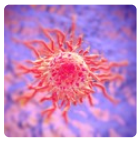


GeeshKopp

Comparing K-Means Clustering vs GMM

Using Gaussian Mixture Models to Explore Differences in Clustering

I decided to approach this problem from a more unsupervised learning method.

When considering K-means clustering often one of the pitfalls can be the shape of the clusters. When considering the number of dimensions that the data has it seemed intuitive that spherical clusters would be the least likely. These are the following steps in my approach

- Researched Guassian Mixture models
- Researched comparable metrics
- Combined several approaches to the data cleaning, modeling and metric scoring It must be noted that only the combination of these analytical methods are my own. I must give credit to:
  - Kajot for the data and first part of the script for the data processing and cleaning
  - Kam Sen and Prabhath Nanisetty from their Q&A on stats.stackexchange.com http://stats.stackexchange.com /questions/90769/using-bic-to-estimate-the-number-of-k-in-kmeans (http://stats.stackexchange.com/questions /90769/using-bic-to-estimate-the-number-of-k-in-kmeans)
  - Sklearns gaussian mixture model example

Use Bayesian Information Criterion to Compare K-means and Gaussian Mixture Model

Guassian Mixture Models and K-means use different metrics for comparing the best clusters. I used BIC for both forms of clustering so that I could compare the approaches

Results

- The optimal number of clusters using BIC score and GMM is 5 and this has a "full" geometry parameter
- The optimal number of clusters using BIC score and K-means was 4.
- The GMM provided a lower BIC score than K-mean.

Next Steps

Calculate silouhette score and inertia for GMM and compare to K-means Play around with different imputation methods to see if that makes a difference

Thoughts

I can't access the paper behind the paywall so I cannot use their methods or compare their methods to mine on how to find the