KSKアナリティクス会社紹介

データサイエンス本部　ビジネス推進部　部長

シニアデータソリューションプランナー

高木宏明

KSKアナリティクス会社紹介

◇会社名：株式会社KSKアナリティクス

◇設立：2006 年 8 月

◇従業員数： 40人

◇Webサイト：https://www.ksk-anl.com/

◇所在地：

（本社オフィス）

-大阪市西区江戸堀1-18-35 肥後橋IPビル6F

（東京オフィス）

-東京都中央区築地2-7-10 築地シティプラザ6F

◇事業内容：

-統計・機械学習を活用したデータ分析サービス

-分析コンサルティング、分析ソフトウェアの販売

-分析業務基盤構築、導入支援、サポート

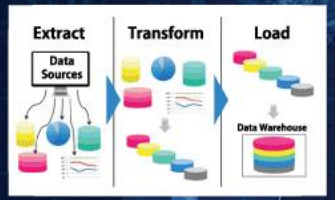
-分析教育プログラムの提供

◇主要取引先（敬称略）：

アサヒビール、味の素ゼネラルフーヅ、NRIシステムテクノ、ＮＴＴデータグループ、NTTドコモ、ＮＴＴ西日本、カルチュア・コンビニエンス・クラブ、キッセイ薬品工業、クボタ、KDDI、住友金属鉱山、ソフトバンクテレコム、トヨタ自動車グループ、日本ＨＰ、野村総合研究所、パイオニア、パナソニック、PFU、日立製作所、ファーストリテイリング、富士通、本田技術研究所、三菱電機、村田製作所、リクルートホールディングス、他

ビジネステクノロジー事業
データ蓄積・統合、可視化、BI

ビッグデータ基盤
（データレイク）構築



データ統合
ETL 作成

ビジネス
インテリジェンス構築

データサイエンス事業
予測、最適化、機械学習・AI

ディープラーニング
コンサルティング

機械学習
コンサルティング

データ分析教育
分析者育成

最近のトピック

TensorFlowについて、雑誌に寄稿

弊社アナリストがTensorFlowを解説

『初めてのTensorFlow』出版

数式なしでDeep Learningを解説・実践！

青山学院大学「データマイニング」講師

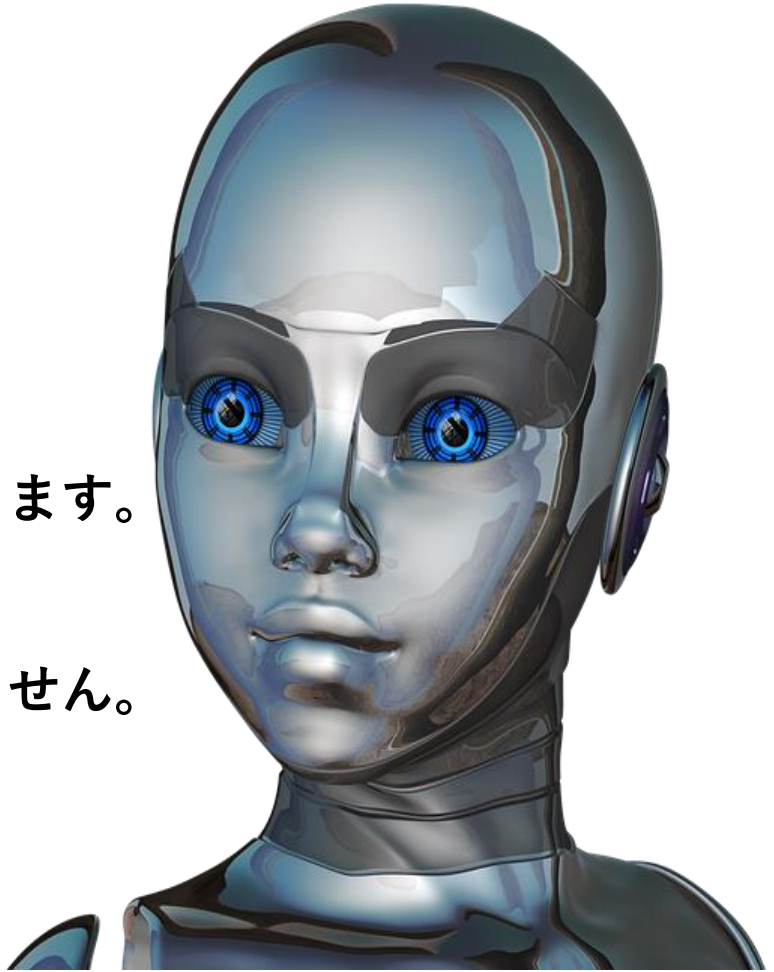
弊社社員が非常勤講師として「データマイニング」、
「データマイニング演習」を担当

