# 機械学習入門

松本　博士

## 1　機械学習とは

機械学習は，現代の技術で非常に重要な役割を果たしているが，その歴史は意外に古い．機械学習の歴史は20世紀初頭に遡り，進化し始めた統計学が後の発展の基盤となった．1940年代，アラン・チューリングがコンピューターの概念を開発し，将来的にコンピューターが人間のように思考できるかを考察した[1]．1950年代，彼はマシンが人間のように思考できるかを判定する「チューリングテスト」を提唱した．1952年，IBMの研究者アーサー・サミュエルが，チェッカー（アメリカンチェス）をプレイするコンピュータープログラムを開発した[2]．これは，コンピューターが性能を改善するために学習する最初の例とされている．1960年代から1970年代，機械学習の研究はいくつかの異なるアプローチに分かれた．一部の研究者は人間の脳の動作に焦点を当て，脳の神経細胞（ニューロン）を模倣したニューラルネットワークモデルを開発した[3)4]．他方，一部の研究者は統計学や確率論に基づいたアプローチに焦点を当てた．1980年代，コンピューターの計算能力が向上し，大量のデータを扱えるようになったことで，機械学習のアルゴリズムの実用化が進んだ．1990年代から2000年代，インターネットの普及によりデータ量が急増し，機械学習の発展を加速させる要因となった．特に，2000年代後半から2010年代，機械学習はスパムメールのフィルタリング，ウェブ検索，広告のターゲティング，音声認識，画像認識など，多くの実用的なアプリケーションに取り入れられた[5)~8)]．機械学習の歴史は，統計学，コンピュータサイエンス，人工知能，情報工学など，多くの分野と深く結びついている．本稿では，初めて機械学習を学ぶ研究者を対象にその概念や基本的な技術を紹介した後，応用的な技術について記した．そして分析領域への応用例と，よく用いられるツール・プラットフォーム等についても記したので，興味を持った人は試していただきたい．

## 2　基本的なアルゴリズム

ここでは機械学習で一般的に用いられる基本的なアルゴリズムについて紹介する．これらのアルゴリズムを理解することで，機械学習の様々な問題に対して効果的な解を見つけることが可能となる．

### 2･1　回　帰

回帰とは，データのパターンを解析し，そのパターンを用いて未知のデータを予測する手法の一つである．例えば，家の広さや築年数などの情報から，その家の価格を予測することが可能である．回帰の根本的な考え方は，データ間に存在する「関係」を特定することである．例えば，家の価格は，その家の広さ，築年数，場所などによって変動する．これらの要因と家の価格の間には，特定の「関係」が存在する．回帰分析を用いることで，これらの要因の影響を数値化し，その数値を使って新しいデータの家の価格を予測することが可能になる．

回帰分析には，いくつかの種類があるが，最も基本的で一般的なのが「線形回帰」である．線形回帰は，データの関係が直線的であると仮定し，その直線の方程式を求める方法である．この直線の方程式を用いることで，新しいデータの予測値を計算することができる．ただし，現実のデータは，必ずしも直線的な関係になるとは限らない．そのため，線形回帰だけでは解決できない問題も多く存在する．そのような場合には，決定木，サポートベクターマシン（support vector machine, SVM），そしてニューラルネットワーク（neural network, NN）を用いることができる．

決定木はデータを特定の基準で分割し，枝分かれさせていき，最終的に葉に分類する．これにより，データの中に潜んでいるパターンや関連性を見つけることができる（図1）．このパターンや関連性を用いて，新しいデータの予測が可能になる．適切な分割基準を選び，過学習に注意することが重要である．

SVMは，データを分析し，分類や回帰の問題を解くための手法である．SVMは，あるデータの集団について，あるグループと別のグループを最もうまく分ける境

Introduction to Machine Learning.

