機械学習　実践編

2024年12月4日　ARIMアカデミー　データ構造化ワークショップ（２）

物質・材料研究機構
マテリアル先端リサーチインフラセンターハブ

松波　成行

ARIM Japan

１．機械学習

機械学習の流れ

①探索的データ分析（EDA）	②データ可視化	③機械学習
概要統計量の算出 ペアプロットの作図 相関係数の計算	頻度分布 単回帰	線形重回帰 決定木 非線形回帰

ライブラリ
pandas
matplotlib
seaborn

ライブラリ
scikit-learn
matplotlib

ライブラリ
scikit-learn
matplotlib

ARIM Japan

機械学習の手順（予測）

① データ行列

説明変数（X, Xb, Xc）と目的変数(Y)の設定

Xa	Xb	Xc	Y
3.297	4.356	0.03129	2
4.267	4.118	0.03129	2
4.088	4.763	0.03337	5
4.338	4.556	0.03337	5
4.732	5.138	0.03551	10
4.9	4.941	0.03551	10

② データ行列の分割

学習データとテストデータへの分割

Xa	Xb	Xc	Y
3.297	4.356	0.03129	2
4.267	4.118	0.03129	2
4.088	4.763	0.03337	5
4.338	4.556	0.03337	5
4.732	5.138	0.03551	10
4.9	4.941	0.03551	10

③ 学習データでMLモデル構築

Xa	Xb	Xc	Y
3.297	4.356	0.03129	2
4.267	4.118	0.03129	2
4.088	4.763	0.03337	5
4.338	4.556	0.03337	5

MLモデル

Y = f(Xa, Xb, Xc)

④ モデル評価

予測値 Y'		実測値 Y
2		2
2	vs	2
5		5
4		5

⑤ テストデータによる予測

Xa	Xb	Xc
4.732	5.138	0.03551
4.9	4.941	0.03551

⑥ モデル評価

予測値 Y'		実測値 Y
10	vs	10
10		10

学習データ（Training Data）

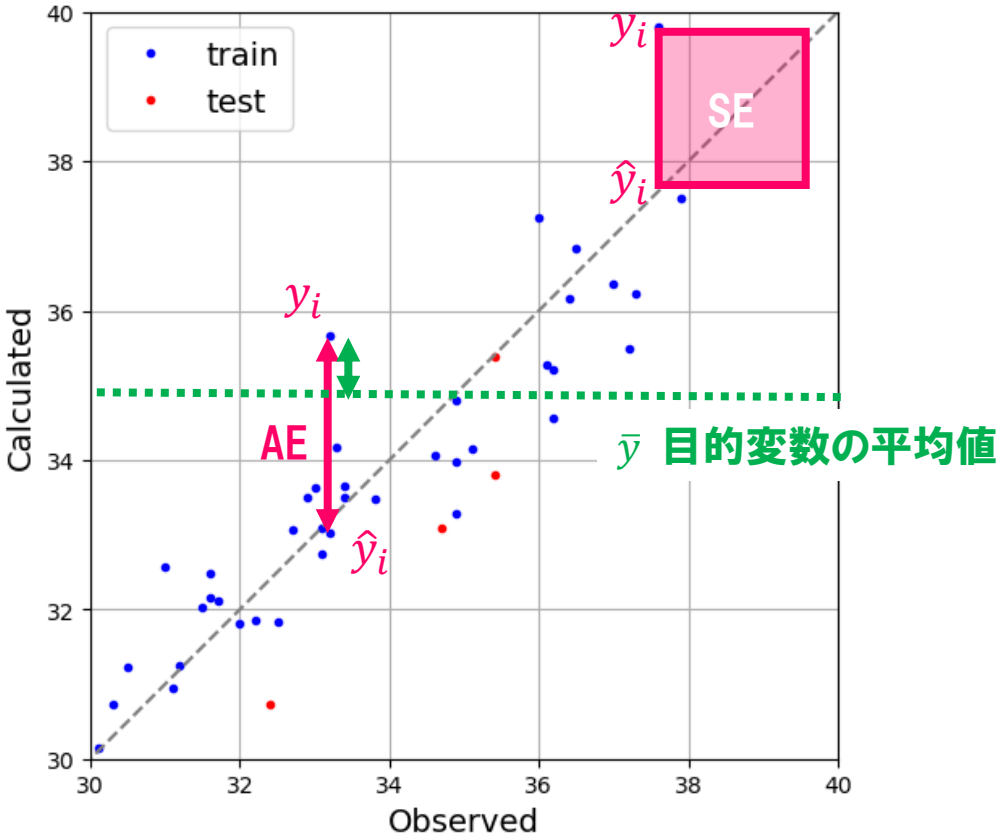
テストデータ（Test Data）

⑦ モデル決定（選定）　④ モデル評価 ≈ ⑥ モデル評価

ARIM Japan

モデルの評価指標(メトリクス)について

✓ モデルは、評価関数(損失関数、目的関数)=残差変動(SST)が最小となるようにFitting係数(パラメータ)が最適化される。



$$MAE = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i|$$

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2$$

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{N}(y_i - \bar{y})^2}$$

残差変動(SSE)

全変動(SST)

残差平方値$(y_i - \hat{y}_i)^2$は外れ値や異常値に引きずられやすい
→ ばらつきの多い系においてはMAEで確認する

ARIM Japan

Scikit-learnでのポイント

```
# モデルライブラリ(クラス)の読み込み
from sklearn.linear_model import LinearRegression

#予測器モデルの設定 （インスタンス化）
model = LinearRegression()

#予測器の構築
model.fit(X_train, y_train)
```

機械学習モデルを構築するのはたった3行

ARIM Japan

サンプルコード

本演習でのデータセットであるBoston Housingデータセットには、住宅価格の中央値（MEDV）をはじめ、犯罪率、部屋数、地域ごとの特性など、多岐にわたる特徴量が含まれています。

- **さまざまな回帰モデルの構築と比較**
  線形回帰、ランダムフォレストなど、代表的な回帰モデルを構築し、それぞれのモデルの予測性能を比較します。各モデルの特性を理解し、どのモデルがボストン市の住宅価格を最も正確に予測できるかを実験的に検証します。

- **モデル評価**
  構築したモデルの性能を評価するために、平均二乗誤差（MSE）や決定係数（R2スコア）などの評価指標を用います。これらの指標に基づいて、モデルの精度を定量的に評価し、より良いモデルへと改善するための指針を得ます。

- **特徴量の重要度分析**
  モデルの学習過程で、各特徴量が予測にどの程度貢献しているのかを分析します。これにより、住宅価格に最も影響を与える要因を特定し、モデルの解釈性を高めることができます。また、特徴量選択の際に役立つ情報を得ることも可能です。

【入門編の続き】
https://colab.research.google.com/github/ARIM-Academy/Advanced_Tutorial_1/blob/main/Scikit-learn-0.ipynb

ARIM Japan

２．機械学習の実践的利用例

NEDO IoTを活用した新産業モデル創出基盤整備事業

外面腐食データと気象データを統合した大気腐食モデルの立案と検証

要約

気象モニタリングデータと腐食データから**機械学習による大気腐食モデルの構築**および**GISによる腐食環境の1kmメッシュ可視化**

土木学会論文集A1(構造・地震工学), Vol. 75, No. 2, 141-160, 2019.

