第4回目の補足資料

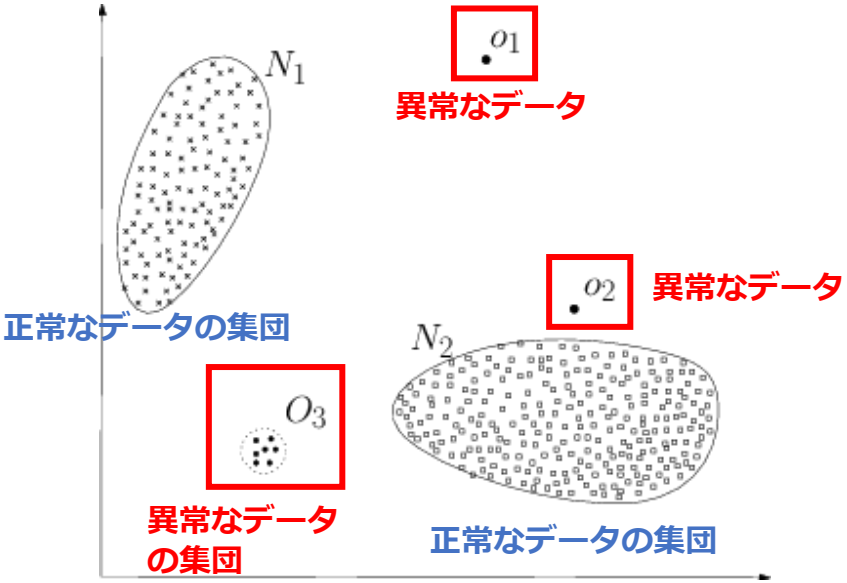
異常検知

参考・引用

- A Unifying Review of Deep and Shallow Anomaly Detection、 https://arxiv.org/pdf/2009.11732.pdf
- https://en.wikipedia.org/wiki/Anomaly_detection
- Pythonによる異常検知
- 入門 機械学習による異常検知 Rによる実践ガイド

異常検知とは？

- 異常検知とは、正常なデータのから乖離しているデータを検出すること



https://www.kaggle.com/code/matheusfacure/semi-supervised-anomaly-detection-survey

実際の現場で行う異常検知の例

- 金融
  - クレジットカード等の利用の不正検知
- 製造業
  - センサーの時系列データからの異常の検知（予防保全）
- マーケティング
  - Twitterのバースト検知（トレンドの検知）
- IT
  - サーバアクセスの異常の検知
- 医療
  - 心電図データの異常検知

非時系列データにおける異常検知

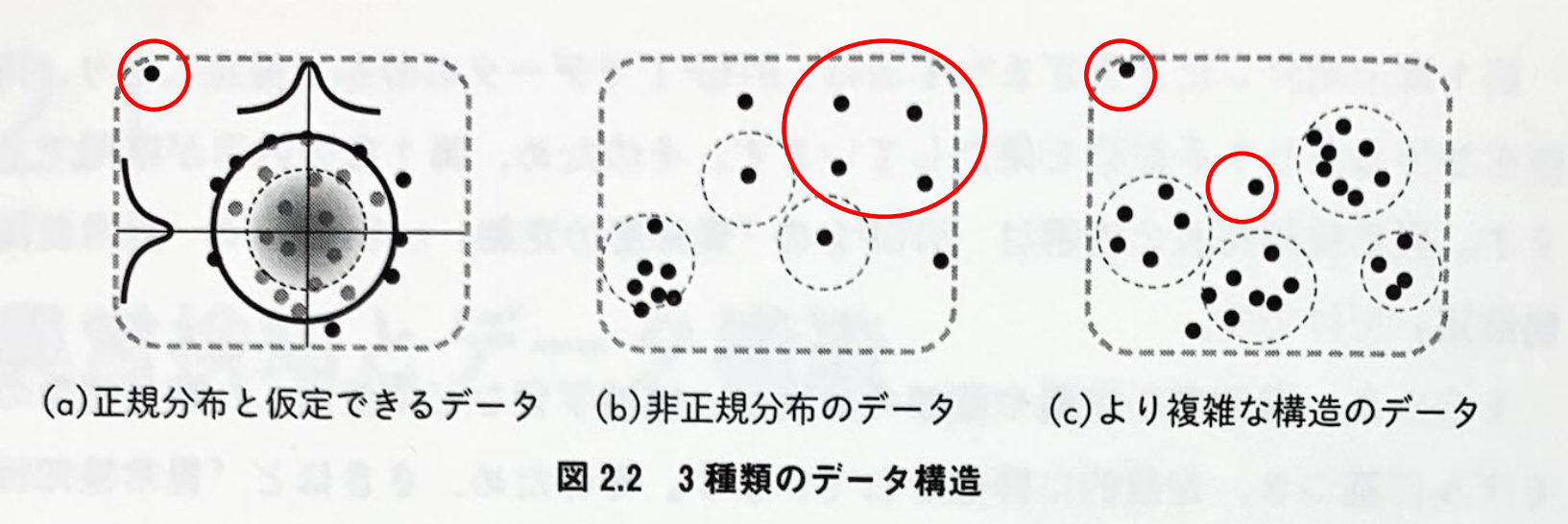
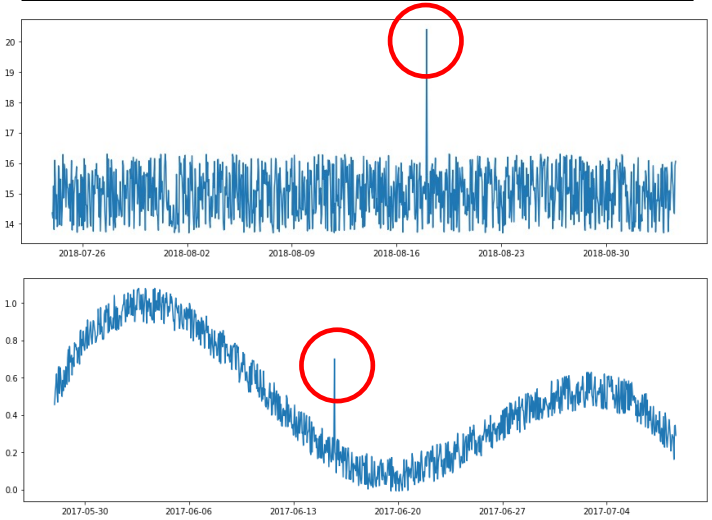


