Homework 2
By: Jaron Ho (1005011)

Answers rounded to 5 s.f. where required.

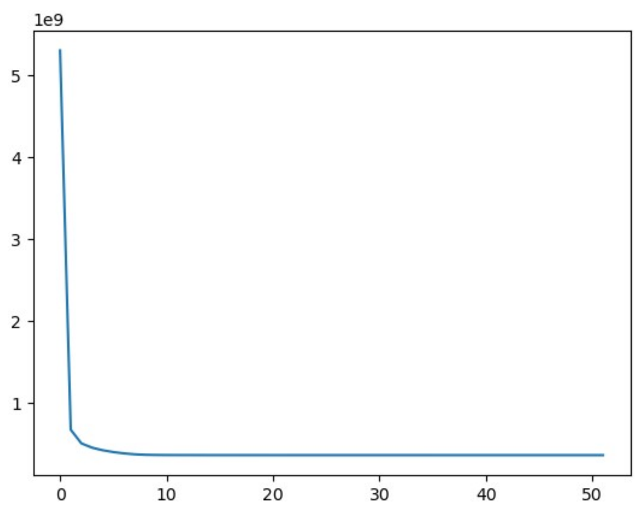
Q1
(a)
Since one of the clusters is empty at the end,
Number of clusters = 7

(b)
Final centroids:
$z^{(1)}$ = [225.99, 213.51, 213.16]
$z^{(2)}$ = [209.10, 160.04, 129.69]
$z^{(3)}$ = [162.41, 116.66, 95.228]
$z^{(4)}$ = [104.77, 73.808, 61.029]
$z^{(5)}$ = [91.246, 93.539, 162.39]
$z^{(6)}$ = [44.487, 33.809, 32.922]
$z^{(7)}$ = [5.7179, 5.1135, 17.640]

(The empty cluster has the centroid [0, 255, 0].)

(c)
In the same order as the centroids from (b),
$|C_1|$ = 24434
$|C_2|$ = 82757
$|C_3|$ = 64453
$|C_4|$ = 59740
$|C_5|$ = 3553
$|C_6|$ = 86905
$|C_7|$ = 488158

(d)



Left:
Graph of sum of squared Euclidean distance against number of iterations.

Below:
The image with every pixel replaced with its nearest centroid. The full image file is produced as `centroid_only.jpg` after running the code in `HW2.ipynb`.

Q2

Since there are 3 examplars out of 4 points and there are 3 clusters in each figure, it means that there is only one non-medoid point for all 3 figures. This non-medoid point will belong to the cluster that minimises the cost according to the distance measure used, which can be determined by computing the distance between each pair of points using all 3 distance measures:

Distance between	l_1	l_2	l_∞
x_0, x_1	|4-0| + |4-0| = 8	$\sqrt{(4-0)^2+(4-0)^2} \approx 5.6569$	max(|4-0|, |4-0|) = 4
x_0, x_2	|0-0| + |-6-0| = 6	$\sqrt{(0-0)^2+(-6-0)^2} = 6$	max(|0-0|, |-6-0|) = 6
x_0, x_3	|-5-0| + |2-0| = 7	$\sqrt{(-5-0)^2+(2-0)^2} \approx 5.3852$	max(|-5-0|, |2-0|) = 5
x_1, x_2	|0-4| + |-6-4| = 14	$\sqrt{(0-4)^2+(-6-4)^2} \approx 10.770$	max(|0-4|, |-6-4|) = 10
x_1, x_3	|-5-4| + |2-4| = 11	$\sqrt{(-5-4)^2+(2-4)^2} \approx 9.2195$	max(|-5-4|, |2-4|) = 9
x_2, x_3	|-5-0| + |2-(-6)| = 13	$\sqrt{(-5-0)^2+(2-(-6))^2} \approx 9.4340$	max(|-5-0|, |2-(-6)|) = 8

The minimum distance using each norm is highlighted in yellow.

Since figure A groups x_0 and x_1 together in the same cluster, it uses the l_∞ norm as this pair of points gives the minimum distance under the l_∞ norm. Similarly, we can conclude that figure B uses the l_2 norm, while figure 3 uses the l_1 norm.