海塩輸送シミュレーションと気象情報を用いた
機械学習に基づく大気腐食量評価モデル開発と
高精細腐食環境地図の作成

松波　成行[1]・柳生　進二郎[2]・篠原　正[3]・片山　英樹[4]・
須藤　仁[5]・服部　康男[6]・平口　博丸[7]

[1]非会員　(国研)物質・材料研究機構統合型材料開発・情報基盤部門
(〒305-0047 茨城県つくば市千現1-2-1)
E-mail: MATSUNAMI.Shigeyuki@nims.go.jp

[2]非会員　(国研)物質・材料研究機構国際ナノアーキテクトニクス研究拠点（同上）
E-mail: YAGYU.Shinjiro@nims.go.jp

[3]非会員　(国研)物質・材料研究機構構造材料研究拠点（同上）
E-mail: SHINOHARA.Tadashi@nims.go.jp

[4]正会員　(国研)物質・材料研究機構構造材料研究拠点（同上）
E-mail: KATAYAMA.Hideki@nims.go.jp

[5]非会員　(一財)電力中央研究所地球工学研究所（〒270-1194 千葉県我孫子市我孫子1646）
E-mail: suto@criepi.denken.or.jp

[6]非会員　(一財)電力中央研究所地球工学研究所（同上）
E-mail: yhattori@criepi.denken.or.jp

[7]正会員　(一財)電力中央研究所地球工学研究所（同上）
E-mail: hiromaru@criepi.denken.or.jp

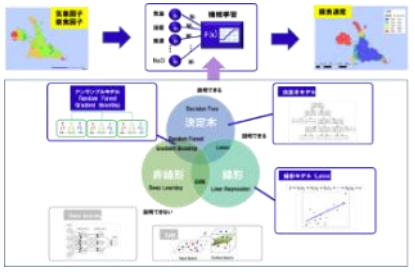
土木学会論文集A1,vol. 75, p141-160, 2019
https://doi.org/10.2208/jscejseee.75.141

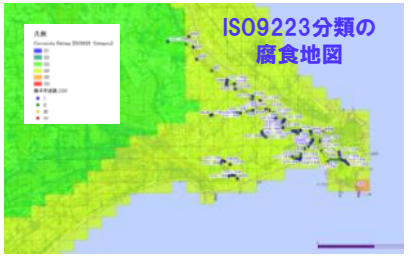
成果概略

手　法　Random Forest（腐食速度 ← 気象9因子）

成　果　機械学習による1kmメッシュ腐食環境地図

① ISO9223分類可視化(国際標準化に対応した腐食環境図)
② 月単位の腐食の特徴量:「最高気温」、「風速」、および「NaCl」





期待効果

IoT気象モニタリングデータと橋梁腐食モニタリングのデータを増やすことで、インフラの予防保全にかかるピンポイント地点の経年劣化が推定可能に。

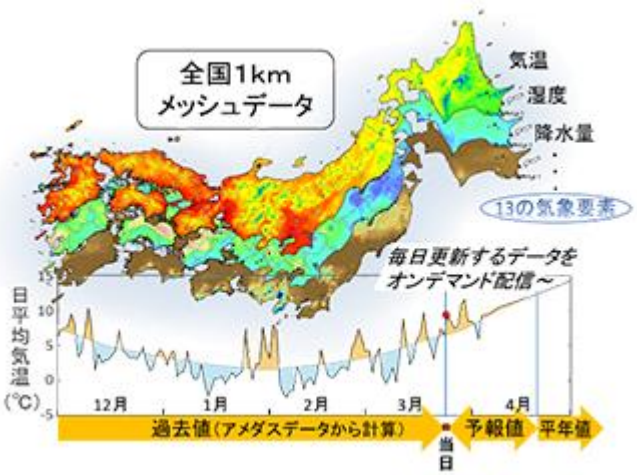
より確度の高い予防保全管理の優先順位指標の提供

農研機構のメッシュ気象データ（気象解析値）

■ 農研機構が提供する「メッシュ農業気象データシステムから、過去気象データについて下記の日別の気象要素を1kmメッシュ単位で取得できる。

表１．システムが作成・配信する農業気象データの一覧

気象要素	単位	過去値	予報値	平年値
日平均気温	℃	1980年1月1日～前日	当日～26日先	2011年～2020年
日最高気温	℃	1980年1月1日～前日	当日～26日先	2011年～2020年
日最低気温	℃	1980年1月1日～前日	当日～26日先	2011年～2020年
降水量	mm/day	1980年1月1日～前日	当日～26日先	2011年～2020年
日照時間	h/day	1980年1月1日～前日	なし	2011年～2020年
全天日射量	J/m²/day	1980年1月1日～前日	なし	2011年～2020年
下向き長波放射量	J/m²/day	2008年1月1日～前日	なし	なし
日平均相対湿度	%	2008年1月1日～前日	当日～9日先	なし
日平均風速	m/s	2008年1月1日～前日	当日～9日先	なし
積雪深	cm	2008年1月1日～前日	なし	なし
積雪相当水量	mm	2008年1月1日～前日	なし	なし
日降雪相当水量	mm/day	2008年1月1日～前日	なし	なし
予報気温の確からしさ*	℃	2011年1月1日～前日	当日～26日先	なし



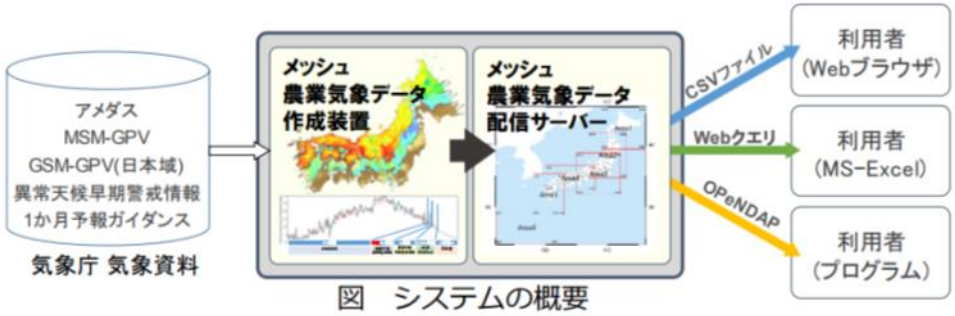
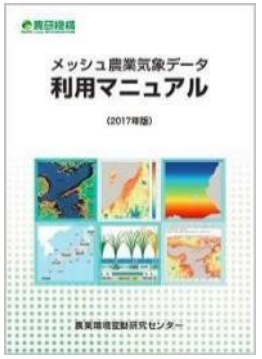