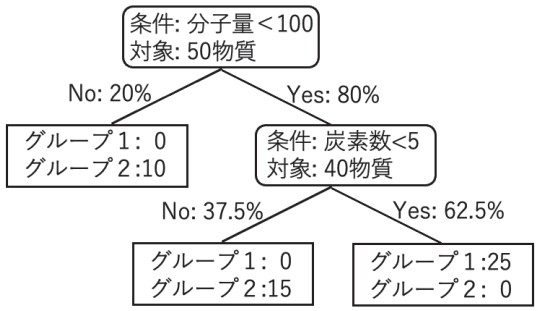


**図 1　決定木で特徴的なパターンを見いだした例**

実験結果に二つの傾向がみられた構造の異なる 50 種の物質（傾向ごとにグループ 1,2 と命名し，各グループ 25 物質で構成）について各物質の特徴をもとに解析．グループ 1 は分子量＜100 かつ炭素数 <5 の特徴を持つ物質からなることが示された．

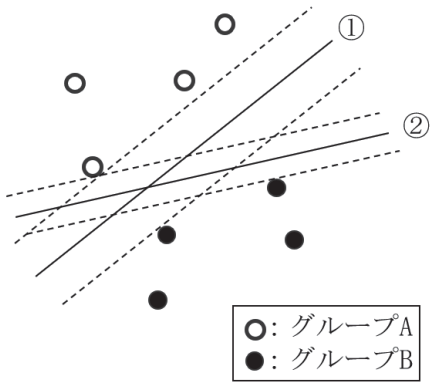


**図 2　SVM による境界線の例**

①と②は 2 グループ間で取りうる境界線を示し，各境界のマージンは破線で示した．②に比べて①のマージンの方が広いため①の境界線が最終的に採用される．

界線（または境界面）を見つける．SVM の特徴は，この分ける線を見つける際に「マージン」を最大化することである．マージンとは，分ける線から最も近いデータポイントまでの距離を指す．このマージンを最大化することで，新しいデータに対しても，うまく分類ができるモデルを作成することができる（図 2）．SVM は，線形だけでなく非線形のデータにも適用することができる．これは，特定の計算（カーネルトリック）を行うことで，非線形データを高次元空間に変換し，2 次元では非線形のデータも高次元の空間では線形の境界を見つけることで実現される．

NN は，人間の脳が情報を処理する方法にインスパイアされたコンピュータープログラムの一種である．人間の脳は，たくさんの神経細胞である「ニューロン」で構成されている．これらのニューロンが互いにつながり，情報を送受信することで私たちが考えたり，学んだりすることができる．このように，ニューロンからニューロンへとデータが流れることで，最終的にはデータ間の複雑な関係を把握し，予測を行うことができる（図 3）．

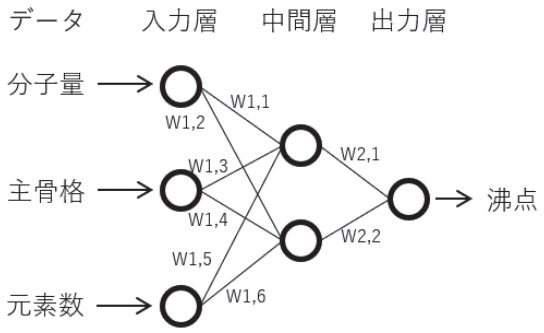


**図 3　沸点を予測する NN モデルの例**

図中の丸印はニューロンを現わず入力したデータ（分子量，主骨格，元素数）を基に最適な重み（W）を決定し，沸点を出力する．

回帰分析は，ビジネスや科学研究，経済学など，様々な分野で使われている．例えば，企業は，過去の売上データから未来の売上を予測するために回帰分析を使う．また，研究者は，気温や湿度などのデータから，農作物の収穫量を予測するために回帰分析を使うことがある．回帰分析は，データのパターンを見つけ出し，未来を予測するための強力なツールである．ただし，使う際には，データの特性や問題の性質に合わせて，適切な種類の回帰分析を選ぶことが大切である．

### 2･2　分　類

分類とは，データがどのグループに属するかを予測する手法である．事前に大量のラベル付きデータ（どのグループに属するか事前にわかっているデータ）を使って分類モデルを作成し，そのモデルを使って新しいデータのグループを予測する．例えば，メールがスパムかそうでないか，画像が犬か猫か，などである．よく用いられる手法としては，決定木，サポートベクターマシン，ナイーブベイズ，k 近傍法，ランダムフォレストなどがある．

