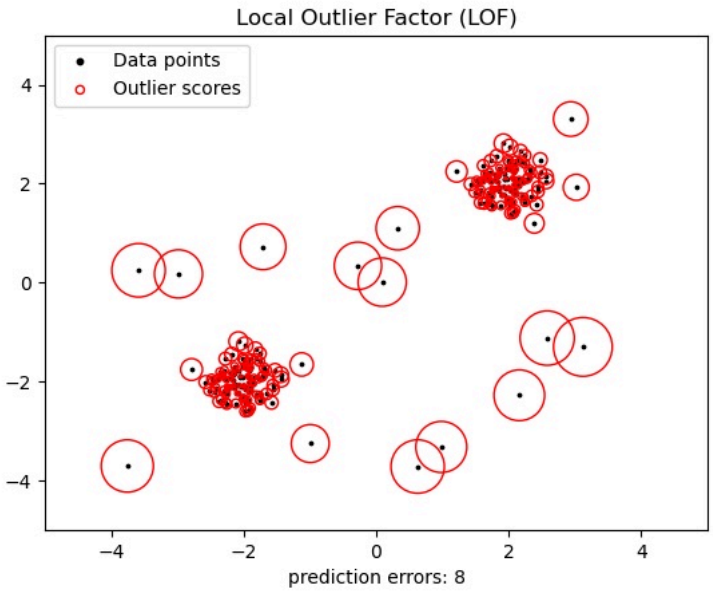
Local Outlier Factor

Add comment

Unsupervised Outlier Detection using the Local Outlier Factor (LOF).

The anomaly score of each sample is called the Local Outlier Factor. It measures the local deviation of the density of a given sample with respect to its neighbors. It is local in that the anomaly score depends on how isolated the object is with respect to the surrounding neighborhood. More precisely, locality is given by k-nearest neighbors, whose distance is used to estimate the local density. By comparing the local density of a sample to the local densities of its neighbors, one can identify samples that have a substantially lower density than their neighbors. These are considered outliers.



Function

```python
def mark_outliers_lof(dataset, columns, n=20):
    """Mark values as outliers using LOF

    Args:
        dataset (pd.DataFrame): The dataset
        col (string): The column you want apply outlier detectio
        n (int, optional): n_neighbors. Defaults to 20.

    Returns:
        pd.DataFrame: The original dataframe with an extra boole
        indicating whether the value is an outlier or not.
    """

    dataset = dataset.copy()

    lof = LocalOutlierFactor(n_neighbors=n)
    data = dataset[columns]
    outliers = lof.fit_predict(data)
    X_scores = lof.negative_outlier_factor_

    dataset["outlier_lof"] = outliers == -1
    return dataset, outliers, X_scores
```

Links

sklearn.neighbors.LocalOutlierFactor
Examples using sklearn.neighbors.LocalOutlierFactor: Comparing anomaly detection ...
scikit-learn.org/stable/modules/generated/sklearn.neighbors.LocalOutlierFactor.html

Outlier detection with Local Outlier Factor (LOF)
The Local Outlier Factor (LOF) algorithm is an unsupervised anom...
scikit-learn.org/stable/auto_examples/neighbors/plot_lof_outlier_detection.html