図　システムの概要



全国の気象観測値からデータ同化され1kmメッシュ化された本データセットを用いることができれば、迅速に腐食マップが作成可能

ARIM Japan

腐食マップのデータ構造

■ 説明変数： アメダス（地域気象観測システム）で観測可能な気象データとした

サンプル（観測地）	月	説明変数（1kmメッシュの月単位の気象データ）				目的変数
		平均気温	平均最高気温	・・・	平均風速の2乗（NaCl相当）	腐食速度（月単位）
つくば	1					
	2					
	3					
	・ ・					
	12					
宮古島	1					
	2					
	・ ・					
	12					

6地点×12か月＝72

9変数（※）

※製造などのプロセスインフォマティクスでは9つの製造条件と考えてください

ARIM Japan

実データを使ったサンプルコード

この演習では、鉄の大気腐食量を気象データから予測するための『大気腐食データセット』を活用し、予測モデルにかかる機械学習の基礎を習得します。

- **予測アルゴリズムの習得:**
  このデータセットは、6か所で月次に測定された腐食量データを含んでいます。これを用いて、線形回帰やランダムフォレストなどの予測アルゴリズムを学び、腐食量の予測を実践します。

- **特徴選択と次元削減の理解:**
  気象因子を特徴量とするこのデータセットを分析し、適切な特徴量の選択や次元削減を行うことで、モデルのパフォーマンスを向上させる方法を理解します。

- **モデル評価とパフォーマンス指標の理解:**
  データセットを使用してトレーニングしたモデルを評価し、R^2などのパフォーマンス指標を用いて予測精度を評価するスキルを習得します。

【予測モデル】
https://colab.research.google.com/github/ARIM-Academy/Advanced_Tutorial_1/blob/main/Scikit-learn-2.ipynb

ARIM Japan

3.　分類モデル

機械学習の手順（分類）

① データ行列

説明変数（X, Xb, Xc）と目的変数(Y)の設定

Xa	Xb	Xc	Y
3.297	4.356	0.03129	A
4.267	4.118	0.03129	A
4.088	4.763	0.03337	B
4.338	4.556	0.03337	B
4.732	5.138	0.03551	A
4.9	4.941	0.03551	B

② データ行列の分割

学習データとテストデータへの分割

Xa	Xb	Xc	Y
3.297	4.356	0.03129	A
4.267	4.118	0.03129	A
4.088	4.763	0.03337	B
4.338	4.556	0.03337	B
4.732	5.138	0.03551	A
4.9	4.941	0.03551	B

学習データ（Training Data）

テストデータ（Test Data）

③ 学習データでMLモデル構築

Xa	Xb	Xc	Y
3.297	4.356	0.03129	A
4.267	4.118	0.03129	A
4.088	4.763	0.03337	B
4.338	4.556	0.03337	B

⑤ テストデータによる予測

Xa	Xb	Xc
4.732	5.138	0.03551
4.9	4.941	0.03551

MLモデル

Y = f(Xa, Xb, Xc)

④ モデル評価

予測値 Y'		実測値 Y
A		A
A	VS	A
B		B
A		B

⑥ モデル評価

予測値 Y'		実測値 Y
A	VS	A
A		B

⑦ モデル決定（選定）　**④ モデル評価** ≈ **⑥ モデル評価**

ARIM Japan

モデルの評価指標： 混同行列(Confusion Matrix)について

✓ モデル評価は二値の場合、2行2列の表で、真陽性、真陰性、偽陽性、偽陰性からなる混同行列(Confusion Matrix)で集計・評価される。

実測クラス	予測クラス: Positive	予測クラス: Negative
Positive	**真陽性** (True Positive)	**偽陰性** (False Negative)
Negative	**偽陽性** (False Positive)	**真陰性** (True Negative)

○再現率・感度

$$recall(sensitivisty) = \frac{TP}{TP + FN}$$

○誤検知率(偽陽性率)

$$\mathrm{FPR} = \frac{FP}{FP + TN}$$

○適合率(陽性的中率)