Q3

Kernel 0 accuracy = 84.375%
Kernel 1 accuracy = 81.25%
Kernel 2 accuracy = 90.625%
Kernel 3 accuracy = 43.75%

Q4
(a)
Let **w** = [a c]

$y^{(2)}\boldsymbol{w}\cdot\boldsymbol{x}^{(2)}\geq 1$
$-1[a \quad c]\cdot[1 \quad 0]^T\geq 1$
$-a+0\geq 1$ - (1)
Hence, $a\leq -1$

$y^{(1)}\boldsymbol{w}\cdot\boldsymbol{x}^{(1)}\geq 1$
$1[a \quad c]\cdot[1 \quad 1]^T\geq 1$
$a+c\geq 1$
$c\geq 1-a\geq 1+1$ (from (1))
Hence, $c\geq 2$

$$\frac{1}{2}\|\boldsymbol{w}\|^2=\frac{1}{2}(a^2+c^2)$$

To minimise $\frac{1}{2}\|\boldsymbol{w}\|^2$, we need to minimise a^2 and c^2. As a becomes more negative, b becomes more positive, so the magnitude of both a and c (and in turn a^2 + c^2) increases. Hence, the minimum $\frac{1}{2}\|\boldsymbol{w}\|^2$ is achieved when a is the least negative value possible.

Hence, a = -1 and c = 2, so **w** = [-1 2].

$$\gamma^{(1)}(\boldsymbol{w})=\frac{y^{(1)}(\boldsymbol{w}\cdot\boldsymbol{x}^{(1)})}{\|\boldsymbol{w}\|}$$
$$\gamma^{(1)}(\boldsymbol{w})=\frac{1(-1+2)}{\sqrt{((-1)^2+2^2)}}$$
$$\gamma^{(1)}(\boldsymbol{w})=\frac{1}{\sqrt{5}}$$

Sanity check to ensure margin is the same on both sides of the decision boundary:
$$\gamma^{(2)}(\boldsymbol{w})=\frac{y^{(2)}(\boldsymbol{w}\cdot\boldsymbol{x}^{(2)})}{\|\boldsymbol{w}\|}$$
$$\gamma^{(2)}(\boldsymbol{w})=\frac{-1(-1+0)}{\sqrt{5}}$$
$$\gamma^{(2)}(\boldsymbol{w})=\frac{1}{\sqrt{5}}$$

Hence, $\gamma=\frac{1}{\sqrt{5}}$

(b)
If b is non-zero, it means we need to minimise
$\frac{1}{2}\|\boldsymbol{w}\|^2$ subject to $y^{(t)}(\boldsymbol{w}\cdot\boldsymbol{x}^{(t)}+b)\geq 1$

$y^{(2)}(\boldsymbol{w}\cdot\boldsymbol{x}^{(2)}+b)\geq 1$
$-a+0-b\geq 1$
$-a\geq 1+b$ - (2)
Hence, $a\leq -1-b$

$y^{(1)}(\boldsymbol{w}\cdot\boldsymbol{x}^{(1)}+b)\geq 1$
$a+c+b\geq 1$
$c\geq 1-a-b\geq 1+1+b-b$ (from (2))
Hence, $c\geq 2$

As in (a), we want to minimise the magnitude of both a and c. $c\geq 2$ regardless of what b is, so the c which has the lowest magnitude is c = 2.

As for a, the lowest magnitude possible is 0, which is achieved when a = 0. Thus,
-1 – b = 0
b = -1

Hence, **w** = [0 2] and b = -1.

$$\gamma^{(1)}(\boldsymbol{w},b)=\frac{y^{(1)}(\boldsymbol{w}\cdot\boldsymbol{x}^{(1)}+b)}{\|\boldsymbol{w}\|}$$
$$\gamma^{(1)}(\boldsymbol{w},b)=\frac{1(0+2-1)}{\sqrt{(0^2+2^2)}}$$
$$\gamma^{(1)}(\boldsymbol{w},b)=\frac{1}{\sqrt{4}}$$
$$\gamma^{(1)}(\boldsymbol{w},b)=\frac{1}{2}$$

Sanity check:
$$\gamma^{(2)}(\boldsymbol{w},b)=\frac{y^{(2)}(\boldsymbol{w}\cdot\boldsymbol{x}^{(2)}+b)}{\|\boldsymbol{w}\|}$$
$$\gamma^{(2)}(\boldsymbol{w},b)=\frac{-1(0+0-1)}{2}$$
$$\gamma^{(2)}(\boldsymbol{w},b)=\frac{1}{2}$$

