Clustering: Anomaly Detection with Local Outlier Factor

DS 5230-Summer 2020

Joshua Galloway

Overview *(Slide 1)*

Anomaly/Outlier Detection [3]

Problem: The main objective of a clustering algorithm is to find clusters, and not to optimize outlier detection due to being more globally perceptive.

Local Outlier Factor (LOF) [1]

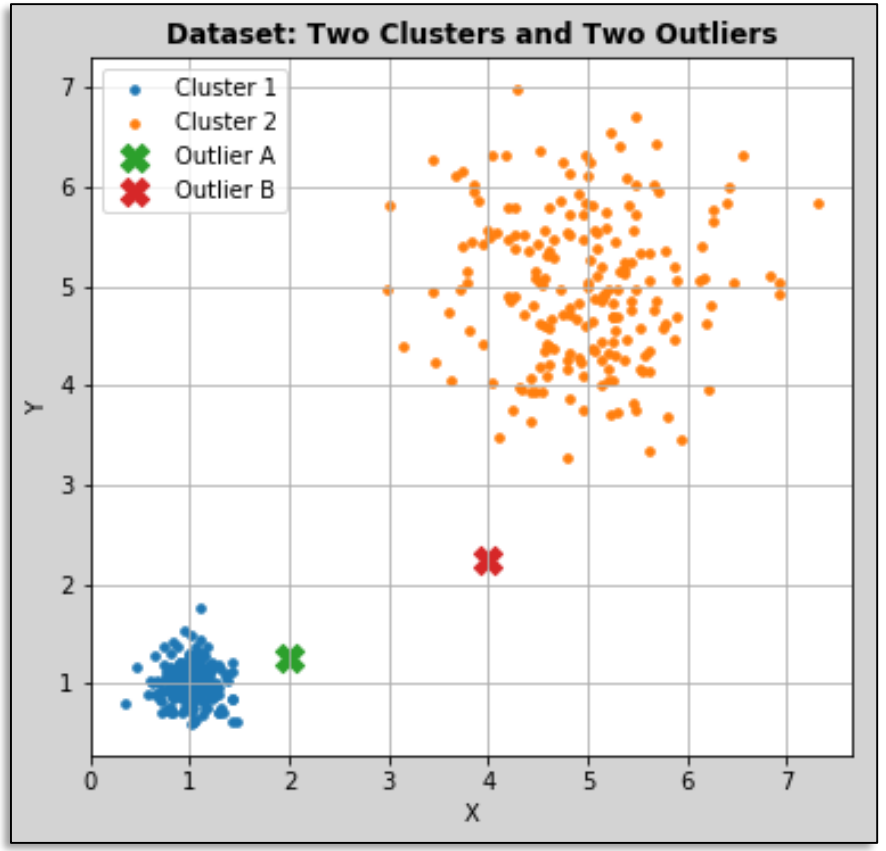
- LOF is a density-based algorithm, like DBSCAN or OPTICS, which is able to adjust for local variations in dataset density.
- A distance-based system would have trouble finding 'Outlier A' in the graphic due to relative proximity issues.
- Similarly, Centroid and Distribution Methods are likely to include the outliers as part of a nearby cluster.

Outlier (Hawkins-Outlier) [4]

- An outlier is an observation that deviates so much from other observations as to arouse suspicion that it was generated by a different mechanism.

Applications [2]

- Data Cleaning
- Intrusion Detection Systems
- Fraud Detection Systems
- Novelty Detection



Model Development *(Slide 2)*

Basic Idea: Compare local density of a point with the densities of its k-nearest neighbors. Points with a lower density are outliers [2]

For a given data point, $\overline{X}$, let $D_k(\overline{X})$ be its distance to the k-nearest neighbor. Let $L_k(\overline{X})$ be the set of points with k-nearest neighbor distance of $\overline{X}$.

