ is the feature of the r-th datapoint and $y^{(r)} \in \{0, 1\}$ is its label. For the classification task, suppose the linear logistic classifier (LLC) is used and learnt, which has the form

$$h(\mathbf{x}; \boldsymbol{\theta}, \theta_0) = \sigma(\boldsymbol{\theta}^{\mathsf{T}}\mathbf{x} + \theta_0) \tag{1}$$

where $\boldsymbol{\theta}$, θ_0 are unknown parameters, $\sigma(a) = \frac{1}{1+e^{-a}}$.

(a) The output of (1) is in the interval $(0, 1)$. Explain the meaning of the output and write the classification rule for LLC. (6 points) If the original label $\tilde{y}^{(r)} \in \{-1, 1\}$, give a linear transform f to map the label $\tilde{y}^{(r)}$ to the label $y^{(r)}$, i.e., $f(\tilde{y}^{(r)}) = \begin{cases} 0, & \text{if } \tilde{y}^{(r)} = -1 \\ 1, & \text{if } \tilde{y}^{(r)} = 1 \end{cases}$. (1 points)

(b) The LLC uses negative log-likelihood as the loss function. Write the expression of the loss $\mathcal{L}$ and derive the gradients $\frac{\partial \mathcal{L}}{\partial \boldsymbol{\theta}}$, $\frac{\partial \mathcal{L}}{\partial \theta_0}$. (10 points)

Problem 2. (17 points) *[Dimensionality reduction]*

(a) Given the data which has 2 dimensions and 2 classes in the following picture, draw the PCA direction and LDA direction on it, and explain the reason. (6 points)

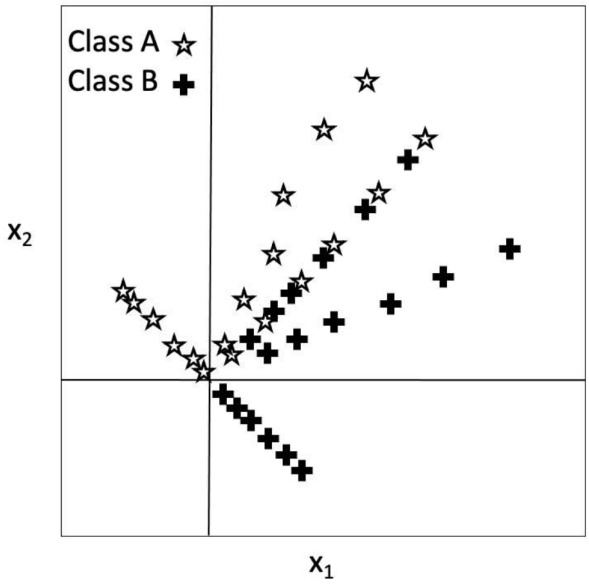


Figure 1: Problem 2(a)

(b) Suppose we have the data which has D dimensions and C classes , what is the range of the dimension of new data after applying LDA? Why? (11 points)

Problem 3. (16 points) *[Multi-layer Perceptron, RNN]*

(a) Consider a single output MLP. After injecting an input $\mathbf{x}^{(n)} \in \mathbb{R}^J$ into the MLP, its output is given by

$$y^{(n)} = \sum_{h=1}^{H} v_h z_h^{(n)} + v_0,$$

where

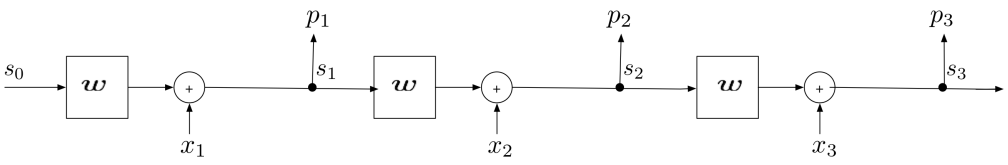
$$z_h^{(n)} = \sigma(\mathbf{w}_h^{\mathsf{T}} \mathbf{x}^{(n)}) = \sigma(\sum_{j=1}^{J} w_{hj} x_j^{(n)}),$$

with $\sigma(a) = e^a$ being the active function for some reasons. Given sample $\{\mathbf{x}^{(n)}, r^{(n)}\}_{n=1}^{N}$, the loss function in learning is defined as

$$\mathcal{L} = \frac{1}{2} \sum_{n=1}^{N} (r^{(n)} - y^{(n)})^2.$$

Compute $\frac{\partial \mathcal{L}}{\partial w_{hj}}$. (8 points)

(b) Consider a simple recurrent neural network (RNN), as shown in the following picture.



