Solution to Written Assignment 2
SENG 474/CSC 578D

Question 1

(a) Let $x_1 = (\begin{bmatrix}1\\2\end{bmatrix}, -1), x_2 = (\begin{bmatrix}2\\1\end{bmatrix}, -1), x_3 = (\begin{bmatrix}3\\6\end{bmatrix}, 1), x_4 = (\begin{bmatrix}4\\7\end{bmatrix}, 1)$. We have:

$$d(x_1, L) = \frac{|4 \cdot 1 + 3 \cdot 2 - 12|}{\sqrt{4^2 + 3^2}} = \frac{2}{5} \tag{1}$$

Similarly, we have $d(x_2, L) = \frac{1}{5}, d(x_3, L) = \frac{18}{5}, d(x_4, L) = \frac{25}{5} = 5$. Thus, the margin of (L) is:

$$\min(\frac{2}{5}, \frac{1}{5}, \frac{18}{5}, 5) = \frac{1}{5} \tag{2}$$

(b) Look at two points x_1, x_3. The line (L) separating this two points must go through the midpoint $(\begin{bmatrix}2\\4\end{bmatrix})$ and perpendicular to the vector $\vec{x_1x_3} = (\begin{bmatrix}2\\4\end{bmatrix})$. This line is:

$$2(x - 2) + 4(y - 4) = 0 \Leftrightarrow x + 2y - 10 = 0 \tag{3}$$

The distances of x_1, x_3 to (L) are the same: $\sqrt{5}$ and this is also the minimum distance of the point set $\{x_1, x_2, x_3, x_4\}$ to (L). Thus (L) must be the SVM line.

Question 2

(a) Let the line be $y = ax + b$. We must find (a, b) to minimize:

$$f(a, b) = (a + b - 1)^2 + (a + b - 2)^2 + (2a + b - 2)^2 + (2a + b - 3)^2. \tag{4}$$

Then compute the partial derivatives of $f(a, b)$ w.r.t a and b, we have:

$$\frac{\partial f(a, b)}{\partial a} = 20a + 12b - 26 \qquad \frac{\partial f(a, b)}{\partial b} = 12a + 8b - 16 = 0 \tag{5}$$

Then a, b can be obtained by solving the system of equations:

$$\begin{cases} 20a + 12b - 26 = 0 \\ 12a + 8b - 16 = 0 \end{cases}$$

(b) We have:

$$\begin{aligned} \nabla_B \mathrm{Tr}(C \nabla_A \mathrm{Tr}(AB)) &= \nabla_B \mathrm{Tr}(CB^T) \\ &= (\nabla_{B^T} \mathrm{Tr}(CB^T))^T \\ &= (C^T)^T = C \\ &= \begin{bmatrix} 1 & -1 \\ 2 & 1 \end{bmatrix} \end{aligned} \tag{6}$$

Question 3

(a) Observe that: among all frequent item that have a prime factor $z \neq 2$, the smallest would be $2^2 z$.

From that, we have $2^2 z \leq 100 \Rightarrow z \leq 25$. Now consider each prime number smaller than 25:
$z = 23 : 2^2 \cdot 23 = 92$ is a frequent item.
$z = 19 : 2^2 \cdot 19$ is a frequent item.
$z = 17 : 2^2 \cdot 17$ is a frequent item.
$z = 13 : 2^2 \cdot 13, 2 \cdot 3 \cdot 13 = 78$ are a frequent items.
$z = 11 : 2^2 \cdot 11, 3^2 \cdot 11, 2 \cdot 3 \cdot 11$ are frequent items.
$z = 7 : 2^2 \cdot 7, 2^3 \cdot 7, 3^2 \cdot, 2 \cdot 3 \cdot 7, 2^2 \cdot 3 \cdot 7, 2 \cdot 5 \cdot 7$ are frequent items.
$z = 5 : 2^2 \cdot 5, 2^3 \cdot 5, 2^4 \cdot 5, 3^2 \cdot 5, 2 \cdot 3 \cdot 5, 2^2 \cdot 3 \cdot 5, 2 \cdot 5^2, 3 \cdot 5^2, 2^2 \cdot 5^2$ are frequent items.
$z = 3 : 2^2 \cdot 3, 2^3 \cdot 3, 2^4 \cdot 3, 2^5 \cdot 3, 2 \cdot 3^2, 2^2 \cdot 3^2, 2^3 \cdot 3^2, 2 \times 3^3$ are frequent items.
and lastly,
$2^4, 2^5, 2^6$ are frequent items.

(b) Observe that (x, y) is a frequent item pair if and only if $\gcd(x, y)$ is a frequent item. Furthermore, since $2 \gcd(x, y) < \max(x, y) \leq 100$, we have $\gcd(x, y) \leq 50$.

Thus, the way to list all frequent itempairs is to find all frequent items less than 50 to be $\gcd(x, y)$ and then form (x, y). For example,

Frequent items when $z \geq 13$ in part (a) cannot be $\gcd(x, y)$ because they are all bigger than 50. Let consider the frequent item $2^2 \times 11 = 44$ (when $z = 11$). The only possible frequent item pair is $(44, 44 \times 2) = (44, 88)$.

Consider frequent items that has $z = 5$ as a factor. The only candidates for $\gcd(x, y)$ are $2^2 \cdot 5 = 20, 2^3 \cdot 5 = 40, 2 \cdot 3 \cdot 5 = 30$. For 20, the corresponding frequent itempairs are $(20, 40), (20, 60), (20, 80), (20, 100)$. For 40, the only corresponding frequent itempairs is $(40, 80)$. For 30, the corresponding frequent itempairs are $(30, 60), (30, 90)$.

(c) $\gcd(12, 60, 8) = \{1, 2, 4\}$ and $\gcd(12, 60) = \{1, 2, 3, 4, 6, 12\}$. Thus, the confidence of $\{12, 60\} \rightarrow 8$ is

$$\frac{3}{6} = \frac{1}{2} \tag{7}$$

Question 4

(a) The probability that two pairs are hashed to the same location is:

$$1 - (1 - 0.6^4)^{10} = \frac{3}{4} \tag{8}$$

The total number of distanct pairs is $\frac{nm}{2}$. Thus, the number of expected pairs found by MinHash is:

$$\frac{mn}{2} \frac{3}{4} = \frac{3mn}{8} \tag{9}$$

(b) The probability that for a given p, there is at least one similar question is put to the table in the same location with p is:

$$1 - \frac{1}{4}^m = 1 - \frac{1}{n} \tag{10}$$

Thus, *any* question p, the probability that there is at least one similar question is put to the table in the same location with p is:

$$(1 - \frac{1}{n})^n \sim e^{-1} \tag{11}$$

Thus, the probability that there exists at least one questions that the MinHash returns no similar question is:

$$1 - e^{-1} \sim 0.63 \tag{12}$$