ナイーブベイズは，学習データを用いて事前確率と条件付き確率を計算し，それらの情報をもとに新しいデータが属する確率が最も高いグループを予測する．ナイーブベイズは，そのシンプルな仕組みにより計算量が少なく，大量のデータを高速に処理することができる．そのため，スパムメールのフィルタリング，文章の分類，顧客のセグメンテーションなど，様々な分野で広く使われている．

k 近傍法は，特定のデータがどのカテゴリーに属するかを判定するための，シンプルで直感的な分類手法である．最初に「k」という値を決める．この「k」は，学習データをプロットした空間において予測したいデータの周辺で最も近い距離に位置するデータの個数を表す．例えば，k が 3 の場合，予測したいデータの周辺で最も近い三つのデータを指す．次に，予測したいデータが入

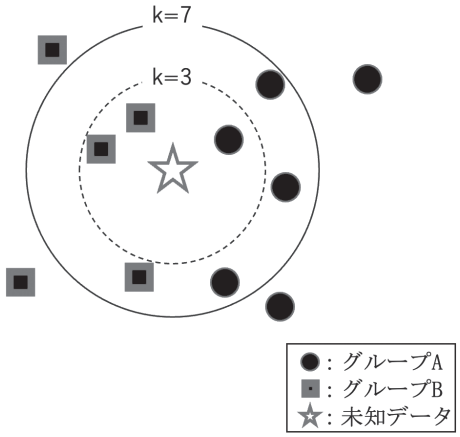


**図 4　k 近傍法で未知データを分類した結果**
k＝3 の時はグループ B，k＝7 の時はグループ A に分類された．

力された時，そのデータの周辺で最も近い「k」個のデータを，既存の学習データの中から見つける．そして，その「k」個のデータの中で最も多く存在するカテゴリーを，予測したいデータのグループとして決定する．例えば，k が 3 で，最も近い三つのデータが，「犬」が二つ，「猫」が一つであれば，予測したいデータのグループは「犬」となる．k 近傍法は，そのシンプルさと直感性から，分類問題を解く手法として広く使われている．ただし，データの量が多くなると計算量が増え，計算に時間がかかる欠点がある．また，「k」の値の選択がモデルの性能に大きな影響を与えるため，適切な「k」の値を選ぶことが重要である（図 4）．

ランダムフォレストは，決定木を多数作成し，それぞれの決定木でデータを分類する．そして，最も多くの決定木が選んだグループを，最終的な予測結果とする．例えば，ランダムフォレストで生成した 100 個の決定木のうち，60 個の決定木が「晴れ」と予測し，40 個の決定木が「雨」と予測した場合，最終的な予測結果は「晴れ」となる．ランダムフォレストのメリットは，多数の決定木を使うことで，一つ一つの決定木の過ちを補うことができ，全体として高い予測精度を得ることができる点である．また，ランダムフォレストは，各決定木の作成において，データの一部をランダムに選び，また，分類に使う特徴もランダムに選ぶため，過学習を防ぐことができる．ただし，本手法は多数の決定木を作成するため，計算量が大きく，計算に時間がかかるという欠点がある．

### 2･3　クラスタリング

大量のデータをいくつかのグループに分ける方法としてクラスタリングがある．これは，似たようなデータを一緒にグループ化することで，データの中に隠れたパターンを見つけるのに役立つ．スーパーマーケットの販売データを例に考える．販売データには，それぞれの商品を買ったお客さんの情報が記録されている．このデータをクラスタリングすることで，お客さんをいくつかのグループ，例えば，「子供のいる家庭」，「健康志向の人」，「お菓子好きの人」などのグループに分けることができる．この解析によって，スーパーマーケットは，それぞれのグループに対して適した商品を提供したり，効果的なマーケティングを行ったりすることができる．

クラスタリングは，教師なし学習の一種である．教師なし学習とは，データにラベル（答え）が付いていない場合の学習方法である．クラスタリングのアルゴリズムが，データの特徴を分析し，データを自動的にグループに分ける．

