Local Outlier Factor (LOF)

明治大学 理工学部 応用化学科
データ化学工学研究室　金子 弘昌

Local Outlier Factor (LOF) とは？

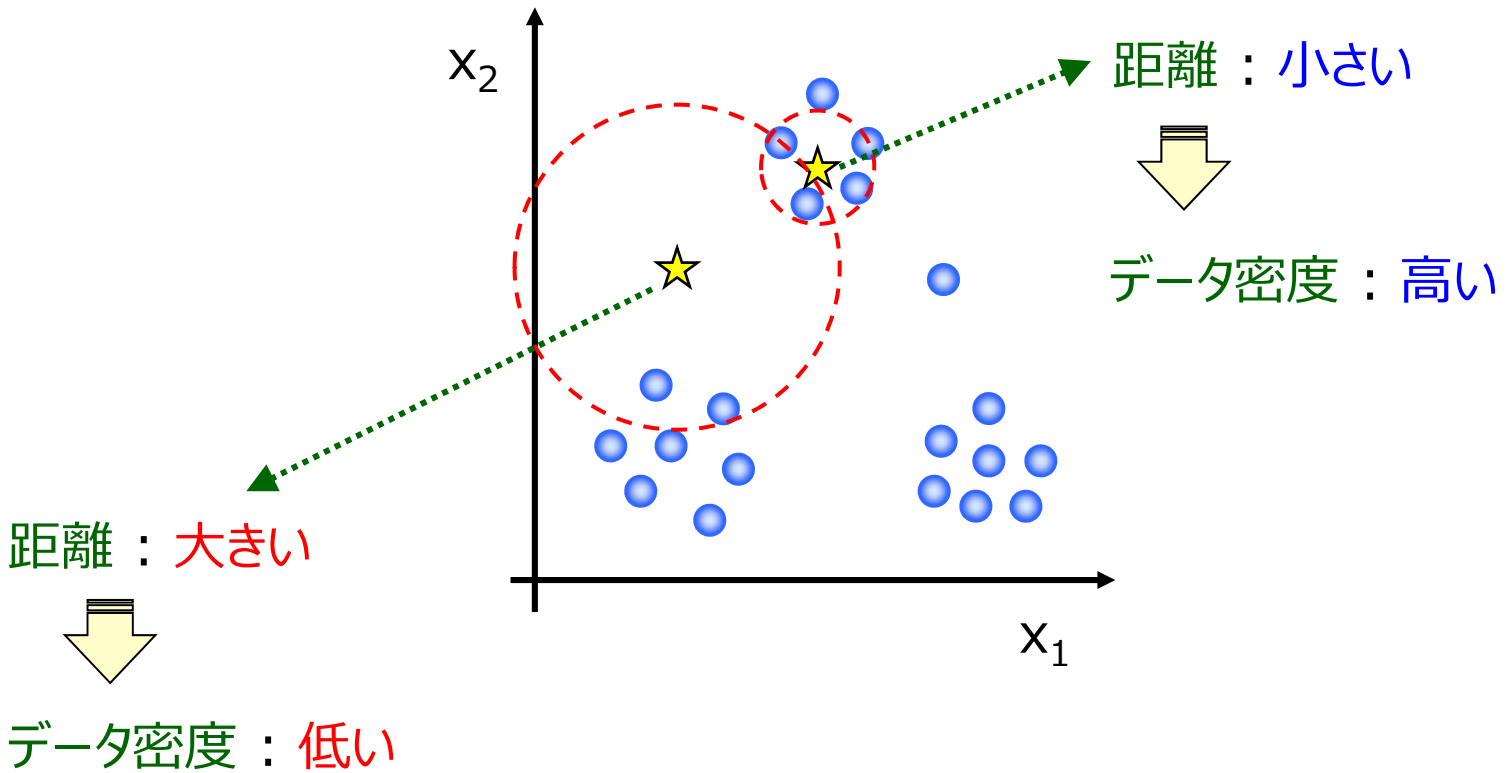
✓データ密度を推定する手法

✓k 最近傍法 (k-Nearest Neighbor, k-NN) による密度推定と比べて、
データ分布における局所的なデータ密度の違いを考慮可能

✓LOF の結果から外れサンプル検出や (装置やプロセスなどの)
異常検出が可能

復習）k-NN によるデータ密度の指標

例) $k = 3$

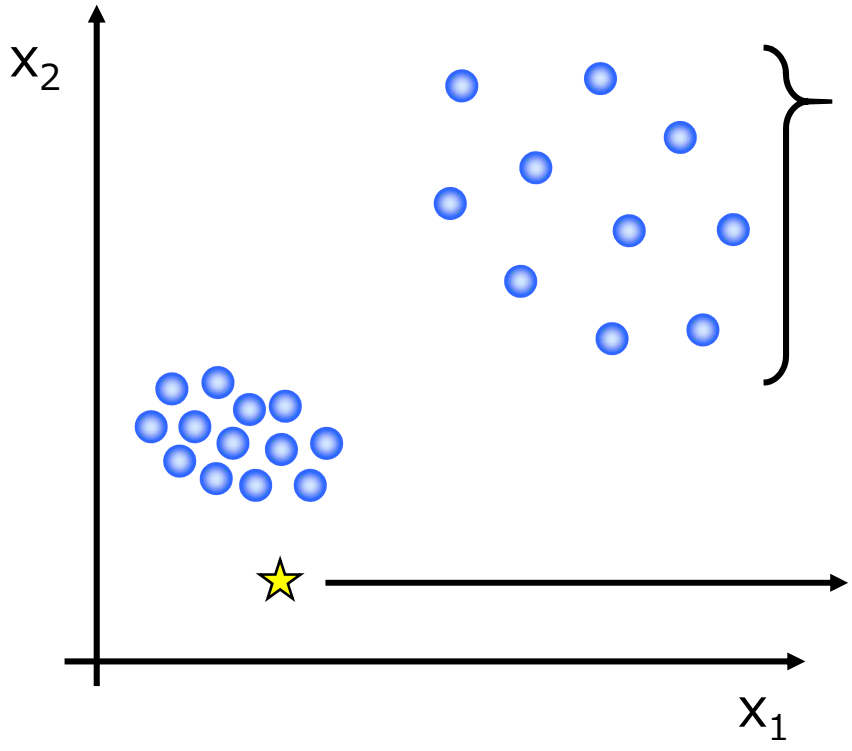


k 個の距離の平均をデータ密度の指標とする

指標の値が小さいほど、データ密度が高い

k-NN で何が問題か？

データがまんべんなく分布していればよいが、領域によって
データ分布に違いがあると、

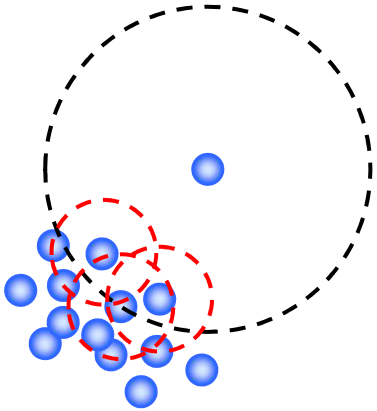


このあたりのサンプルのばらつきが
大きく、「距離のしきい値を
大きくしても」いいのかな、となり、

