Assignment 1 - AIT 512 Maths for ML

Due on **September 7, 2025**

August 31, 2025

1. Consider the vectors $\vec{x}_1, \cdots, \vec{x}_L,\ L > 1,\ \vec{x}_i \in \Re^{n\times 1}\ \forall\, i$ and the centroid vector $\vec{\overline{x}} = \frac{1}{L}\left(\vec{x}_1 + \cdots + \vec{x}_L\right)$. Consider another vector $\vec{z} \in \Re^{n\times 1}$.

   - Show that it is only for $\vec{z} = \vec{\overline{x}}$ does the sum $J = \sum_{i=1}^{L} \|\vec{x}_i - \vec{z}\|^2$ have the smallest value? Note that $\|\cdot\|^2$ is the square of the 2-norm of any vector.
   - In which application would this sum be used?

   Hints: To each of the elements in the sum J, add and subtract $\vec{\overline{x}}$. Simplify the resulting expression and argue that $\vec{z} = \vec{\overline{x}}$ minimises the sum. There is no need to apply any differential calculus.

2. Let the vectors $\vec{x}_1, \cdots, \vec{x}_m \in \Re^{n\times 1}$ be such that they can be partitioned into exactly *two* clusters, G_1 and G_2, whose representative vectors are $\vec{z}_1, \vec{z}_2$, respectively. Let some vector $\vec{x}_i \in G_1$. What would be the sign of the difference $\|\vec{x}_i - \vec{z}_1\|^2 - \|\vec{x}_i - \vec{z}_2\|^2$? What would be the sign of this difference if $\vec{x}_i \in G_2$?

   Can you now state a classification rule, of the form $\vec{w}^T \vec{x}_k + v > 0$ (or < 0), to determine if the vector $\vec{x}_k$ belongs to G_1 or G_2?

3. Consider the vectors $\vec{x}_1, \cdots, \vec{x}_L,\ L > 1,\ \vec{x}_i \in \Re^{n\times 1}\ \forall\ i$ and another vector $\vec{z} \in \Re^{n\times 1}$. The vector $\vec{x}_i$ is the *distance* nearest neighbour to $\vec{z}$ if $\|\vec{z} - \vec{x}_i\|^2 \le \|\vec{z} - \vec{x}_j\|^2,\ i \ne j$. Similarly $\vec{x}_i$ is the *angle* nearest neighbour to $\vec{z}$ if $\angle(\vec{x}_i, \vec{z}) \le \angle(\vec{x}_j, \vec{z}),\ i \ne j$. Note that the angle between any vectors can be computed as

$$\angle(\vec{x}_i, \vec{z}) = \gamma_i = \cos^{-1}\left(\frac{\vec{x}_i^T \vec{z}}{\|\vec{x}_i\|\,\|\vec{z}\|}\right).$$

   - Give a simple specific numerical example where the *distance* nearest neighbor is not the same as the *angle* nearest neighbour. Vectors with 2 elements can be considered.
   - Suppose the vectors $\vec{x}_i$ are normalized. Show that in this case, the *distance* nearest neighbour and the *angle* nearest neighbour are always the same. **Hint**: $\cos^{-1}$ is a decreasing function, that is, for any scalars a and b, where $-1 < a < b < 1$, the relation $\cos^{-1}(a) > \cos^{-1}(b)$ holds.

4. Consider the n vectors $\vec{x}_i \in \Re^{n\times 1},\ i = 1, \cdots, n$, where the vectors $\vec{x}_i$ are of the type

$$\vec{x}_1 = \begin{bmatrix}1\\0\\0\\\vdots\\0\end{bmatrix},\ \vec{x}_2 = \begin{bmatrix}1\\1\\0\\\vdots\\0\end{bmatrix},\ \cdots,\ \vec{x}_n = \begin{bmatrix}1\\1\\1\\\vdots\\1\end{bmatrix}.$$

   Using the Gram-Schmidt algorithm, examine if the vectors $\vec{x}_i$ form an independent set. Suppose the algorithm does not terminate early, what other property do these vectors have?

5. A questionnaire in a magazine has 30 questions, broken into two sets of 15 questions. Someone taking the questionnaire answers each question with ‘Rarely’, ‘Sometimes’, or ‘Often’. The answers are recorded as a 30-vector $\overrightarrow{a}$, with $a_i = 1, 2, 3$ if question i is answered ‘Rarely’, ‘Sometimes’, or ‘Often’, respectively.

   The total score on a completed questionnaire is found by adding up 1 point for every question answered ‘Sometimes’ and 2 points for every question answered ‘Often’ on questions 1–15, and by adding 2 points and 4 points for those responses on questions 16–30. (Nothing is added to the score for Rarely responses.) Express the total score s in the form $s = \overrightarrow{w}^T \overrightarrow{a} + v$, where $\overrightarrow{w}$ is a 30-vector and v is a scalar.

   Hints: For each type of response, find a function that will yield the number of points for that response; do this separately for the set 1-15 and for the set 16-30. Next, add the outputs of these functions to yield the total score.