COT 5615 Math for Intelligent Systems, Fall 2021

Homework 2

3.2 (10 pts)

(a) The RMS value of the vector $x = (a_1, \dots, a_n, b_1, \dots, b_m)$ is

$$\begin{aligned}
\mathbf{rms}(x) &= \left(\frac{a_1^2 + \cdots + a_n^2 + b_1^2 + \cdots + b_m^2}{n+m} \right)^{1/2} \\
&= \left(\frac{n(a_1^2 + \cdots + a_n^2)/n + m(b_1^2 + \cdots + b_m^2)/m}{n+m} \right)^{1/2} \\
&= \left(\frac{n \ \mathbf{rms}(a)^2 + m \ \mathbf{rms}(b)^2}{n+m} \right)^{1/2}.
\end{aligned}$$

(b) The average of the vector $x = (a_1, \dots, a_n, b_1, \dots, b_m)$ is

$$\mathbf{avg}(x) = \frac{a_1 + \cdots + a_n + b_1 + \cdots + b_m}{n+m} = \frac{n \ \mathbf{avg}(a) + m \ \mathbf{avg}(b)}{n+m}.$$

3.5 (20 pts) First the 1-norm.

(a) *Homogeneity.*

$$\begin{aligned}
\|\beta x\|_1 &= |\beta x_1| + |\beta x_2| + \cdots + |\beta x_n| \\
&= |\beta||x_1| + |\beta||x_2| + \cdots + |\beta||x_n| \\
&= |\beta|(|x_1| + |x_2| + \cdots + |x_n|) \\
&= |\beta|\|x\|_1.
\end{aligned}$$

(b) *Triangle inequality.*

$$\begin{aligned}
\|x + y\|_1 &= |x_1 + y_1| + |x_2 + y_2| + \cdots + |x_n + y_n| \\
&\le |x_1| + |y_1| + |x_2| + |y_2| + \cdots + |x_n| + |y_n| \\
&= \|x\|_1 + \|y\|_1.
\end{aligned}$$

(c) *Nonnegativity.* Each term in $\|x\|_1 = |x_1| + |x_2| + \cdots + |x_n|$ is nonnegative.

(d) *Definiteness.* $\|x\|_1 = 0$ only if $|x|_1 = \cdots = |x_n| = 0$.

Next the ∞-norm.

(a) *Homogeneity.*

$$\begin{aligned}
\|\beta x\|_\infty &= \max\{|\beta x_1|, |\beta x_2|, \dots, |\beta x_n|\} \\
&= \max\{|\beta||x_1|, |\beta||x_2|, \dots, |\beta||x_n|\} \\
&= |\beta| \max\{|x_1|, |x_2|, \dots, |x_n|\} \\
&= |\beta|\|x\|_\infty.
\end{aligned}$$

(b) *Triangle inequality.*

$$\begin{aligned}\|x+y\|_\infty &= \max\{|x_1+y_1|, |x_2+y_2|, \ldots, |x_n+y_n|\}\\ &\leq \max\{|x_1|+|y_1|, |x_2|+|y_2|, \ldots, |x_n|+|y_n|\}\\ &\leq \max\{|x_1|, |x_2|, \ldots, |x_n|\} + \max\{|y_1|, |y_2|, \ldots, |y_n|\}\\ &= \|x\|_\infty + \|y\|_\infty.\end{aligned}$$

(c) *Nonnegativity.* $\|x\|_\infty$ is the largest of n nonnegative numbers $|x_k|$.

(d) *Definiteness.* $\|x\|_\infty = 0$ only if $|x_k| = 0$ for $k = 1, \ldots, n$.

3.16 (10 pts)

(a) First we find the average of $\alpha x + \beta\mathbf{1}$:

$$\mathbf{avg}(\alpha x + \beta\mathbf{1}) = \mathbf{1}^T(\alpha x + \beta\mathbf{1})/n = (\alpha\mathbf{1}^T x + \beta\mathbf{1}^T\mathbf{1})/n = \alpha\mathbf{avg}(x) + \beta$$

where we use $\mathbf{1}^T\mathbf{1} = n$.

