K-means Clustering and Principal Component Analysis

- K-means algorithm

Input:

$-K(number\ of\ clusters)$;

$-Training\ set\ \{x^{(1)}, x^{(2)}, \ldots, x^{(m)}\}$.

$x^{(i)} \in R^n$(drop $x_0 = 1$ convention).

Randomly initialize K cluster centroids $\mu_1, \mu_2, \ldots, \mu_K \in R^n$

Repeat {

For $i = 1\ to\ m$

$c^{(i)}$:= index(from 1 to K) of cluster centroid closest to $x^{(i)}$

For $k = 1\ to\ K$

μ_k := average(mean) of points assigned to cluster k

}

- K-means optimization objective

$c^{(i)}$ = index of cluster (1, 2, …, K) to which example $x^{(i)}$ is currently assigned

μ_k = cluster centroid k ($\mu_k \in R^n$)

$\mu_{c^{(i)}}$ = cluster centroid of cluster to which example $x^{(i)}$ has been assigned.

Optimization objective:

$$J\left(c^{(1)}, \ldots, c^{(m)}, \mu_1, \ldots, \mu_K\right) = \frac{1}{m}\sum_{i=1}^{m}\left\|x^{(i)} - \mu_{c^{(i)}}\right\|^2$$

$$\underbrace{\min}_{\substack{c^{(1)},\dots,c^{(m)} \\ \mu_1,\dots,\mu_K}} J(c^{(1)}, \dots, c^{(m)}, \mu_1, \dots, \mu_K)$$

- Random initialization

Should have $K < m$;

Randomly pick K training examples;

Set $\mu_1, \dots, \mu_K$ equal to these K examples.