KSK ANALYTICS

機械学習による分類モデル作成

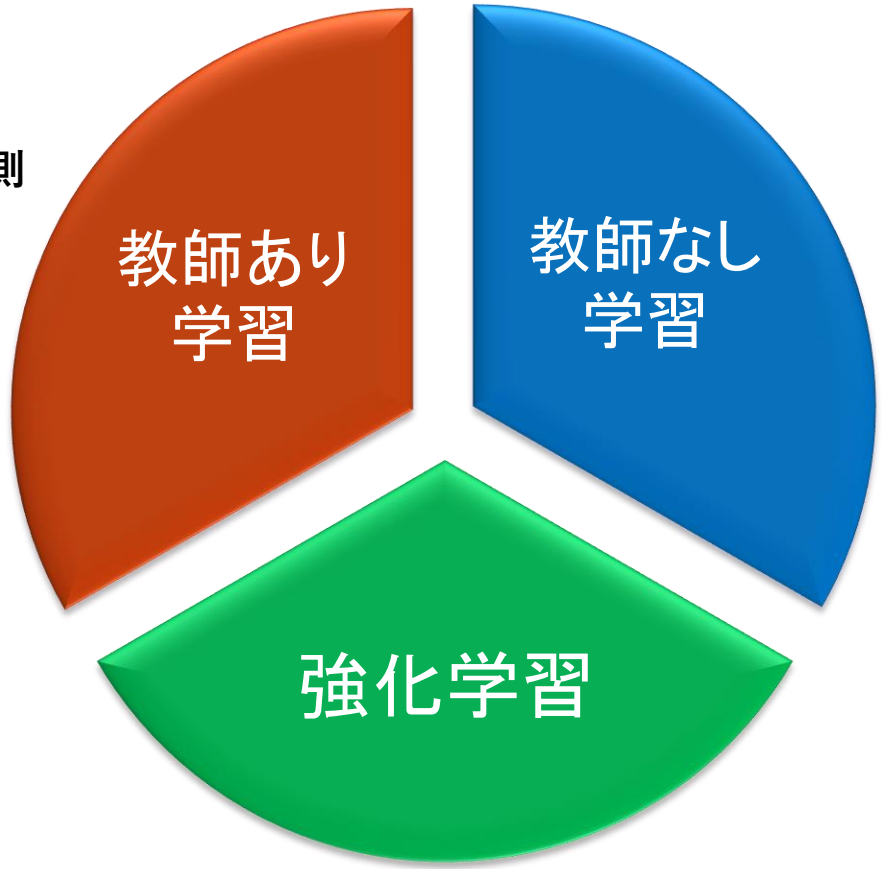
Why 機械学習？

- 人間が苦手な複雑さに対応
  -数値の計算にとても強いです。
- 速くて正確
  -定められたロジック通りに計算します。
- 自動化できる
  -その分析、あなたの手を煩わせません。



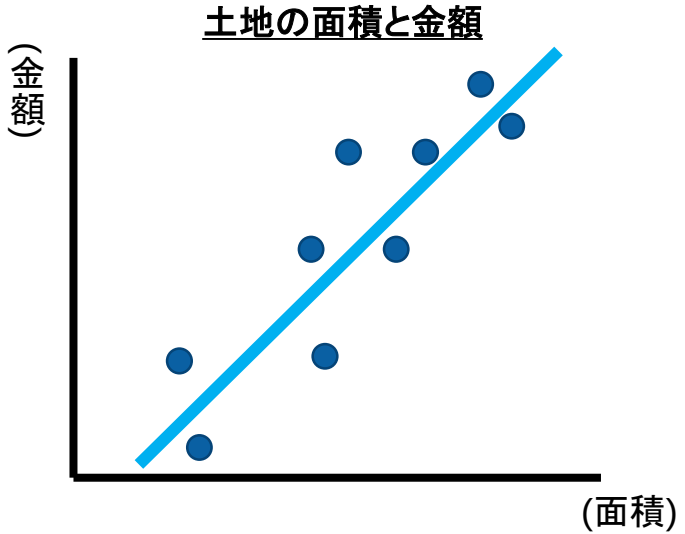
機械学習の種類

- 教師あり学習
  - 状態や結果がわかっているデータから予測を行う
- 教師なし学習
  - 漠然とデータのみ存在
    ⇒新たな知見の発見
- 強化学習
  - 試行を積み重ねてよいやり方、悪いやり方を学んでいく
    成功　⇒　報酬
    失敗　⇒　ペナルティ



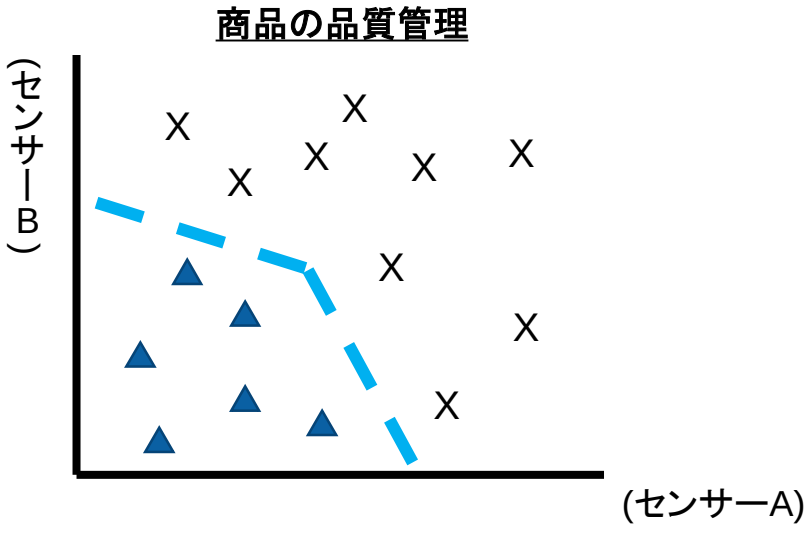
教師あり学習のイメージ

◆回帰系(数値をあてる)



(例)土地の面積　⇒　地価
　　天候・気温　⇒　ビールの売上

◆分類系(A or B)



(例) センサーデータ　⇒　正常 or 故障
　　購買履歴　⇒　買う or 買わない

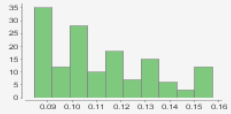
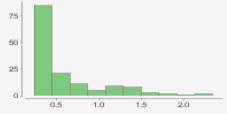
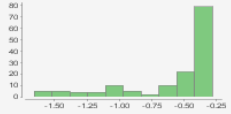
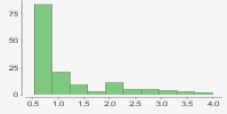
データ理解

