Tech Review of EM algorithm

Yuchen Zeng
University of Illinois at Urbana-Champaign
CS410: Text Information Systems
yuchenz8@illinois.edu

Introduction

EM algorithm is short for expectation maximization algorithm. It is an iterative algorithm to try finding a local maximum likelihood. There are two steps in the algorithm, one is expectation which will augment data by predicting values of hidden variables (Zhai, 2021). The other step is called maximization which will exploit the data to improve estimate of parameters (Zhai, 2021). This algorithm is very popular in text clustering. Then we will have to mention another related algorithm called k-means. K-means clustering is a method of vector quantization (Wiki 2021). We will compare these two methods to see how they perform in text clustering mission.

History

EM algorithm was first introduced to the world in a 1977 classic paper. The authors of the paper claimed that the method is inherited from gene-counting method by Cedric Smith. In 1977, the authors generalize the method and publish a convergence analysis on it in order for a wider usage(Wiki 2021). On the other side, k-means was first introduced to the world in an earlier time, 1967. It was first used as a technique for pulse-code modulation(Wiki 2021).

Body

In EM algorithm, we will first initialize probability with random values. Then we will do E-step and M-step iteratively until it converges. After the likelihood converges, we will stop, and the likelihood will be the final result. The formula for E-step and M-step is showed below.

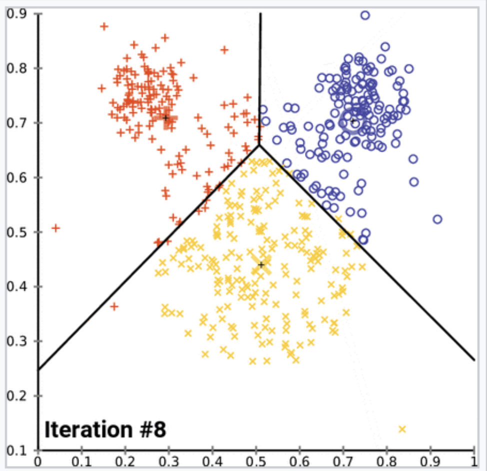
$$p^{(n)}\underset{\uparrow}{(}z=0\,|\,w)=\frac{p(\theta_d)p^{(n)}(w\,|\,\theta_d)}{p(\theta_d)p^{(n)}(w\,|\,\theta_d)+p(\theta_B)p(w\,|\,\theta_B)} \qquad \text{E-step}$$

Fig1 from Zhai 2021

$$p^{(n+1)}(w\,|\,\theta_d)=\frac{c(w,d)p^{(n)}(z=\overset{\downarrow}{0}\,|\,w)}{\sum_{w'\in V}c(w',d)p^{(n)}(z=0\,|\,w')} \qquad \text{M-step}$$

Fig2 from Zhai 2021

In k-means algorithm, it will represent the text objext as a term vector and assume a similarity function defined on two objects (Zhai, 2021). As the initialization, k-means will randomly select points and assign them as the centroids of a cluter. Then the step is really closed to the EM step. It will assign the vector again by calculating the distance between centroid and vector again. After reassigning all the vector, it will re-calculate the centroid of each cluster. This will be also a iterative process until the function converges. From here we will notive that k-means and EM-algorithm are pretty similar in some aspects.



Convergence of *k*-means

Fig3 wiki,2021

Conclusion

We already notice the similarity between EM-algorithm and k-means. They are both iterative method which will stop when the function converges. However, there are still differences between the two methods in doing text clustering. The first one is that k-means will assign data point to cluster on convergence manually (Stackoverflow 2021). While EM will softly assign a point to cluster. The second one is that k-means is highly dependent on the L2 norm optimizing distance calculations. While EM does not based on L2 norm distance but on the expectation step. In conclusion, EM algorithm and k-means method are both useful and popular methods in text clustering.

Reference

Probabilistic Topic Models: Expectation-Maximization Algorithm, lecture notes, University of Illinois at Urbana-Champaign (ChengXiang Zhai, 2021, p.9)

Probabilistic Topic Models: Expectation-Maximization Algorithm, lecture notes, University of Illinois at Urbana-Champaign (ChengXiang Zhai, 2021, p.8)

Text clustering: Similarity-based approaches, lecture notes, University of Illinois at Urbana-Champaign (ChengXiang Zhai, 2021, p.10)

En.wikipedia.org.2021. k-means clustering-Wikipedia. Available at:https://en.wikipedia.org/wiki/K-means_clustering [Accessed 7 Nov 2021].

En.wikipedia.org.2021. Expectation-maximization algorithm-Wikipedia. Available at: https://en.wikipedia.org/wiki/Expectation%E2%80%93maximization_algorithm#History [Accessed 7 Nov 2021].

Stackoverflow.2021 https://stats.stackexchange.com/questions/76866/clustering-with-k-means-and-em-how-are-they-related