

National University of Singapore
 School of Computing
 CS3244: Machine Learning
 Solution to Tutorial 10
Unsupervised Learning

1. ***K*-means convergence:** We have learned the iterative algorithm for *K*-means. Let's look into the algorithm again.

Algorithm 1 *K*-means Clustering

```

1: for  $c = 1$  to  $k$  do
2:    $\mu_c \leftarrow$  some random location
3: while Not Converged do
4:   for  $j = 1$  to  $m$  do
5:      $y^{(j)} \leftarrow c = \arg \min_c \|x^{(j)} - \mu_c\|^2$  ▷ A. assign example
6:      $x^{(j)} \leftarrow S_c$ 
7:   for  $c = 1$  to  $k$  do
8:      $\mu_c \leftarrow \frac{1}{|S_c|} \sum_{x \in S_c} x$  ▷ B. re-estimate center
9: return  $y$ 

```

The algorithm leads to convergence when the clustering quality L_{clust} is minimized (*i.e.*, when cluster centers or assignments stop changing).

$$L_{clust} = \sum_{c=1}^k \sum_{x \in S_c} \|x - \mu_c\|^2 \quad (1)$$

- (a) Show that each data assignment step (Line 5 in Algorithm 1) minimizes L_{clust} , given fixed cluster centers.
 - (b) Show that each cluster center update step (Line 8 in Algorithm 1) minimizes L_{clust} , given fixed data assignments.
 - (c) Despite the guarantee of convergence, the clustering result varies depending on initialization of the centroids. Sub-optimal clustering may also occur when the clusters are of different sizes and densities. Suggest two solutions to overcome this issue.
2. **Auto-Encoder:** An auto-encoder is a neural network that is trained to attempt to reconstruct the input to the output. Autoencoder has an encoder E and a decoder D . The encoder $\mathbf{z} = E(\mathbf{x})$ consumes the input \mathbf{x} and produce an intermediate representation \mathbf{z} which is fed into the decoder $\hat{\mathbf{x}} = D(\mathbf{z})$ to reconstruct the input. Figure 1 shows this.

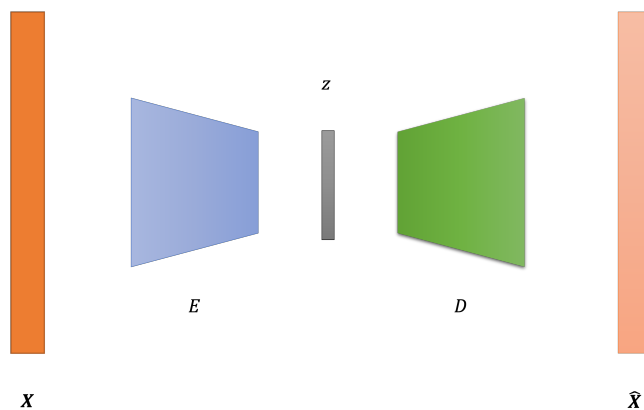


Figure 1: Autoencoder

- (a) Given a large image dataset without label (dataset P) and a small image dataset (dataset Q) with label, both datasets come from the same domain. Propose a way to improve classifier's performance on the small dataset.
- (b) Autoencoder can compress the original input into a lower dimensional encoding. However, it is hardly used in practice for image compression. List the disadvantages of autoencoder when used for compression.