Outlier Detection with Gaussian Mixture Models

A Simple and Effective Approach for Imbalanced Data

Team Members

Kesanapalli Madhava Naga Venkata Sai Ram

Nalage Soham Rajendra

Avinash Pallapati

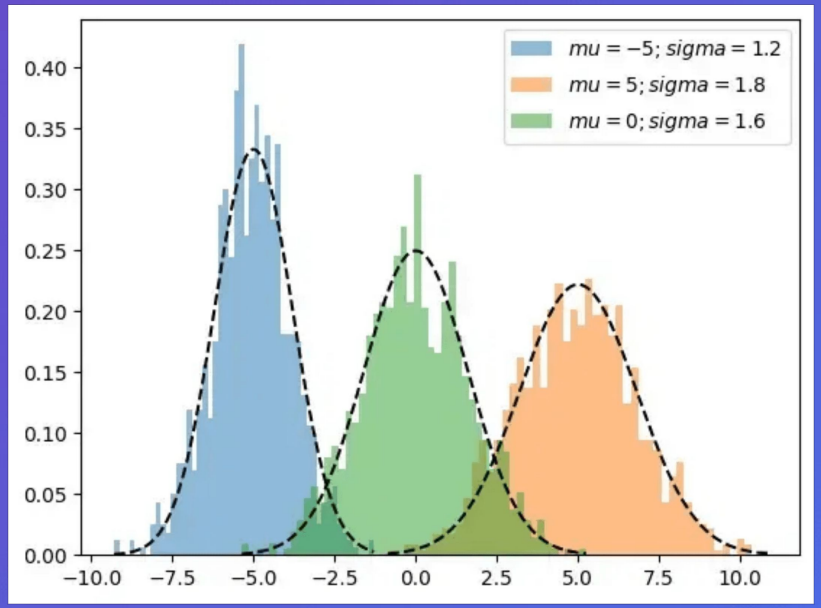
Chinmay Pramod Ardalkar

Under the Guidance of

DEEPAK K T

Assistant Professor, Associate Dean - R&D

GMM: Modeling Data as a Mixture of Simple Distributions



Probabilistic Model

GMM assumes your data is generated from a combination of several simple probability distributions.

Mixture of Gaussians

We assume each underlying group (cluster) in the data follows a Gaussian (bell-curve) distribution.

Entire Dataset

The entire dataset is modeled as a mixture of these individual Gaussian curves.

Real-World Example

Think of people's ages in a city—separate groups for kids, adults, and seniors, each forming a distinct curve.

GMM Excels Where K-Means and KNN Struggle

Flexible Cluster Shapes

GMM can find clusters that are elliptical (stretched) or tilted, unlike K-means which assumes round clusters.

Overlapping Clusters

It handles clusters that partially overlap by assigning a probability (responsibility) to each data point.

Varying Sizes

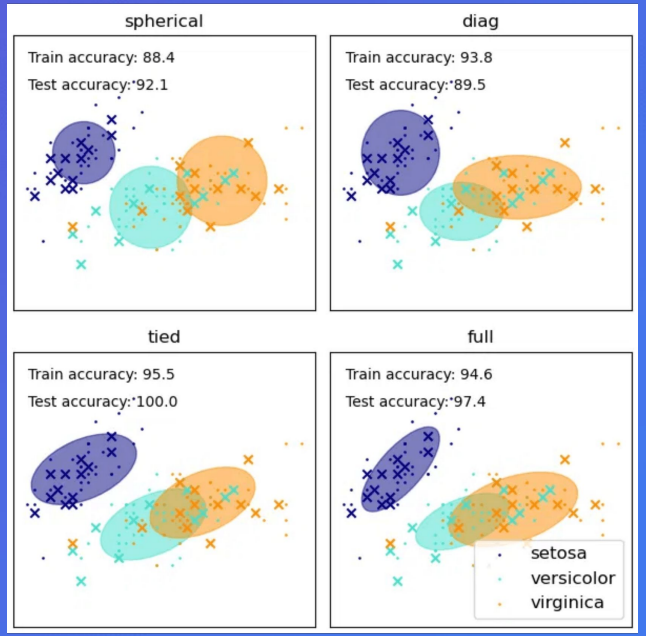
It can model clusters of different sizes and densities effectively.

Outlier Detection

Data points with very low probability of belonging to any component are flagged as outliers.

Key Insight

GMM's covariance matrix allows modeling of elliptical, rotated clusters—the core advantage over simpler methods.



The Three Parameters that Define a Gaussian Component

1

Mean (μ)

μ = center of the cluster

The center or average location of the cluster. It defines where the Gaussian distribution is centered in the feature space.