クラスタリングの一般的な方法の一つは，k-means 法である．k-means 法では，最初に「k」という値を決める．この「k」は，学習データをいくつのグループに分けるかを表す．例えば，k が 3 の場合，データは三つのグループに分けられる．次に，学習データの中からランダムに「k」個のデータを選び，それぞれをグループの中心とする．そして，残りのデータは，三つの中心のどれに最も近いかを計算し，その中心のグループに分類する．これを繰り返し行い，最終的にデータがグループに分類される．

以上のように，クラスタリングは，似たようなデータを一緒にグループ化することで，データの中に隠れたパターンを見つけるのに役立つ方法である．ただし，適切な「k」の値を選ぶことが重要であり，また，クラスタリングの結果は，データの特徴によって異なるため，結果の解釈に注意が必要である．

### 2･4　次元削減

次元削減は，データの特徴を減らす方法である．これは，大量のデータを扱う際に，データを簡単に分析しやすくするために使われる．あるお店の売上データを例に考える．このデータには，商品の色，サイズ，価格，販売数，季節など，たくさんの特徴が含まれている．しかし，これらの特徴の中でも，売上に影響を与える重要な特徴と，それほど重要でない特徴がある．次元削減を使うことで，データの特徴の中でも，最も重要な特徴だけを取り出し，分析の際に使う特徴を減らすことができる．次元削減の一般的な方法の一つは，主成分分析（principal component analysis, PCA）である．PCA では，データの特徴の中でも，最も重要な特徴を取り出す．これによって，データの特徴を減らすことができる．また，PCA は，データの特徴が多く，データの分布が複雑な場合にも，データを簡単に理解するのに役立つ．ただし，次元削減を行う際には，重要な特徴を取り除かないよう注意が必要である（図 5）．

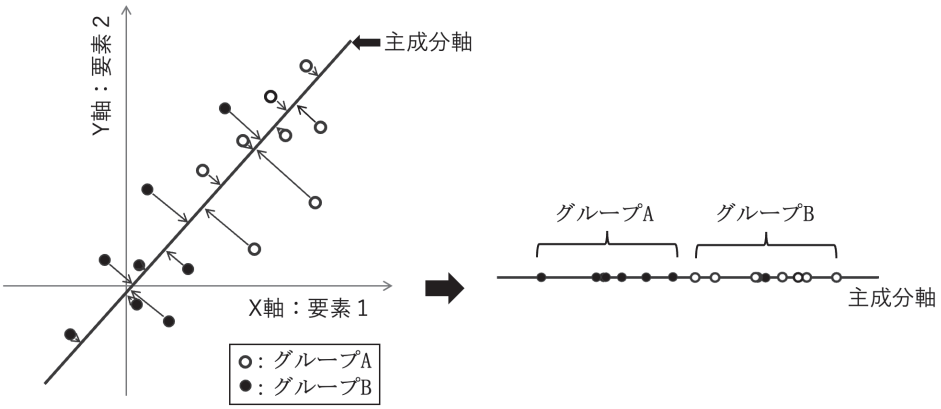


**図５　PCA の解析結果例**
要素１と２からなる二次元データを主成分軸に写し，１次元に次元を削減したまた，次元が減っても元のデータが持つ特徴（ここではグループ A と B の分布）は保存されている．

### 2･5　最適化

最適化とは，予測モデルを最良の状態にするための方法を見つけることである．前述した機械学習のアルゴリズム（今後モデルと呼ぶ）は，色々な設定値を持っている（k 近傍法や k-means 法の k もその一つ）．これらの設定値をうまく選ぶことで，モデルの性能が良くなる．最適化の目的は，これらの設定値をうまく選び，モデルの性能を最も良くすることである．最適化は，以下のように進行する．まず，モデルのパラメータを初期設定する．次に，パラメータの値を少しずつ変えながら，モデルの性能が改善されるかどうかを確認する．性能が改善される場合は，そのパラメータで最適化の検討を進める．性能が改善されない場合は，他のパラメータを試す．そして，モデルの性能が十分に良くなったと判断された場合，最適化を終了する．このプロセスを通じて，私たちはモデルの性能を向上させ，より良い予測結果を得ることができる．

