SUTD 2021 50.007 Homework 2

James Raphael Tiovalen 1004555

1. Clustering: K-Means and K-Medoids

Question 1.1

a) False. For unsupervised learning, we use unlabeled data. If only some of the data is labeled, it is called semi-supervised learning.

b) True. The choice of distance metric can affect how the clustering algorithm will perform since it is one of the most meaningful numerical measure that said algorithm can use a benchmark and check against (as well as how much to punish/reward similarity/dissimilarity).

c) False. K-means is actually more sensitive to outliers relative to k-medoids since the averages/means of cluster data points could potentially be skewed by anomalous data points.

d) False. While it is true that each iteration of the k-means and k-medoids algorithms would lower the cost (by definition), they are not guaranteed to always converge to the global optimal solution (since they might converge to only the local minimum).

e) False. Different initialization values might lead to different clustering outputs for k-medoids, which might vary the quality of said clusterings.

Question 1.2

a) The centroid of C_1 would be:

$$\mu_1 = \frac{3 + 4 + 11 + 12}{4} = 7.5$$

b) The centroid of C_2 would be:

$$\mu_2 = \frac{2 + 10}{2} = 6$$

c) The new cluster formed by μ_1 is $D_1 = \{10, 11, 12\}$, while the new cluster formed by μ_2 is $D_2 = \{2, 3, 4\}$.

d) The centroid of D_1 would be:

$$\mu_1' = \frac{10 + 11 + 12}{3} = 11,$$

while the centroid of D_2 would be:

$$\mu_2' = \frac{2 + 3 + 4}{3} = 3$$

e) Yes, this clustering is stable since the k-means algorithm will converge. With the new centroid values μ_1' and μ_2', the cluster assignments will remain the same as the D_1 and D_2 clusters (and hence, the centroid values as well) found previously for any further iterations of the algorithm.

2. Support Vector Machines

Question 2.1

a) The kernel defined by the mapping stated in the question would be:

$$\begin{aligned}
K(\mathbf{x},\mathbf{y}) &= \varphi^T(\mathbf{x})\varphi(\mathbf{y}) \\
&= \begin{bmatrix} 1 & x_1^2 & \sqrt{2}x_1x_2 & x_2^2 & \sqrt{2}x_1 & \sqrt{2}x_2 \end{bmatrix} \begin{bmatrix} 1 \\ y_1^2 \\ \sqrt{2}y_1y_2 \\ y_2^2 \\ \sqrt{2}y_1 \\ \sqrt{2}y_2 \end{bmatrix} \\
&= 1 + x_1^2y_1^2 + 2x_1x_2y_1y_2 + x_2^2y_2^2 + 2x_1y_1 + 2x_2y_2 \\
&= \left(1 + \mathbf{x}^T\mathbf{y}\right)^2
\end{aligned}$$

b) With $\mathbf{x} = \begin{bmatrix} 1 & 2 \end{bmatrix}^T$ and $\mathbf{y} = \begin{bmatrix} 3 & 4 \end{bmatrix}^T$, we can obtain the value of the kernel via simple substitution:

$$\begin{aligned}
K(\mathbf{x},\mathbf{y}) &= 1 + x_1^2y_1^2 + 2x_1x_2y_1y_2 + x_2^2y_2^2 + 2x_1y_1 + 2x_2y_2 \\
&= 1 + 1^2 \cdot 3^2 + 2 \cdot 1 \cdot 2 \cdot 3 \cdot 4 + 2^2 \cdot 4^2 + 2 \cdot 1 \cdot 3 + 2 \cdot 2 \cdot 4 \\
&= 1 + 9 + 48 + 64 + 6 + 16 \\
&= 144
\end{aligned}$$

Question 2.2

1) Let α_i and β_i be the Lagrange multipliers. Let us first re-write the constraint $d_i(\mathbf{w}^T\mathbf{x}_i + b) - 1 + \xi_i \geq 0$ in the standard form as:

$$-d_i(\mathbf{w}^T\mathbf{x}_i + b) + 1 - \xi_i \leq 0$$

Then, using the generalized Lagrangian function, we can write the Lagrangian function:

$$\begin{aligned}
L(\mathbf{w}, b, \xi, \boldsymbol{\alpha}, \boldsymbol{\beta}) &= \frac{1}{2}\mathbf{w}^T\mathbf{w} + C\sum_{i=1}^{N}\xi_i - \sum_{i=1}^{N}\alpha_i\left(d_i(\mathbf{w}^T\mathbf{x}_i + b) - 1 + \xi_i\right) - \sum_{i=1}^{N}\beta_i\xi_i \\
&= \frac{1}{2}\mathbf{w}^T\mathbf{w} + C\sum_{i=1}^{N}\xi_i - \sum_{i=1}^{N}\alpha_i d_i\mathbf{w}^T\mathbf{x}_i - b\sum_{i=1}^{N}\alpha_i d_i + \sum_{i=1}^{N}\alpha_i - \sum_{i=1}^{N}\alpha_i\xi_i - \sum_{i=1}^{N}\beta_i\xi_i
\end{aligned}$$

