1 Clustering

1.1 K Means

a. Definition

Given a set of observations(x_1, x_2, , x_n), where each observation is a dimensional real vector, k-means clustering aims to partition the n observations into k($\leq$n) sets S=$\{S_1,\ S_2,,\ S_k\}$ so as to minimize the within cluster sum of squares. Formally, the objective is to find:

$$argmin_S \sum_{i=1}^{k} \sum_{x \in S_i} ||x - u_i||^2 = argmin_S \sum_{i=1}^{k} |S_i| Var S_i$$

b. Algorithm

1) Give the initial guess of k means m_1,,m_k
2) Assign each observation to the cluster whose mean has the least squared Euclidean distance.
3) Calculate the new means to be the centroids of the observations in the new clusters.
4) $m_i^{(t+1)} = \frac{1}{S_i^{(t)}} \sum_{s_j in S_i^{(t)}} x_j$

c. Time Complexity

O(nkdi), where n is the number of d dimensional vectors, k is the number of clusters and i is the number of iterations need till convergence.

2 Gaussian Mixture

a. Idea and Definition

1) In K means clustering, one sample point exclusively belongs to one cluster. In other words, we assign a sample point to a cluster with probability 1. In Mixture model, we assign sample point i to a cluster k with the probability r_{ik}, with

$$\sum_k r_{ik} = 1$$

The r_{ik} also follows the fact

$$\sum_i \sum_k r_{ik} = \sum_i 1 = N$$

By changing the order of summation

$$\sum_i \sum_k r_{ik} = \sum_k \sum_i r_{ik}$$

Define the weight of cluster: $w_k = \sum_i r_{ik}/N = \sum_k \omega_k * N = N$
So

$$\sum_k w_k = 1$$

We can also interpret w_k as a prior distribution of a sample point being assigned to cluster k.
2) And for each cluster k, we define the probability of having a sample point i at x_i use a normal distribution $N(x_i|u_k, \Sigma_k)$
Diagram:
3) The $r_{ik}\pi_k$ and $N(x_i|u_k, \Sigma_k)$ are connected with Bayesian rule

$$
\begin{aligned}
P(A|B) &= \frac{P(A)P(B|A)}{P(B)} \\
&= \frac{P(A)P(B|A)}{\sum_c P(C)|P(B|C)}
\end{aligned}
$$

According this rule, we have the following

$$
\begin{aligned}
&P(X_i = x_i \text{ and } X_i \text{ in cluster k}) \\
=&P(X_i \text{ in cluster k})P(X_i = x_i \text{ given } X_i \text{ in cluster k}) \\
=&P(X_i \text{ in cluster k}|X_i = x_i)P(X_i = x_i) \\
\text{So}& \\
&P(X_i \text{ in cluster k}|X_i = x_i) \\
&P(X_i \text{ in cluster k})P(X_i = x_i \text{ given } X_i \text{ in cluster k})/(X_i = xi)
\end{aligned}
$$

Namely,

$$r_{ik} = \frac{\pi_k N(x_i|u_k, \Sigma_k)}{\sum_j \pi_j N(x_i|u_j, \Sigma_j)}$$

4) Our goal is the find u_k, Σ_k, w_k.

b. Cost function and Minimization
For a given point x_i, the likelihood function is

$$p(x_i) = \sum_k \pi_k N(x_i|u_k, \Sigma_k)$$

The likelihood function for the whole sample is

$$\Pi_{i=1}^N p(x_i) = \Pi_{i=1}^N \sum_k \pi_k N(x_i|u_k, \Sigma_k)$$

The goal is to minimize the negative of Log Likelihood

$$L = -\sum_{i=1}^N ln(\sum_k \pi_k N(x_i|u_k, \Sigma_k))$$

1)Take the derivative with respect to u_k

$$dL/du_k = \sum_i \frac{\pi_k N(x_i|u_k, \Sigma_k)}{\sum_j \pi_j N(x_i|u_j, \Sigma_j)} \Sigma^{-1}(x_i - u_k))$$

We found that the term

$$\frac{\pi_k N(x_i|u_k, \Sigma_k)}{\sum_j \pi_j N(x_i|u_j, \Sigma_j)}$$

is exactly r_{ik}
Let the derivative equal to zero, we have

$$u_k = \frac{1}{N_k} \sum_i r_i k x_i (N_k = \sum_i r_{ik})$$

2) Taking the derivative with respect to Σ_k gives

$$\Sigma_k = 1/N_k \sum_i r_{ik}(x_i - u_k)(x_i - u_k)^T$$

3) Taking the derivative with respect to π_k gives

$$\pi_k = \frac{N_k}{N}$$

We see u_k, Σ_k, w$_k$,r$_{ik}$ are mutually dependent, therefore we need to solve this iteratively.