### 2･6　画像認識

画像認識はコンピューターが写真や映像の中の物体や人物，場面などを認識する技術で，スマートフォンのカメラアプリで顔を認識してピントを合わせる機能や，セキュリティカメラが不審な動きを検出する機能などに使われている．一方，機械学習はコンピューターに大量のデータを与え，そのデータの中にあるパターンを学ばせる方法で，コンピューターは与えられたデータから学んだ知識を使って，新しいデータに対しても適切な判断や予測を行うことができる．画像認識の技術はこの機械学習の方法を使って進化しており，コンピューターに大量の写真や映像を与え，その中にある物体や人物，場面などのパターンを学ばせることで，新しい写真や映像の中の物体や人物，場面などを認識することができる．

## 3　データの前処理

機械学習モデルの性能は，供給されるデータの質に大きく依存する．そのため，データの前処理は，機械学習を用いた研究において非常に重要なステップである．以下に，データの前処理の重要性についていくつかのポイントを挙げる．

### 3･1　欠損値の取り扱い

欠損値は，データセットの特定の項目やフィールドが空白または不在の場合に発生する．これは，データ収集過程でのエラーや，情報が利用できなかったためなど，様々な理由で発生する．

データの前処理の重要なステップの一つである欠損値の取り扱いには，いくつかの一般的な方法がある．最も簡単な方法は，特定の行または列に欠損値が含まれている場合，その行または列を完全に削除することである．ただし，この方法ではデータが大量に失われる可能性がある．次に，欠損値をその列の平均値，中央値，またはモード（最頻値）で置き換える方法がある．これは数値データに対して一般的に使用される．また，前方埋め（前の行の値で欠損値を埋める）または後方埋め（次の行の値で欠損値を埋める）を使用して，欠損値を埋めることもできる．さらに，欠損値を含む列の他の値から，欠損値を予測するための統計的手法を使用する補間の方法がある．その他，欠損値をゼロまたは特定の値で置き換える方法もある．これらの方法はそれぞれ長所と短所を持っているため，データの性質，欠損値の発生パターン，データセットのサイズによって，適切な方法を選択するための慎重な検討が必要である．

### 3･2　ノイズの除去

ノイズの除去は，データをクリーニングし，不要な変動やエラーを取り除く過程である．ノイズは，センサー

エラー，人為的なミス，データの変換や伝送中のエラーなど，多くの理由でデータに含まれる．ノイズの除去は，データの質を向上させ，機械学習モデルのパフォーマンスを向上させるために重要である．ノイズの除去にはいくつかの手法がある．一つ目は，フィルタリングである．これは，データの平滑化を行う過程で，突出した値や不規則な変動を取り除く．例えば，移動平均フィルタは，データポイントの周囲の値の平均を取ることで，ノイズを減らす．二つ目は，データ変換である．これは，データを別の形式に変換することで，ノイズを取り除く方法である．例えば，対数変換は，データのスケールを変更し，ノイズを減らす効果がある．ノイズの除去の後は，データの品質が向上したことを確認するため，可視化や統計的な分析を行うことが重要である．

### 3･3　スケーリング

スケーリングとは，データの範囲を変更し，データの各特徴量が同じ尺度で比較されることを保証するプロセスである．機械学習の多くのアルゴリズムは，データの特徴量が同じ尺度であることを前提としている．例えば，k 近傍法，サポートベクターマシン，ニューラルネットワークなどがそれにあたる．データの特徴量が異なる尺度で混在すると，尺度が大きい特徴量が，モデルの学習において過度に影響を与えることになる．例えば，ある特徴量の範囲が 0 から 1 で，他の特徴量の範囲が 0 から 1000 である場合，範囲が 0 から 1000 の特徴量が，モデルの学習において，過度に影響を与えることになる．主な手法として正規化と標準化がある．正規化はデータの範囲を 0 から 1 に変更する．具体的には，データの各特徴量から，その特徴量の最小値を引き，最大値と最小値の差で割る．標準化はデータの平均を 0，標準偏差を 1 に変更する．具体的には，データの各特徴量から，その特徴量の平均を引き，標準偏差で割る．スケーリングの後は，データの分布を確認し，適切にスケーリングされたことを確認する．スケーリングは，データの前処理の一部であり，他の前処理の手順と組み合わせて行われることが一般的である．

