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

1:30PM - 3:30PM, Thursday, Jan. 5th, 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)

(a) *[Bayesian Decision Theory]* Consider a binary classification problem. Suppose the two classes C_0 and C_1 can be roughly distinguished by one feature $x > 0$, and the class-conditional densities are given as follows:

$$p(x \mid C_0) = \alpha e^{-\alpha x} \quad \text{and} \quad p(x \mid C_1) = \beta e^{-\beta x},$$

where α, β $(0 < \alpha < \beta)$ are known parameters.

Assume that both classes C_0, C_1 are equally probable and that all the decision errors have the same penalty. Please define the Bayes' classifier for this problem and find the decision regions of the classifier. (10 points)

(b) *[Parameter Estimation]* Let $\mathbf{x}^t \in \mathbb{R}^2$ $(t = 1, \cdots, n)$ be i.i.d. data sampled from a uniform distribution over a disc of radius θ and the probability density function is

$$p(\mathbf{x}; \theta) = \begin{cases} \frac{1}{\pi\theta^2} & \text{if } \|\mathbf{x}\|_2 \leq \theta \\ 0 & \text{otherwise} \end{cases}.$$

Please find the maximum likelihood estimate of θ. (7 points)

Problem 2. (15 points) *[Dimensionality Reduction]*

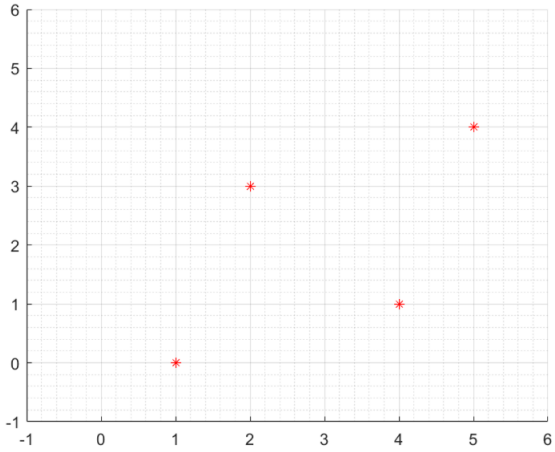
Consider the following sample matrix

$$\mathbf{X} = \begin{bmatrix} 4 & 1 \\ 2 & 3 \\ 5 & 4 \\ 1 & 0 \end{bmatrix},$$

with each row one data point. We turn to principal component analysis (PCA) to represent the data in only one dimension.

(a) Compute all the unit-length principal components of $\mathbf{X}$, and point out which one the PCA algorithm would choose if we request only one principal component. (5 points)

(b) The following plot depicts all the data points in $\mathbf{X}$. If we want a one-dimensional representation of the data, please draw the principal component direction (as a line) and the projections of all the four data points. (Label each projected point with its principal coordinate value and give the principal coordinate values exactly.) (5 points)



(c) Suppose we can also use the linear discriminant analysis (LDA) for dimension reduction of $\mathbf{X}$ and discuss the similarities and differences between PCA and LDA. (5 points)

Problem 3. (18 points)

(a) *[Multi-layer Perceptrons (MLP)]* A single output MLP is considered with its output given by

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

where

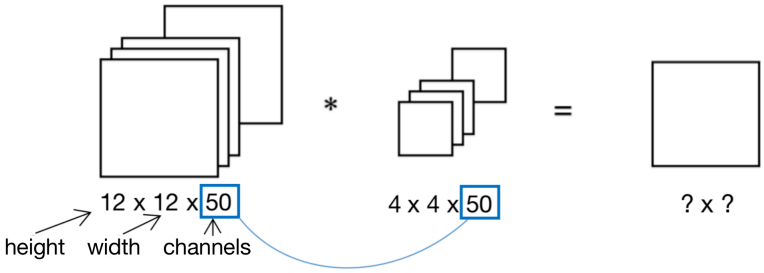
$$z_h^t = \sigma(\mathbf{w}_h^T \mathbf{x}^t) = \sigma(\sum_{j=1}^{K} w_{hj} x_j^t),$$

with $\sigma(a) = \frac{1}{1+e^{-a}}$. Given sample $\{\mathbf{x}^t, r^t\}_t$, the loss function in learning is defined as

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

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

(b) *[Deep Learning]* Consider a multichannel 2D convolutional neural network, where a convolutional layer C is followed by a max pooling layer P. The input of layer C has 50 channels, each of which is of size 12×12. We use a filter that is of size 4×4 to convolve with the input, with padding $= 1$ and stride $= 2$. The convolution of the input with the filter leads to an output. Here is the illustration:

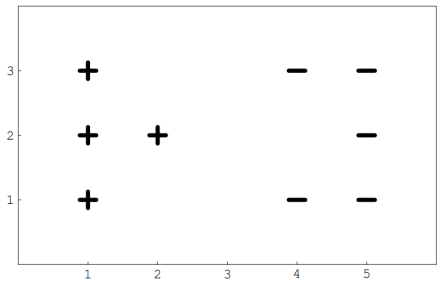


Layer P performs max pooling over the C's output, with the pooling filter is of size 3×3, and stride $= 1$. What is size of each of layer P's output?

