K-means clustering

Ayaan Alam

June 2024

1 Introduction

Suppose we are given m training examples $x^{(1)}, x^{(2)}, ..., x^{(m)}$ where each

$$x^{(i)} = (x_1^{(i)}, x_2^{(i)}, ..., x_n^{(i)}) \in \mathbb{R}^n$$

i.e. each training example has n features, and each of the features is continuous (not categorical).
Aim: group all training examples into k clusters

2 The Algorithm

1. Randomly initialize k cluster centroids. $\mu_1, \mu_2, ..., \mu_k$
   Note that each $\mu_i \in \mathbb{R}^n$, like the training examples.
2. Repeat (for a large number of iterations, until convergence)
   (a) Assign each of the m training examples to the nearest cluster centroid. Let $c^{(i)}$ = index of the cluster centroid nearest to the i^{th} training example. So $c^{(i)} \in \{1, ..., k\}$.
   (b) Shift (assign) each of the k centroids to the mean of the points which were assigned to it in the previous step, i.e. which have that centroid as their closest centroid.

3 Correctness

Why does the clustering converge? Or why does the assignment of points to the clusters, and position of cluster centroids, become stable after some number of iterations?
The cost function being minimised by the algorithm is

$$J(c^{(1)}, ..., c^{(m)}, \mu_1, ..., \mu_k) = \frac{1}{m} \sum_{i=1}^{m} ||x^{(i)} - \mu_{c^{(i)}}||^2$$

i.e. the mean squared distance of a point to its nearest cluster centroid. Here $||.||$ is the l^2 norm. In each iteration J decreases as:

1. In (a), J decreases as each point is assigned to the centroid nearest to it currently (we change the $c^{(i)}$'s keeping $\mu'_j s$ constant).

2. In (b), J decreases further as each centroids is assigned to the mean* of the points currently assigned to it (we change $\mu'_j s$ keeping the $c^{(i)}$'s constant).

Also, J is bounded below by 0. Thus, after a large number of iterations, J converges**.

*For a given set of points $x^{(1)}, ..., x^{(m)} \in \mathbb{R}^n$, which point $p \in \mathbb{R}^n$ minimises

$$J = \frac{1}{m} \sum_{i=1}^{m} ||x^{(i)} - p||^2 \ ?$$

As $||x^{(i)} - p||^2 = \sum_{j=1}^{n} (x_j^{(i)} - p_j)^2$,

$$J = \frac{1}{m} \sum_{i=1}^{m} \sum_{j=1}^{n} (x_j^{(i)} - p_j)^2 = \frac{1}{m} \sum_{j=1}^{n} \sum_{i=1}^{m} (x_j^{(i)} - p_j)^2$$

Now, $\sum_{i=1}^{m} (x_j^{(i)} - p_j)^2 = mp_j^2 - 2(\sum_{i=1}^{m} x_j^{(i)})p_j + \sum_{i=1}^{m} (x_j^{(i)})^2$ is a quadratic function in p_j. Comparing with the standard equation of parabola $ax^2 + bx + c$, here $a = m > 0$, so it is upward facing and has a minimum at $x = -b/2a$ where $b = -2(\sum_{i=1}^{m} x_j^{(i)})$, i.e. it attains minimum value at $p_j = \frac{\sum_{i=1}^{m} x_j^{(i)}}{m}$.
Thus, J is minimum at

$$p = (\frac{\sum_{i=1}^{m} x_1^{(i)}}{m}, ..., \frac{\sum_{i=1}^{m} x_n^{(i)}}{m}) = \frac{1}{m} (\sum_{i=1}^{m} x_1^{(i)}, ..., \sum_{i=1}^{m} x_n^{(i)})$$

$$= \frac{1}{m} \sum_{i=1}^{m} (x_1^{(i)}, ..., x_n^{(i)}) = \frac{1}{m} \sum_{i=1}^{m} x^{(i)}$$

which is the mean (centroid) of the set of points.

4 Choosing the right k

If $k = m$, then the cluster centroids would converge at the points themselves. Obviously, k should not be $\geq m$. In general, J would decrease with increase in k.

5 **Obtaining the lowest J (best clustering)

Running the algorithm from some initialization of the centroids would give a set of clusters which correspond to a local minimum of J. To obtain the clustering which would give the lowest J, run the algorithm many times with different random initialization of the centroids, compute J for each, and the one with the lowest J would likely be the global minimum.