図 2.2　3 種類のデータ構造

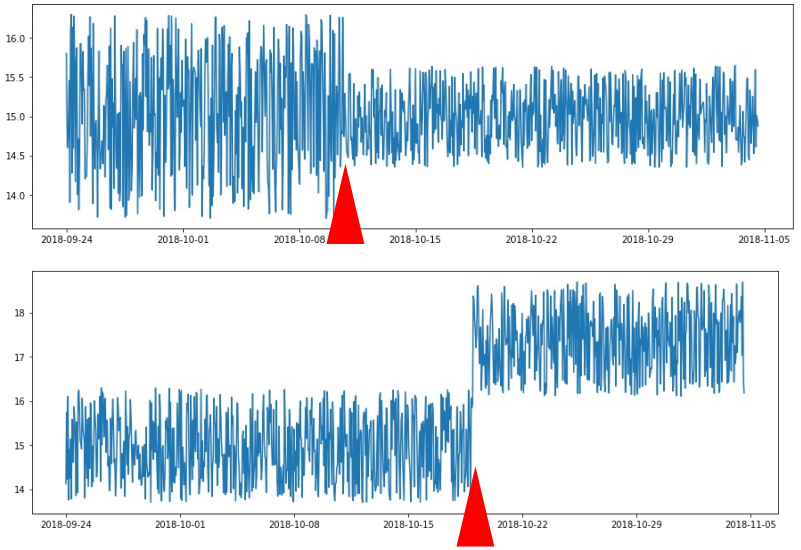
(書籍)Pythonによる異常検知

時系列データにおける異常検知 (何を検知するか？)

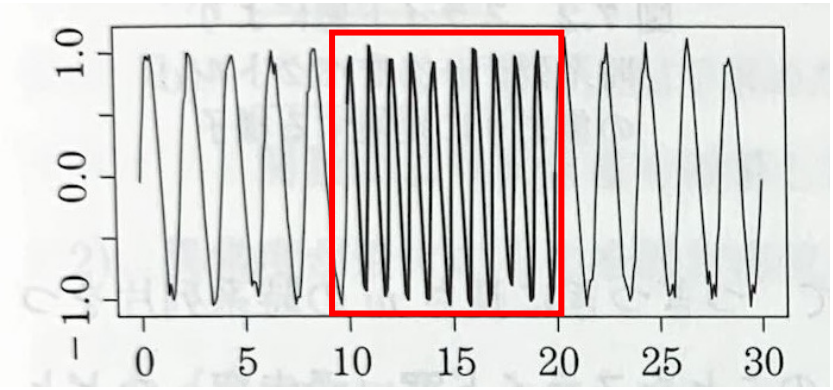
外れ値検知（データ点を検出）



変化点検知（切り替え点）



異常部位検知（部分時系列）



参考・引用)

- https://www.albert2005.co.jp/knowledge/machine_learning/anomaly_detection_basics/anomaly_detection_time
- https://practicaldatascience.co.uk/machine-learning/how-to-create-ecommerce-anomaly-detection-models
- (書籍)入門 機械学習による異常検知

異常検知の手法～考え方 1

- 大きく分けて 3 つのカテゴリがある

- 1）教師あり学習による検出
  - 「正常」と「異常」のラベルが付けられたデータ から、正常と異常を判別する学習器を学習する
  - 一般的に、異常データはごく少数で不均衡データになるため（正例が3個、負例が1000個など）、それだけから境界面を学習することは難しい

- 2）教師なし学習による検出
  - 正常なデータから正常性のモデル（統計モデルも含む）を学習し、モデルからの逸脱によって異常を検知する。逸脱度合いは「異常スコア (Anomaly score)」などと言われる。一般的に多く用いられる

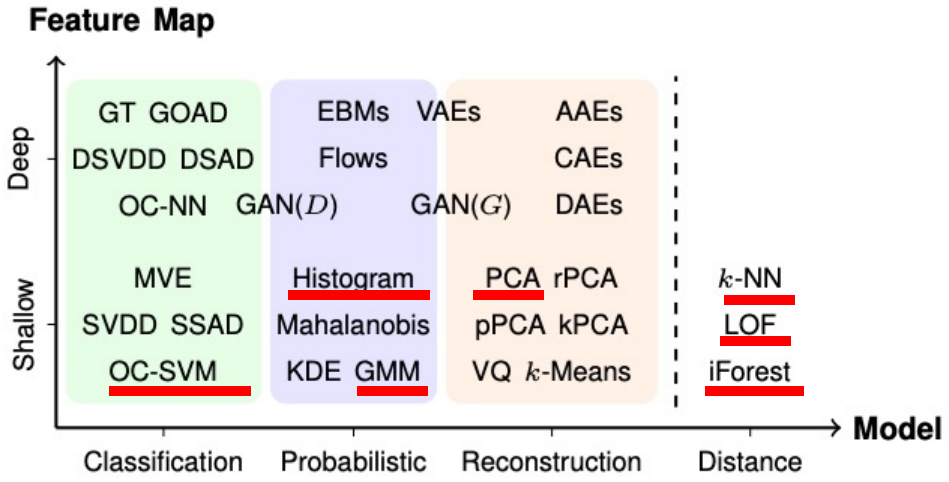
- 3）半教師あり学習による検出
  - データの一部にラベルが付与されたデータを学習する（後述）。あまり一般的ではない