Then, **Reachability Distance**, $R_k(\overline{X},\overline{Y})$, of object $\overline{X}$ with respect to $\overline{Y}$ is defined as the greater of the distance between $\overline{X}$ and $\overline{Y}$, $dist(\overline{X},\overline{Y})$, and the k-nearest neighbor distance of $\overline{Y}$. So,

$$R_k(\overline{X},\overline{Y}) = max\left[dist(\overline{X},\overline{Y}), D_k(\overline{Y})\right]$$

Now we define **Average Reachability Distance**, $AR_k(\overline{X})$, of datapoint $\overline{X}$ with respect to its neighborhood, $L_k(\overline{X})$, to be the average of its reachablility distances to all objects in its neighborhood. Or,

$$AR_k(\overline{X}) = MEAN_{\overline{Y}\in L_k(\overline{X})}\left[R_k(\overline{X},\overline{Y})\right]$$

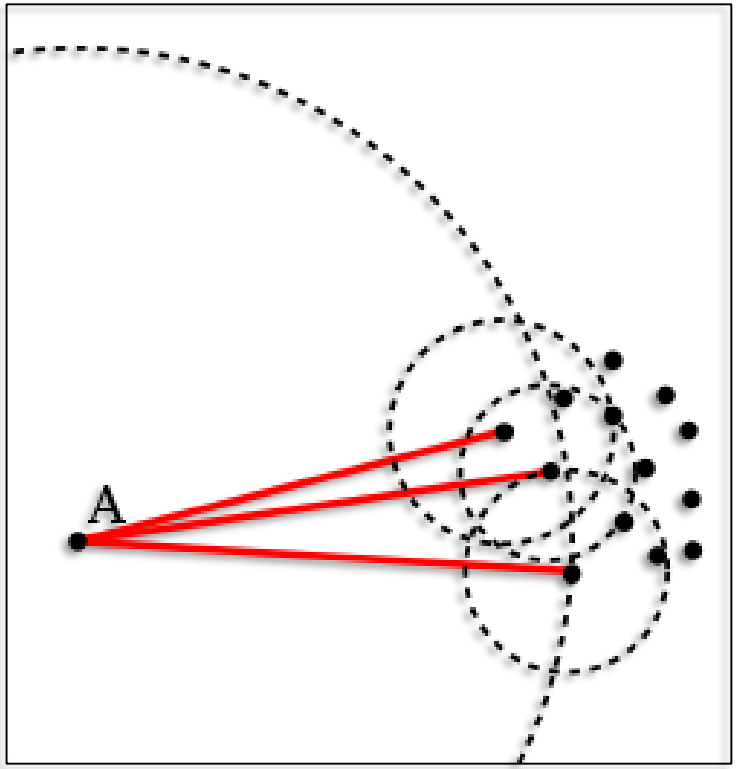
Finally, we define **Local Outlier Factor**, $LOF_k(\overline{X})$, to simply be the mean ratio of $AR_k(\overline{X})$ to the corresponding values of all points in the k-neighborhood of $\overline{X}$.

$$LOF_k(\overline{X}) = MEAN_{\overline{Y}\in L_k(\overline{X})}\left[\frac{AR_k(\overline{X})}{AR_k(\overline{Y})}\right]$$

[1]

$$LOF_k(\overline{X}) \approx 1, \text{Inlier}$$

$$LOF_k(\overline{X}) > 1, \text{Outlier}$$



Point ‘A’ has a lower local density than that of its *k* neighbors and is thus an outlier as intuition would suggest. [2]

Local Outlier Factor Algorithm *(Slide 3)*

Small Scale Example with $k = 3$

Point $(2, 2)$ has a neighbors $(5, 4.5), (4.5, 5.5), (5, 4.5)$ with euclidean distances from $(2, 2)$ of $3.91, 4.3, 5.0$, respectively. So, $D_3\big((2, 2)\big) = 5.0$ since $(5, 4.5)$ is the 3^{rd} nearest neighbor.

Calculating similarly for the other points shows they have D_3 of $1.80, 1.58, 1.58$, respectively.

