# ML Lab 12

## Why is PCA used before applying the Gaussian model? Explain the role of PCA in this context. What issues might arise if PCA is not used before estimating the multivariate Gaussian distribution?

**Role of PCA**

PCA (Principal Component Analysis) reduces the dimensionality of the dataset while retaining most of the variance.

* Removes noise and multicollinearity - Many features in high-dimensional data may be correlated, which can lead to unstable covariance matrix estimates in a multivariate Gaussian model.
* Prevents overfitting - High-dimensional data (28+ features in this dataset) can lead to overfitting when estimating the covariance matrix due to the curse of dimensionality.
* Speeds up computation - PCA reduces the number of features, making PDF calculations and matrix inversions faster and more stable.

**Issues without PCA**

* The covariance matrix may become singular or ill-conditioned, making it hard to compute its inverse a requirement in Gaussian density estimation.
* Overfitting can occur due to high variance in noisy or redundant features.
* Numerical instability during computation of multivariate normal probabilities.

## What does the threshold (epsilon) represent in anomaly detection? Discuss how selecting different epsilon values affects precision, recall, and the number of detected anomalies.

Epsilon (𝜖) is the decision boundary on the probability score from the Gaussian model. It is the threshold below which a data point is classified as an anomaly.

**Effect of Different Epsilon Values**

|  |  |  |
| --- | --- | --- |
| **Epsilon Value** | **Prediction Behavior** | **Effect on Metrics** |
| Very Low | Few anomalies detected | High precision, low recall |
| Very High | Many anomalies detected | High recall, low precision |
| Optimal (Chosen) | Balanced anomalies | Maximizes F1 score (harmonic mean) |

In the lab 12 notebook,

* Best epsilon - 6.34e-08
* F1 Score - 0.0240, which is low, likely due to extreme class imbalance or a non-ideal fit of the Gaussian model to real fraud behavior.

## Why is a separate validation set used to select the threshold instead of using the full dataset?

**Why use validation?**

* To simulate unseen data, ensuring the model generalizes well.
* Prevents overfitting the training data when choosing epsilon.
* Mimics real-world deployment where the threshold is applied to new, unseen data.

**Train-validation split benefits**

* Separates parameter estimation (training set) from threshold tuning (validation set).
* Allows objective evaluation of performance metrics like F1 score.
* Using the full dataset would risk data leakage, where the threshold is optimized for the same data used to fit the model, leading to inflated performance estimates.

## What assumptions does the Gaussian model make about the data? Explain whether these assumptions are realistic in the context of fraud detection. When might this model fail?

**Key Assumptions**

* Multivariate normality - The data (after PCA) follows a multivariate Gaussian distribution.
* Independent and identically distributed (i.i.d) - Assumes each data point is drawn independently from the same distribution.
* Linear relationships - Implicit due to covariance-based modeling.

**In the context of fraud detection**

These assumptions may not hold well

* Fraudulent transactions are rare and non-Gaussian, often following non-linear, complex patterns.
* Normal transactions may not be perfectly Gaussian either, even after PCA.

**When might the model fail**

* Non-Gaussian feature distributions, even post-PCA.
* Fraudsters evolve techniques over time, introducing concept drift.
* High class imbalance can severely bias probability estimates.