Hence, $\gamma=\frac{1}{2}$

Q5

(a)

$K(\boldsymbol{x},\boldsymbol{z})=K_1(\boldsymbol{x},\boldsymbol{z})K_2(\boldsymbol{x},\boldsymbol{z})$

Let $K_1(\boldsymbol{x},\boldsymbol{z})=\phi_1(\boldsymbol{x})\cdot\phi_1(\boldsymbol{z})$, $K_2(\boldsymbol{x},\boldsymbol{z})=\phi_2(\boldsymbol{x})\cdot\phi_2(\boldsymbol{z})$ and $K=\phi_3(\boldsymbol{x})\cdot\phi_3(\boldsymbol{z})$.

$K_1(\boldsymbol{x},\boldsymbol{z})K_2(\boldsymbol{x},\boldsymbol{z})=(\phi_1(\boldsymbol{x})\cdot\phi_1(\boldsymbol{z}))(\phi_2(\boldsymbol{x})\cdot\phi_2(\boldsymbol{z}))$

Let *i* denote the *M*-dimensional feature vector of K_1 and *j* denote the *N*-dimensional feature vector of K_2.

$K_1(\boldsymbol{x},\boldsymbol{z})K_2(\boldsymbol{x},\boldsymbol{z})=(\sum_{i=0}^{M} f_i(\boldsymbol{x})f_i(\boldsymbol{z}))(\sum_{j=0}^{N} g_j(\boldsymbol{x})g_j(\boldsymbol{z}))$

$K_1(\boldsymbol{x},\boldsymbol{z})K_2(\boldsymbol{x},\boldsymbol{z})=\sum_{i=0}^{M}\sum_{j=0}^{N} f_i(\boldsymbol{x})f_i(\boldsymbol{z})g_j(\boldsymbol{x})g_j(\boldsymbol{z})$

$K_1(\boldsymbol{x},\boldsymbol{z})K_2(\boldsymbol{x},\boldsymbol{z})=\sum_{i=0}^{M}\sum_{j=0}^{N} (f_i(\boldsymbol{x})g_j(\boldsymbol{x}))(f_i(\boldsymbol{z})g_j(\boldsymbol{z}))$

$K_1(\boldsymbol{x},\boldsymbol{z})K_2(\boldsymbol{x},\boldsymbol{z})=\sum_{i=0}^{M}\sum_{j=0}^{N} h_{ij}(\boldsymbol{x})h_{ij}(\boldsymbol{z})$, where $h_{ij}(\boldsymbol{x})=f_i(\boldsymbol{x})g_j(\boldsymbol{x})$

$K_1(\boldsymbol{x},\boldsymbol{z})K_2(\boldsymbol{x},\boldsymbol{z})=\phi_3(\boldsymbol{x})\cdot\phi_3(\boldsymbol{z})=K(\boldsymbol{x},\boldsymbol{z})$, where h_{ij} denotes the $M\cdot N$-dimensional feature vector of *K*.

Since *K* can be expressed as an inner product of feature vectors, *K* is a valid kernel by definition.

Reference: https://stats.stackexchange.com/questions/48509/proof-of-closeness-of-kernel-functions-under-pointwise-product

(b)

$K(\boldsymbol{x},\boldsymbol{z})=aK_1(\boldsymbol{x},\boldsymbol{z})-bK_2(\boldsymbol{x},\boldsymbol{z})$

Let $K_1(\boldsymbol{x},\boldsymbol{z})=\phi_1(\boldsymbol{x})\cdot\phi_1(\boldsymbol{z})$ and $K_2(\boldsymbol{x},\boldsymbol{z})=\phi_2(\boldsymbol{x})\cdot\phi_2(\boldsymbol{z})$.

