Q1 Fig. Q1 shows the decision plane (solid line) and the two hyperplanes (dashed lines) that maximize the margin of separation d between the two classes with training samples represented by filled circles ($\bullet$) and hollow circles ($\circ$), respectively. Denote $\mathbf{x}_i \in \Re^D$ and $y_i \in \{-1, +1\}$ as the i-th training sample and its label, respectively, where $i = 1, \ldots, N$. Then, the training set is denoted as $\mathcal{T} = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_N, y_N)\}$.

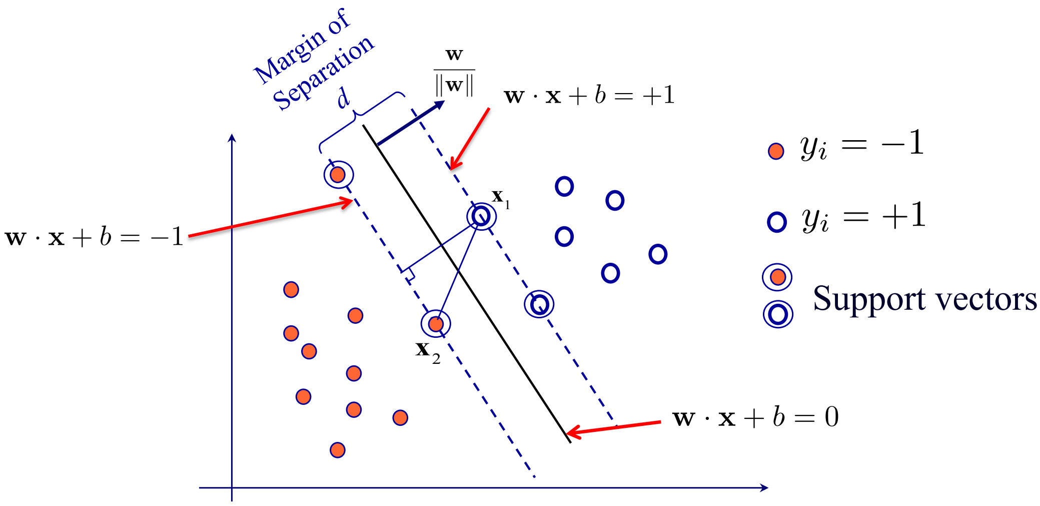

Fig. Q1

(a) Show that

$$d = \frac{2}{\|\mathbf{w}\|},$$

where $\|\mathbf{w}\|$ is the norm of the vector $\mathbf{w}$.

(5 marks)

(b) Given the training set $\mathcal{T}$, the optimal solution of $\mathbf{w}$ can be found by maximizing d subject to the following constraints:

$$\mathbf{x}_i \cdot \mathbf{w} + b \geq +1 \quad \text{for} \quad i \in \{1, \ldots, N\} \text{ where } y_i = +1$$
$$\mathbf{x}_i \cdot \mathbf{w} + b \leq -1 \quad \text{for} \quad i \in \{1, \ldots, N\} \text{ where } y_i = -1.$$

(i) Based on your answer in Q1(a), explain why the solution of $\mathbf{w}$ can be obtained by solving the following constrained optimization problem:

$$\begin{array}{ll} \min & \frac{1}{2}\|\mathbf{w}\|^2 \\ \text{subject to} & y_i(\mathbf{x}_i \cdot \mathbf{w} + b) \geq 1 \quad \forall i = 1, \ldots, N. \end{array}$$

(10 marks)

(ii) Given that the Lagrangian function of this problem is

$$L(\mathbf{w}, b, \{\alpha_i\}) = \frac{1}{2}\|\mathbf{w}\|^2 - \sum_{i=1}^{N} \alpha_i [y_i(\mathbf{x}_i \cdot \mathbf{w} + b) - 1],$$

where α_i's are Lagrange multipliers. State the constraints for minimizing $L(\mathbf{w}, b, \{\alpha_i\})$.

(10 marks)

(iii) Show that the Wolfe dual of this constrained optimization problem is

$$\max_{\alpha} \sum_{i=1}^{N} \alpha_i - \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} \alpha_i \alpha_j y_i y_j (\mathbf{x}_i \cdot \mathbf{x}_j)$$

$$\text{subject to} \sum_{i=1}^{N} \alpha_i y_i = 0 \quad \text{and} \quad \alpha_i \geq 0, i = 1, \ldots, N.$$

(15 marks)

Q2 The K-means algorithm aims to divide a set of training data $\mathcal{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$ into K disjoint sets $\{\mathcal{X}_1, \ldots, \mathcal{X}_K\}$ such that

$$\mathcal{X} = \bigcup_{k=1}^{K} \mathcal{X}_k \quad \text{and} \quad \mathcal{X}_i \cap \mathcal{X}_j = \emptyset \ \forall i \neq j.$$

This is achieved by minimizing the sum of the squared error:

$$E = \sum_{k=1}^{K} \sum_{\mathbf{x} \in \mathcal{X}_k} \|\mathbf{x} - \boldsymbol{\mu}_k\|^2.$$

(a) Show that E is minimal when $\boldsymbol{\mu}_k = \frac{1}{N_k}\sum_{\mathbf{x}\in\mathcal{X}_k} \mathbf{x}$, where N_k is the number of samples in $\mathcal{X}_k$.