(b) Using the definition of $\mathbf{std}(x)$ and part (a),

$$\begin{aligned}\mathbf{std}(\alpha x + \beta\mathbf{1}) &= \mathbf{rms}(\alpha x + \beta\mathbf{1} - (\alpha\mathbf{avg}(x) + \beta)\mathbf{1}))\\ &= \mathbf{rms}(\alpha x - \alpha\mathbf{avg}(x)\mathbf{1}))\\ &= |\alpha|\mathbf{rms}(x - \mathbf{avg}(x)\mathbf{1}))\\ &= |\alpha|\mathbf{std}(x).\end{aligned}$$

3.26 (24 pts)

(a) $R(0)$ is the correlation coefficient between x and x, which is always one.
To find $R(\tau)$, we consider the vectors $(\mu\mathbf{1}_\tau, x)$ and $(x, \mu\mathbf{1}_\tau)$. They each have mean μ, so their de-meaned versions are $(0_\tau, x - \mu\mathbf{1})$ and $(x - \mu\mathbf{1}, 0_\tau)$. The two vectors have the same norm $\|x - \mu\mathbf{1}\|$. Therefore their correlation coefficient is

$$R(\tau) = \frac{(0_\tau, x - \mu\mathbf{1})^T(x - \mu\mathbf{1}, 0_\tau)}{\|x - \mu\mathbf{1}\|^2}$$

For $\tau \geq T$, the inner product in the numerator is zero, since for each i, one of the two vectors in the inner product has ith entry zero. For $\tau = 0, \cdots, T-1$, the expression for $R(\tau)$ reduces to

$$R(\tau) = \frac{\sum_{t=1}^{T-\tau}(x_t - \mu)(x_{t+r} - \mu)}{\sum_{t=1}^{T}(x_t - \mu)^2}.$$

(b) We express the formula above as the inner product

$$\begin{aligned}R(\tau) &= \left(\frac{(0_\tau, x - \mu\mathbf{1})}{\|x - \mu\mathbf{1}\|}\right)^T \left(\frac{(x - \mu\mathbf{1}, 0_\tau)}{\|x - \mu\mathbf{1}\|}\right)\\ &= \left(\frac{(0_\tau, x - \mu\mathbf{1})}{\sqrt{T}\mathbf{std}(x)}\right)^T \left(\frac{(x - \mu\mathbf{1}, 0_\tau)}{\sqrt{T}\mathbf{std}(x)}\right)\\ &= \frac{1}{T}(0_\tau, z)^T(z, 0_\tau).\end{aligned}$$

In this inner product we do not need to include the sum over the first τ entries (since the first vector has zero entries), and we do not need to sum over the last τ entries (since the second vector has zero entries). Summing over the remaining indices we get

$$R(\tau) = \frac{1}{T}\sum_{t=1}^{T-\tau} z_t z_{t+\tau}.$$

(c) The time series $x = (+1, -1, \cdots, +1, -1)$ has mean zero and norm $\sqrt{T}$. For $\tau = 0, \cdots, T-1$, the auto-correlation is

$$\begin{aligned} R(\tau) &= \frac{1}{T}\sum_{t=1}^{T-\tau} x_t x_{t+\tau} \\ &= \frac{1}{T}\sum_{t=1}^{T-\tau} (-1)^{t+1}(-1)^{t+\tau+1} \\ &= \frac{T-\tau}{T}(-1)^{\tau} \\ &= \begin{cases} 1-\tau/T & \tau \textit{ even} \\ -1+\tau/T & \tau \textit{ odd.} \end{cases} \end{aligned}$$

(d) $R(7)$ large (*i.e.*, near one) means that x_t and x_{t+7} are often above or below the mean value together. x_{t+7} is the number of meals served exactly one week after x_t, so this means that, for example, Saturdays are often above the mean together, and Tuesdays are often below the mean together.

A3.2 (8 pts) The minimum distance is 3.1622776601683795, which is from a to x_1. x_4 makes the smallest angle with a. The angle is 0.24256387409548533.

A3.5 (8 pts)

$$\begin{aligned} \|\alpha a + \beta b + \gamma c\|^2 &= (\alpha a + \beta b + \gamma c)^T(\alpha a + \beta b + \gamma c) \\ &= (\alpha a)^T(\alpha a) + (\alpha a)^T(\beta b) + (\alpha a)^T(\gamma c) \\ &\quad + (\beta b)^T(\alpha a) + (\beta b)^T(\beta b) + (\beta b)^T(\gamma c) \\ &\quad + (\gamma c)^T(\alpha a) + (\gamma c)^T(\beta b) + (\gamma c)^T(\gamma c) \\ &= \alpha^2(a^T a) + \beta^2(b^T b) + \gamma^2(c^T c) \\ &= \alpha^2\|a\|^2 + \beta^2\|b\|^2 + \gamma^2\|c\|^2. \end{aligned}$$