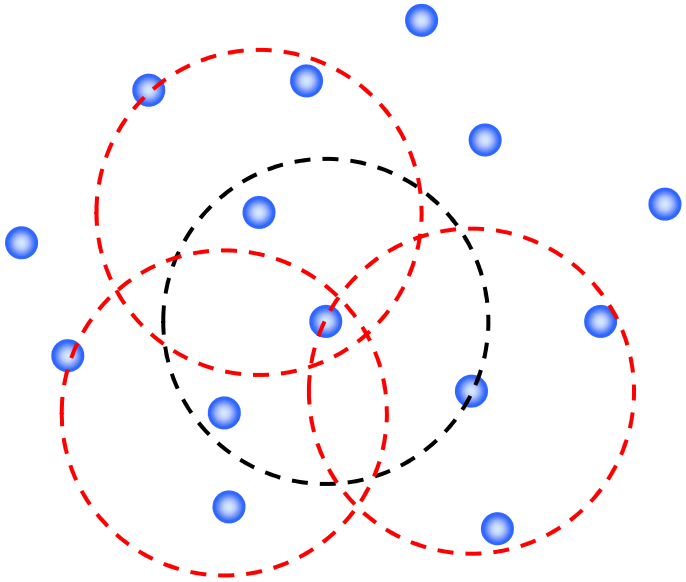
サンプルが密に固まっている領域に
おいて、このような少し外れたところに
あるサンプルを外れサンプル (データ
密度が低い) として検出できない

LOF ではどうするか？

最も近い k 個のサンプルとの距離だけでなく、その k 個のサンプルに
最も近い k 個のサンプルとの距離も考慮する！



区別できる！



あるサンプルの LOF をどう計算するか？

あるサンプル A と最も距離の近い k 個のサンプル郡を $N_k(A)$ とする

A とサンプル B との距離を $d(A, B)$ とする

A に対して k 番目に近いサンプルと A との距離を k-distance(A) とする

A と B との間の reachability distance という距離 reach-dist$_k$(A,B) を以下のように定義する

$$\text{reach-dist}_k(A, B) = \max(d(A, B), k\text{-distance}(B))$$

あるサンプルの LOF をどう計算するか？ 1/4

reach-dist$_k$(A, B) = max(k-distance(B), d(A, B))