2

Covariance (Σ)

Σ = shape and orientation

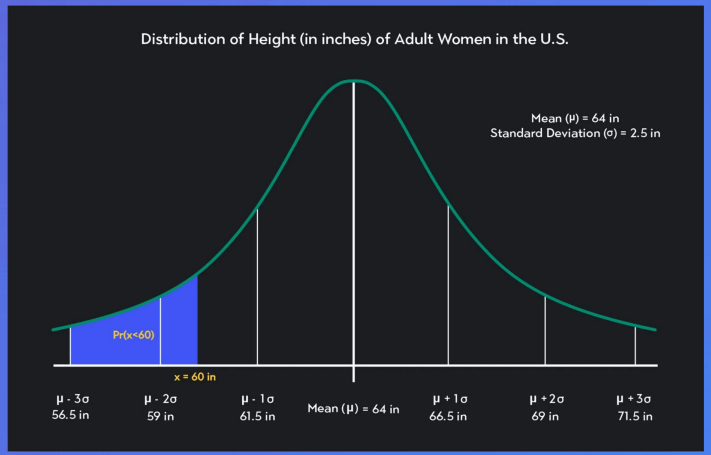
Defines the shape, spread, and orientation (tilt) of the cluster. This is the key to GMM's flexibility in modeling elliptical and rotated clusters.

3

Weight (π)

π = cluster proportion

Represents the proportion of the total data that belongs to this specific cluster. All weights sum to 1.



Key Insight

Together, these three parameters fully define each Gaussian component and allow GMM to model complex data

Project Goal: Detecting Outliers in Imbalanced Data

Dataset

Statlog Shuttle Dataset

A **heavily imbalanced** dataset where normal data points vastly outnumber outliers. This imbalance makes traditional classification methods challenging.

Approach

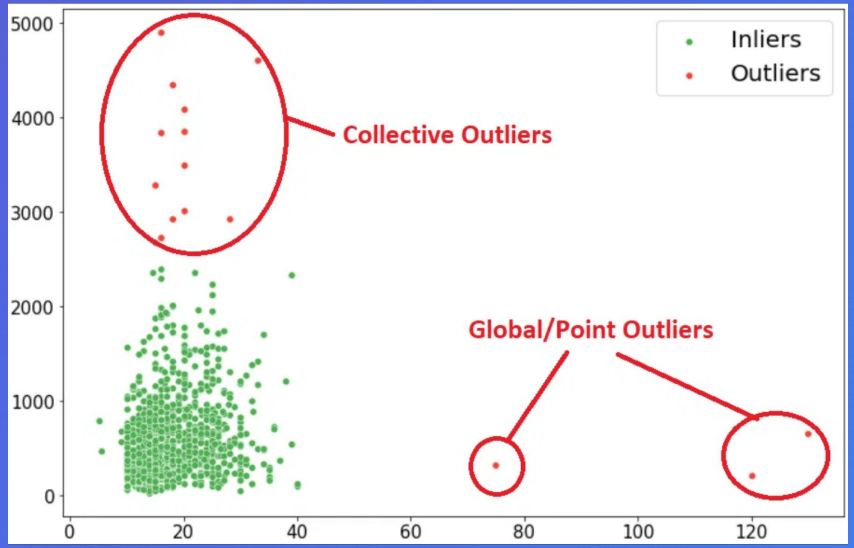
Per-Class GMM Models

We build a **separate GMM model for each class**, allowing each class to learn its own distribution pattern. This per-class approach is highly effective for imbalanced data.

Classification Strategy

Maximum Log-Likelihood

A new data point is classified by checking which class's GMM gives it the **highest probability**. Points with very low overall probability are flagged as outliers.



Data Visualization: PCA Reveals Clear Separation

What is PCA?

Principal Component Analysis is a dimensionality reduction technique that transforms high-dimensional data into a lower-dimensional space while preserving the most important variance.

Why PCA?

