Mathematical Foundations of Machine Learning
Assignment 1

David Felipe Alvear Goyes

1. Minimizing mean square distance to a set of vectors.
Let $x_1, \ldots, x_L$ be a collection of vectors. In class, when describing the k-means algorithm, we asserted that the vector z which minimizes the sum-square distance to the vectors,

$$J(z) = \|x_1 - z\|^2 + \cdots + \|x_L - z\|^2,$$

is the average or centroid of the vectors, $\bar{x} = (1/L)\sum_i x_i$. You will show this property in this exercise.

- Explain why, for any z, we have

$$J(z) = \sum_{i=1}^{L} \|x_i - \bar{x} - (z - \bar{x})\|^2 = \sum_{i=1}^{L} \left(\|x_i - \bar{x}\|^2 - 2(x_i - \bar{x})^T(z - \bar{x})\right) + L\|z - \bar{x}\|^2.$$

- Explain why $\sum_{i=1}^{L}(x_i - \bar{x})^T(z - \bar{x}) = 0$. *Hint.* Write the left-hand side as $\left(\sum_{i=1}^{L}(x_i - \bar{x})\right)^T(z - \bar{x})$, and argue that the left-hand vector is 0.
- Combine the two results above to get $J(z) = \sum_{i=1}^{L}\|x_i - \bar{x}\|^2 + L\|z - \bar{x}\|^2$. Explain why for any $z \neq \bar{x}$, we have $J(z) > J(\bar{x})$. This shows that the choice $z = \bar{x}$ minimizes $J(z)$.

Solution

→ Explain why for any z:

$$J(z) = \sum_{i=1}^{L} \|x_i - \bar{x} - (z - \bar{x})\|^2 = \sum_{i=1}^{L} (\|x_i - \bar{x}\|^2 - 2(x_i - \bar{x})^T(z - \bar{x})) + L\|z - \bar{x}\|^2$$

we have: $J(z) = \|x_1 - z\|^2 + \cdots + \|x_L - z\|^2$

and $\bar{x} = \frac{1}{L}\sum_{i=1}^{L} x_i$ } Average of centroids of the vectors

→ $J(z) = \sum_{i=1}^{L} \|x_i - z\|^2$ → we can add: $-\bar{x} + \bar{x}$

$$J(z) = \sum_{i=1}^{L} \|x_i - \bar{x} - z + \bar{x}\|^2 = \sum_{i=1}^{\infty} \|\underline{x_i - \bar{x}} - \underline{(z - \bar{x})}\|^2$$

Using Cauchy schwartz inequality $\|a+b\|^2 = \|a\|^2 + a^Tb + \|b\|^2$

$$J(z) = \sum_{i=1}^{L} \|x_i - \bar{x}\|^2 - 2(x_i - \bar{x})^T(z - \bar{x}) + \underline{\|z - \bar{x}\|^2}$$

term not depends on i

$$\boxed{J(z) = \sum_{i=1}^{L} \left(\|x_i - \bar{x}\|^2 - 2(x_i - \bar{x})^T(z - \bar{x}) \right) + L\|z - \bar{x}\|^2}$$

→

- Explain why $\sum_{i=1}^{L}(x_i - \bar{x})^T(z - \bar{x}) = 0$. *Hint.* Write the left-hand side as $\left(\sum_{i=1}^{L}(x_i - \bar{x})\right)^T (z - \bar{x})$, and argue that the left-hand vector is 0.

$$\sum_{i=1}^{L}(x_i - \bar{x})^T(z - \bar{x}) = 0 \rightarrow \left[\sum_{i=1}^{L}(x_i - \bar{x})\right]^T(z - \bar{x}) = 0$$

$$= \left[\left(\underline{\sum_{i=1}^{L} x_i}\right) - L\bar{x}\right]^T(z - \bar{x}) = 0$$

→ we have that: $\bar{x} = \frac{1}{L}\sum_{i=1}^{L} x_i \rightarrow L\bar{x} = \sum_{i=1}^{L} x_i$ Replace in eq.

$$\boxed{\begin{aligned} &= [L\bar{x} - L\bar{x}](z - \bar{x}) = [\emptyset](z - \bar{x}) \\ &\qquad = \emptyset \end{aligned}}$$

- Combine the two results above to get $J(z) = \sum_{i=1}^{L} \|x_i - \bar{x}\|^2 + L\|z - \bar{x}\|^2$. Explain why for any $z \neq \bar{x}$, we have $J(z) > J(\bar{x})$. This shows that the choice $z = \bar{x}$ minimizes $J(z)$.

→ Joining the results

we had → $J(z) = \sum_{i=1}^{L} \left(\|x_i - \bar{x}\|^2 - 2(x_i - \bar{x})^T(z - \bar{x}) \right) + L\|z - \bar{x}\|^2$

$$J(z) = \sum_{i=1}^{L} \|x_i - \bar{x}\|^2 - \sum_{i=1}^{L} 2(x_i - \bar{x})^T(z - \bar{x}) + L\|z - \bar{x}\|^2$$

but we found → $\sum_{i=1}^{L}(x_2 - \bar{x})^T(z - \bar{x}) = 0$

then → $\boxed{J(z) = \sum_{i=1}^{L} \|x_i - \bar{x}\|^2 + L\|z - \bar{x}\|^2}$

why → For any $z \neq \bar{x}$ we have $J(z) > J(\bar{x})$?