At time t, the true label is denoted as y_t, and the progress of the network is

$$s_t = ws_{t-1} + x_t$$

$$p_t = s_t,$$

where s_t, x_t, and p_t is the current state, input, and output, respectively, and w is a scalar parameter. Assume that $s_0 = 0$, and the loss at a single time step is $\mathcal{L}_t(p_t, y_t) = \frac{1}{2}(p_t - y_t)^2$. Considering three steps ($t = 1, 2, 3$), the loss function is $\mathcal{L} = \sum_{t=1}^{3} \mathcal{L}_t$. Derive $\frac{\partial \mathcal{L}}{\partial w}$. (8 points)

Problem 4. (15 points) *[Support vector machine]*

(a) Please discuss the differences between the hard-margin SVMs and the soft-margin SVMs. (7 points)

(b) Given $K(x_1, x_2) = x_1^T x_2, \forall x_1, x_2 \in \mathbb{R}^m$, prove that K is a kernel. (8 points)

Problem 5. (20 points) *[Clustering and Mixture Models, Nonparametric Methods]*

(a) Consider a mixture distribution of the form

$$p(\mathbf{x}) = \sum_{k=1}^{K} \pi_k p(\mathbf{x} \mid k)$$

where the elements of $\mathbf{x}$ could be discrete or continuous or a combination of these. Denote the mean and covariance of $p(\mathbf{x} \mid k)$ by $\boldsymbol{\mu}_k$ and $\boldsymbol{\Sigma}_k$, respectively. Derive that the mean and covariance of the mixture distribution are given by

$$\mathbb{E}[\mathbf{x}] = \sum_{k=1}^{K} \pi_k \boldsymbol{\mu}_k$$

$$\text{cov}[\mathbf{x}] = \sum_{k=1}^{K} \pi_k \left(\boldsymbol{\Sigma}_k + \boldsymbol{\mu}_k \boldsymbol{\mu}_k^T\right) - \mathbb{E}[\mathbf{x}]\mathbb{E}[\mathbf{x}]^T$$

(10 points)

(b) We are given the location of each object and type of object (triangle, circle or square) along with an identifying name (e.g., T1, C1, S1). The data is tabulated below. Try to classify the object located in (2,-1) using k-nearest neighbors estimator with $k = 3$, where the distance measure is computed by

$$|x_1 - x_2| + |y_1 - y_2|$$

for two points (x_1, y_1) and (x_2, y_2). (10 points)

Shape	ID & Location		
Triangle	T1: (-3,-2)	T2: (3,3)	T2: (0,3)
Circle	C1: (3,2)	C2: (-2,-1)	
Square	S1: (7,4)	S2: (4,5)	S3: (3,0)

Problem 6. (15 points) *[Bayesian Decision Theory]* Suppose we have a two-class recognition problem with salmon ($w = 1$) and sea bass ($w = 2$). Suppose we have two features $\mathbf{x} = (x_1, x_2)$ and the two class-conditional densities,

$p(x \mid w = 1)$ and $p(x \mid w = 2)$ are 2D Gaussian distributions centered at points $(4, 11)$ and $(10, 3)$ respectively with the same covariance matrix $\mathbf{\Sigma} = 3\mathbf{I}$. Suppose the priors are $p(w = 1) = 0.6$ and $p(w = 2) = 0.4$.

(a) Write the two discriminant functions $g_1(\mathbf{x})$ and $g_2(\mathbf{x})$ using Bayesian decision theory. (5 points)

(b) Derive the equation for the decision boundary. (5 points)

(c) Based on the above results, classify the point$(6, 6)$. (5 points)

(end of exam paper)