So,

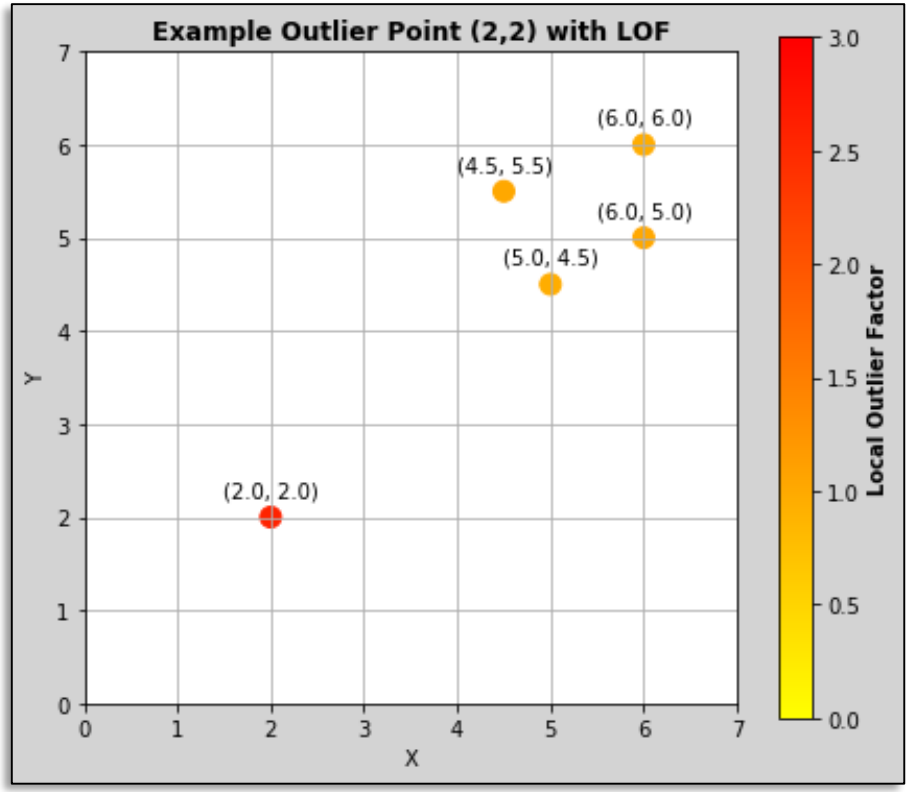
$$AR_3\big((2, 2)\big) = \frac{max(3.91, 1.8) + max(4.3, 1.58) + max(5, 1.58)}{3} = \frac{3.91 + 4.3 + 5}{3} = 4.4$$

Now, $(2, 2)$'s neighbors have AR_3 of $1.66, 1.73, 1.73$ and this is used to calculate its **Local Outlier Factor** thusly,

$$LOF_3\big((2, 2)\big) = \frac{\frac{4.4}{1.66} + \frac{4.4}{1.73} + \frac{4.4}{1.73}}{3} = \frac{2.66 + 2.55 + 2.55}{3} = \boxed{2.58}$$

This shows $(2, 2)$ is an outlier since $2.58 > 1$. Filling in the other points gives the following table. Notice that the other points in the dataset all are close to one another and have an $LOF_3 \approx 1$.

	(x, y)	Neighborhood	K Nearest	ARk	LOF
0	[2.0, 2.0]	[3, 1, 2]	2.0	4.40	2.58
1	[4.5, 5.5]	[3, 2, 4]	4.0	1.73	1.03
2	[6.0, 5.0]	[4, 3, 1]	1.0	1.73	1.03
3	[5.0, 4.5]	[1, 2, 4]	4.0	1.66	0.97
4	[6.0, 6.0]	[2, 1, 3]	3.0	1.66	0.97



