Gaussian Mixture Model Report-Cinnamon AI Bootcamp 2020

Student Name:Nguyen Trong Tung

February 3^{rd} 2019

1 Overview about Gaussian Mixture Model

Gaussian Mixture Model is a similar approach as K-Mean Clustering Algorithm to solving unsupervised machine learning problem in which data points are separated into a specific cluster. The only difference is that Gaussian Mixture Model assigns "soft" label in which data points are assigned to cluster with an appropriate probability in stead of "hard" label. Given a set of data points and number of clusters, GMM will find the best parameters for all clusters specified before(mean, co-variance and coefficient) which maximize the likelihood of occurrence of training data, then use these optimized parameters to predict whether an incoming data point has the same property as training samples. Therefore, GMM model is good to detect anomaly data point if the point's score evaluated by such model is below some threshold. Each cluster k of the model is constrained by three parameters:

+Coefficient: determines the confidence score of cluster k

+Mean: determines weighted average of all data points with corresponding posterior probability

+Co-variance: determines weighted average of variance among data points

It has been stated in Pattern Recognition Book(written by Bishop) that due to the singularity property, the maximum log likelihood function doesn't have closed form solution. The author also provide an alternative solution by using gradient-based techniques known as Expectation Maximization algorithm. The progress of this algorithm is described below: Formally, given each observation x, we first can calculate its likelihood of specific cluster k which is gaussian distribution: gauss(x|μ_k, σ_k) represented by (μ_k, σ_k) .

Then in E step we calculate probability of belonging to cluster k given x by:

$b_k = p(k|x) = \frac{p(x|k).p(k)}{p(x)} = \frac{p(x|k).p(k)}{\sum_{k=1}^{K} p(x|k).p(k)} = \frac{gauss(x|\mu_k,\sigma_k).p(k)}{\sum_{k=1}^{K} gauss(x|\mu_k,\sigma_k).p(k)}$

$M step will update parameters of each cluster by using appropriate coefficient above.$

The progress is iteratively calculated until it reaches convergence.

2 Implementation of GMM

2.1 Experimenting with dataset

The dataset given in this problem is Cardiotocography classification dataset consists of measurements of fetal heart rate (FHR) and uterine contraction (UC) features on cardiotocograms classified by expert obstetricians. The outliers is downsampled to 176 points and the remaining is inliers. This ratio will be used to split the dataset into training and testing where training(normal) will be fitted into the Gaussian Mixture Model and testing will be used to predict anomaly property. Each sample contain 22 values where 21 first values representing for different features and final one is anomaly class type.

2.2 Fitting Multivariate Gaussian Mixture Model

Due to the presence of many features in the dataset, Multivariate GMM is utilized instead of univariate GMM. The Gaussian Model is implemented following progress described in previous section except the initializing steps have to be handled carefully. By applying random initialization method, the co-variance matrix can be failed to be invertible to calculate the determinant needed in Multivariate Gaussian Distribution. A simple step to resolve this problem is multiplying the transpose of random matrix A with itself provide a positive semi-definite matrix which is always invertible. Moreover, the total sum of confident score of all cluster must equal to one.

2.3 Anomaly prediction

In this phase, testing set is used to predict the anomaly property. An approach to calculating the threshold is setting the lowest example's probability score as threshold. This formula, therefore, can consider those observation whose score is lower than threshold as anomaly data points.

3 References

1/Pattern Recognition Book-Christopher M.Bishop.
2/http://odds.cs.stonybrook.edu/cardiotocogrpahy-dataset/