### 3･4　特徴選択

特徴選択は，データの中から重要な情報だけを取り出す作業である．これを行うことで，計算量を減らし，且つ予測モデルの性能を良くすることができる．

再帰的特徴消去（recursive feature elimination, RFE）は，データの中から重要な特徴だけを選び出す方法の一つである．学習データにはたくさんの特徴が含まれているが，その中でも特定の特徴だけが目的の予測に重要で，他はあまり重要でない，ということがよくある．しかし，どの特徴が重要で，どの特徴が重要でないかを人間が一つ一つ確認するのは非常に大変である．そこで，RFE を使う．RFE は，コンピューターにデータの特徴を選ばせる方法の一つである．まず，RFE は，学習データのすべての特徴を使ってモデルを作成する．そして，そのモデルの中で，それぞれの特徴の重要度を計算する．次に，一番重要度が低い特徴をデータから取り除く．そして，残った特徴だけを使って，再度モデルを作成する．そして，またそれぞれの特徴の重要度を計算し，一番重要度が低い特徴を取り除く．これを繰り返し，指定した数の特徴だけを残すまで行う．そして，その選ばれた特徴だけを使って，最終的なモデルを作成する．RFE は，人間が特徴を選ぶ手間を省き，コンピューターに自動的に選ばせることができるため，非常に便利な方法である．ただし，どの特徴が重要であるかは，データやモデルによって異なるため，RFE を使った結果をそのまま信じるのではなく，必ず他の方法とも比較して，最終的な判断を下すことが大切である．

LASSO 回帰は，モデルの中で，予測に用いる各特徴の重要度に応じて係数を変化させ，重要ではない特徴の影響を小さくすることで，モデルの計算から不要な特徴を取り除く．デメリットとしては，LASSO 回帰に入力するパラメータにより，各特徴の重要度が変化するため，選択される特徴が変わってしまう点が挙げられる．また，学習データ中に同じ傾向を示す特徴が複数存在する場合，本手法はその中から一つしか選ばない．これは，本当は重要な特徴が取り除かれる可能性があることを意味する．

## 4　モデルの評価とチューニング

### 4･1　モデルの評価

モデルの評価は，作成したモデルがどれほど良い予測をするかを確認するための非常に重要な段階である．モデルを評価する際，まず，モデルの学習に使用していないデータ，つまり，テストデータを使用してモデルの性能を評価する．これは，モデルが未知の新しいデータである必要がある．次に，評価指標を使用してモデルの性能を数値化する．様々な評価指標があるが，問題の性質や目的によって，最適な評価指標が異なる．たとえば，分類問題の場合，一般的な評価指標には，正確度，適合率，再現率，F1 スコアなどがある．一方，回帰問題の場合，平均絶対誤差，平均二乗誤差，R2 スコア（決定係数）などがよく使用される．さらに，交差検証という手法を使用して，データを複数のサブセットに分割し，それぞれのサブセットをテストデータとして使用してモデルの性能を評価することもある．たとえば，k 分割交差検証では，データを k 個のサブセットに分割し，k 回モデルの評価を行い，k 回の評価結果の平均を取ってモデルの性能を評価する．モデルの評価は，モデルの性能を適切に評価し，過学習や未学習を避けるために重要である．