Local Outlier Factor Algorithm *(Slide 4)*

Applying To Our Original Dataset

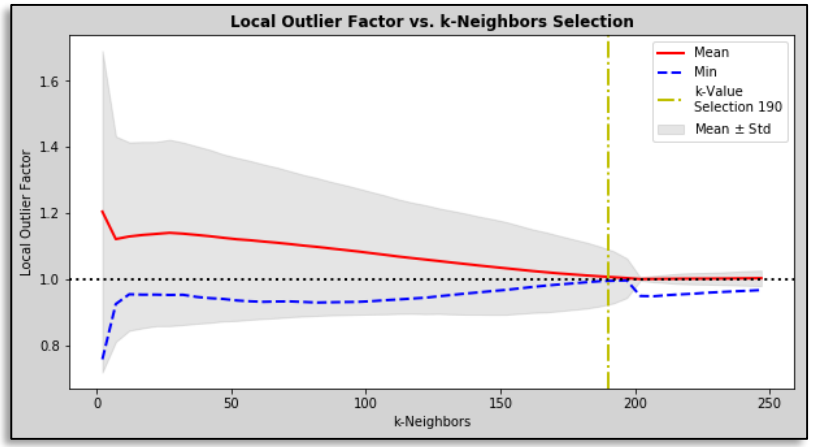
- The data set is comprised of two clusters and two outliers.

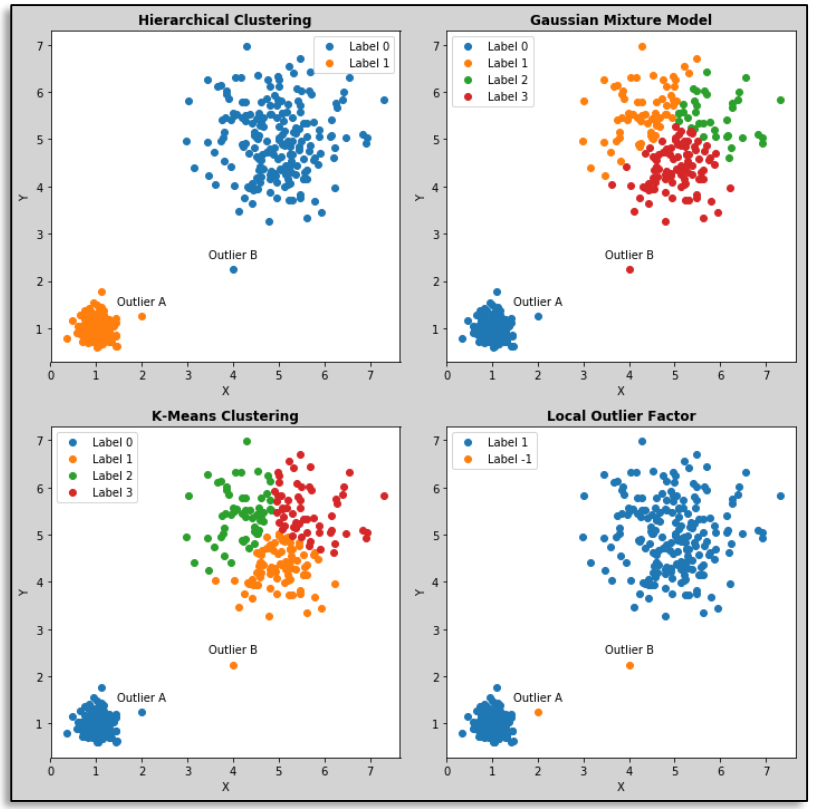
LOF Hyperparameter Tuning [3]

- *K*-Neighbors is the main tuning parameter.
- Selection of *k* can be regarded as the Minimum Number of Points a cluster must contain.
- Plot the Mean, Minimum, and Standard Deviation of LOF vs *k* and the point at which the Minimum and Mean converge to 1 is the approximate value best suited for *k.*

Comparing Results

- Standard hyperparameter selection methods applied to Hierarchical, Gaussian Mixture Model (GMM) and K-Means Algorithms were Applied
- Distance-based clustering, Hierarchical, simply includes the outliers as part of the two clusters.
- Distribution-based clustering, GMM, and Centroid-based clustering, K-Means, found erroneous extra clusters and included the outliers.
- LOF is uniquely designed to find the outliers and so does even with ‘Outlier A’ being relatively close to the adjoining cluster.

