Local Outlier Factor Tutorial

September 21, 2025

1 Introduction

Local Outlier Factor (LOF) detects anomalies by comparing the local density of a point to that of its neighbors. Points with significantly lower density than their neighbors receive large LOF scores, signalling potential outliers. LOF excels at identifying anomalies in data with varying densities where global distance-based thresholds may fail.

2 Theory and Formulas

2.1 k-Distance and Reachability

For a point $\mathbf{x}$ and neighborhood size k, the k-distance $d_k(\mathbf{x})$ is the distance to the k-th nearest neighbor. The reachability distance from $\mathbf{x}$ to neighbor $\mathbf{y}$ is defined as

$$\mathrm{reach}_k(\mathbf{x}, \mathbf{y}) = \max\{d_k(\mathbf{y}), \|\mathbf{x} - \mathbf{y}\|_2\}. \tag{1}$$

2.2 Local Reachability Density and LOF

The local reachability density (LRD) of $\mathbf{x}$ is the inverse of the average reachability distance to its neighborhood $N_k(\mathbf{x})$:

$$\mathrm{lrd}_k(\mathbf{x}) = \left(\frac{1}{|N_k(\mathbf{x})|} \sum_{\mathbf{y} \in N_k(\mathbf{x})} \mathrm{reach}_k(\mathbf{x}, \mathbf{y}) \right)^{-1}. \tag{2}$$

The LOF score compares lrd_k at $\mathbf{x}$ to the densities of its neighbors:

$$\mathrm{LOF}_k(\mathbf{x}) = \frac{1}{|N_k(\mathbf{x})|} \sum_{\mathbf{y} \in N_k(\mathbf{x})} \frac{\mathrm{lrd}_k(\mathbf{y})}{\mathrm{lrd}_k(\mathbf{x})}. \tag{3}$$

Values near 1 indicate density comparable to neighbors, while scores significantly above 1 suggest anomalies.

2.3 Hyperparameters and Practical Notes

Choosing k controls the locality of the detector; small k may overreact to noise, whereas large k blurs local structure. Distance metrics, feature scaling, and handling of categorical variables also influence performance. LOF scores are relative, so a decision threshold must be selected based on application-specific tolerances or contamination estimates.

3 Applications and Tips

- **Network intrusion detection**: surface unusual traffic patterns in high-dimensional log data.
- **Industrial monitoring**: detect sensor drifts or faults that reduce local density in feature space.
- **Fraud analytics**: identify customer behavior that deviates sharply from peer groups.
- **Best practices**: standardize features, experiment with multiple k values, inspect LOF score distributions, and combine with domain knowledge to validate flagged points.

4 Python Practice

The script `gen_lof_figures.py` synthesizes a dataset with dense clusters and scattered anomalies, fits scikit-learn's `LocalOutlierFactor`, and exports a score heatmap and histogram illustrating the separation between inliers and outliers.

Listing 1: Excerpt from gen$_lof_figures.py$

```python
from sklearn.neighbors import LocalOutlierFactor

lof = LocalOutlierFactor(n_neighbors=35, contamination=0.08)
pred = lof.fit_predict(points)
scores = -lof.negative_outlier_factor_
```