A と B が離れていると (そして B の近くにサンプルがあると)、
reach-dist$_k$(A, B) は単純に A と B の距離

A と B が十分近くにいると (そして B の近くにサンプルがあると)、
reach-dist$_k$(A, B) は k-distance(B) に置き換わる

注) reach-dist$_k$(A, B) には対称性がないため、数学的には
距離ではない

あるサンプルの LOF をどう計算するか？ 2/4

A と $N_k(A)$ の間の reachability distance の平均を
$\text{mean-reach-dist}_k(A, N_k(A))$ とすると、

$$\text{mean-reach-dist}_k\left(\text{A}, \text{N}_k\left(\text{A}\right)\right) = \frac{\sum_{C \in \text{N}_k(\text{A})}^{k} \text{reach-dist}_k\left(A, C\right)}{k}$$

A の local reachability density ($\text{lrd}_k(A)$) を以下のように定義する

$$\text{lrd}_k\left(\text{A}\right) = \frac{1}{\text{mean-reach-dist}_k\left(\text{A}, \text{N}_k\left(\text{A}\right)\right)}$$

距離 distance が小さいと、密度 density が大きい

あるサンプルの LOF をどう計算するか？ 3/4

A と $N_k(A)$ の間の reachability distance の平均を
$\text{mean-reach-dist}_k(A, N_k(A))$ とすると、

$$\text{mean-reach-dist}_k\left(\text{A}, \text{N}_k\left(\text{A}\right)\right) = \frac{\sum_{C \in \text{N}_k(\text{A})}^{k} \text{reach-dist}_k\left(A, C\right)}{k}$$

A の local reachability density ($\text{lrd}_k(A)$) を以下のように定義する

$$\text{lrd}_k\left(\text{A}\right) = \frac{1}{\text{mean-reach-dist}_k\left(\text{A}, \text{N}_k\left(\text{A}\right)\right)}$$

