Unit 4 - lecture 13: Clustering 1

Objectives:

- Understand the definition of clustering
- Understand clustering cost with different similarity measures
- Understand the K-means algorithm

Introduction to Clustering:

classification : $S_m = \{ (x^{(i)}, y^{(i)}) \mid i = 1, \dots m \}$

The goal of the algorithm was to learn the mapping from the feature vectors to the corresponding label, and the training data was the main source for learning such a mapping

In unsupervised learning, we will still have a training set, to some extent but we will not have a label.

unsupervised learning : $S_m = \{ x^{(i)} \mid i = 1, \dots m \}$

We are trying to learn some meaningful structure which would represent this set.

We can clearly see that even if we don't know the labels, there is a very clear structure in our training data, and it will be meaningful for us to automatically identify this structure

Clustering Example: Image Quantization

typical picture : 1024 × 1024 pixels   each pixel : 24 bits → $\underset{\text{Red}}{\underline{8}}\ \underset{\text{Blue}}{\underline{8}}\ \underset{\text{green}}{\underline{8}}$

Size : $1024 \times 1024 \times 24 \approx 3M$

let's say I want to compress it to use much less space to get high-quality image.

I'm going to limit my self by 32 colors $= 2^5$ to encode them

Size : $1024 \times 1024 \times 5 + \underbrace{32}_{} \times 24 \approx 640k$

dictionary which remember how we can translate each one of these colors to our original encoding in 24-bit representation

We have 2 goals : compress the image as much as possible and preserve the quality

And we have controle on it, by deciding how many clusters do we have (number of clusters K)

Clustering Definition:

The first way of thinking about clustering is thinking about it as partitioning.

Clustering input : $S_m = \{ \underbrace{x^{(i)} | i=1, \dots m}_{\text{set of feature vectors}} \}, \underbrace{k}_{\text{number of clusters}}$

Clustering Output : $\underbrace{C_1, \dots C_K}_{\text{partitions}}$; $\cup C_j = \{1, \dots m\}, C_i \cap C_j = \emptyset \ (i \neq j)$

We can think about clustering as selecting representatives: $\underbrace{z^{(1)}, \dots z^{(k)}}_{}$ vectors which represent every single partitioning

Similarity Measures - Cost functions:

How to define the clustering cost?

looking a these points, let's say I'm forced to divide it into two clusters, I can divide them in many ways. How do I know which one of them is better?

$cost(C_1, \dots C_k) = \sum_{j=1}^{k} cost(C_j)$ There are many ways to define the cost of a specific cluster

cost (C)?

- diameter
- average distance
- $Cost(C, z) = \sum_{i \in C} \underline{distance(x^{(i)}, z)}$ We will use this one.

$$\cos(x^{(i)}, x^{(j)}) = \frac{x^{(i)} \cdot x^{(j)}}{\|x^{(i)}\| \cdot \|x^{(j)}\|}$$

$$dist(x^{(i)}, x^{(j)}) = \| x^{(i)} - x^{(j)} \|^2$$

I want to define the cost of the partitioning in terms of the distance, in this case, Euclidean squared distance between the elements of the cluster and the representative z.

$$Cost(C_1, \dots C_K, z^{(1)}, \dots z^{(k)}) = \sum_{j=1}^{k} \sum_{i \in C_j} \| x^{(i)} - z^{(j)} \|^2$$

The K-means Algorithm: The Big Picture

We need an Algorithm which would tell us how to navigate in this space to get to the right positions. That's the K-Mean Algorithm.

It will take a selection of points, the whole space of points, and then randomly assigned the representatives.

Then every representative will draw itself the participants which are closest to it.

We Randomly assigned the representative, and then every point was looking for the best one for itself. And then we find, we devoted the representative according to that cloud.

1. Randomly select $z^{(1)}$ $z^{(k)}$

2. Iterate on 2 steps :

   a. Given $z^{(1)} \,..\, z^{(k)}$, assign x's to the closest z.

   $$\text{cost}(z^1, \dots z^{(k)}) = \sum_{i=1}^{m} \min_{j=1\,k} \| x^{(i)} - z^{(j)} \|^2$$

   b. Given $C_1 \dots C_k$ find the best representatives z.

   $$\text{cost}(C_1, \dots C_k) = \min_{z^1 \dots z^k} \sum_{j=1}^{k} \sum_{i \in C} \| x^{(i)} - z^{(j)} \|^2$$

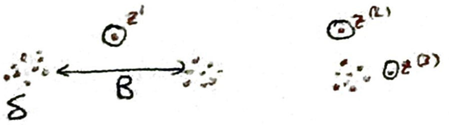
We will look at a specific cluster, let's say j:

specific j: find Z's which will minimize this particular seem: $\sum_{i \in C_j} \| x^{(i)} - \underline{z^{(j)}} \|^2$

$$\frac{\partial}{\partial z^{(j)}} \sum_{i \in C_j} \| x^{(i)} - z^{(j)} \|^2 = 0$$

$$\Longrightarrow z^{(j)} = \frac{\sum x^{(i)}}{|C_j|}$$ (We will find the representative which is in the center of the cluster)

Impact of initialization: let's say we have three tiny clusters, with a radius δ



The cost will be in order of $\Theta(m\delta^2)$

$\delta << B$

If I do bad initialization, the following will happen: the points between z_1 will not move

$\Longrightarrow$ the cost $\Theta(mB^2)$ because we started in the wrong place, your algorithm converges to very suboptimal solution with a high cost.