$aK_1(\boldsymbol{x},\boldsymbol{z})-bK_2(\boldsymbol{x},\boldsymbol{z})=a(\phi_1(\boldsymbol{x})\cdot\phi_1(\boldsymbol{z}))-b(\phi_2(\boldsymbol{x})\cdot\phi_2(\boldsymbol{z}))$

$aK_1(\boldsymbol{x},\boldsymbol{z})-bK_2(\boldsymbol{x},\boldsymbol{z})=(\sqrt{a}\,\phi_1(\boldsymbol{x}))\cdot(\sqrt{a}\,\phi_1(\boldsymbol{z}))+(\sqrt{-b}\,\phi_2(\boldsymbol{x}))\cdot(\sqrt{-b}\,\phi_2(\boldsymbol{z}))$

$aK_1(\boldsymbol{x},\boldsymbol{z})-bK_2(\boldsymbol{x},\boldsymbol{z})=\begin{bmatrix}\sqrt{a}\,\phi_1(\boldsymbol{x})\\ \sqrt{-b}\,\phi_2(\boldsymbol{x})\end{bmatrix}\cdot\begin{bmatrix}\sqrt{a}\,\phi_1(\boldsymbol{z})\\ \sqrt{-b}\,\phi_2(\boldsymbol{z})\end{bmatrix}$

Let $\phi_3(\boldsymbol{x})=\begin{bmatrix}\sqrt{a}\,\phi_1(\boldsymbol{x})\\ \sqrt{-b}\,\phi_2(\boldsymbol{x})\end{bmatrix}$.

Then, $aK_1(\boldsymbol{x},\boldsymbol{z})-bK_2(\boldsymbol{x},\boldsymbol{z})=\phi_3(\boldsymbol{x})\cdot\phi_3(\boldsymbol{z})=K(\boldsymbol{x},\boldsymbol{z})$

However, since $b > 0$ and *b* is real, $\sqrt{(-b)}$ is not real.

Thus, *K* is not a kernel since its values are complex.

Alternatively, let $K_3(\boldsymbol{x},\boldsymbol{z})=aK_1(\boldsymbol{x},\boldsymbol{z})$ and $K_4(\boldsymbol{x},\boldsymbol{z})=-bK_2(\boldsymbol{x},\boldsymbol{z})$.

Using rule 3 from the notes (sum of valid kernels is a valid kernel),

$K(\boldsymbol{x},\boldsymbol{z})=K_3(\boldsymbol{x},\boldsymbol{z})+K_4(\boldsymbol{x},\boldsymbol{z})$ would be a valid kernel if K_3 and K_4 are valid kernels.

By Mercer's theorem, K_1 and K_2 are positive semidefinite.

Since $a > 0$, K_3 is also positive semidefinite.

However, since $b > 0$, $-b < 0$. Thus, K_4 becomes negative semidefinite, so it is no longer a valid kernel.

Since *K* is not a sum of valid kernels, *K* is not a valid kernel.

(c)

$K(\boldsymbol{x}, \boldsymbol{z}) = \tanh(\alpha K_1(\boldsymbol{x}, \boldsymbol{z}) + \boldsymbol{C})$

By Mercer's theorem, since K_1 is a valid kernel, it is positive semidefinite.

Since $\alpha > 0$ and $\boldsymbol{C} > 0$, $\alpha K_1(\boldsymbol{x}, \boldsymbol{z}) + \boldsymbol{C}$ is also positive semidefinite.

Since $\tanh(x) > 0$ for all $x > 0$, $\tanh(\alpha K_1(\boldsymbol{x}, \boldsymbol{z}) + \boldsymbol{C})$ is positive semidefinite.

Hence, K is a valid kernel by Mercer's theorem.

(d)

$K(\boldsymbol{x}, \boldsymbol{z}) = f(\boldsymbol{x}) f(\boldsymbol{z})$

$K(\boldsymbol{x}, \boldsymbol{z}) = f(\boldsymbol{x})(1) f(\boldsymbol{z})$

Let $K_1(\boldsymbol{x}, \boldsymbol{z}) = 1$, which is a valid kernel according to rule 1 from the notes.

$K(\boldsymbol{x}, \boldsymbol{z}) = f(\boldsymbol{x}) K_1(\boldsymbol{x}, \boldsymbol{z}) f(\boldsymbol{z})$

By rule 2, K is a valid kernel.

Alternatively, let $K(\boldsymbol{x}, \boldsymbol{z}) = \phi(\boldsymbol{x}) \phi(\boldsymbol{z})$ and $\phi(\boldsymbol{x}) = f(\boldsymbol{x})$.

$K(\boldsymbol{x}, \boldsymbol{z}) = \phi(\boldsymbol{x}) \phi(\boldsymbol{z}) = f(\boldsymbol{x}) f(\boldsymbol{z})$

Thus, K is a valid kernel.