ExampleSet (146 examples, 2 special attributes, 20 regular attributes)

Filter (146 / 146 examples): all

Class	3H-Mean	3H-Stdev	3H-Max	3H-Min	3H-Range	3V-Mean	3V-Stdev	3V-Max	3V-Min	3V-Range	4H-Mean	4H-Stdev	4H-Max	4H-Min	4H-Range
1	0.004	0.098	0.931	-1.281	2.212	-0.000	0.168	0.884	-0.900	1.783	0.002	0.116	0.639	-0.334	0.974
1	0.004	0.103	2.335	-1.658	3.993	-0.001	0.172	0.978	-1.591	2.569	0.000	0.115	0.550	-0.342	0.893
1	0.004	0.106	1.142	-1.106	2.249	-0.001	0.204	1.350	-1.283	2.634	0.004	0.116	0.611	-0.325	0.935
1	0.003	0.096	0.922	-1.226	2.148	-0.001	0.147	0.899	-1.058	1.957	0.003	0.116	0.648	-0.348	0.996
1	0.004	0.098	1.157	-1.009	2.165	-0.001	0.152	0.784	-1.084	1.868	-0.000	0.116	0.555	-0.349	0.904
1	0.004	0.100	0.815	-0.894	1.709	-0.000	0.180	1.009	-0.895	1.904	0.002	0.115	0.582	-0.335	0.917
1	0.005	0.098	1.458	-1.441	2.900	0.000	0.152	1.336	-1.394	2.731	0.002	0.117	0.657	-0.342	0.999
1	0.003	0.095	1.182	-0.987	2.169	0.001	0.139	0.842	-1.413	2.255	0.001	0.118	0.604	-0.333	0.936
1	0.003	0.105	1.446	-1.563	3.009	0.001	0.174	1.371	-1.298	2.670	0.002	0.116	0.591	-0.317	0.908
1	0.001	0.100	0.813	-0.845	1.657	0.000	0.174	0.902	-1.034	1.936	0.001	0.115	0.601	-0.322	0.923
1	0.001	0.101	1.097	-1.099	2.195	-0.001	0.175	1.043	-1.028	2.071	0.001	0.116	0.604	-0.300	0.905
1	0.002	0.105	1.855	-1.308	3.163	-0.000	0.197	1.258	-1.088	2.347	0.002	0.116	0.525	-0.314	0.840
2	0.002	0.089	0.621	-0.531	1.153	0.001	0.122	0.408	-0.440	0.848	0.004	0.117	0.591	-0.338	0.928
2	0.001	0.088	0.761	-0.707	1.468	0.001	0.120	0.392	-0.420	0.812	0.001	0.115	0.532	-0.366	0.898
2	-0.001	0.088	0.687	-0.534	1.221	0.001	0.118	0.588	-0.445	1.033	0.002	0.117	0.553	-0.342	0.895
2	-0.003	0.087	0.376	-0.360	0.736	0.001	0.125	0.440	-0.472	0.912	0.001	0.118	0.655	-0.350	1.005
2	-0.003	0.089	0.516	-0.469	0.985	0.001	0.127	0.460	-0.483	0.943	0.001	0.116	0.618	-0.320	0.938
2	-0.003	0.089	0.586	-0.542	1.128	0.001	0.126	0.397	-0.475	0.873	0.002	0.115	0.548	-0.321	0.869
2	-0.001	0.090	0.439	-0.455	0.895	0.000	0.127	0.440	-0.435	0.875	0.002	0.117	0.662	-0.310	0.972
2	0.001	0.088	0.522	-0.481	1.004	0.001	0.123	0.463	-0.420	0.883	0.003	0.117	0.591	-0.330	0.921
2	0.001	0.089	0.475	-0.472	0.947	0.000	0.131	0.468	-0.426	0.893	0.002	0.115	0.647	-0.314	0.961
2	0.002	0.089	0.716	-0.635	1.351	0.001	0.129	0.439	-0.559	0.999	0.001	0.115	0.596	-0.329	0.925
2	0.003	0.090	0.670	-0.622	1.292	0.002	0.128	0.429	-0.456	0.885	0.002	0.117	0.593	-0.320	0.912