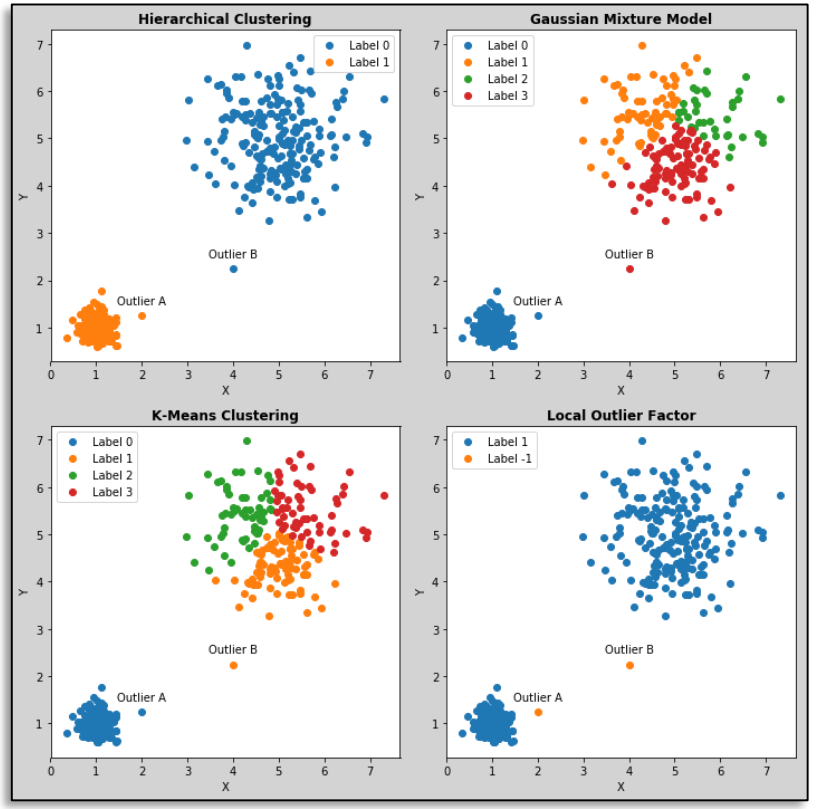




Summary *(Slide 5)*

Anomaly/Outlier Detection [3]

- LOF fills a gap left by other clustering methods for outlier detection.
- LOF takes into account the local density of each point and assigns a continuous value indicating the degree to which each point is an outlier.



LOF Algorithm[1]

- Given *k*, Computer the **Reachability Distance** to the point's *k*-neighbors and take the average of that set to get the **Average Reachability Distance**
- Take the ratio of the point's average reachability distance to that of all its *k*-neighbors and average this set to obtain the **Local Outlier Factor**
- Points with an LOF near unity are in a cluster, and points with LOF higher than unity are outliers.

Strengths [2]

- Due to the local approach, LOF is able to identify outliers in a data set that would not be outliers in another area of the data set.

Weaknesses

- Threshold of LOF for being an outlier may vary depending on dataset [2]
- Algorithm is susceptible to the Curse of Dimensionality [5]

Recommended Further Reading

- References 1, 3, and 5

References

1. Aggarwal, C. C. (2013). Chapter 4: Proximity-Based Outlier Detection. In *Outlier Analysis* (pp. 117-131). New York, NY: Springer.
2. Local outlier factor. (2020, May 22). Retrieved August 12, 2020, from https://en.wikipedia.org/wiki/Local_outlier_factor
3. Breunig, M., & Kreigel, H. (2000). LOF: Identifying Density-based Local Outliers. In *Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data: 2000, Dallas, Texas, United States, May 15- 8, 2000* (pp. 93-104). New York, NY: Association for Computing Machinery.
4. Hawkins, D.: "*Identification of Outliers*", Chapman and Hall, London, 1980.
5. Zimek, A., Schubert, E., & Kriegel, H. (2012). A survey on unsupervised outlier detection in high-dimensional numerical data. *Statistical Analysis and Data Mining, 5*(5), 363-387. doi:10.1002/sam.11161