1 Week 7 Unsupervised Learning

1.1 K-means Clustering

- In this this exercise, you will implement the K-means algorithm and use it for image compression.
- You will first start on an example 2D dataset that will help you gain an intuition of how the K-means algorithm works. After that, you wil use the K-means algorithm for image compression by reducing the number of colors that occur in an image to only those that are most common in that image.
- You will be using ex7.m for this part of the exercise.

1.2 Implementing K-means

- The K-means algorithm is a method to automatically cluster similar data examples together.
- The intuition behind K-means is an iterative procedure that starts by guessing the initial centroids, and then refines this guess by repeatedly assigning examples to their closest centroids and then recomputing the centroids based on the assignments.
- The inner-loop of the algorithm repeatedly carries out two steps:
  (i) Assigning each training example to its closest centroid
  (ii) Recomputing the mean of each centroid using the points assigned to it.
- The K-means algorithm will always converge to some final set of means for the centroids. Note that the converged solution may not always be ideal and depends on the initial setting of the centroids.
- Therefore, in practice the K-means algorithm is usually run a few times with different random initializations.
- One way to choose between these different solutions from different random initializations is to choose the one with the lowest cost function value (distortion).
- Random initialization The initial assignments of centroids for the example dataset in `ex7.m` were designed so that you will see the same gure as in Figure 1.
- In practice, a good strategy for initializing the centroids is to select random examples from the training set.

Question 1.

For which of the following tasks might K-means clustering be a suitable algorithm? Select all that apply.

Question 2.

Suppose we have three cluster centroids μ_1=$[12]^T$, μ_2=$[-30]^T$ and μ_3=$[42]^T$. Furthermore, we have a training example x(i)=$[31]^T$. After a cluster assignment step, what will $c^{(i)}$ be?

Question 3.

K-means is an iterative algorithm, and two of the following steps are repeatedly carried out in its inner-loop. Which two?

Question 4.

Suppose you have an unlabeled dataset $\{x(1), \ldots, x(m)\}$. You run K-means with 50 different random initializations, and obtain 50 different clusterings of the data. What is the recommended way for choosing which one of these 50 clusterings to use?

Question 5.

Which of the following statements are true? Select all that apply.

Anomaly Detection

- Suppose you are developing an anomaly detection system to catch manufacturing defects in airplane engines.
- Your model uses
$$p(x) = \Pi_{j=1}^{n} p(x_j; \mu_j, \sigma_j^2)$$
- You have two features x_1 = ***vibration intensity***, and x_2 = ***heat generated***.
- Both x_1 and x_2 take on values between 0 and 1 (and are strictly greater than 0), and for most "normal" engines you expect that $x_2 \approx x_2$.
- One of the suspected anomalies is that a flawed engine may vibrate very intensely even without generating much heat (large x_1, small x_2), even though the particular values of x_1 and x_2 may not fall outside their typical ranges of values.
- What additional feature x_3 should you create to capture these types of anomalies:

Solution Options

- $x_3 = x_1 + x_2$ This could take on large or small values for both normal and anomalous examples, so it is not a good feature.