データ理解

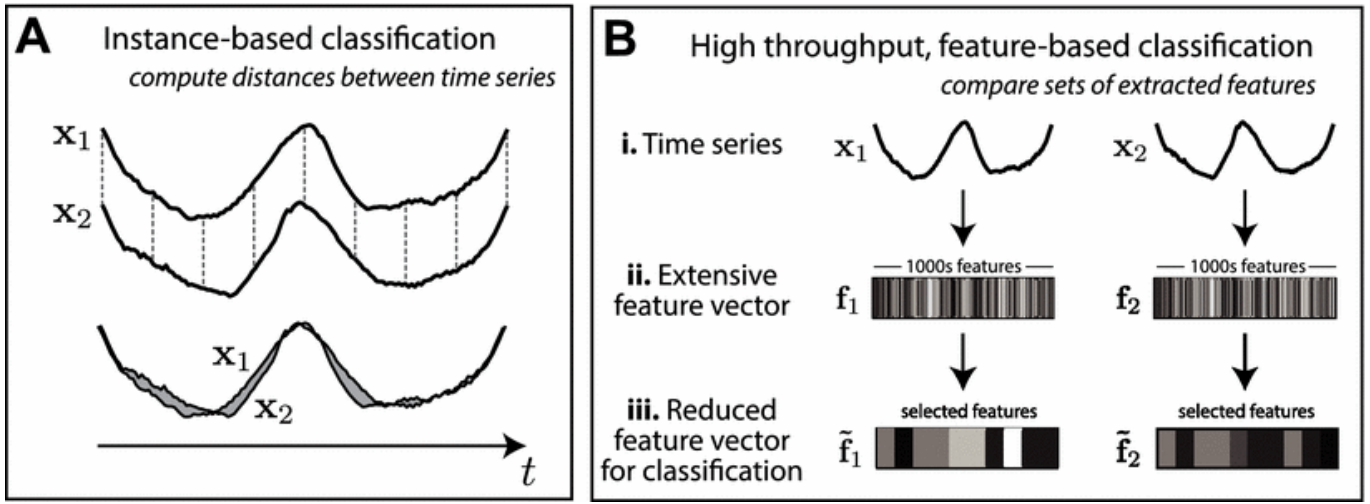
Name	Type	Missing	Statistics				
3H-Mean	Real	0	 Open chart	Min -0.005	Max 0.009	Average 0.002	Deviation 0.003
3H-Stdev	Real	0	 Open chart	Min 0.084	Max 0.157	Average 0.112	Deviation 0.021
3H-Max	Real	0	 Open chart	Min 0.243	Max 2.335	Average 0.602	Deviation 0.456
3H-Min	Real	0	 Open chart	Min -1.659	Max -0.278	Average -0.588	Deviation 0.374
3H-Range	Real	0	 Open chart	Min 0.533	Max 3.993	Average 1.190	Deviation 0.817

参考：時系列データからの特徴量抽出

A. Instance based classification - 時系列データを直接比較し、その距離（乖離）を元に分類をする。

B. Feature based classification - 時系列データから特徴量を抽出、選択し、

その特徴量の距離（乖離）を元に分類をする。



参考：B. D. Fulcher, N. S. Jones, Highly Comparative Feature-Based Time-Series Classification, arXiv:1401.3531v2

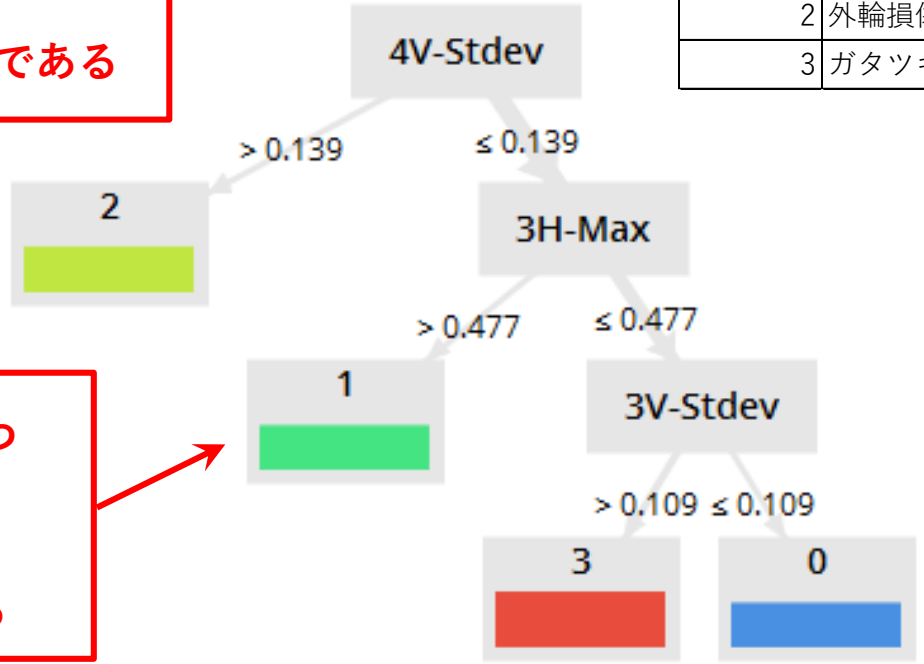
Christ Maximilian, Kempa-Liehr Andreas W., Feindt Michael, Distributed and parallel time series feature extraction for industrial big data applications,arXiv:1610.07717v3