$$precision = \frac{TP}{TP + FP}$$

○陰性的中率

$$NPV = \frac{TN}{FN + TN}$$

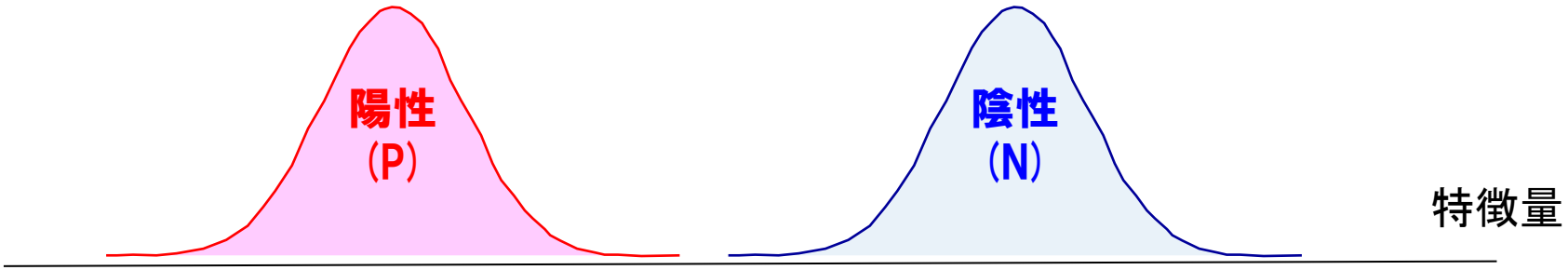
○正解率

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

ARIM Japan

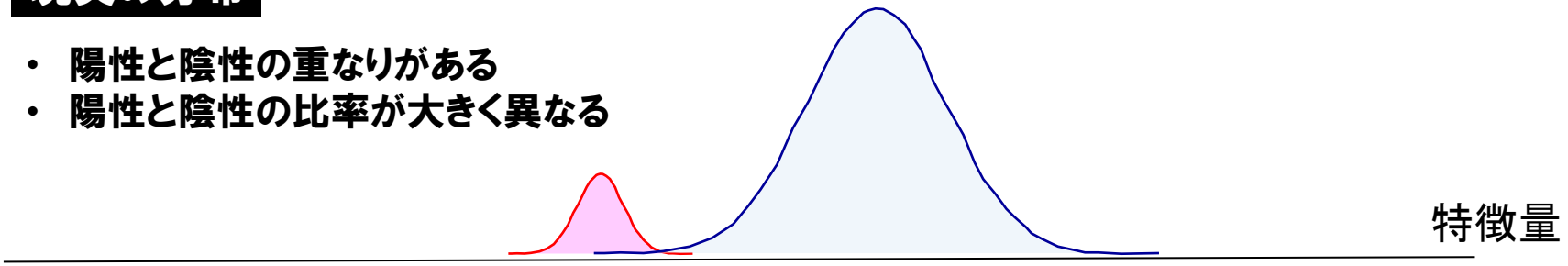
分類モデルにおける指標の解釈のイメージ（１変数の場合）

理想の分布



現実の分布

- 陽性と陰性の重なりがある
- 陽性と陰性の比率が大きく異なる



誤検知重視

機器メーカの視点
（医療製造側）

適合率重視

機器利用の視点
（医者側）

Scikit-learnでのポイント

```
# モデルライブラリ(クラス)の読み込み （ここでは決定木）
from sklearn.tree import DecisionTreeClassifier

#予測器モデルの設定 （インスタンス化）
model = DecisionTreeClassifier()

#予測器の構築
model.fit(X_train, y_train)
```

分類モデルも予測モデルと構文は同じです。

ARIM Japan

分類の機械学習モデルの練習 （お茶の元素分析データセット）

概要

『茶の元素分析データセット』は、ブラックセイロン (BC) 、ブラックトルコ (BT) 、グリーンセイロン (GC) 、グリーントルコ (GT) の4種類の茶葉について、3つの濃度(1%、2%、3%)で抽出した液体の元素分析データです。誘導結合プラズマ発光分光分析法(ICP-OES)で分析された元素は、Al, Ca, Cd, Cr, Cu, Hg, Fe, K, Mg, Mn, Na, Pb, Znです。[1]

[1] Durmus, Y., Atasoy, A.D. & Atasoy, A.F. Mathematical optimization of multilinear and artificial neural network regressions for mineral composition of different tea types infusions. Sci Rep 14, 18285 (2024). https://doi.org/10.1038/s41598-024-69149-1

金谷茶と旧東海道

誘導結合プラズマ発光分析装置(NM-203)

https://nanonet.mext.go.jp/facility.php?mode=detail&code=25

ARIM Japan

お茶のミネラル抽出のプロセスデータ（データ行列）

Sci Rep 14, 18285 (2024).
https://doi.org/10.1038/s41598-024-69149-1

ICP-OES観測値											
Al	Ca	Cu	Fe	K	Mg	Mn	Na	Zn	tea	Concentration	time
3.297	4.356	0.03129	0.067	99.06	3.531	1.455	0.541	0.131	BT	1	2
4.267	4.118	0.03129	0.079	106.5	3.378	1.542	0.603	0.126	BT	1	2
4.088	4.763	0.03337	0.084	114	4.763	1.838	1.058	0.156	BT	1	5
4.338	4.556	0.03337	0.091	122.6	5.005	2.269	0.958	0.162	BT	1	5
4.732	5.138	0.03551	0.11	132.4	5.626	2.998	1.51	0.165	BT	1	10
4.9	4.941	0.03551	0.101	144.4	6.114	3.172	1.525	0.172	BT	1	10
4.848	5.357	0.03382	0.122	151.2	7.455	3.814	1.582	0.187	BT	1	20
5.023	5.268	0.03382	0.132	163.4	7.79	3.986	1.568	0.188	BT	1	20
4.769	5.368	0.02792	0.136	179	9.158	4.269	1.594	0.206	BT	1	30
5.193	5.357	0.02792	0.148	169.2	8.987	4.429	1.638	0.196	BT	1	30
4.732	5.417	0.02753	0.149	181.9	9.953	4.989	1.741	0.222	BT	1	45
5.047	5.5795	0.02753	0.157	183.3	9.852	5.032	1.708	0.211	BT	1	45
5.061	5.4611	0.02693	0.165	186.1	11.09	5.351	1.974	0.227	BT	1	60
4.828	5.794	0.02693	0.173	190.3	10.898	5.489	1.898	0.244	BT	1	60
7.561	5.307	0.064	0.093	172.7	7.293	3.369	1.501	0.138	BT	2	2
8.128	5.2	0.057	0.103	186.3	6.712	3.981	1.583	0.136	BT	2	2
9.169	5.597	0.072	0.137	214.7	9.829	4.546	1.889	0.175	BT	2	5
8.989	5.364	0.062	0.11	228.3	10.123	4.832	1.933	0.194	BT	2	5
8.751	5.943	0.071	0.163	238.7	13.91	5.947	2.373	0.217	BT	2	10
9.663	6.015	0.072	0.155	246.5	14.21	5.736	2.249	0.219	BT	2	10

実データを使ったサンプルコード

本演習では『**茶の元素分析データセット**』を用いて教師なし機械学習の分類技術を学びます。

- **データの可視化と解釈**:
  散布図やペアプロットを用いて、特徴量の分布や相関関係を視覚的に把握します。

- **分類アルゴリズムの実装**:
  元素分析値からお茶の銘柄を分類することに挑戦します。決定木、ランダムフォレスト、サポートベクターマシーンらの分類アルゴリズムを実装し、それぞれの特徴を理解します。

- **モデル評価** :
  混同行列を用いてモデルの予測結果と正解ラベルを比較し、性能を評価する方法を習得します。

【分類モデル】
https://colab.research.google.com/github/ARIM-Academy/Advanced_Tutorial_1/blob/main/Scikit-learn-3.ipynb

ARIM Japan

4. 次元削減

機械学習における次元削減の活用方法

高次元空間から低次元空間へデータを変換しながら、低次元表現が元データの何らかの意味ある特性を保持すること

① 特徴選択（Feature Selection）:
高次元の特徴空間から重要な特徴を選択し、それに基づいてモデルを構築することで、計算効率の向上やモデルの性能が向上。

② 特徴抽出（Feature Extraction）:
高次元のデータを低次元の新たな特徴空間に射影することで、データの表現を簡素化し、計算効率やモデルの性能が向上。