異常検知の手法～考え方２

手法	概要	長所	短所
Probabilistic	正常データの確率分布を推定し、生起確率が低い事象を異常とみなすアプローチ	データが仮定したモデルに従っていれば非常に高精度を達成	分布モデル等多くの仮定が必要。また高次元への対応も困難
Distance	データ間の距離に基づき異常検知を行うアプローチ。異常データは正常データから距離的に離れているという発想	単純なアルゴリズムで出力への説明可能性も高い	計算量が大きい（O(n2)など）
Classification	データの正常・異常を分ける境界を直接推定するアプローチ	分布の形状を仮定する必要がない。また、SVMを用いた手法においては比較的少数データで学習可能	正常データが複雑な構造をしていると上手く動作しない場合がある
Reconstruction	正常データを再構成するようにモデルを学習しそのモデルが上手く再構成できないデータは異常であるというアプローチ	オートエンコーダ(AE)やGAN等の手法が適応できる	あくまでも次元圧縮や再構成精度を高めるために目的関数がデザインされているため、必ずしも異常検知に特化した手法ではない

- **A Unifying Review of Deep and Shallow Anomaly Detection**
- **https://qiita.com/toucan/items/c3343de3cfa236df3bda**

- Probabilistic
  - GMM：ガウス混合モデル
- Distance
  - k-NN：k近傍法
  - LOF：局所外れ値因子法
- Classification
  - OC-SVM：One-Class SVM
  - iForest： IsolationForest
- Reconstruction
  - PCA



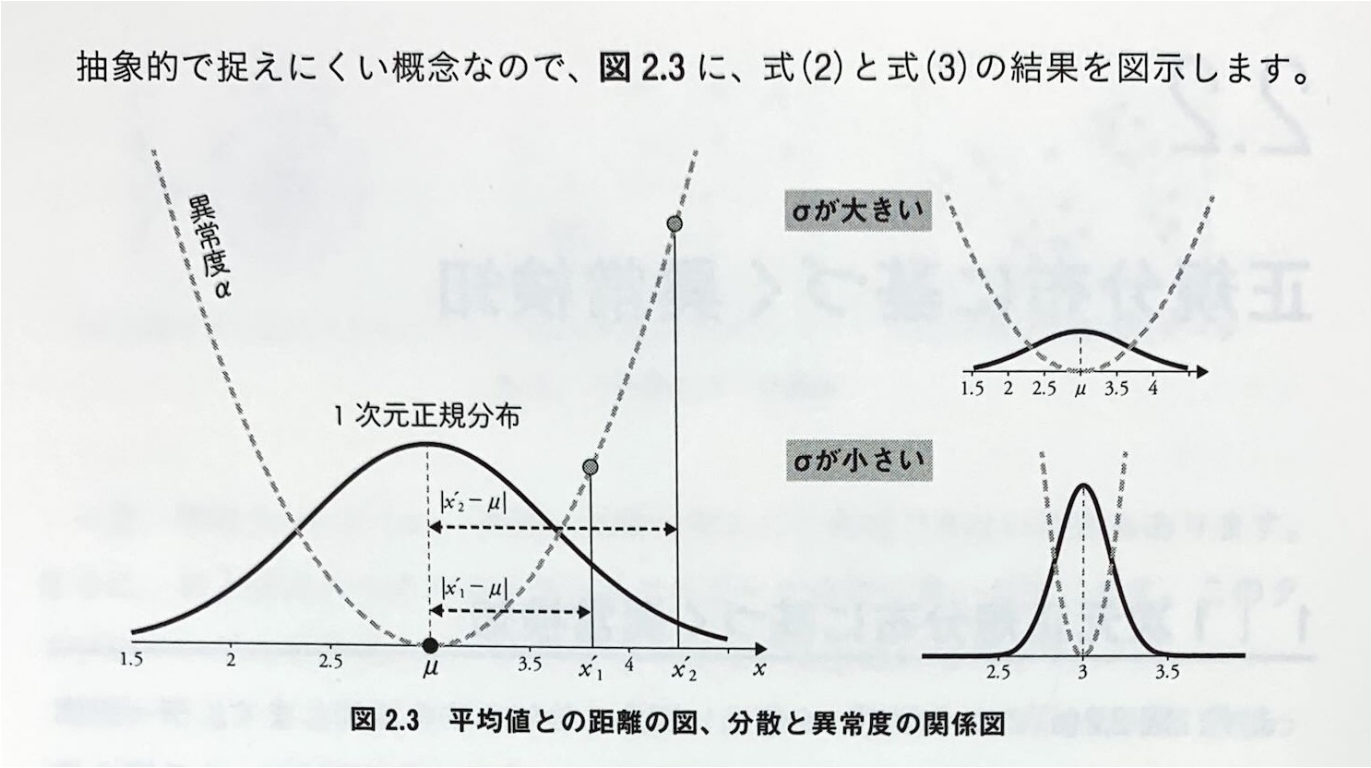
A Unifying Review of Deep and Shallow Anomaly Detection