決定木(Decision Tree)によるクラス分類

class	名称
0	正常状態
1	外輪損傷３
2	外輪損傷４
3	ガタツキ

変数 4V-Stdevが0.139超

なら状態“2(外輪損傷4)”である

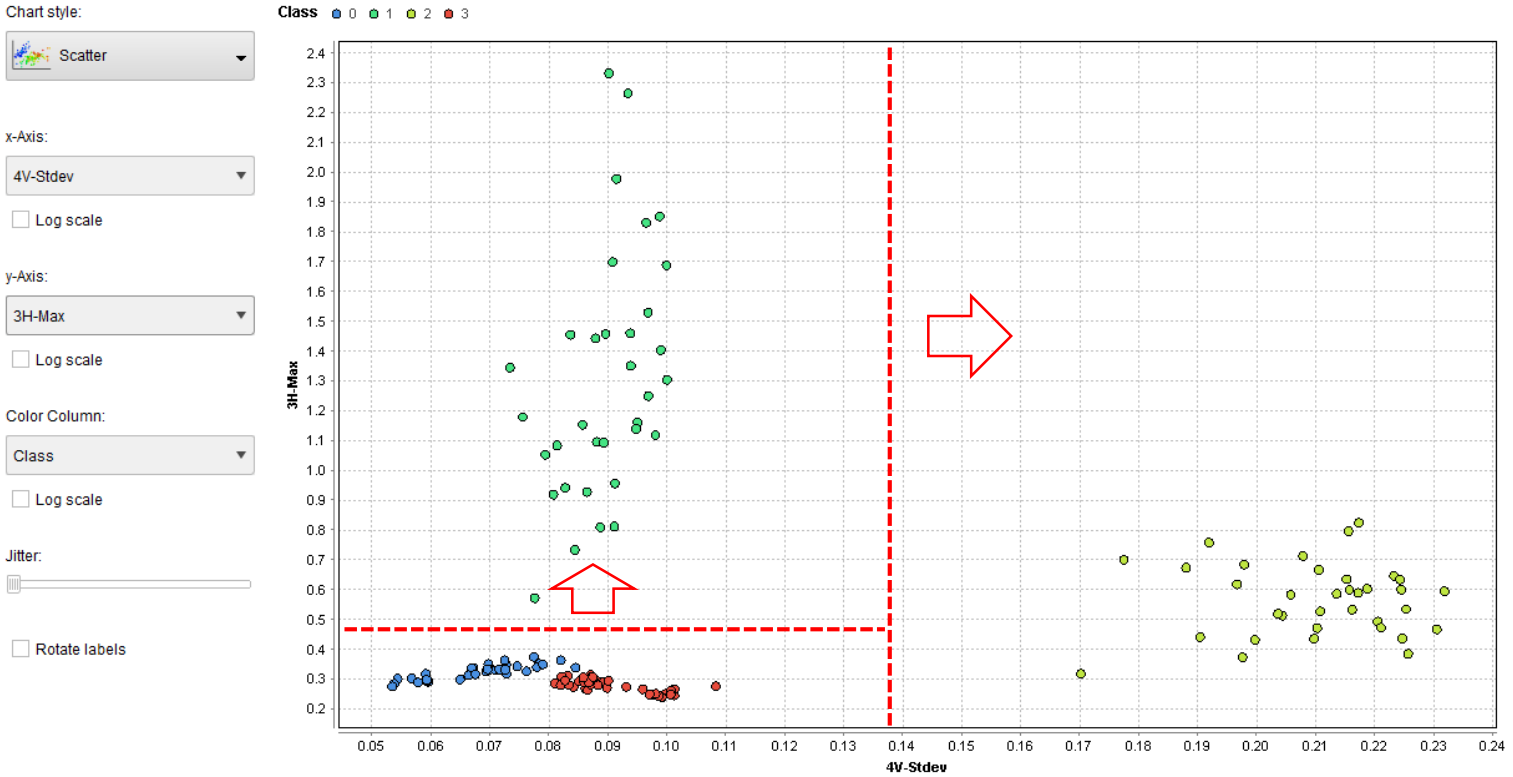


変数 4V-Stdevが0.139以下かつ

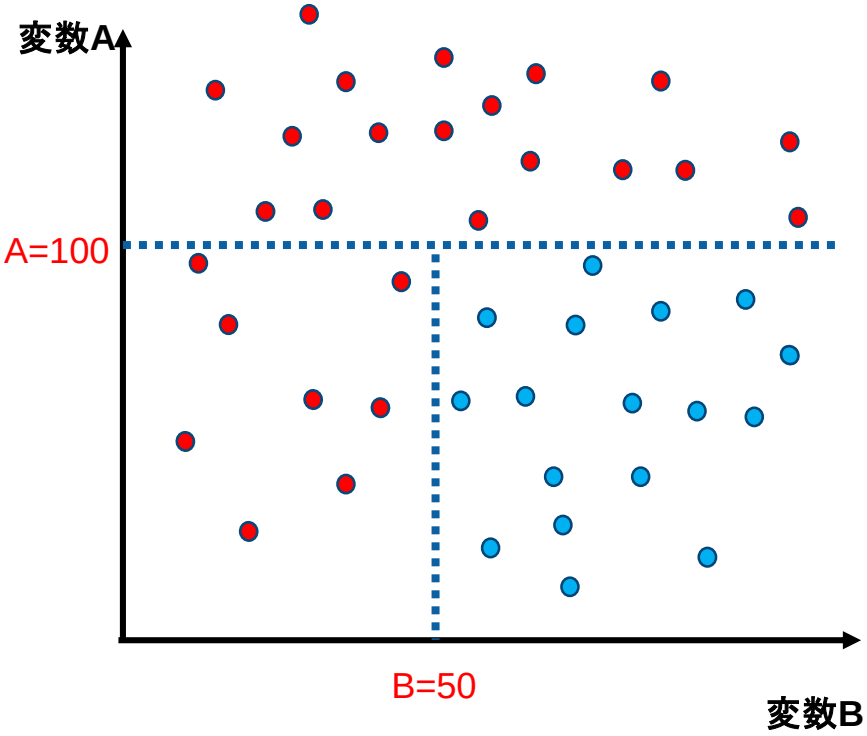
変数3H-Maxが0.477超

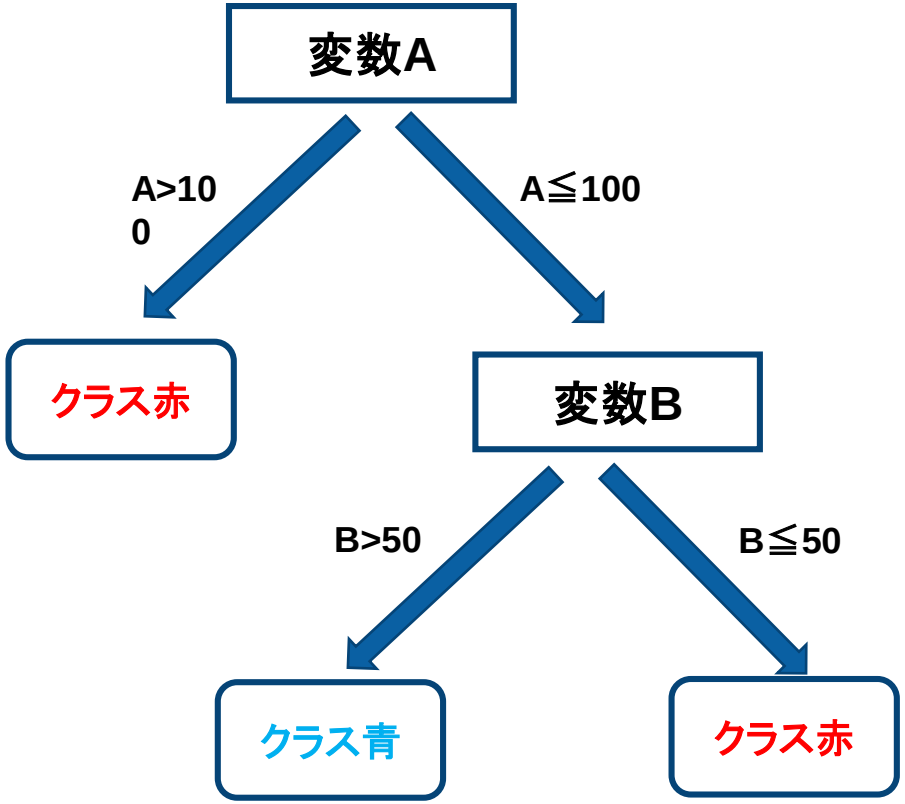
なら状態“1(外輪損傷3)”である

変数“4V-Stdev”と変数”3H-Max”のデータ分布をプロットすると



決定木(Decision Tree) 分割条件（閾値）計算の考え方





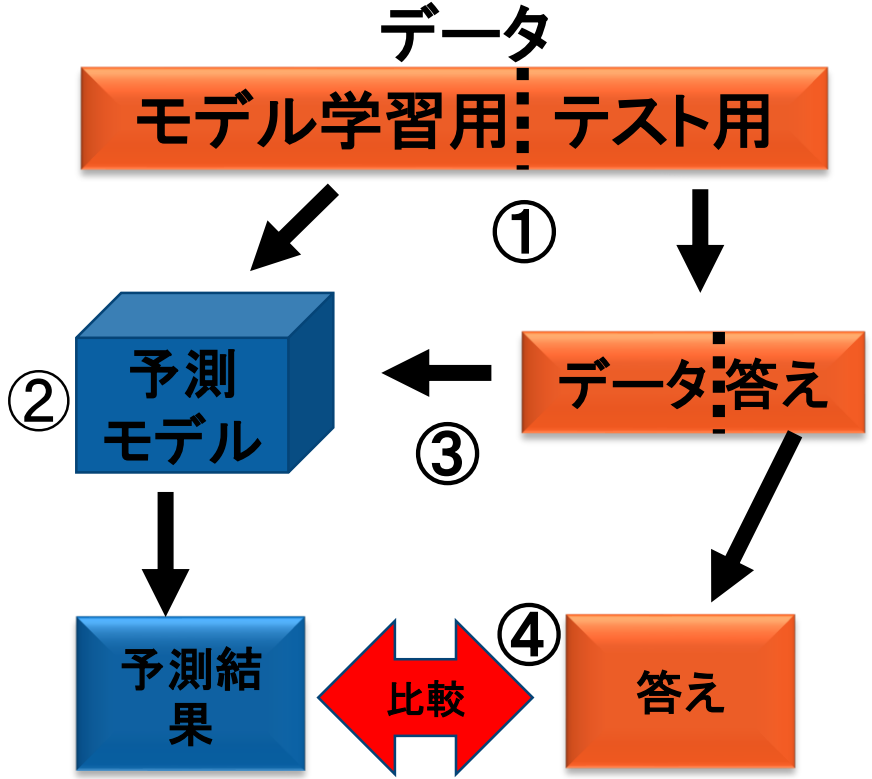
スプリッドバリデーション　～予測モデルの精度確認手法～

①データをモデル学習用・テスト用に分割

②モデル学習用データで予測
　モデルを構築

③テスト用データの答えを隠して、
　作成したモデルで予測

④「予測結果」と「テスト用データの答え」
　を比較して精度確認



予測精度検証：コンフュージョンマトリックス

- 決定木による予測モデルの検証結果

accuracy: 96.57% +/- 3.43% (mikro: 96.58%)

	true 0	true 1	true 2	true 3	class precision
pred. 0	33	0	0	0	100.00%
pred. 1	0	33	0	0	100.00%
pred. 2	1	1	33	1	91.67%
pred. 3	0	0	2	42	95.45%
class recall	97.06%	97.06%	94.29%	97.67%	

Appendix
～モデルの精度向上～

- 学習データ量を増やす
- 特徴量設計
- 複雑なモデルを用いる
  （アンサンブル学習など）

欠損値の補完

欠損があるデータ。欠損値が発生メカニズムには以下３パターンが原因として考えられる

MCAR
(Missing Completely At Random)

いわゆる欠損値が完全にランダムに生じているようなケース
（欠損値の有無が、その他の変数や欠損値のある変数自体の値とは無関係）

MAR
(Missing At Random)