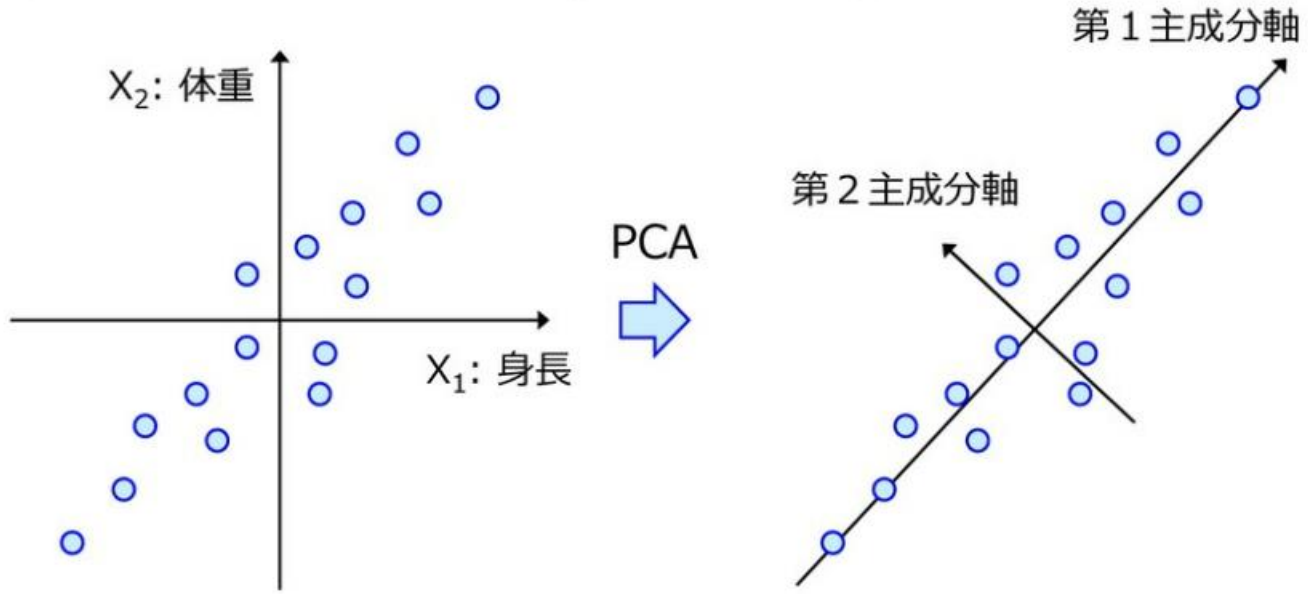
③ データ可視化（Visualization）
データを低次元に射影することで、データのクラスタリングや分布の構造を可視化し、洞察を得ることが可能。

主成分分析（PCA）

https://datachemeng.com/principalcomponentanalysis/

PCAの図解 2

例）１５人の身長・体重データ (多次元のデータ)



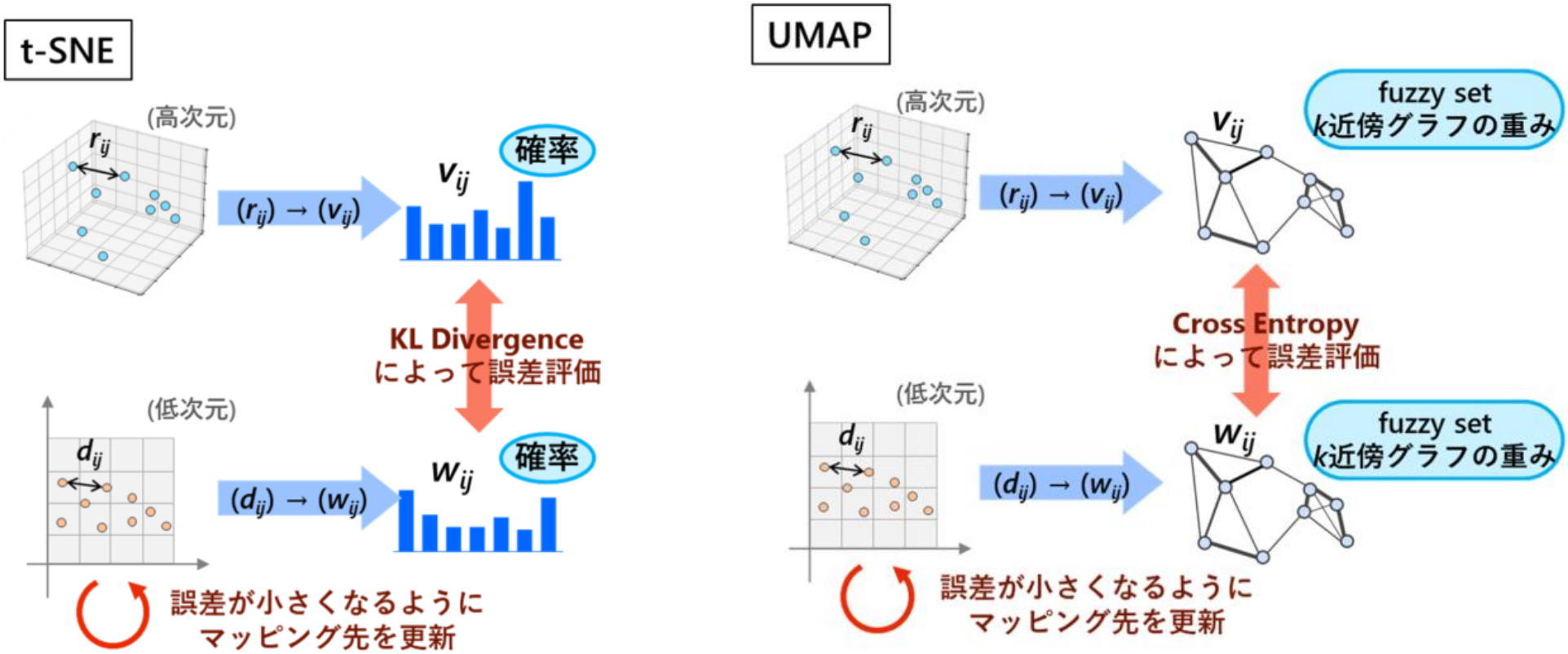
第１主成分だけでも、１５人のだいたいの情報はおさえられる

代表的な次元削減の手法

手　　法		定　　義
PCA （principal component analysis）	線形	元の多次元空間データの情報をできるだけ保持しつつ、データの変動を最大限に捉える主成分空間への射影する手法。
t-SNE （t-Distributed Stochastic Neighbor Embedding）	非線形	高次元データを低次元空間に埋め込むことで、データ間の類似性やクラスタ構造をできるだけ保持しつつ、可視化を行う手法。
UMAP （Uniform Manifold Approximation and Projection）	非線形	高次元データを低次元のマニフォールド（滑らかな多様体）上に確率的な近似に基づいて埋め込みの最適化を行う手法。