距離 distance が小さいと、密度 density が大きい

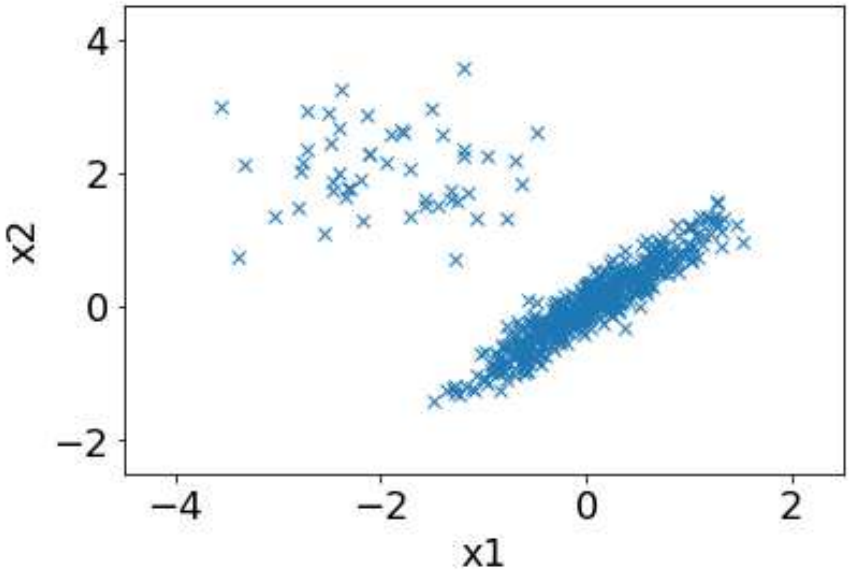
あるサンプルの LOF をどう計算するか？ 4/4

LOF の値 ($LOF_k(A)$) は以下のように計算される

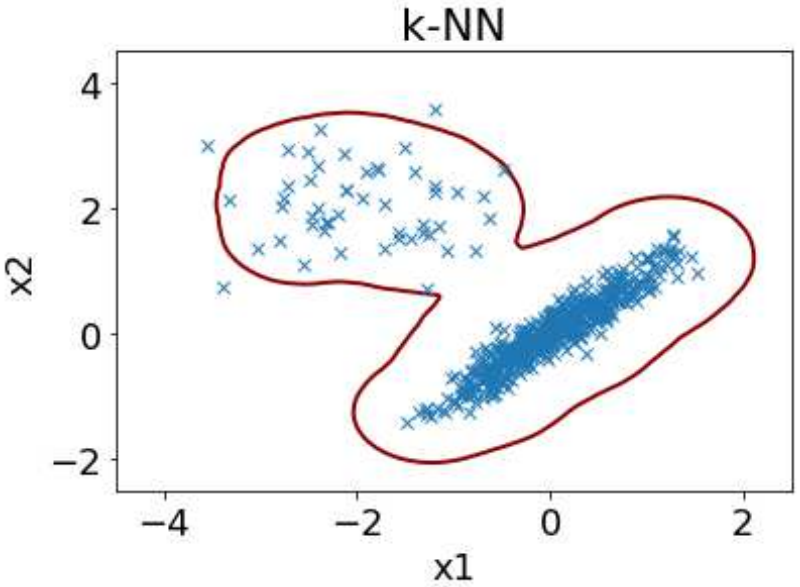
$$\mathrm{LOF}_k(\mathrm{A}) = \frac{\sum_{C \in \mathrm{N}_k(\mathrm{A})} \mathrm{lrd}_k(\mathrm{C})}{k} \frac{1}{\mathrm{lrd}_k(\mathrm{A})}$$

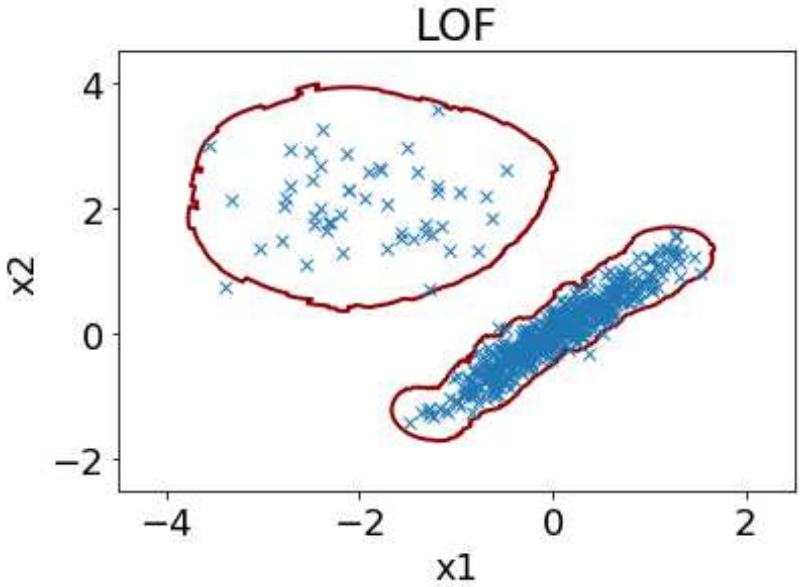
$N_k(A)$ の local reachability density を平均し、
A の local reachability density で標準化している

実行例



左のようなサンプルにおいて、k-NN と LOF をそれぞれ実行し、外れサンプルの割合を 1% として、正常なサンプル (外れサンプルでないサンプル) の領域を赤線で囲むと、以下の通り。k-NN では右下のデータが密集した領域の正常領域が広くなっているが、LOF では適切な領域を正常としている





Python コード例

こちら https://github.com/hkaneko1985/dcekit にある

demo_lof.py で「実行例」と同じことができます

scikit-learn を使うときの注意点

- ✓scikit-learn の sklearn.neighbors.LocalOutlierFactor が便利
  https://scikit-learn.org/stable/modules/generated/sklearn.neighbors.LocalOutlierFactor.html
- ✓トレーニングデータから外れサンプルを探すときには (anomaly detection)、デフォルトの設定 (novelty=False) でよいが 、トレーニングデータが存在するときの、テストデータのデータ密度を計算するときは (novelty detection)、novelty=True とする
- ✓contamination で外れサンプルの割合を設定する必要がある

k-NN, OCSVM との比較・注意点

✓k-NN
- データ分布に違いがあると対応が難しい
- 次元の呪いの影響を受ける

✓OCSVM (One-Class Support Vector Machine)
- データ分布に違いがあると対応が難しい
- 次元の呪いの影響を受けにくい

✓LOF
- データ分布に違いがあっても対応できる
- 次元の呪いの影響を受ける

参考文献

- ✓ M. M. Breunig, H. P. Kriegel, R. T. Ng, J. Sander, LOF: identifying density-based local outliers, Proceedings of the 2000 ACM SIGMOD international conference on Management of data, 93-104, 2000. DOI: 10.1145/342009.335388 https://dl.acm.org/citation.cfm?id=335388

- ✓ https://en.wikipedia.org/wiki/Local_outlier_factor

- ✓ J. Lee, B. Kang, S. H. Kang, Integrating independent component analysis and local outlier factor for plant-wide process monitoring, Journal of Process Control, 21, 1011-1021, 2011. DOI: 10.1016/j.jprocont.2011.06.004 https://www.sciencedirect.com/science/article/pii/S095915241100144