欠損値の有無は、ほかの変数の値と関係しているが、その変数の値とは無関係であるケース

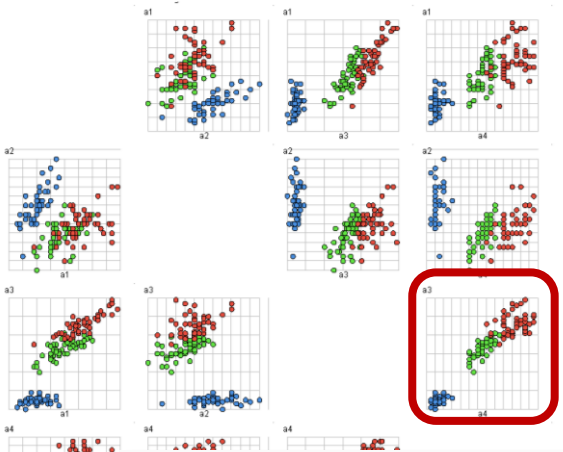
例：身体測定で体重が欠損しているが、女性に多く欠損が見られる場合など

MNAR
(Missing Not At Random)

分析に含まれる他の変数を統制した後でも、欠損値の有無が欠損値を持つ変数自身と関係を持つケース

例：IQテストの回答欄に空白が多く、IQテスト結果と欠損発生頻度で相関がある場合など

多重共線性の排除(Multicollinearity、通称マルチコ)



Correlation Matrix (Correlation Matrix)

Attributes	a1	a2	a3	a4
a1	1	-0.109	0.872	0.818
a2	-0.109	1	-0.421	-0.357
a3	0.872	-0.421	1	0.963
a4	0.818	-0.357	0.963	1

- **多重共線性：**
  相関性が高い2つ以上の説明変数が存在する場合に起こる問題

  -回帰の係数などモデルが解釈困難となる
  （本来の意味合いと異なる方向へと係数の符号がつく等）

  -モデルの予測結果が不安定になる
  （極端な予測値がでる場合がある）

  -予測精度が低下する場合がある

- **例：身長と座高、雨天日数と総雨量**

不均衡データへの対応

・各labelサンプルデータ数が偏りがある状態

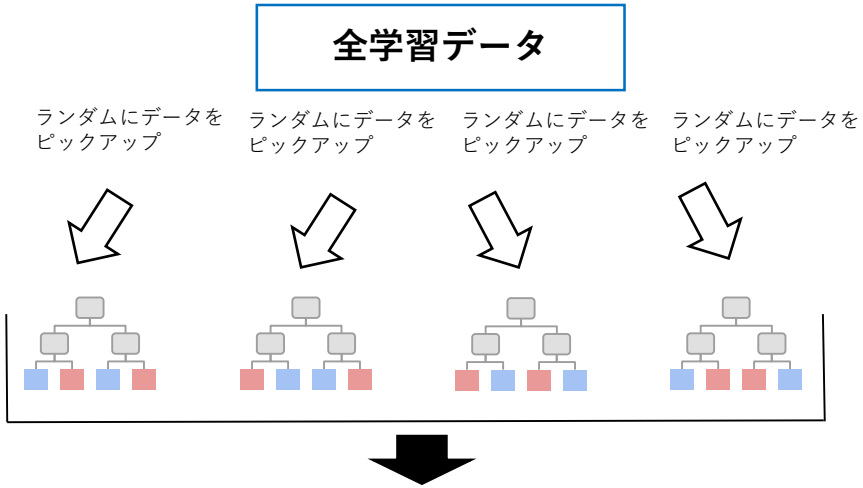
例：機械の故障予測問題では、正常稼働データは大量に取得できるが
故障はめったに発生せず故障データは非常に少数

・下図の場合、すべて「正常」と予測するモデルの判別精度(Accuracy)は99%となる
（このような判別モデルは適切か？）

アンサンブル学習の例：決定木の発展系アルゴリズム

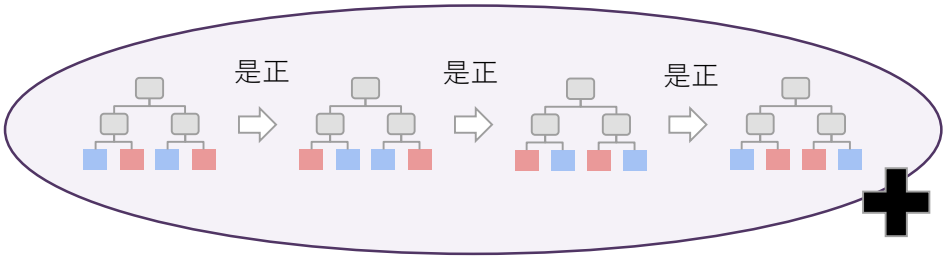
・複数の決定木を合成（アンサンブル）する手法が考案されている
・決定木メリットである幅広いデータ分布の型にも対応しながら、”高い予測精度(コンテストでよく利用)”
・デメリットとしては、複数のTreeを合成することで予測の根拠が”ブラックボックス化”