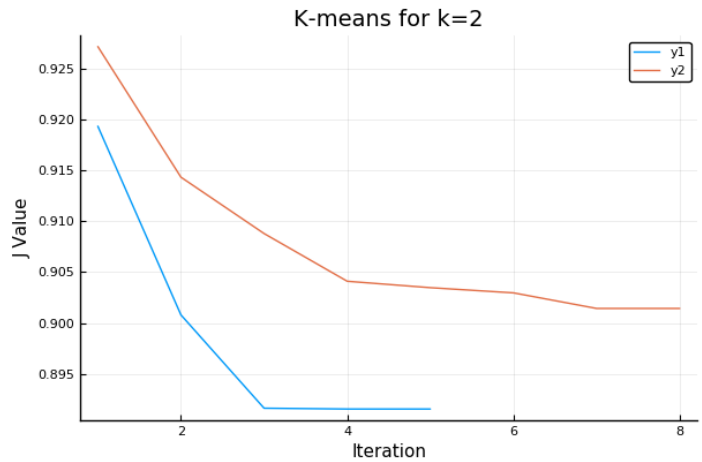
So $\|\alpha a + \beta b + \gamma c\| = \sqrt{\alpha^2\|a\|^2 + \beta^2\|b\|^2 + \gamma^2\|c\|^2}$.

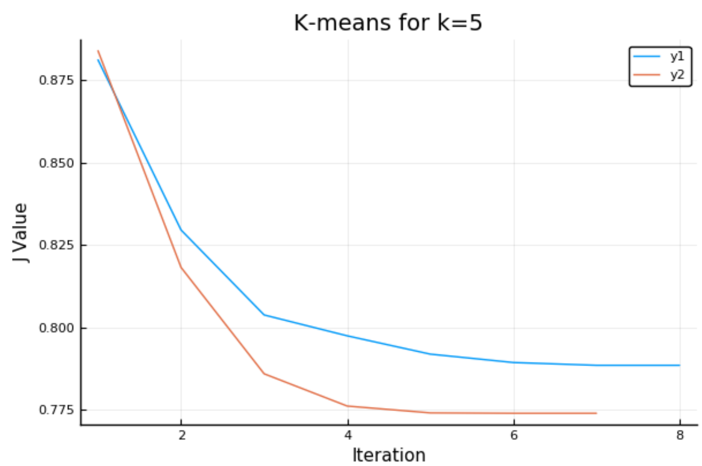
A4.2 (20 pts) For this question the graphs for everyone is expected to be different, due to the randomness involved in creating the graphs. So the grading will be accordingly.

(a) Some observations might include:

- In each run, the objective function is always monotonically non-increasing.
- They don't necessarily converge to the same objective value, which is expected since k-means is a heuristic—there is no guarantee that the output of the algorithm actually minimizes the objective.

- Typically, a larger k results in a smaller objective value at the end; this again not surprising, as we can assign all the points to less than k centoids, and having the choice of more centroids centainly could not hurt.





(b) When $k = 2$ the topics may not make too much sense, so it is better approach to pick $k = 5$. You can find the words that are clustered together in the file 'Articles.txt'. When $k = 5$, the articles are clustered into 5 distinguishable classes.

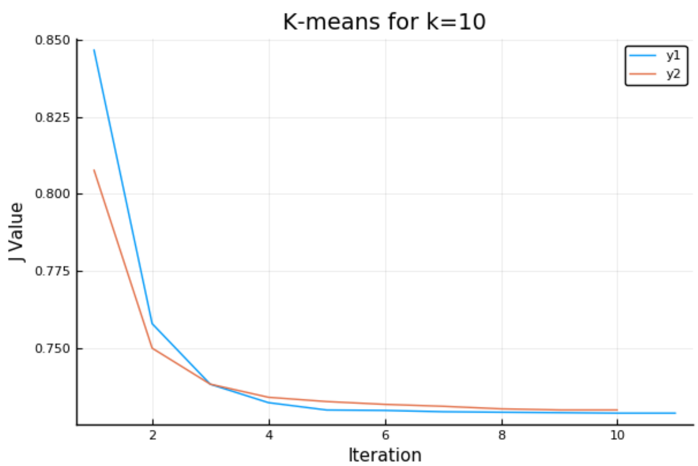
Most common 3 words for Centroid 1 are nations, international and member. Hence, Centroid 1 represents the organizations.

For Centroid 2, the most common 3 words are paintings, art and artists. Hence, Centroid 2 represents the articles related to art.

For Centroid 3, the most common 3 words are radio, signal and frequency. Hence, Centroid 3 represents the articles related to communication.

For Centroid 4, the most common 3 words are pokemon, game and player. Hence, Centroid 4 represents the articles about Pokemon.

For Centroid 5, the most common 3 words are are weather, pressure and wind. And hence, Centroid 5 represents the articles about weather and climate.


K-means for k=10
J Value
Iteration
0.850
0.825
0.800
0.775
0.750
2
4
6
8
10
y1
y2