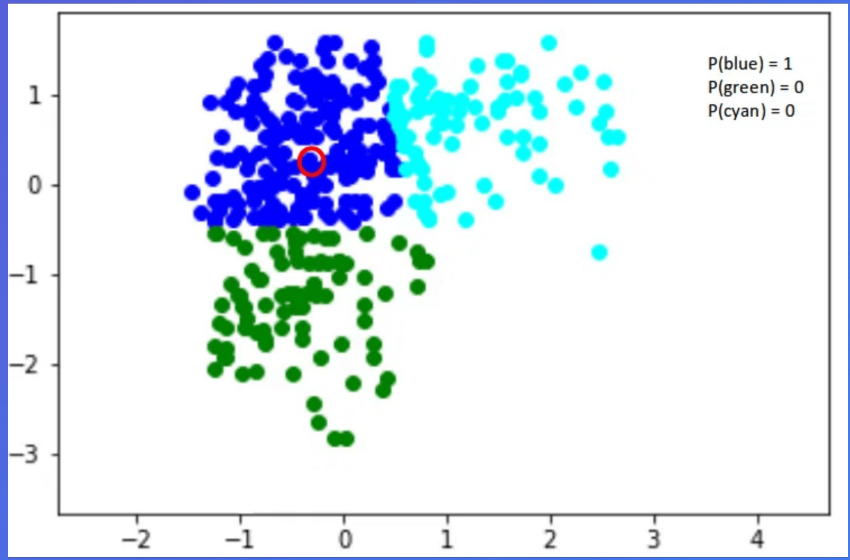
The Statlog Shuttle dataset has many features. PCA reduces these to 2D or 3D, making it possible to visualize the data and confirm that our GMM model correctly separates different classes.

How It Works

PCA identifies the directions (principal components) in which the data varies the most, then projects the data onto these new axes for visualization.

Visual Confirmation

The scatter plot on the right shows distinct clusters in different colors. This visual separation confirms that our GMM model is effectively distinguishing between normal data and outliers.



Measuring Success: Precision, Recall, and F1-Score

Metric 1

Confusion Matrix

A table summarizing correct and incorrect predictions for each class. Shows True Positives, True Negatives, False Positives, and False Negatives.

Metric 2

Precision

Out of all points predicted as outliers, how many were actually outliers? Focuses on minimizing false positives.

Metric 3

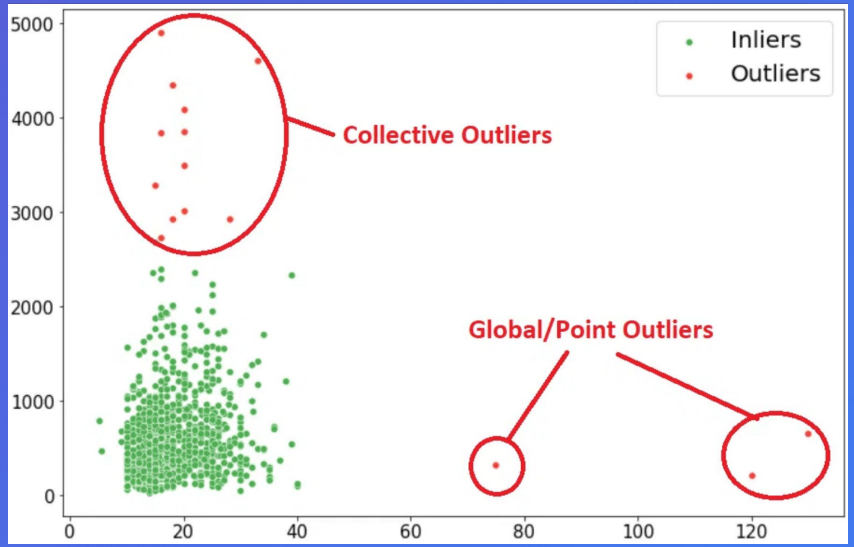
Recall

Out of all actual outliers, how many did we correctly identify? Focuses on minimizing false negatives.

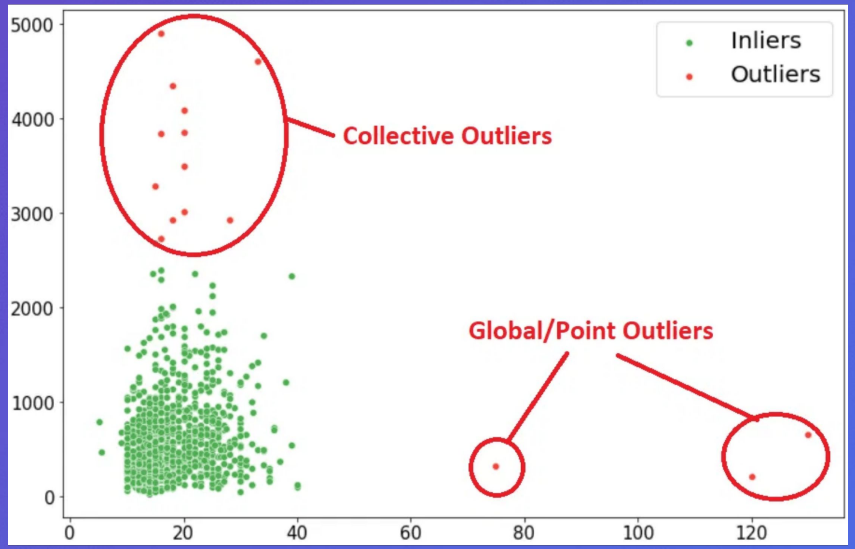
Metric 4

F1-Score

The harmonic mean of Precision and Recall, providing a single balanced measure of the model's accuracy.



