School of Engineering and Applied Science (SEAS), Ahmedabad University

B.Tech (CSE Semester VI)/M.Tech/PhD:
Machine Learning (CSE 523)

Project Submission #4: Gaussian Mixture Models

Submission Deadline: April 12, 2020 (11:59 PM)

- **Group No.: 21**
- **Project Area: Environment and Climate Change**
- **Project Title: Air Quality Prediction Using ML**
- **Name of the group members :**
  1. Vishal Saha (1741004)
  2. Rushil Shah (1741009)
  3. Muskan Matwani (1741027)
  4. Mohit Vaswani (1741039)

- Using the Gaussian Mixture Model we are trying to visually interpret the makeup of our data-set and divide our data-set into two category so that we can train our classifier to correctly classify unseen data and predict future quality of air.As GMM have soft boundaries, where data points can belong to multiple cluster at the same time but with different mixture weight corresponding to each cluster. e.g. a data point can have a 60% of belonging(mixture weight) to cluster 1, 40% of belonging(mixture weight) to cluster 2.As we can compute the likelihood of each point being in each cluster, the points with a likelihood less than the threshold (threshold value is determined by us) can be labeled as outliers. So, we can also remove outliers from the data-set before training our classifier. Also provide the comments on updating mean, variance and mixture weights.In this work we are categorizing data-set into two clusters.
- **Implementation code Link:** For GMM code, click here

  For dataset, click here

- **Inference:** This section should include three paragraph's.

We are using GMM to categorize our data-set into two categories

1. Harmful specimen of air.
2. Favourable specimen of air.

– **What have we done and why is it important**

Our actual data-set contains 4-dimensional data and each dimension corresponds to a pollutant concentration in the specimen taken into consideration.The pollutants are :so2,no2,rspm,spm .Data being 4-dimensional it is quite difficult to interpret the construction of our data-set so using GMM we are providing a visual interpretation of the data-set.Through this implementation of GMM we can use the mean and variance calculated corresponding to the categories to train different classifier without manually labelling the data-set. GMM also allows us to filter out the outliers before we train the classifier.

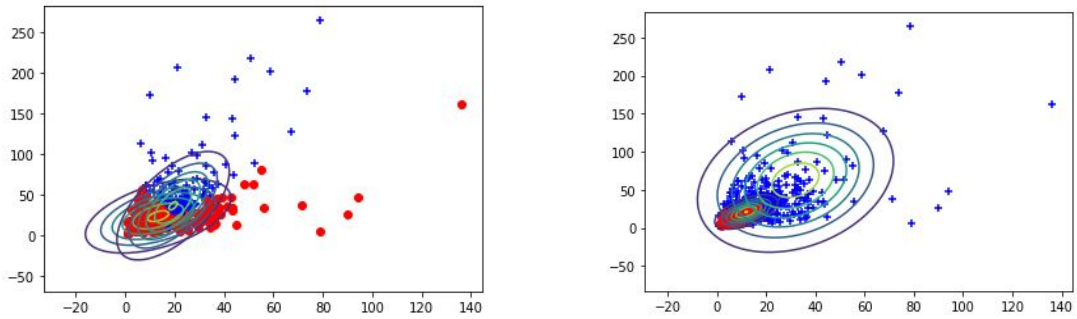
– **Implementation**

To simplify the visualization we are converting our data-set into two dimensional data.Example: (so2 , no2), (no2,spm) instead of the original form of data (so2,no2,rspm,spm) so that we can visualize in 2-dimensional space.Therefore ,there are C_2^4=6 possible plots out of which we have produced 3: (so2, no2),(no2,spm),(so2, spm).Corresponding to each of these plots we have applied GMM on this converted form of data-set and each of these iteration of GMM gives us a mean and variance metric corresponding to the categories to implement a classifier.We are clustering our data-set into two clusters.