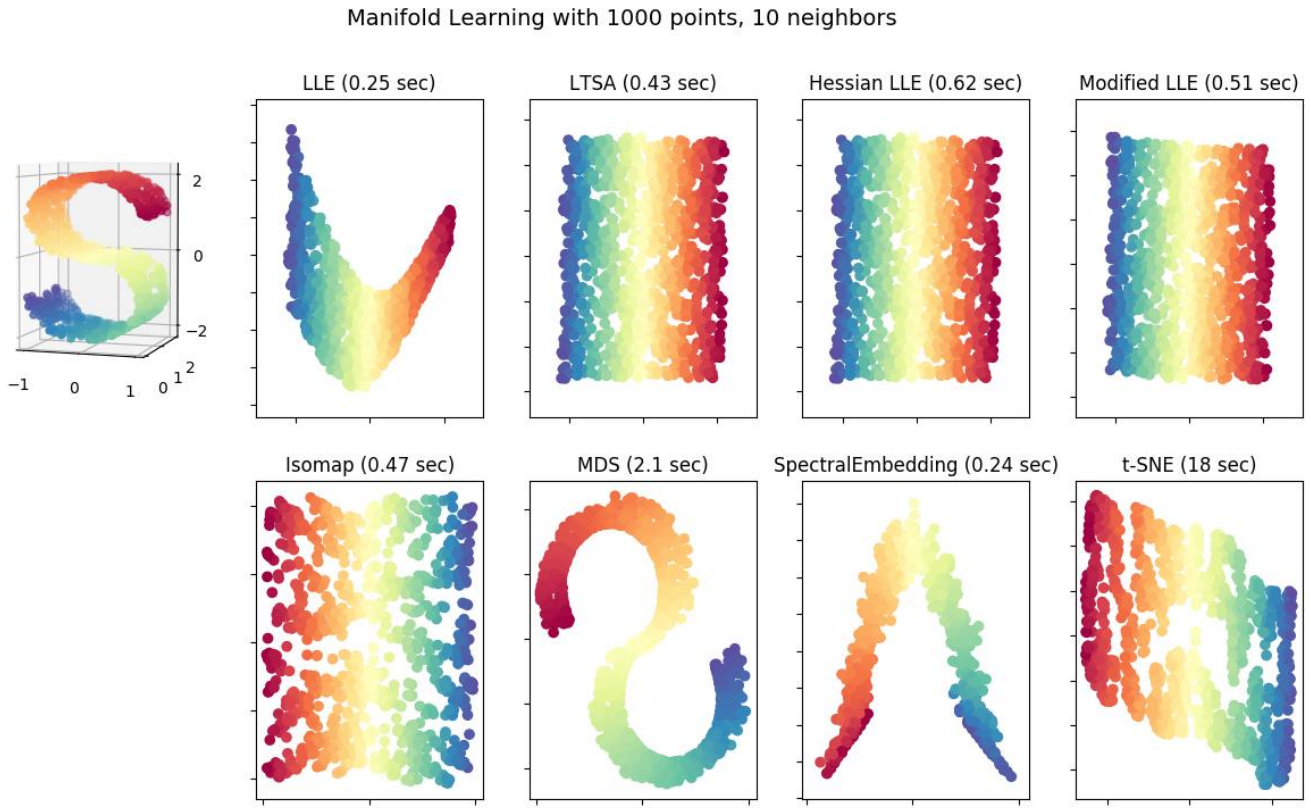
非線形手法による次元削減のイメージ

https://kntty.hateblo.jp/entry/2020/12/14/070022



様々な次元削減方法

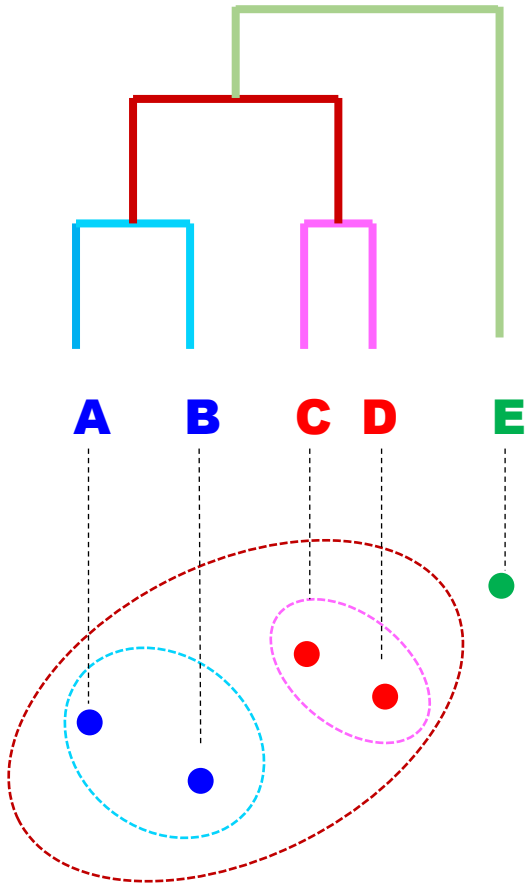
https://blogger.googleusercontent.com/img/b/R29vZ2xl/AVvXsEgWMkZrfH7C6ZbeDX7h2FeYkr9lcIBjfRI-JMhd0puIR_4nPVAeGgkhZfB9T5bVnNZs11QCTwIO18Nxt97yVNLFecM-HVIMMqxwrj2JK91Qp4V_hAUIUHcCmIt20YLMRGIXYsHkz47YfLke/s1600/sphx_glr_plot_compare_methods_0011.png



代表的なクラスター法

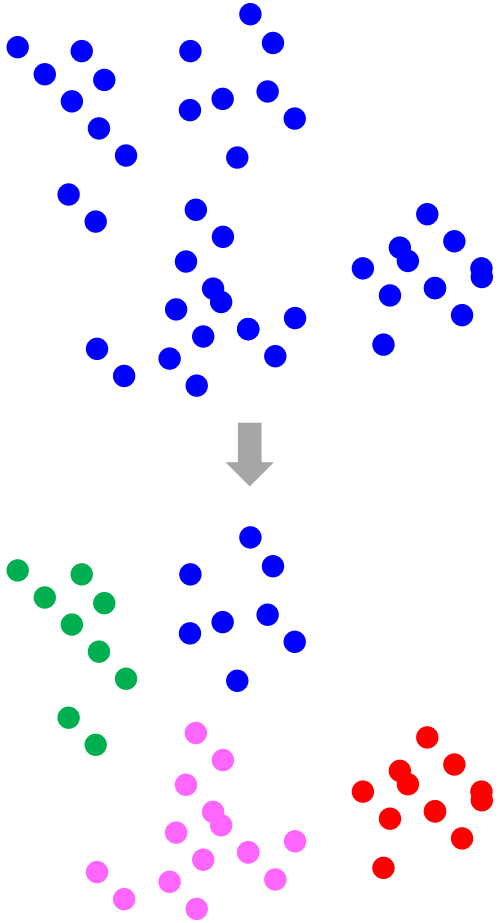
□ 階層クラスターと非階層クラスターに大別される。双方ともテキスト解析では多用される。