5 Result

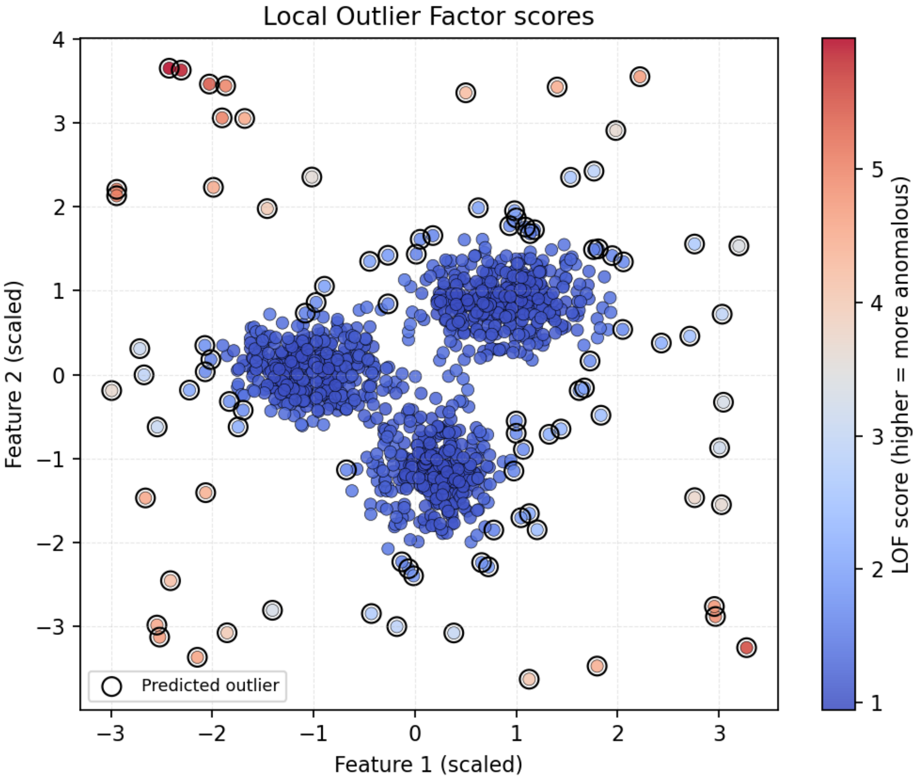

Figure 1: LOF scores across the feature space; higher scores highlight sparser regions

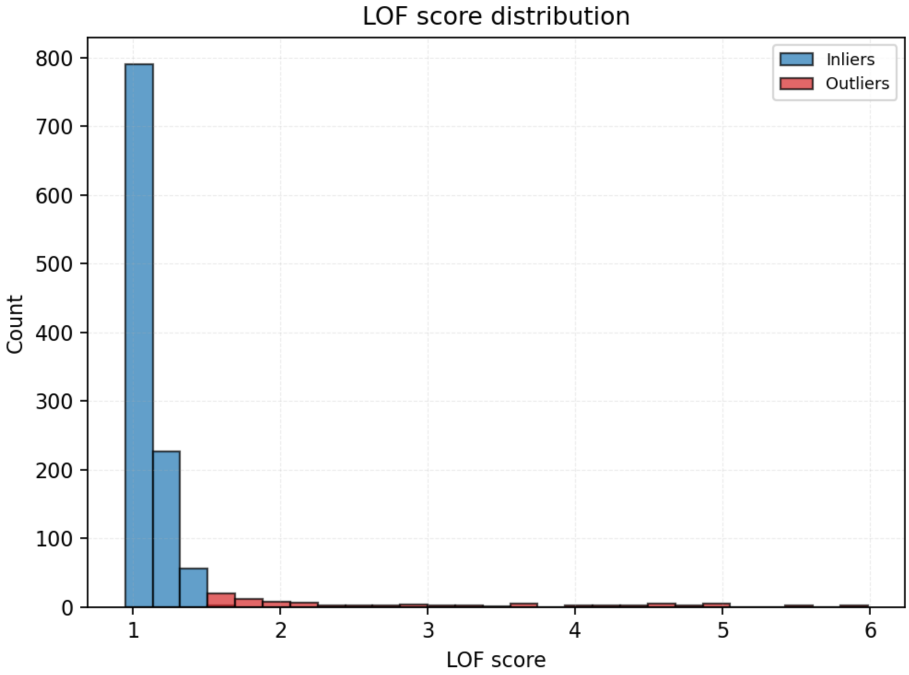

Figure 2: Histogram of LOF scores comparing predicted inliers and outliers

6 Summary

LOF evaluates relative local density to flag anomalies, making it effective on data with heterogeneous cluster structure. Proper scaling, neighborhood tuning, and score visualization enable robust deployments in security, industrial, and financial monitoring. The synthetic example demonstrates how LOF separates sparse points from dense clusters and how score distributions guide threshold selection.