Math 330 Linear Algebra Homework 6

4.1 *Minimizing mean square distance to a set of vectors.* Let $x_1, \ldots, x_L$ be a collection of n-vectors. In this exercise you will fill in the missing parts of the argument to show 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, $\overline{x} = (1/L)(x_1 + \cdots + x_L)$. (This result is used in one of the steps in the k-means algorithm. But here we have simplified the notation.)

(a) Explain why, for any z, we have

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

(b) Explain why $\sum_{i=1}^{L}(x_i - \overline{x})^T(z - \overline{x}) = 0$. *Hint.* Write the left-hand side as

$$\left(\sum_{i=1}^{L}(x_i - \overline{x})\right)^T (z - \overline{x}),$$

and argue that the left-hand vector is 0.

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

(a) By the definition of the norm and expanding

$$\|x_i - \overline{x} - (z - \overline{x})\|^2$$

$$= (x_i - x - (z - \overline{x}))^T (x_i - x - (z - \overline{x}))$$

$$= (x_i - x)^T(x_i - x) - (z - \overline{x})^T(x_i - x) - (x_i - x)^T(z - \overline{x}) + (z - \overline{x})^T(z - \overline{x})$$

$$= \|x_i - x\|^2 - 2(x_i - x)^T(z - \overline{x}) + \|z - \overline{x}\|^2$$

Therefore

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

doesn't depend on the index i

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

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

(b) Since $\bar{x} = (1/L)(x_1 + \cdots + x_L)$ then

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

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

$$= (L\bar{x} - L\bar{x})^T (z - \bar{x}) = 0.$$

(c) Combining the results of (a) and (b) gives

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

Now, if $z = \bar{x}$ then

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

On the other hand if $z \neq \bar{x}$, then

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

Therefore, subtracting yields

$$J(z) - J(\bar{x}) = L \| z - \bar{x} \|^2 > 0$$

when $z \neq \bar{x}$, or equivalently

$$J(z) > J(\bar{x}).$$

This shows the choice of $z = \bar{x}$ minimized the function $J(z)$.