・Random Forest



・Gradient Boosted Trees

（XGBoost ：eXtreme Gradient BoostingはGBTの実装例）



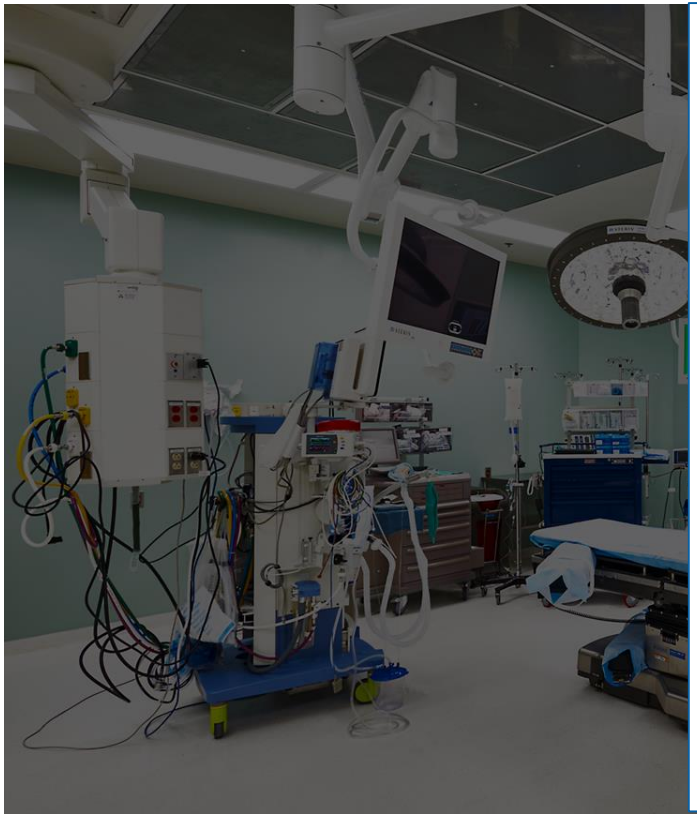
①決定木モデルを作る

②次のモデルはそれまでの誤差を少しだけ是正する

③最後に全部のモデル出力を合算する

KSK ANALYTICS

分析事例

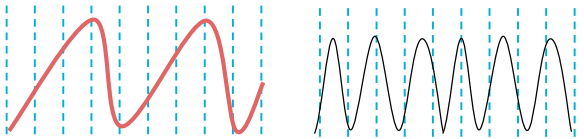


Case:
センサーデータによる検査機器の故障予測

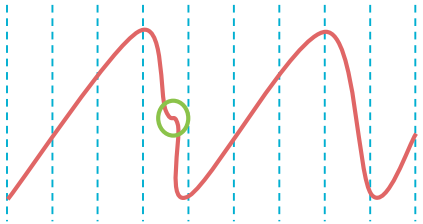
機器に設置した圧力計のセンサーデータを利用し、異常・故障を100%の精度で検知したい。
(1説明変数-波形分析)

対象データセット

■対象データセット



正常時の波形



異常/故障時の波形

「データの特徴」

-圧力計データ(1説明変数)

-時系列データ(波形データ)

ただし周波数は設定により変動

(FFT変換の手法は使用不可)

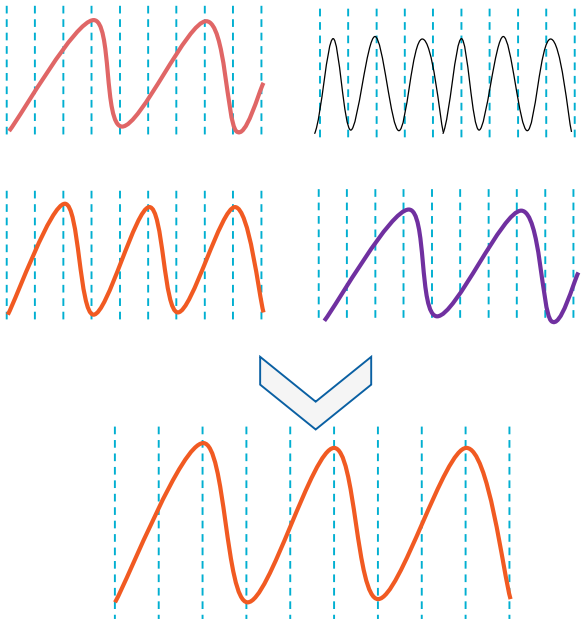
-故障(異常)時は波形に”ぶれ”発生

なお正常稼働データは大量にあるが

異常時データは極少(不均衡データ)

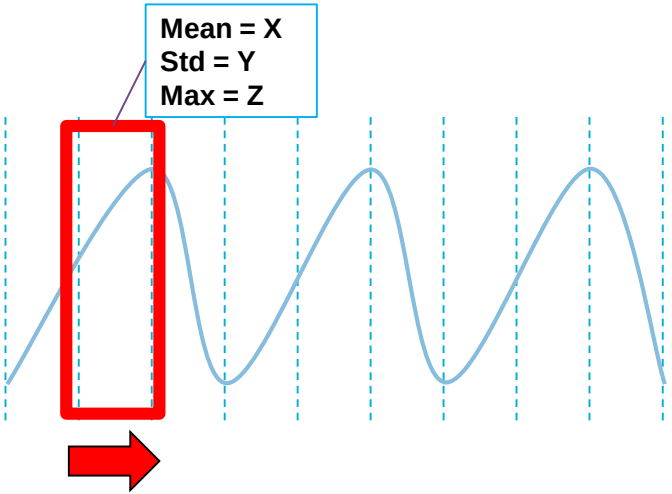
データ分析テクニック

■センサパターンをどう統一するか？



・機器モーター出力より**周期を標準化**

■波形の特徴量をどう抽出するか？



・波形毎にSliding Windowにより
要約統計量を算出し、特徴量とした
・サポートベクターマシンアルゴリズム
により高速処理かつ**予測精度100％**達成

機械学習ソフトRapidMinerご紹介

1万行まで無料
で使える！！

RapidMiner Free版 ダウンロード申請サイト
http://www.rapidminer.jp/download

RapidMinerブログ
http://www.rapidminer.jp/blog

KSKアナリティクス　イベント情報
（ご紹介セミナー・トレーニング　東京・大阪で定期開催）
http://www.ksk-anl.com/event

データ分析についてお気軽にご相談下さい

KSKアナリティクス　ビジネス推進部
コーポレートサイト: http://www.ksk-anl.com/
製品サイト: https://www.rapidminer.jp/
メールアドレス: sales@ksk-anl.com
東京オフィス:03-6228-4932　大阪本社オフィス:06-6131-6656

KSK ANALYTICS