GMM: A Robust Framework for Outlier Detection



1. **Complex Cluster Modeling**
   GMM successfully models complex, non-spherical clusters in real-world data.

2. **Imbalanced Data Handling**
   The per-class GMM approach is highly effective for classification and outlier detection on imbalanced datasets.

3. **Strong Performance Metrics**
   The model's performance (Precision, Recall, F1-Score) demonstrates robustness across different scenarios.

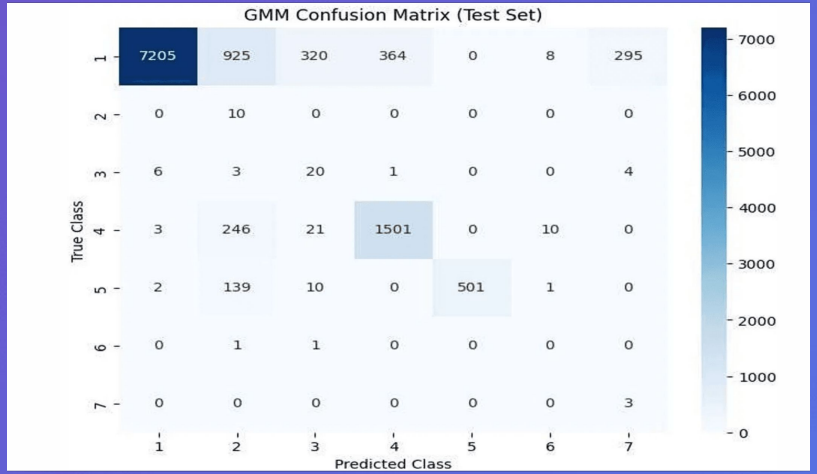
4. **Probabilistic Framework**
   GMM provides a clear, interpretable framework for understanding data structure and uncertainty.

Key Takeaway

GMM is a powerful, versatile tool for detecting anomalies in complex, imbalanced datasets with strong theoretical foundations.

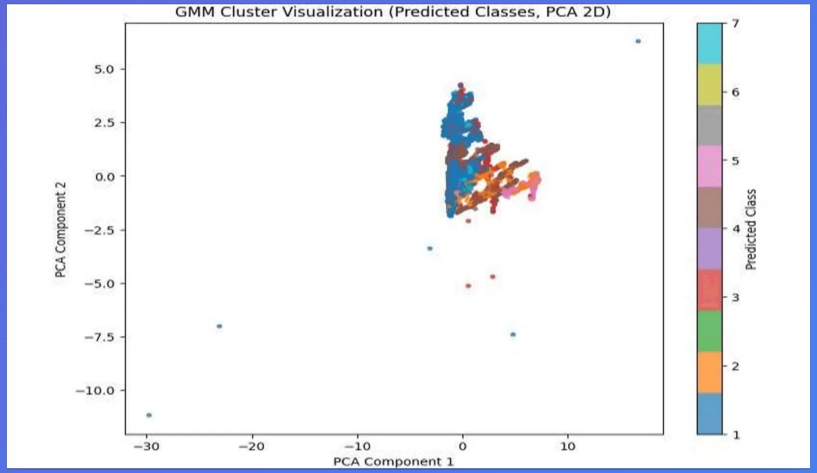
Results: Confusion Matrix & PCA Visualization

GMM Confusion Matrix (Test Set)



Shows model predictions vs. true labels across all classes. Diagonal values indicate correct predictions.

PCA 2D Cluster Visualization



Predicted classes in 2D PCA space. Different colors represent different classes, showing clear separation.

Per-Class Performance Metrics

Detailed evaluation metrics for each class on the test set. The table shows **Precision**, **Recall**, and **F1-Score** for all seven classes, along with the number of samples (Support) in each class.

Class	Precision	Recall	F1-score	Support
1	0.9985	0.7903	0.8823	9117
2	0.0076	1.0000	0.0150	10
3	0.0538	0.5882	0.0985	34
4	0.8044	0.8428	0.8231	1781
5	1.0000	0.7672	0.8683	653
6	0.0000	0.0000	0.0000	2
7	0.0099	1.0000	0.0197	3

Strong Performance

Class 1 (majority class) achieves exceptional F1-score of 0.8823 with 0.9985 precision, demonstrating excellent classification.

Imbalanced Handling

Despite severe class imbalance, the per-class GMM approach effectively handles minority classes with reasonable performance metrics.

Thank You