Given $x_1, x_2, \cdots, x_n$ all scalars, we assume that a scalar multiplication $x_i \cdot x_j$ or a scalar addition $x_i + x_j$ accounts for one floating-point operation (FLOP), and a max operation $\max\{x_1, x_2, \cdots, x_n\}$ accounts for $n-1$ FLOPs. During one forward pass, how many FLOPs do layer C and layer P conduct in total? (10 points)

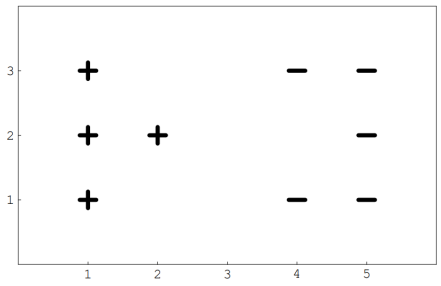
Problem 4. (20 points) *[Support Vector Machine (SVM)]*

(a) Considering the following data set (composed of two classes "+" and "-"), suppose we have a linear SVM with some large C value. Please draw the decision boundary of the linear SVM and give a brief explanation why it is the boundary.



(6 points)

(b) In the following figure, circle the points such that, after removing them from the training set and retraining the SVM, we would get a different decision boundary from training on the full sample. Please give a brief explanation why you choose these data points.



(6 points)

(c) Suppose we have a kernel $K(\cdot,\cdot)$ such that there is an implicit high-dimensional feature map $\phi : \mathbb{R}^d \to \mathbb{R}^D$ ($D > d$) that satisfies $\forall \mathbf{x}, \mathbf{z} \in \mathbb{R}^d,\ K(\mathbf{x}, \mathbf{z}) = \phi(\mathbf{x}) \cdot \phi(\mathbf{z})$, where $\phi(\mathbf{x}) \cdot \phi(\mathbf{z}) = \sum_{i=1}^{D} \phi_i(\mathbf{x})\phi_i(\mathbf{z})$ is the dot product in the D-dimensional space. Please show based on the kernel how to calculate the Euclidean distance in the D-dimensional space

$$\|\phi(\mathbf{x}) - \phi(\mathbf{z})\| = \sqrt{\sum_{i=1}^{D} (\phi_i(\mathbf{x}) - \phi_i(\mathbf{z}))^2}$$

without explicitly calculating the values using the D-dimensional vectors (you should provide a formal proof) (8 points)

Problem 5. (15 points) *[Clustering and Mixture Models, Nonparametric Methods]*

(a) Suppose we have five data points as shown in the table below. Try to cluster them into two clusters using the k-means clustering algorithm (choose initial points as B and D). (5 points)

A	B	C	D	E
(0,0)	(0,1)	(3,2)	(2,0)	(1,3)

(b) It is well-known that the k-means clustering is a special case of expectation-maximization (EM) applied to Gaussian mixture model (GMM). Discuss under which conditions that GMM satisfies the k-means algorithm will reduce to the EM algorithm? (6 points)

(c) Based on the result in (a), try to classify the data point $F = (2, 2)$ using a k-nearest neighbor estimator with $k = 3$. (4 points)

Problem 6. (15 points) *[Ensemble Learning]*

(a) AdaBoost can be understood as an optimizer that minimizes an exponential loss function given as follows:

$$E = \sum_{t=1}^{N} \exp\left(-y^t f\left(x^t\right)\right),$$

where $y = +1$ or -1 is the class label, x is the data, and $f(x)$ is the weighted sum of weak learners. Show that the loss function E is an upper bound of the $0-1$ loss function

$$L = \sum_{t=1}^{N} \mathbf{1}\left(y^t f\left(x^t\right) < 0\right).$$

(Hint: $\mathbf{1}(\cdot)$ is a step function that assigns value 1 if the classifier predicts correctly and 0 otherwise.) (7 points)

(b) Two problems may occur to the AdaBoost algorithm. Answer the following questions regarding these.

1) Show mathematically why a weak learner with $< 50\%$ predictive accuracy presents a problem to AdaBoost. (4 points)

2) AdaBoost is susceptible to outliers. Suggest a simple strategy that relieves this. (4 points)

(end of exam paper)