### 4･2　チューニング

モデルのチューニングとは，モデルの性能を向上させるために，モデルのパラメータを調整することである．モデルには，学習時にデータから学習されるパラメータと，事前に設定されるハイパーパラメータの2種類のパラメータがある．モデルのチューニングでは，主にハイパーパラメータを調整する．ハイパーパラメータは，モデルの学習方法を制御するためのパラメータで，例えば，決定木の深さ，ニューラルネットワークの層の数やノードの数，サポートベクターマシンの正則化項の強さなどがある．ハイパーパラメータの値によって，モデルの性能は大きく変わる．ハイパーパラメータの最適な値を探すために，いくつかの方法がある．グリッドサーチは，あらかじめ設定したハイパーパラメータの値の組み合わせをすべて試し，最も性能の良いハイパーパラメータの組み合わせを選ぶ．ランダムサーチは，ハイパーパラメータの値をランダムに選び，性能を評価し，最も性能の良いハイパーパラメータの組み合わせを選ぶ．ベイズ最適化は，ハイパーパラメータの値を順次選び，それまでの評価結果から次に試すハイパーパラメータの値を決定し，最も性能の良いハイパーパラメータの組み合わせを選ぶ．モデルのチューニングは，モデルの性能を最大化するために欠かせない作業であるが，計算リソースや時間がかかるため，適切な方法を選び，効率的に進めることが大切である．

## 5　機械学習のフレームワークとツール

機械学習のフレームワークとツールは，機械学習モデルの開発，トレーニング，評価を効率的に行うためのソフトウェアツールである．一般的な機械学習のフレームワークとツールについて紹介する．

### 5･1　TensorFlow

Google が開発したオープンソースの機械学習ライブラリである．機械学習のモデルを構築，トレーニング，実用環境へ展開するための多くのツールが用意されている．

### 5･2　Keras

TensorFlow の上に構築された，NN のライブラリである．モデルの構築，トレーニング，評価をシンプルに行える．NN モデルを作成するための多くのビルディングブロック（例：層，損失関数，最適化ツールなど）を提供しており，これらを自由に組み合わせることができる．

### 5･3　PyTorch

Facebook が開発したオープンソースの機械学習ライブラリである．動的計算グラフを使用しており，デバッグやモデルの変更が容易になる．

### 5･4　Scikit-learn

Scikit-learn は，Python のオープンソースライブラリで，機械学習の分類，回帰，クラスタリング，次元削減，モデル選択，前処理といったさまざまなアルゴリズムを実装し，効率的に利用できるようになっている．

### 5･5　Pandas

Python のデータ解析ライブラリで，多くの便利な機能を提供している．まず，データフレームという2次元のラベル付きデータ構造を中心に，データの読み書き，前処理，結合，統計解析を効率的に行うことができる．具体的には，CSV，エクセル，SQL データベース，JSON など様々なファイル形式のデータを簡単に読み込み，それらの形式でデータを書き出すことができる．また，データの欠損値を補完したり，データをフィルタリング，並び替え，集計したりといった前処理も簡単に行える．さらに，異なるデータソースから得られたデータフレーム同士をキーによって結合し，効率的に分析することができる．平均，標準偏差，最大値，最小値などの基本的な統計量の計算やヒストグラムの作成など，統計解析も行える．

### 5･6　NumPy

Python で数値計算を効率的に行うためのライブラリである．このライブラリを使うことで高速で効率的な数値計算が可能となる．NumPy は，数値データを効率的に扱うための多次元配列（ndarray）と，これに対する数学的な演算を用意している．たとえば，ベクトルや行列のような数値データを配列として作成し，これらの配列に対して，加算，減算，乗算，除算といった基本的な演算だけでなく，行列の積，逆行列，固有値，統計量の計算，ソートなどの高度な演算も効率的に行うことができる．また，ランダムな数の生成，フーリエ変換，線形代数の計算などの機能も提供されている．

## 6　分析領域への応用例

分析化学の分野においても機械学習を活用した研究例が報告されている．Melnikovr らは，メタボロミクスの分析プラットフォームとして使用されている液体クロマトグラフィー質量分析計（LC-MS）のデータ処理に関連した研究について報告している 9)．LC-MS の生データには何千もの MS スペクトルが含まれているため，手作業によるデータ処理はほぼ不可能である．著者らは LC-MS データのピーク検出と統合の問題を解決するための新しいアルゴリズム peakonly を開発した．このアルゴリズムでは，畳み込みニューラルネットワーク（convolutional neural network, CNN）を用いることで

