CS 273A Machine Learning Homework 3

Chen Li

October 28, 2015

1 Data Generation

Using the same generation method as in previous homework. The data are generated from a mixture Gaussian Distribution. I generate the 100 data points in 10 clusters to test my algorithm.

2 Kmeans

1. I randomly select points from the data to be the initial value of μ.
2. The stop condition is when μ doesn't change anymore, i.e., the function converged.
3. The error measure I use is $error = \frac{\|\mu_{estimated} - \mu_{true}\|}{\|\mu_{true}\|}$.

3 EM

1. Same as in Kmeans, I randomly select points from the data to be the initial value of μ. Also, I set each $\pi = \frac{1}{number of clusters}$ and $\sum$ equal to the identity matrix.
2. The stop condition is when μ's change is less than $1e^{-6}$. Using this approach can largely reduce the EM's running time and not hurt the performance too much.
3. I use the same error measurement as the Kmeans so as to give a comparison.

4 Comparision

I very the number of cluster centers and the number of data points and measure the error rate and time cost of kmeans and EM algorithm. The results are as follows, in which K represents number of clusters, N represents number of data points. It turns out that the Kmeans algorithm only a little worse than the EM algorithm but it is much more faster than the EM algorithm. I think that's why kmeans so popular.

Function	K	N	$Error_{estimated}$	$Time_{estimated}$
Kmeans	10	100	0.935	0.175
	10	150	1.325	0.284
	15	100	1.050	0.647
	15	150	1.101	0.921
EM	10	100	0.924	30.795
	10	150	1.215	60.966
	15	100	1.129	75.085
	15	150	1.093	108.512

5 Implementation

In *hw3.py*, *Kmeans*() method implement the kmeans algorithm, *EM*() implement the EM algorithm. The *genData*() is used for generate mixture Gaussian data while the *initialze*() function is used to generate the initial data for EM algorithm. The test results can be find in the $test_results$ folder.