Anomaly detection using Gaussian Mixture Model

Do Viet Tung

December 2019

1 Introduction

Hi, readers. At this stage, you probably know about K-Means as a simple and intuitive tool to cluster the data. However, K-Means has a strong assumption that the clusters have circular forms. Thus, we need another clustering algorithm for different shapes. As suggested by Central Limit Theorem, as we collect more and more samples from a dataset, they tend to resemble a Gaussian distribution. Gaussian Mixture Model (GMM) is quite an appropriate candidate.

In this assignment, you will have to implement a simple GMM from scratch (using numpy) to detect abnormal data.

2 Gaussian Mixture Model

Similar to K-Means, in GMM, we have to first choose a number of clusters K beforehand. Unlike K-Means, GMM allow a data point to belong to all clusters with the corresponding probabilities.

In GMM, each cluster is an unique Gaussian distribution parameterized by: mean μ, variance σ^2. Note that for the cases of multi-dimensional data, we use covariance instead.

3 Expectation maximization

To learn such parameters, GMMs use the expectation-maximization (EM) algorithm to optimize the maximum likelihood. EM can be simplified in 2 phases: The E (expectation) and M (maximization) steps.

In the E step, we calculate the likelihood of each observation x using the estimated parameters of the k^{th}:

$$f(x|\mu_k, \sigma_k^2) = \frac{1}{\sqrt{2\pi\sigma_k^2}} exp(-\frac{(x-\mu_k)^2}{2\sigma_k^2}) \tag{1}$$

for each $k = 1, 2, 3, ..., K$

Then, we can calculate the likelihood of a given example x to belong to the k^{th} cluster:

$$b_k = \frac{f(x|\mu_k, \sigma_k^2)\phi_k}{\sum_{k=1}^{K} f(x|\mu_k, \sigma_k^2)\phi_k} \tag{2}$$

where ϕ_k models out prior beliefs that an example was drawn from the k^th Gaussian distribution (or cluster).

In M step, we update the parameters as follow:

$$\mu_k = \frac{\sum b_k x}{\sum b_k} \tag{3}$$

$$\sigma_k^2 = \frac{\sum b_k (x - \mu_k)^2}{\sum b_k} \tag{4}$$

$$\phi_k = \frac{1}{N} \sum b_k \tag{5}$$

As a full algorithm, we alternatively run E and M step until converge.

Please prove equation (2) using Bayes' theorem. Hint: which one is the prior probability, which one is the posterior probability

Note that this is just the case of 1-D data, can you provide the formulae for the case of n-D data. Hint: Multivariate Gaussian distribution

4 Anomaly detection

Given a dataset of normal and abnormal samples, you could fit a GMM on the normal dataset. Then for each abnormal sample, calculate its likelihood (this will be easy if you've already proved the equation (2)). If this number is below a threshold calculated from the normal samples, report that sample as abnormal (or outlier).

Which K do you think is suitable for this problem? Can you come up with a threshold formular based on the normal samples? Justify your answer.

5 Implementation

In this section you will have to implement a GMM from scratch to detect abnormal samples from a dataset.

5.1 Data

Please read and download the data from this link: `http://odds.cs.stonybrook.edu/cardiotocogrpahy-dataset/`

5.2 Requirements

You can use tools to model the Gaussian distribution but not the GMM itself (like the case of sklearn).

5.3 Tasks

- Visualize the dataset to see the separation
- Implement a Gaussian Mixture Model
- Train test split (test set ratio could be 0.1 amount of data). Remember that you only fit GMM on the normal samples so split them wisely.
- Fit the model
- Use the fitted model to detect whether a test sample is anomalous or not