Optimization and Machine Learning, Spring 2021
Reference Solutions for Homework 5

1. [10 points] The problem of maximizing margin can be converted into an following equivalent problem

$$\begin{aligned}&\underset{\mathbf{w},b}{\operatorname{argmin}}\ \frac{1}{2}\|\mathbf{w}\|^2\\&\text{subject to } t_n(\mathbf{w}^\top\phi(\mathbf{x}_n)+b)\geq 1,\quad n=1,\cdots,N.\end{aligned}$$

where $\phi(\mathbf{x})$ is a fixed feature-space transformation.

(a) By introducing Lagrange multipliers $\{a_n\}$, please give the Lagrangian function and the dual representation of the maximum margin problem. [5 points]

(b) Please show that the value ρ of the margin for the maximum-margin hyperplane is given by

$$\frac{1}{\rho^2}=\sum_{n=1}^{N}a_n.$$

(Hint: $\{a_n\}$ can be obtained by solving the dual representation of the maximum margin problem.) [5 points]

Solution:

(a) The Lagrangian function is given by

$$L(\mathbf{w},b,\mathbf{a})=\frac{1}{2}\|\mathbf{w}\|^2-\sum_{n=1}^{N}a_n\left\{t_n\left(\mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}\left(\mathbf{x}_n\right)+b\right)-1\right\}. \tag{1}$$

The dual representation of the maximum margin problem is given by

$$\max_{\mathbf{a}}\ \widetilde{L}(\mathbf{a})=\sum_{n=1}^{N}a_n-\frac{1}{2}\sum_{n=1}^{N}\sum_{m=1}^{N}a_na_mt_nt_mk\left(\mathbf{x}_n,\mathbf{x}_m\right) \tag{2}$$

$$\text{subject to } a_n\geq 0,\quad n=1,\cdots,N, \tag{3}$$

$$\sum_{n=1}^{N}a_nt_n=0, \tag{4}$$

where $k(\mathbf{x},\mathbf{x}')=\phi(\mathbf{x})^\top\phi(\mathbf{x}')$.

(b) Let the value of the margin ρ be $1/\|\mathbf{w}\|$ and so $1/\rho^2=\|\mathbf{w}\|^2$. From the KKT conditions of the dual problem which is

$$\begin{aligned}a_n&\geqslant 0\\t_ny\left(\mathbf{x}_n\right)-1&\geqslant 0\\a_n\left\{t_ny\left(\mathbf{x}_n\right)-1\right\}&=0,\end{aligned}$$

we see that, for the maximum margin solution, the second term of (1) vanishes and so we have

$$L(\mathbf{w},b,\mathbf{a})=\frac{1}{2}\|\mathbf{w}\|^2. \tag{5}$$

By setting the derivatives of (1) with respect to $\mathbf{w}$ and b equal to zero, we obtain the following two conditions

$$\mathbf{w} = \sum_{n=1}^{N} a_n t_n \phi(\mathbf{x}_n) \tag{6}$$

$$0 = \sum_{n=1}^{N} a_n t_n. \tag{7}$$

Using (5) together with (6), the dual (2) can be written as

$$\frac{1}{2}\|\mathbf{w}\|^2 = \sum_{n}^{N} a_n - \frac{1}{2}\|\mathbf{w}\|^2,$$

from which the desired result follows.

2. [10 points] Use the k-means algorithm and Euclidean distance to cluster the 8 data points into $K = 3$ clusters. The coordinates of the data points are:

$$x^{(1)} = (2, 8),\ x^{(2)} = (2, 5),\ x^{(3)} = (1, 2),\ x^{(4)} = (5, 8),$$
$$x^{(5)} = (7, 3),\ x^{(6)} = (6, 4),\ x^{(7)} = (8, 4),\ x^{(8)} = (4, 7).$$

Suppose that the Lloyd's algorithm is applied with the initial cluster centers $x^{(3)}$, $x^{(4)}$ and $x^{(6)}$.

(a) Perform one iteration of the k-means algorithm and report the coordinates of the resulting centroids. [3 points]

(b) Calculate the loss function

$$Q(r, c) = \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{K} r_{ij} ||x^{(i)} - c_j||^2, \tag{8}$$

where $r_{ij} = 1$ if $x^{(i)}$ belongs to the j-th cluster and 0 otherwise, before and after the first iteration of k-means. [4 points]

(c) How many more iterations are needed to converge? [3 points]

Solution:

(a) After the first iteration, the centroids become

$$c_1 = \frac{1}{2}(x^{(2)} + x^{(3)}) = (1.5, 3.5)$$
$$c_2 = \frac{1}{3}(x^{(1)} + x^{(4)} + x^{(8)}) = (3.67, 7.67)$$
$$c_3 = \frac{1}{3}(x^{(5)} + x^{(6)} + x^{(7)}) = (7, 3.67).$$

(b) Let's denote the objective values of the initial iteration and the first iteration by Q_0 and Q_1, respectively:

$$Q_0 = \frac{1}{8}(3^2 + 3.1623^2 + 1.4142^2 + 2^2 + 1.4142^2) = 3.375 \tag{9}$$
$$Q_1 = \frac{1}{8}(2.9 + 2.5 + 2.5 + 1.9 + 0.44 + 1.11 + 1.11 + 0.56) = 1.6250. \tag{10}$$

(c) Zero. The assignment of the points is [2,1,1,2,3,3,3,2] and will not change after the first iteration.

3. [10 points] Show that the conventional linear principal component analysis (PCA) can be recovered as a special case of kernel PCA if we choose the linear kernel function given by $k(\mathbf{x}, \mathbf{x}') = \mathbf{x}^T\mathbf{x}'$.

   Solution:

   Suppose $\mathbf{X}$ has been centralized. The kernel function is given by $k(\mathbf{x}, \mathbf{x}') = \mathbf{x}^T\mathbf{x}'$, so we have the matrix form $\mathbf{K} = \mathbf{X}^T\mathbf{X}$. We can represent principal components $\mathbf{v} = \sum_{i=1}^{n} a_i\mathbf{x}_i = \mathbf{X}\alpha$. And to solve the conventional linear PCA, we have

$$
\begin{aligned}
& \frac{1}{n}\mathbf{X}\mathbf{X}^T\mathbf{v} = \lambda\mathbf{v} \\
\Rightarrow \quad & \mathbf{X}^T\frac{1}{n}\mathbf{X}\mathbf{X}^T\mathbf{v} = \mathbf{X}^T\lambda\mathbf{v} \\
\Rightarrow \quad & \frac{1}{n}\mathbf{X}^T\mathbf{X}\mathbf{X}^T\mathbf{X}\alpha = \lambda\mathbf{X}^T\mathbf{X}\alpha \\
\Rightarrow \quad & \frac{1}{n}\mathbf{K}\mathbf{K}\alpha = \lambda\mathbf{K}\alpha \\
\Rightarrow \quad & \frac{1}{n}\mathbf{K}\alpha = \lambda\alpha
\end{aligned}
$$

   Then we show the conventional linear PCA is a special case of kernel PCA with the linear kernel function given by $k(\mathbf{x}, \mathbf{x}') = \mathbf{x}^T\mathbf{x}'$.