We can then write down our KKT conditions:

$$\frac{\partial L}{\partial \mathbf{w}} = \mathbf{w} - \sum_{i=1}^{N} \alpha_i d_i \mathbf{x}_i = \mathbf{0} \qquad d_i(\mathbf{w}^T \mathbf{x}_i + b) - 1 + \xi_i \geq 0$$

$$\frac{\partial L}{\partial b} = -\sum_{i=1}^{N} \alpha_i d_i = 0 \qquad \alpha_i \left(d_i(\mathbf{w}^T \mathbf{x}_i + b) - 1 + \xi_i \right) = 0$$

$$\frac{\partial L}{\partial \xi_i} = C - \alpha_i - \beta_i = 0 \qquad \beta_i \xi_i = 0$$

$$\alpha_i \geq 0 \qquad \beta_i \geq 0$$

Since $\mathbf{w} = \sum_{i=1}^{N} \alpha_i d_i \mathbf{x}_i$, we have:

$$\mathbf{w}^T \mathbf{w} = \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i d_i \alpha_j d_j \mathbf{x}_i^T \mathbf{x}_j$$

$$\sum_{i=1}^{N} \alpha_i d_i \mathbf{w}^T \mathbf{x}_i = \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i d_i \alpha_j d_j \mathbf{x}_i^T \mathbf{x}_j$$

From the KKT conditions, we also have $C = \alpha_i + \beta_i$ and $-b \sum_{i=1}^{N} \alpha_i d_i = 0$. Hence, using all the results that we have found, we can further re-write and simplify our intermediate Lagrangian function to be:

$$\begin{aligned} L(\mathbf{w}, b, \xi, \boldsymbol{\alpha}, \boldsymbol{\beta}) &= \frac{1}{2}\mathbf{w}^T \mathbf{w} + C \sum_{i=1}^{N} \xi_i - \sum_{i=1}^{N} \alpha_i d_i \mathbf{w}^T \mathbf{x}_i - b \sum_{i=1}^{N} \alpha_i d_i + \sum_{i=1}^{N} \alpha_i - \sum_{i=1}^{N} \alpha_i \xi_i - \sum_{i=1}^{N} \beta_i \xi_i \\ &= -\frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i d_i \alpha_j d_j \mathbf{x}_i^T \mathbf{x}_j + \sum_{i=1}^{N} (\alpha_i + \beta_i) \xi_i + \sum_{i=1}^{N} \alpha_i - \sum_{i=1}^{N} \alpha_i \xi_i - \sum_{i=1}^{N} \beta_i \xi_i \\ &= -\frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i d_i \alpha_j d_j \mathbf{x}_i^T \mathbf{x}_j + \sum_{i=1}^{N} \alpha_i \end{aligned}$$

Since $\alpha_i = C - \beta_i$, $\alpha_i \geq 0$, and $\beta_i \geq 0$, we can combine said constraints to form the constraint $0 \leq \alpha_i \leq C$.

Therefore, we can formulate our dual problem with soft margin in finding the optimal hyperplane as:

Given:	$S = \{(\mathbf{x}_i, d_i)\}$
Find:	Lagrange multipliers $\{\alpha_i\}$
Maximizing:	$Q(\boldsymbol{\alpha}) = \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j d_i d_j \mathbf{x}_i^T \mathbf{x}_j$
Subject to:	$\sum_{i=1}^{N} \alpha_i d_i = 0$
	$0 \leq \alpha_i \leq C, \qquad i = 1, \ldots, N$

2) If our data is not linearly separable, we would prefer to use soft margin so as to allow a few points to be classified on the wrong side of the best-possible optimal hyperplane. This is acceptable since sometimes it is impossible to completely and exclusively separate our data (perhaps due to data that is noisy or prone to error). Otherwise, if our data is linearly separable, we would prefer to use hard margin.