各手法の補足

異常度は出現確率と反対の関係

- 平均より距離が遠いほど異常度は高い
- 分散σが大きいと異常度は緩やか上昇に、逆に分散が小さいと急激に上昇する

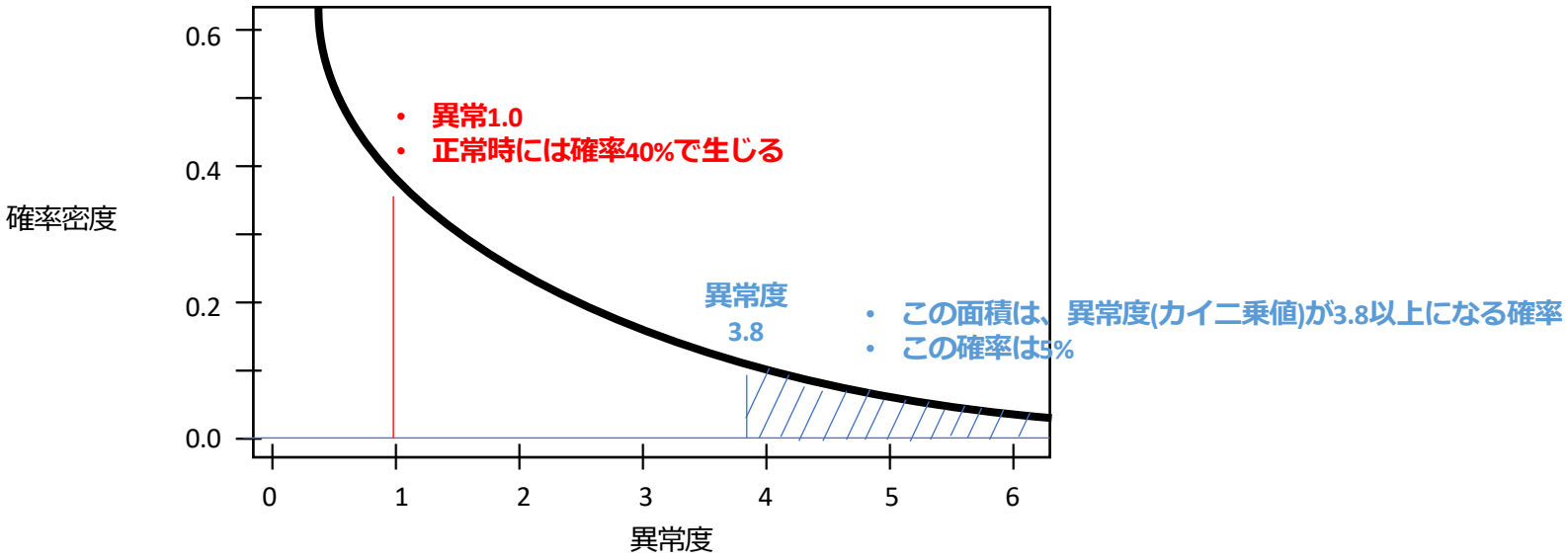
→ |x - μ|^2 / σ



(書籍)Pythonによる異常検知

カイ二乗分布による異常度の閾値の設定

- 観測した正常データが、独立して正規分布に従うとしたときの異常度の確率密度分布は、カイ二乗分布の確率密度分布に従う（ことが証明されている）
  - カイ二乗分布（変数は自由度k）、https://ja.wikipedia.org/wiki/%E3%82%AB%E3%82%A4%E4%BA%8C%E4%B9%97%E5%88%86%E5%B8%83
- 確率分布の面積は確率なので、例えば確率を5%にしたときは「正常時には5%未満でしか生じないデータを異常としている」のように確率的な視点で評価できる。この面積(=正常データで生じる確率)を%で指定すると、逆算してそのときの異常度が分かる。この異常度を閾値をする
  - 正常時には5%未満でしか生じないデータを異常としている＝＞異常度が3.8より大きい値を異常として検出する

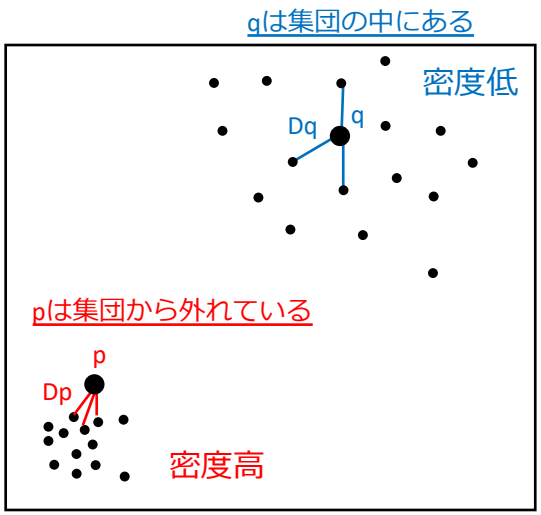


・参考）(書籍)入門 機械学習による異常検知

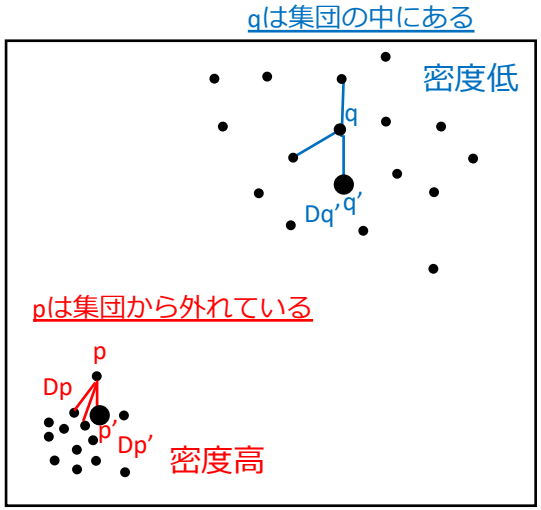
局所外れ値因子法（LOF, Local Outlier factor）

- 通常のk近傍法での問題
  - データの分布に密度差がある場合に最適なkの決定が難しい
  - 下の図の場合でpとqに着目する。pは集団から外れているので異常として検出したい、qは集団の中にあるので正常としたい
  - k=3などで近傍法を適用すると、近傍の3つの点との距離の平均（Dp、Dq）は、pは小さく、qは大きくなるため、qの方が異常サンプルになりやすい
  - 密度の差が原因
- 対策
  - 自身の近傍との距離だけでなく、近傍の点についても同じようにさらに近傍の点との距離を求め（Dp’、Dq’）、自身と近傍とで求めた距離を比べる