階層クラスター



樹形図
（デンドログラム）

非階層クラスター



k-平均法
k = 4

Scikit-learnでのポイント

```
# モデルライブラリ(クラス)の読み込み （ここではPCA）
from sklearn.decomposition import PCA

#次元削減アルゴリズムの設定 （インスタンス化）
pca = PCA(n_components=2)  # 2次元に次元削減

#次元削減の実行
X_pca = pca.fit_transform(X)
```

次元削減も基本は3行。機械学習と構文は同じです。

ARIM Japan

実データを使ったサンプルコード

本演習では『**茶の元素分析データセット**』を用いて、**次元削減とクラスタリング**の技術を使ってデータ分析を学びます。

- **次元削減技術の学習**:
  高次元データを低次元に変換する次元削減手法（主成分分析（PCA）、t-SNE、UMAPなど）を学びます。

- **クラスタリングアルゴリズムの理解**:
  クラスタリング手法（階層クラスタリングやK-means）を使用して、データセット内の類似性、グループやパターンを識別します。

- **次元削減とクラスタリングの統合的活用**:
  次元削減とクラスタリングを組み合わせることで、データの可視化や構造の理解を深めることができます。

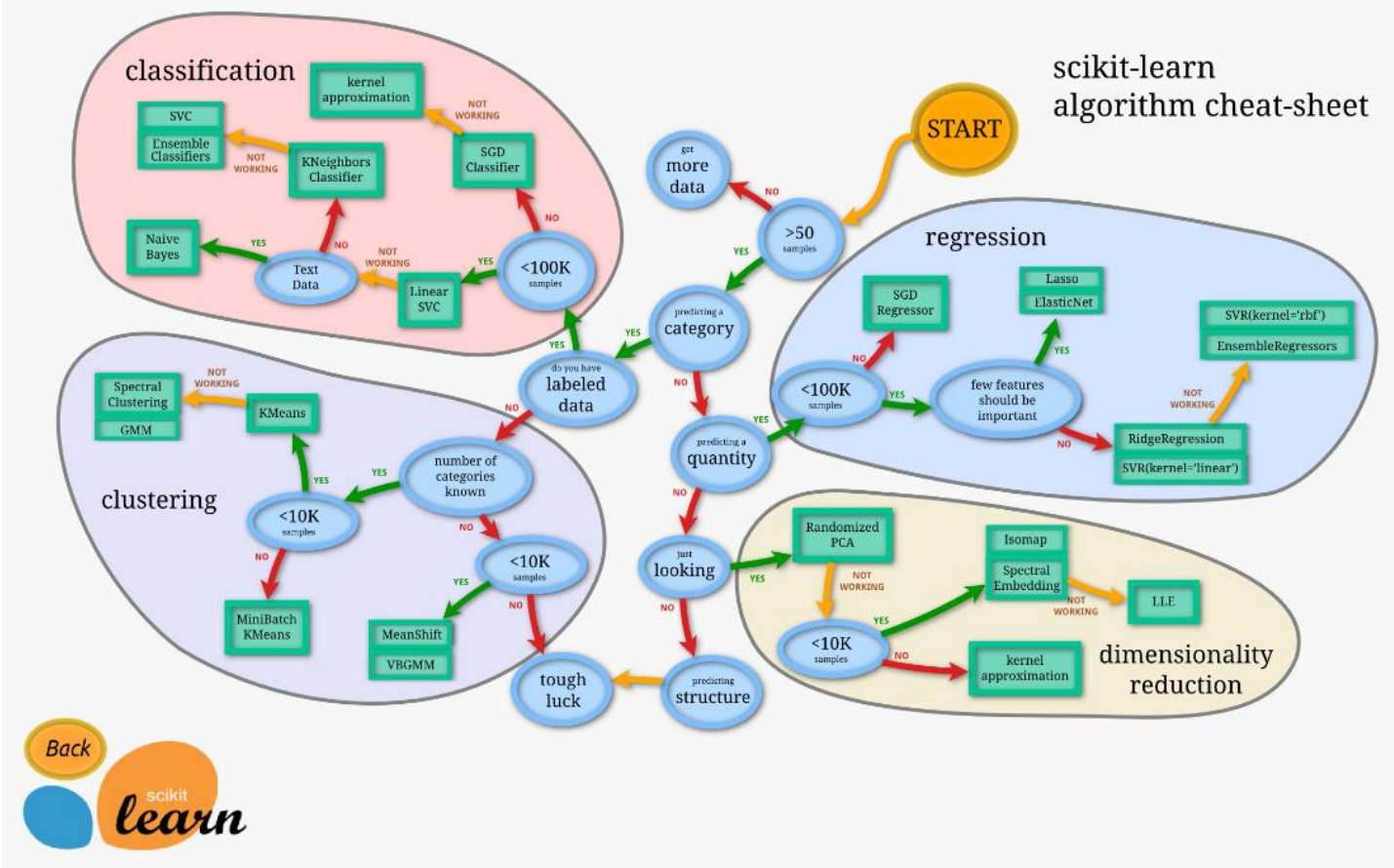
【次元削減とクラスタリング】
https://colab.research.google.com/github/ARIM-Academy/Advanced_Tutorial_1/blob/main/Scikit-learn-4.ipynb

5.　まとめ

機械学習でモデルを作成するにはどれぐらいのサンプルが必要か？

機械学習のモデル選定指針(scikit-learn)

- 最初の分岐点となるのはデータ数(標本数): 目安として50サンプルが示されている。
- 物質創成の過程では、標本数(インスタンス)が50を超えるケースは少ないそのため、多変量解析やその発展としての機械学習が取り入れる素地が小さい。



（上級者編）スモールデータセットの留意点

1. 行列のランク問題

- **行列のランクが低下**： 説明変数が多すぎると、データ行列のランクが低下。行列がフルランクでなくなり、逆行列が求まらず標準的な回帰分析や他の線形代数に基づく手法の適用が難しくなる。

- **解の一意性の欠如**： フルランクでない行列は、線形回帰の解が一意に定まらない。無限に多くの解が存在することになり、モデルの信頼性が低下。

2. 多重共線性

- **共線性の増加**： 説明変数が多いと、多重共線性（複数の説明変数が高い相関を持つ現象）が生じやすなる。変数間の相関が高くなり、個々の変数の影響を正確に評価することがより困難に。

- **回帰係数の不安定性**： 多重共線性が存在する場合、回帰係数の推定値が不安定化。小さなデータの変動が回帰係数に大きな影響を与えるため、モデルの予測性能が低下。