(6 marks)

(b) Explain why the K-means algorithm is not suitable for clustering the samples in Fig. Q2.

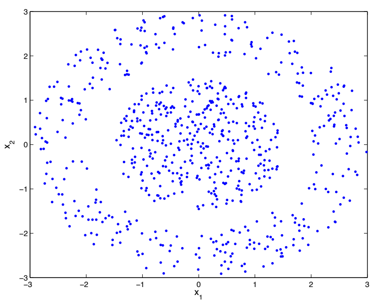

Fig. Q2

(7 marks)

(c) One possible approach to clustering the samples in Fig. Q2 is to map $\mathbf{x}$'s to a high-dimensional feature space using a non-linear function $\boldsymbol{\phi}(\mathbf{x})$, followed by applying the K-means algorithm to cluster the mapped samples in the feature space. Specifically, we aim to find the K disjoint sets $\{\mathcal{X}_1, \ldots, \mathcal{X}_K\}$ that minimize the following objective function:

$$E_{\boldsymbol{\phi}} = \sum_{k=1}^{K} \sum_{\mathbf{x} \in \mathcal{X}_k} \left\| \boldsymbol{\phi}(\mathbf{x}) - \frac{1}{N_k} \sum_{\mathbf{z} \in \mathcal{X}_k} \boldsymbol{\phi}(\mathbf{z}) \right\|^2 . \tag{Q2-a}$$

Show that minimizing Eq. Q2-a is equivalent to minimizing the following objective function

$$E'_{\boldsymbol{\phi}} = \sum_{k=1}^{K} \sum_{\mathbf{x} \in \mathcal{X}_k} \left[\frac{1}{N_k^2} \sum_{\mathbf{z} \in \mathcal{X}_k} \sum_{\mathbf{z}' \in \mathcal{X}_k} \boldsymbol{\phi}(\mathbf{z})^{\mathsf{T}} \boldsymbol{\phi}(\mathbf{z}') - \frac{2}{N_k} \sum_{\mathbf{z} \in \mathcal{X}_k} \boldsymbol{\phi}(\mathbf{z})^{\mathsf{T}} \boldsymbol{\phi}(\mathbf{x}) \right] . \tag{Q2-b}$$

(10 marks)

(d) When the dimension of $\boldsymbol{\phi}(\mathbf{x})$ is very high, Eq. Q2-b cannot be implemented. Suggest a way to solve this problem and write the objective function.

(7 marks)

Q3 Given a set of training vectors $\mathcal{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$ in D-dimensional space, principle component analysis (PCA) aims to find a projection matrix $\mathbf{U}$ that projects the vectors in $\mathcal{X}$ from D-dimensional space to M-dimensional space, where $M \leq D$. The projection matrix can be obtained by solving the following equation:

$$\mathbf{X}\mathbf{X}^{\mathsf{T}}\mathbf{U} = \mathbf{U}\boldsymbol{\Lambda}_M, \tag{Q3-a}$$

where $\boldsymbol{\Lambda}_M$ is an $M \times M$ diagonal matrix and $\mathbf{X}$ is a $D \times N$ centered data matrix whose n-th column is given by $(\mathbf{x}_n - \frac{1}{N}\sum_{i=1}^{N} \mathbf{x}_i)$.

(a) How do the values of the diagonal elements of $\boldsymbol{\Lambda}_M$ related to the variances of the training vectors?

(5 marks)

(b) If D is very large (say 100,000), solving Eq. Q3-a is computationally demanding. Suggest a method to find $\mathbf{U}$ when $M < N \ll D$.

(10 marks)

Q4 A deep neural networks has one input layer, $L - 1$ hidden layers, and one output layers with K outputs.

(a) If the network is to be used for classification, suggest an appropriate activation function for the output nodes. Express the suggested function in terms of the linear weighted sum $a_k^{(L)}$ at output node k, for $k = 1, \ldots, K$.

(5 marks)

(b) For the network to be useful, each neuron in the hidden layers should have a non-linear activation function such as the sigmoid function or the ReLU. Explain why the network will not be very useful if a linear activation function is used for all neurons, including the output neurons. *Hints*: You may answer this question by expressing the output as a function of the input and the weight matrices of the network.

(10 marks)

– END –