通常のk近傍法での距離の算出



LOFでの近傍点でのさらなる距離の算出



- DqとDq’は距離がだいたい同じ
- Dq’ / Dq ～ 1

- DpとDp’は距離が大きく異なる
- Dp’ / Dp << 1

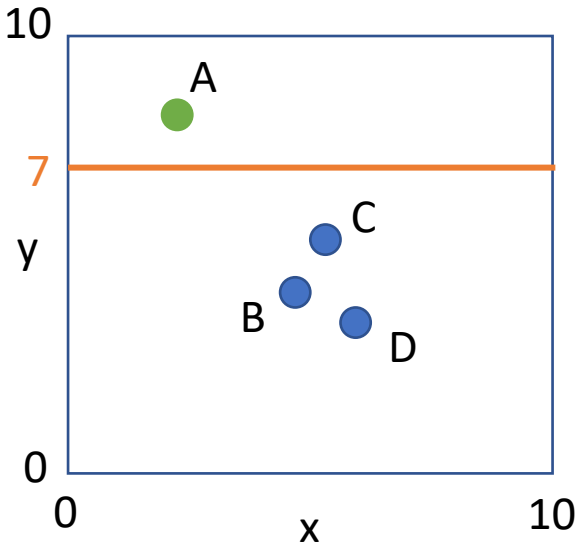
(書籍)入門 機械学習による異常検知

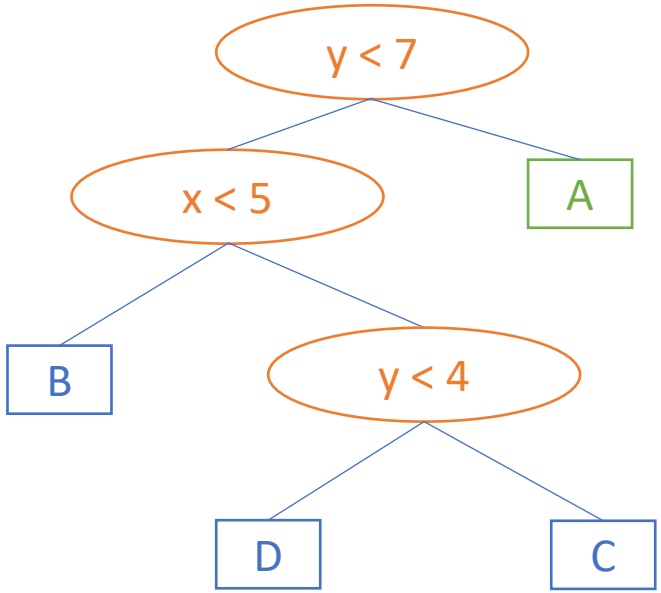
Isolation Forest
アイソレーションフォレスト

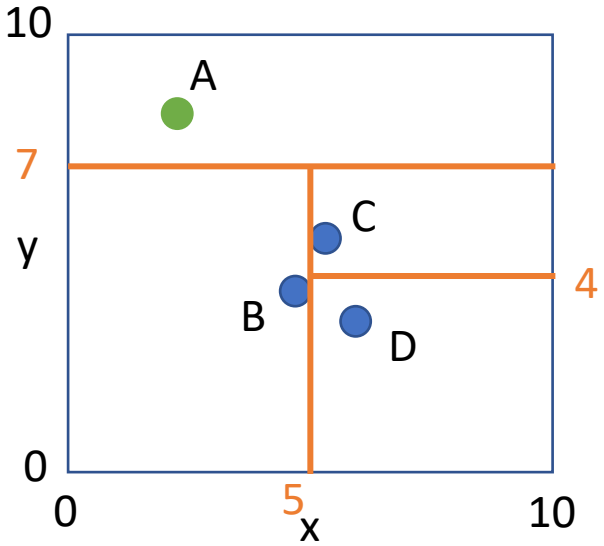
参考・参照

- https://www.slideshare.net/kataware/isolation-forest
- https://towardsdatascience.com/how-to-perform-anomaly-detection-with-the-isolation-forest-algorithm-e8c8372520bc

考え方







- 集団から離れている点ほど決定木を使ったときに上の方に出現しやすく、集団内の点ほど繰り出すために多くの分割を必要とする
- 点ごとの木の深さを異常度の算出に使う。深さが浅い方が異常度が高く、深さが深いほど異常度合いが低い
- 点ごとの深さは、決定木の作り方によって変わる（一番最初にx<6で分割すればDが最も浅い点になる）
- 決定木をたくさん作り、深さの平均を使う
- 分割の条件は、通常の決定木のようなエントロピーやジニ係数等は使わず、ランダムに変数を選択し、変数の最小～最大の中からランダムに数値を選ぶ

決定木を使った異常度合いの定量化

- 異常スコアは、単純な深さの平均ではなく、正規化してある
  - ある点の異常スコア：$2^{-E(h(x))/c(n)}$
  - E(h(x))：
    - h(x)はあるサンプルのある木での深さ
    - E(h(x))は生成した全ての木における深さの平均
  - c(n)：
    - E(h(x))を正規化するための値、０～１の値する
    - c(n) = 2Hn-1 – 2(n-1)/n、Hn = log(n) + 0.57721
  - c(n)について詳しく知りたい人はIsolation Forestの論文や二分木探索について調べてみてください（アルゴリズムとデータ構造）

- 決定木を作る際には、全てのデータは使わずに、ランダムにサンプリングしたサブグループに対して行う。これにより、集団から遠い点が塊になっていたときなどの精度を上げることができる

One-class SVM

サポートベクターマシン

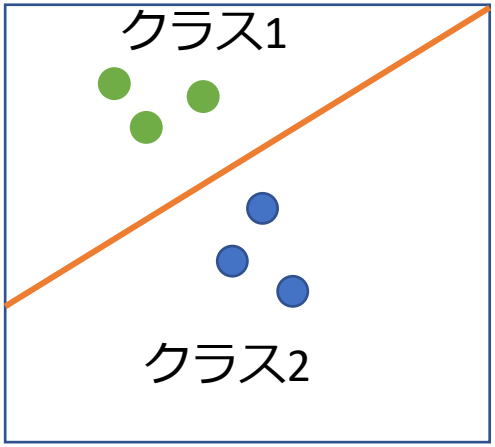
- 以下の2つが特徴のモデル
  - 高次元に写像することでクラスの分離を可能にしている（カーネルトリック）
  - 境界面と境界面の近傍のデータ点との距離を最大する（マージン最大）