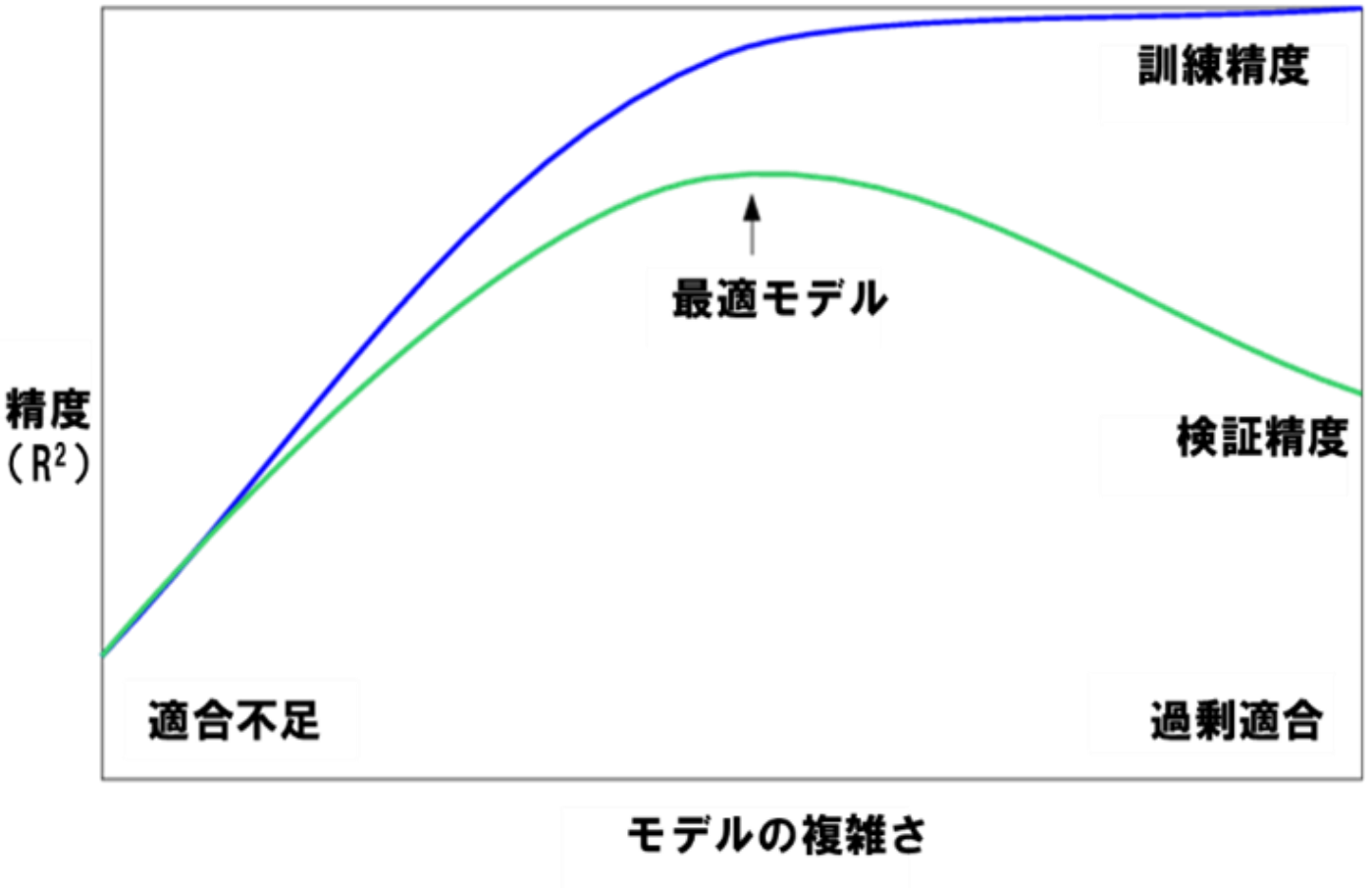
- **解釈の困難さ**： 多重共線性が強いと、変数間の影響を分離できないため、変数の出力に対する寄与の解釈が困難になる。

3. 過学習（オーバーフィッティング）

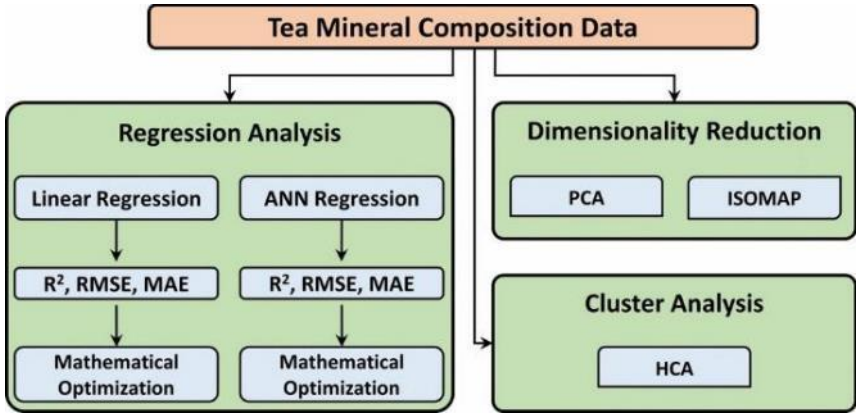
- **汎化性能の低下**： サンプル数より説明変数の数が多い場合、モデルがトレーニングデータに過度に適合してしまい、汎化性能が低下。結果として、新しいデータに対する予測精度が低下する。

過学習（オーバーフィッティング）

■ 実用性のモデルでは複雑なモデル（多変数）は過剰適合になりやすい



お茶のミネラル抽出のプロセスデータの最適化の方法



ミネラルの抽出量の予測モデル（scikit-learn）

束縛条件のもとのベイズ最適化（Optuna）

⑤ データ利活用	
解析手法	茶のサンプル内のミネラル含有量を推定するための ① MLR と ANNによる回帰分析 ② 主成分分析（PCA）と等尺性マッピング（ISOMAP）、階層的クラスター分析 (HCA)による元素分析値のクラスタリング ③ Parzen Estimatorアルゴリズム（tree-structured Parzen estimator (TPE) ）を用いた茶の抽出濃度・時間の最適化
ソフトウェア等	論文ではMinitab → Python (scikit-learn, keras, Optuna)
サンプルコード	**① MLR, ANNによる回帰分析** https://colab.research.google.com/github/ARIM-Usecase/Example_2/blob/main/1_ML_Code-1.ipynb
	② PCA, ISOMAP, HCAによるクラスタリング https://colab.research.google.com/github/ARIM-Usecase/Example_2/blob/main/2_DR_Code-2.ipynb
	③ TPEによる抽出条件の最適化 https://colab.research.google.com/github/ARIM-Usecase/Example_2/blob/main/3_TPE_Code-3.ipynb

ワークショップを終えたあと、是非、試してみてください

ARIM Japan