未処理の LC-MS データにおけるピーク検出と統合の問題を解決した．また，Enders らは，気相 FTIR スペクトル中の官能基の存在を同定するために，CNN を用いた機械学習アルゴリズムにより，一般化可能なモデルを開発した[10]．NIST スペクトルデータベース内の 8728 個の気相有機分子から強度–波数データを取得し，データを画像に変換した．15 種類の官能基モデルにより，未知のスペクトルを効果的に分類し，スペクトルの解釈を容易に行うことが可能となった．

## 7　その他の考慮すべきこと

機械学習の利用には，数々の倫理的な問題が関連している．まず，データのプライバシーは非常に重要なポイントで，人々の個人情報を取り扱う際は，そのデータの安全性とプライバシーを確保することが求められる．また，データのバイアスも重要な問題で，もしデータが偏っていれば，そのデータを基に学習したモデルも偏った結果を出す恐れがある．さらに，モデルが意図しない結果を出す可能性も考慮する必要がある．例えば，自動運転車の場合，事故を起こす可能性があるという問題が挙げられる．また，モデルがどのように結果を導いたのかを人間に分かりやすく説明すること，つまり，モデルの説明可能性も重要なポイントである．そして，最後に，モデルの使用目的についても考慮する必要がある．悪意を持ってモデルが使われる可能性もあるため，そのリスクについても考慮する必要がある．

## 8　最　後　に

近年，機械学習の学習環境は著しく進化している．かつては，機械学習を学びたいと考えても，高性能なコンピューター，大量のデータ，専門知識が必要で，アクセスの障壁が高かったが，現在では YouTube の解説動画を通じて気軽に学べるようになった．さらに，Google や Microsoft などの大手企業も，機械学習に関連する無料教材を公開している[11)12)]．また，Kaggle という実践的な学習ができるプラットフォームも利用可能である[13)]．Kaggle では，機械学習のコンペティションが提供され，実際のデータを使ってモデルを作成し，他の参加者と競いながら学べる．さらに，Google Colaboratory のようなクラウドベースの解析環境も提供されている[14)]．これらの環境を利用すれば，高性能なコンピューターを持っていなくても，インターネットさえあれば，クラウド上で機械学習のモデルを学ぶことや実践することが可能である．

### 文　献


1）A. M. Turing : *Mind*, **59**, 433 (1950).
2）A. L. Samuel : *IBM Journal*, **3**, 211 (1959).
3）F. Rosenhlatt : “Principles of Xenrodynttntics : Perceptrons end the Theory of Bruin Mechanisms”, (1961), (Spartan Books), (Washington, D. C.).
4）H. D. Block : *Rev. mod. Phys.*, **34**, 123 (1962).
5）B. Enrico, A. Bryl : *Artif. Intell. Rev.*, **29**, 63 (2008).
6）A. Lacerda, M. Cristo, M. A. Gonçalves, W. Fan, N. Ziviani, B. Ribeiro-Neto : *In Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, p. 549 (2006).
7）S. K. Gaikwad, B. W. Gawali, P. Yannawar : *Int J. Comput. Appl.*, **10**, 16 (2010).
8）F. Khan, R. K. Krishna : *Int. j. eng. res.gen. sci.*, **3**, 2, (2015).
9）A. D. Melnikov, Y. P. Tsentalovich, V. V. Yanshole : *Anal. Chem.*, **92**, 588 (2020).
10）A. A. Enders, N. M. North, C. M. Fensore, J. Velez-Alvarez, H. C. Allen : *Anal. Chem.*, **93**, 9711 (2021).
11）Tensorflow : 〈https://www.tensorflow.org/〉, (accessed 2023. 10. 06).
12）Microsoft, 〈https://github.com/microsoft/ML-For-Beginners〉, (accessed 2023 10. 06).
13）Kaggle : 〈https://www.kaggle.com/〉, (accessed 2023. 10. 06).
14）Google Colaboratory : <https://colab.google/>, (accessed 2023. 10. 02).



松本 博士（Matsumoto Hiroshi）
ダイキン工業株式会社化学事業部プロセス技術部（〒314–0255 茨城県神栖市砂山 21）．名古屋市立大学大学院医学研究科博士課程修了．博士（医学）．