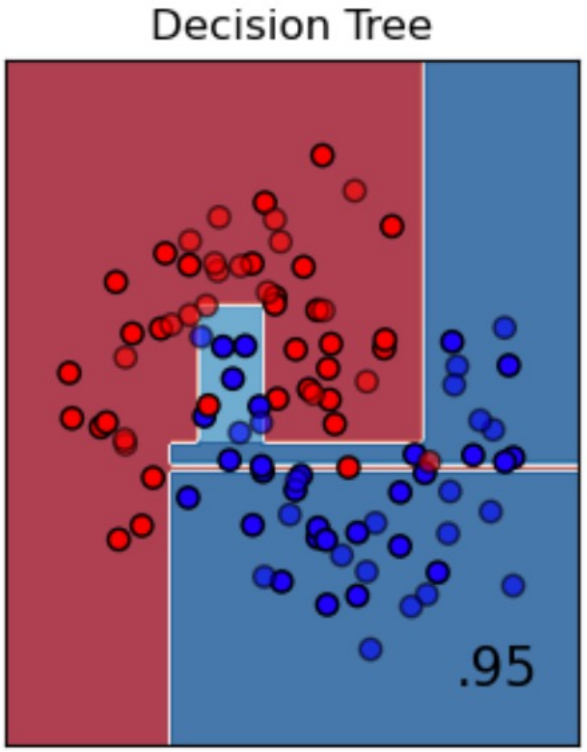
- 使う際のメリットデメリット
  - 汎化性能が高い（とデータマイニング全盛期のときは言われた）
  - パラメータで結果が大きくかわえる
  - 計算コストが高い

カーネルトリックとは～判別モデルの決定境界について

- 判別モデルを作る→境界線（境界面）を引いている

ロジスティック回帰

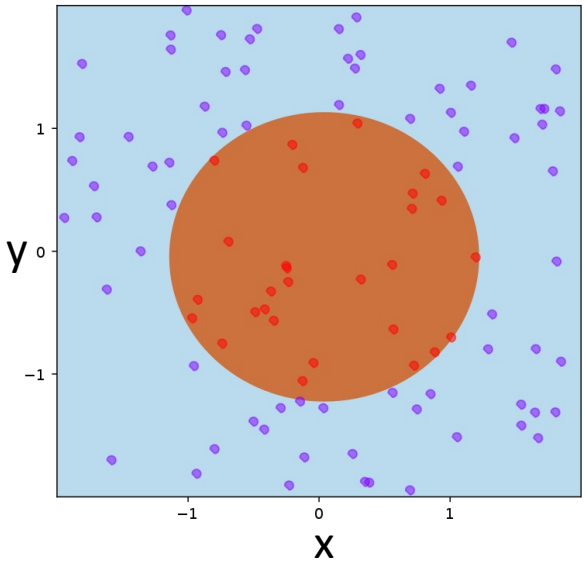


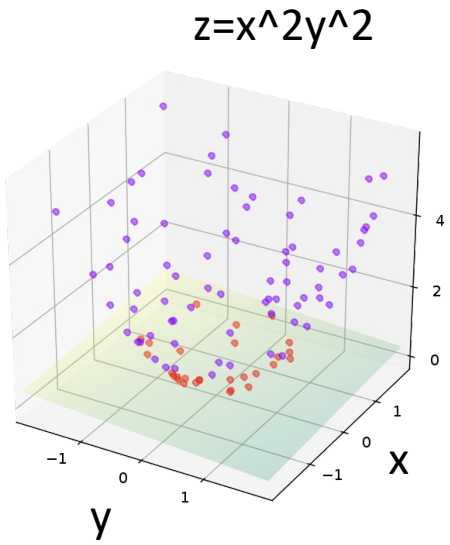


https://scikit-learn.org/stable/auto_examples/classification/plot_classifier_comparison.html#sphx-glr-auto-examples-classification-plot-classifier-comparison-py

カーネルトリックとは～SVMの決定境界について

- xとyの2次元だと赤と青を分割する境界線を引けない
- 新しい変数（z=x^2y^2）を加えて3次元にすると、z=1（くらい？）の面で赤と青のデータを分割できそう
- 新しい変数の作り方は、カーネル関数といったものが使われる





https://en.wikipedia.org/wiki/Support_vector_machine#/media/File:Kernel_trick_idea.svg

No	説明変数 x1	説明変数 x2
1		
2		
・		
N		

x

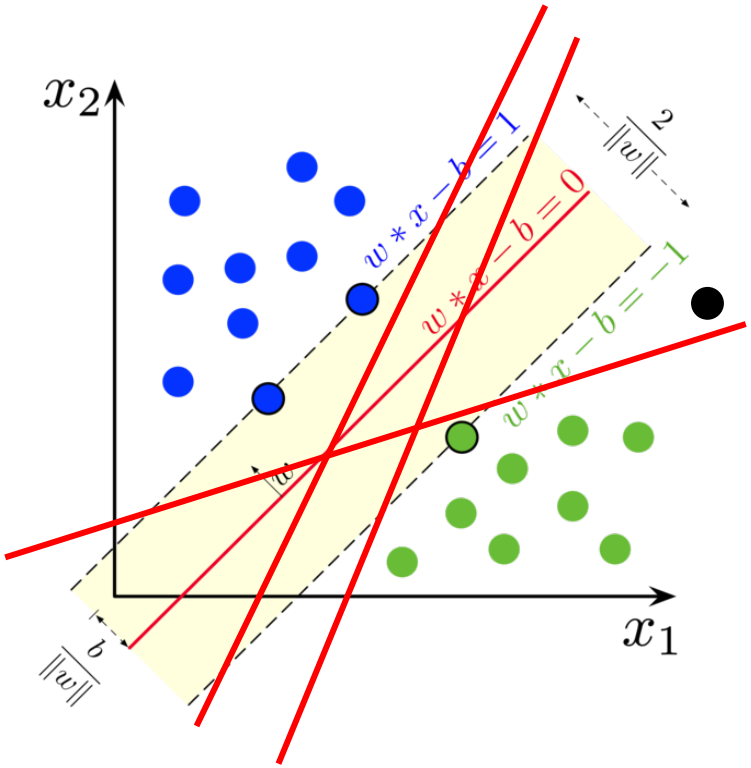
No	特徴量 gx1^2	特徴量 gx2^2	特徴量 √2gx1x2	特徴量 √2grx1	特徴量 √2grx2	特徴量 r
1						
2						
・						
N						