→ Let's check $J(\bar{x}) = \sum_{i=1}^{L} \|x_i - \bar{x}\|^2 + L\cancelto{0}{\|\bar{x} - \bar{x}\|}^2 = \sum_{i=1}^{L} \|x_i - \bar{x}\|^2$

now → $\sum_{i=1}^{L} \cancel{\|x_i - \bar{x}\|^2} + L\|z - \bar{x}\|^2 > \sum_{i=1}^{L} \cancel{\|x_i - \bar{x}\|^2}$

$L\|z - \bar{x}\|^2 > 0$

$\boxed{\|z - \bar{x}\|^2 > 0}$ → this implies that the norm is grefear than zero when $z \neq \bar{x}$, concluding that the minimum value of $J(z)$ occurs when $z = \bar{x}$.

2. Linear separation in 2-way partitioning.
Clustering a collection of vectors into $k = 2$ groups is called 2-way partitioning. Suppose we run k-means, with $k = 2$, on the vectors $x_1, \ldots, x_N$ to obtain two groups G_1 and G_2. Show that there is a nonzero vector w and a scalar b that satisfy

$$w^T x_i + b \geq 0 \text{ for } i \in G_1, \quad w^T x_i + b \leq 0 \text{ for } i \in G_2.$$

In other words, the affine function $f(x) = w^T x_i + b$ is greater than or equal to zero on the first group, and less than or equal to zero on the second group.
Hint. Consider the function $\|x - z_1\|^2 - \|x - z_2\|^2$, where z_1 and z_2 are the group representatives.

$\|x\| - 2x' z_1 + \|z_1\|^2 - \|x\|^2 + 2x' z_2 + \|z_2\|^2$

$2x'(z_1 - z_2) + \|z_1\|^2 + \|z_2\|^2$

↳ $2(z_1 - z_2)' x + \|z_1\|^2 + \|z_2\|^2$

Solution

2-way Partitioning

Vectors = $x_1, x_2 \cdots x_N$
$K = 2$
G_1, G_2

→ show that there exist $w \neq 0$, $b \neq 0$
- $w^T x_i + b \geq 0$ for $i \in G_1$
- $w^T x_i + b \leq 0$ for $i \in G_2$

$$\sum_{i=1}^{N} \|x_i - z_1\|^2 - \|x_i - z_2\|^2 = \sum_{i=1}^{N} \left(\cancel{\|x_i\|^2} - 2x_i^T z_1 + \|z_1\|^2\right) - \left(\cancel{\|x_i\|^2} - 2x_i^T z_2 + \|z_2\|_2\right)$$

$$= \sum_{i=1}^{N} \|z_1\|^2 - \|z_2\|^2 + 2\left(x_i^T z_2 - x_i^T z_1\right) = N\left(\|z_1\|^2 - \|z_2\|^2\right) + 2\sum_{i=1}^{N} x_i^T (z_2 - z_1)$$

$$= N\left(\|z_1\|^2 - \|z_2\|^2\right) + 2N\bar{X}^T (z_2 - z_1) = N\left[\left(\|z_1\|^2 - \|z_2\|^2\right) + 2\bar{X}^T (z_2 - z_1)\right]$$

Generalizing $x \in \{x_1 \cdots x_N\}$

We can see: $\|z_1\|^2 - \|z_2\|^2$ is constant

then call $\|z_1\|^2 - \|z_2\|^2 = b$

and → $2x^T(z_2 - z_1) \longrightarrow \underbrace{2(z_2 - z_1)^T}_{w} x = w^T x$

$$2(z_2 - z_1)^T x + 2\left(\frac{b}{2}\right) \geqslant 0$$

$$(z_2 - z_1)^T x + \frac{b}{2} \geqslant 0$$

↪ we can add $\frac{1}{2}$ to the constant.

$$(z_2 - z_1)^T x + b \geqslant 0 \quad \forall x \in \{x_1, x_2, \dots x_N\}$$

→ Analyze

- $(z_2 - z_1)^T x : x \in G_1$ → Given that $x \in G_1$ then x is closer to z_1, then $z_2^T x > z_1^T x$

For this reason: $\boxed{(z_2 - z_1)^T x + b \geqslant 0 \quad \text{when } x \in G_1}$

- $(z_2 - z_1)^T x : x \in G_2$ → Given that $x \in G_2$ then x is closer to z_2 and far from z_1. then $z_2^T x < z_1^T x$

For this reason: $\boxed{(z_2 - z_1)^T x + b \leq 0 \quad \text{when } x \in G_2}$

3. Clustering the Iris dataset.

You will find on Blackboard the iris data set which consists of 150 training vectors, each consisting of four features (csv file posted on Blackboard, along with a starter file to read the data).

- Write python code to cluster the vectors into $k = 3$ groups. You may want to perform multiple runs, say 10, starting from different random initializations and pick the best one.
- Plot the clustering objective vs iteration count to verify convergence
- Generate a 3d scatter plot (using any 3 features of the data), to visualize the groups generated. Comment.

4. Classification from clustering.

We can use the results of a clustering algorithm to construct a classifier as follows. We first label the k groups generated, and then assign (or classify) *new* vectors to one of the k groups by choosing the nearest group representative. In this exercise, we consider building a (rudimentary) digit classifier which would automatically guess what a written digit is from its image.

- You will find on Blackboard a csv version of the MNIST training data set along with a starter file to read the data. Cluster the data into $k = 20$ groups and display the representative vectors for each group. (As earlier, you may want to perform multiple runs, say 10, starting from different initializations and pick the best one.)
- Label each of the generated groups with one of the digits $0, 1, \ldots, 9$. You may do so by hand from the representative vectors, or you may find the most common label of the vectors in every group. Note that multiple groups will have the same digit associated with them. Also the labels you associate with the groups might be reasonably ambiguous.
- You will also find an MNIST testing data set on Blackboard. Assign each vector in this set to one of the k groups by choosing the nearest group representative. Compare the digit labeling the group to the correct digit, to compute the fraction of the test vectors that are classified correctly.