* Initially we divide our data-set into two equal parts considering each part as a cluster ,then we calculate its mean and variance i.e. mean and variance of the cluster.
* Then we perform ER algorithm
* In every iteration , corresponding to each data point we calculate its likelihood of belonging to any of the two clusters.
* Then we assign points to the cluster which it is most likely to belong.
* After all the points have been assigned to its new cluster,we calculate the new mean and new variance corresponding to each cluster.Thus, we update mean and variance of the cluster.
* At the end of each iteration we calculate likelihood function of the GMM.
* We stop iteraing the ER algorithm when the absolute difference between the newly calculated likelihood function value and previously calculated likelihood function value is less than the threshold value .
* We have set the threshold value epsilon=1e-6.

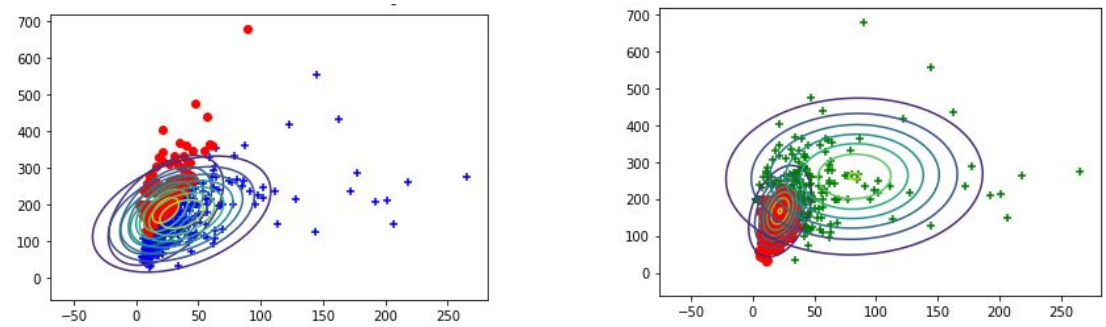
– **Results**

* SO_2 **vs** NO_2 **clusters (before and after GMM)**



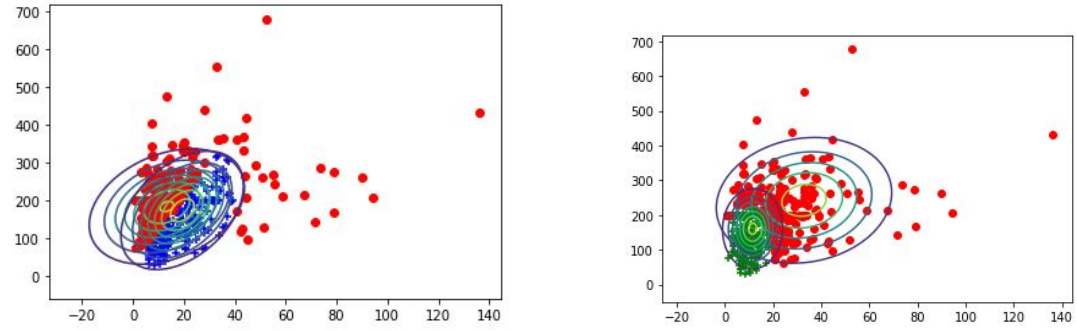
The left graph shows the clusters plotted on the scatter graph SO_2 and NO_2 before applying the GMM model and the right graph are the new reformed clusters after applying the GMM model. Since, our AQI_Binary_Range takes two binary values - 0 or 1, hence these two clusters corresponds to two different groups of AQI values.

* NO_2 **vs** SPM **clusters (before and after GMM)**



The left graph shows the clusters plotted on the scatter graph NO_2 and SPM before applying the GMM model and the right graph are the new reformed clusters after applying the GMM model. Since, our AQI_Binary_Range takes two binary values - 0 or 1, hence these two clusters corresponds to two different groups of AQI values.

* SO_2 **vs** SPM **clusters (before and after GMM)**



The left graph shows the clusters plotted on the scatter graph SO_2 and SPM before applying the GMM model and the right graph are the new reformed clusters after applying the GMM model. Since, our AQI_Binary_Range takes two binary values - 0 or 1, hence these two clusters corresponds to two different groups of AQI values.