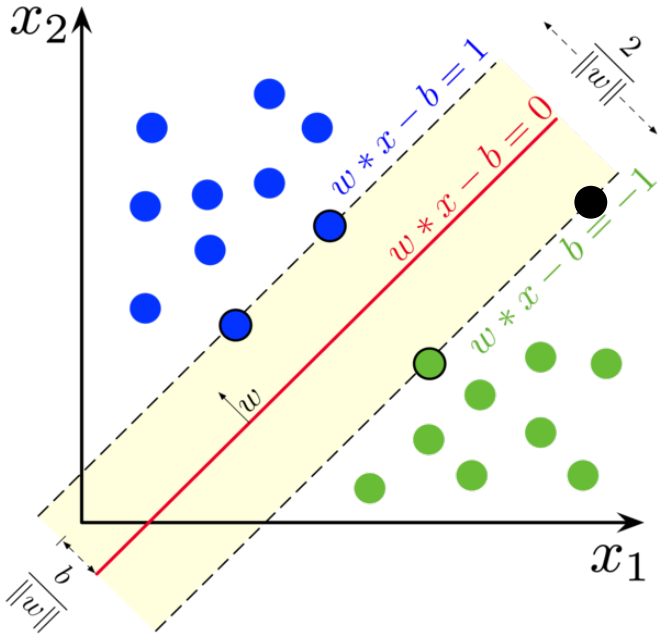
ϕ(X)

- テーブルデータで考えると、入力した変数に対して、カーネル関数を用いて複数の特徴量を生成している
- この特徴量に対して、誤差最小になるような係数を求める

マージン最大化

- 左図）青と緑を分割するだけなら、境界線は無数に引くことができる（学習データでの精度は同じ）
- 右図）境界線に一番近い青と一番近い緑の距離を最大にする（＝マージン最大化）→汎化性を高める（テストデータの黒い点に着目）

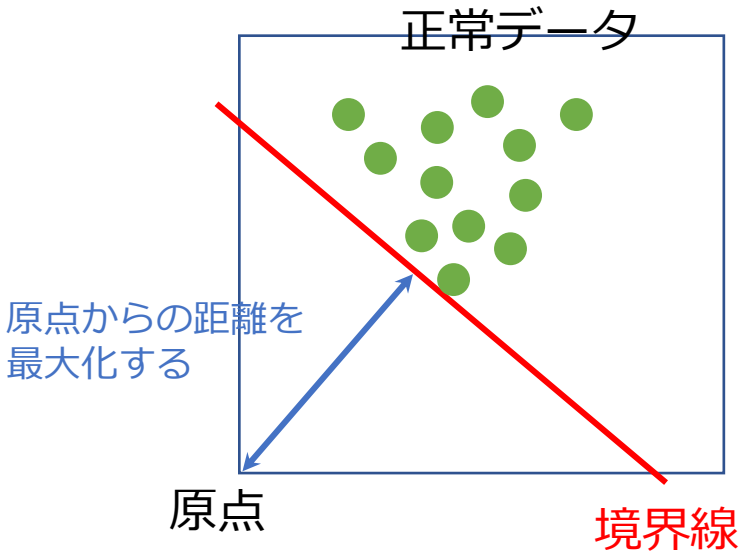


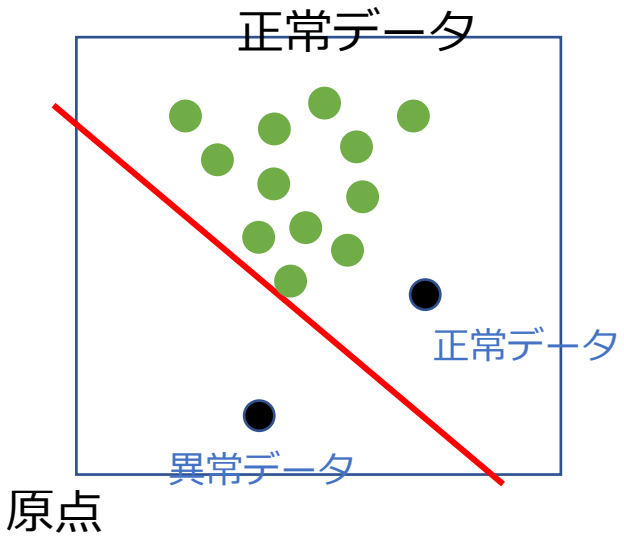


https://en.wikipedia.org/wiki/Support_vector_machine#/media/File:Kernel_trick_idea.svg

One-class SVMとは？

- 1クラスでは、正常データのみから学習を行い、そこから作成したモデルを用いて、正常か？異常か？（＝正常でない）を判別する
- 正例と負例をマージン最大で分割する境界線を求めるのではなく、原点と正常データとをマージン最大で分割する境界線を求める
- テストデータがある際に、境界線よりも正常データ側なら正常データ、原点側なら異常データという風に判別が可能になる



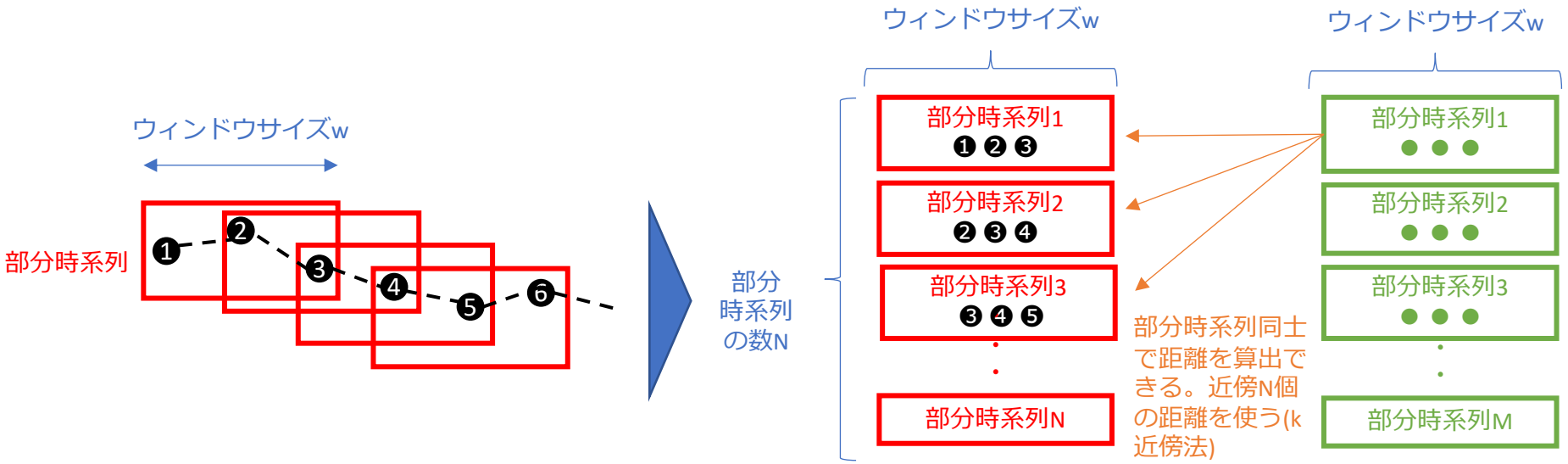


参考
- https://recruit.cct-inc.co.jp/tecblog/machine-learning/one-class-svm/

部分時系列とk近傍法

部分時系列を1つのデータ点とみなしk近傍法を実施

- 時系列データから、一定の幅（＝ウィンドウサイズという）のデータ群を抽出し、データ群が複数集まったものとして分析をする
- 各データ群を部分時系列という。部分時系列の集まりとして分析する



参照
- (書籍)入門 機械学習による異常検知

学習区間 ← テスト区間

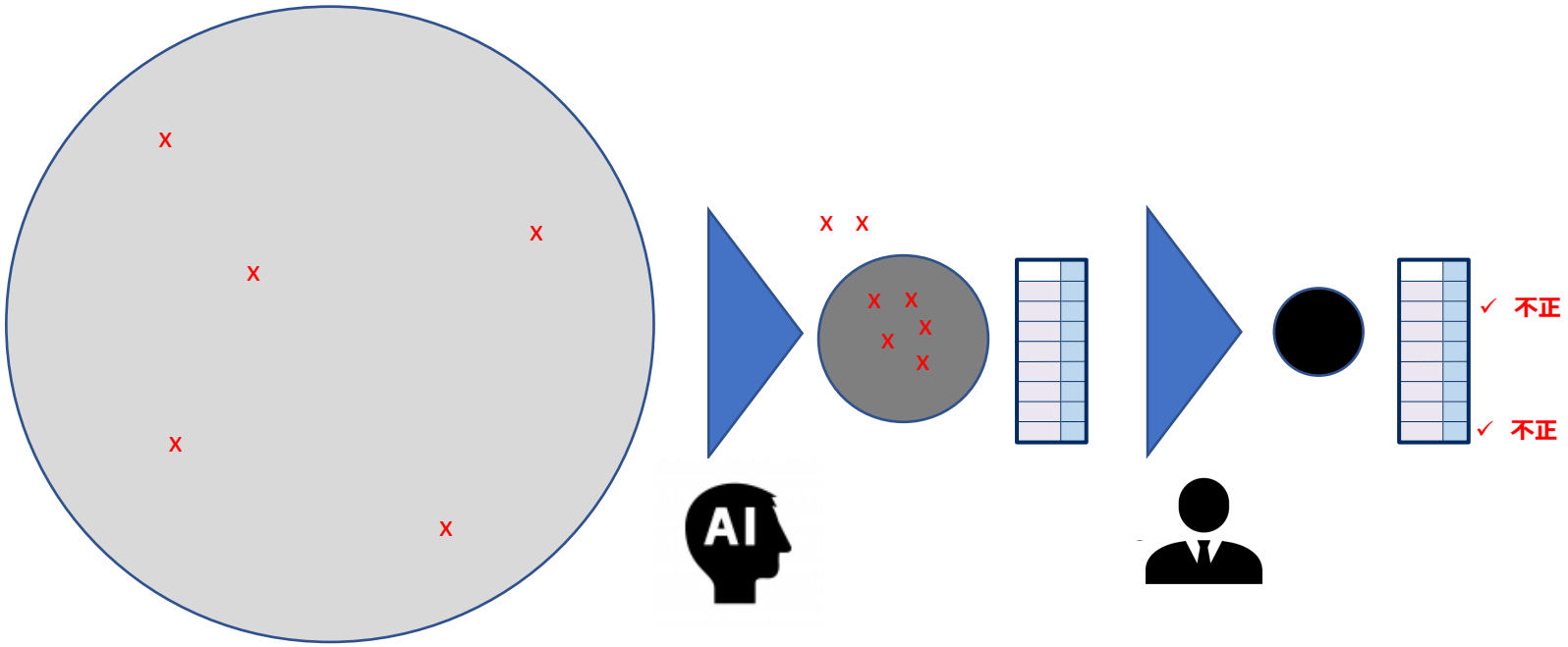
テスト区間の部分時系列kと学習区間の部分時系列1～Nとの近傍の距離を算出して、異常度とする

22

演習の補足説明

スクリーニングとしの機械学習（教師なし）の利用

- 異常の検出など、最終的には人がデータを目で見て総合的に判定すべきケースがある（医療など）
- 候補が大量にあると、全てを人が目で確認することは現実的に不可能になる。そのような場合に、機械学習を用いて怪しいサンプルをある程度スクリーニングすることで、効率的に異常なサンプルを発見できるようになる



全サンプル

数千～数万、
トランザクションだと数億など

候補サンプル
(スクリーニング後)

～100件

確定