

第6講 「機械学習の基本概念」

【学習到達目標】

- ・ 機械学習の仕組みと種類を理解する。
- ・ 学習過程と評価指標を説明できる。
- ・ 過学習とその対策を理解する。

1. 機械学習とは

機械学習とは何か？データからパターンを学び、予測や分類に活用する技術です。教育、医療、金融など幅広い分野で活用可能です。

機械学習を用いて、脳活動から心の中でイメージした内容を画像化することもできます。漫画「ドラえもん」で、眠っている人の頭の中で何を見ているかというようなこと見られる機械が登場しましたが、機械学習を用いて脳活動から心の中でイメージした内容を分類し画像化することができるようになったわけです。

機械学習の枠組みはどうなっているかというと、人間の脳の神経細胞が相互に接続した構造を模倣したものになります。機械学習には、「教師あり学習」、「教師なし学習」、「強化学習」の3つの枠組みがあります。

図1に示すように、それぞれ脳の機能に対応づけられています。

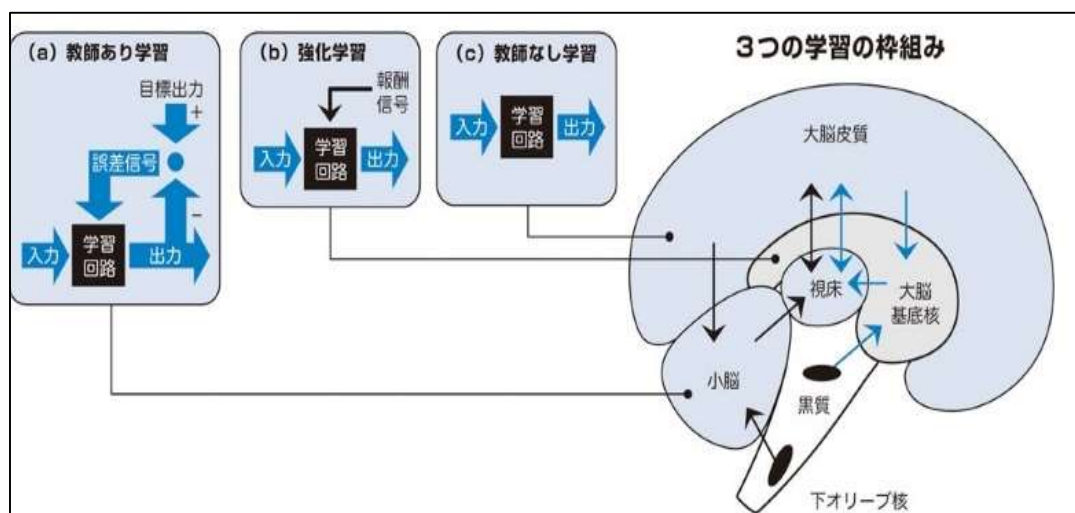


図1 3つの機械学習の枠組み

2. 教師あり学習

教師あり学習は、正解データで学習する手法です。例えば、テスト結果から合否を予測したり、画像から猫か犬かといった正解を教えて学習させたりします。機能的には小脳が担う学習機能です。視覚野で神経活動が活発になっている部位を「fMRI」で測定して、特徴量予測デコーダー（解析器）で予測パターンを認識させると、事前に用意した大規模画像との比較から動物の「ヒョウ」だと分かります。fMRI（磁気共鳴機能画像法）は、医療用の MRI 装置を使って無害に脳活動を調べる方法です。

教師あり学習では、図 1 に示すように正しい入出力の組合せを与え、「目標出力」との誤差信号を学習することで、新規の入力に対し、適切な出力が出せるようになります。

教師あり学習の代表的手法は、誤差信号を入力と逆に伝搬させるので、誤差逆伝播法 (Back Propagation) と言われます。正解、若しくは誤りを入力として、未経験入力に対する意志を決定する決定木 (Decision Tree) の作成なども使われます。

教育の場合、教師あり学習を用いて、数例の模範回答で記述式問題を採点することができます。

図 2 の教師あり学習の例では、数例の模範解答で採点をします。機械学習の立場から「フュー・スロット学習」と言います。ごく少ないデータから AI がモデルを学習して高い精度で予測が行われるようにする機械学習の手法です。例えばプロンプトとして「あなたは有能な AI 教師です。記述式問題「超 AI 時代における教師の役割を説明しなさい。模範回答は以下の通りです」ということで、模範回答を事前に作っておいて教え込みます。

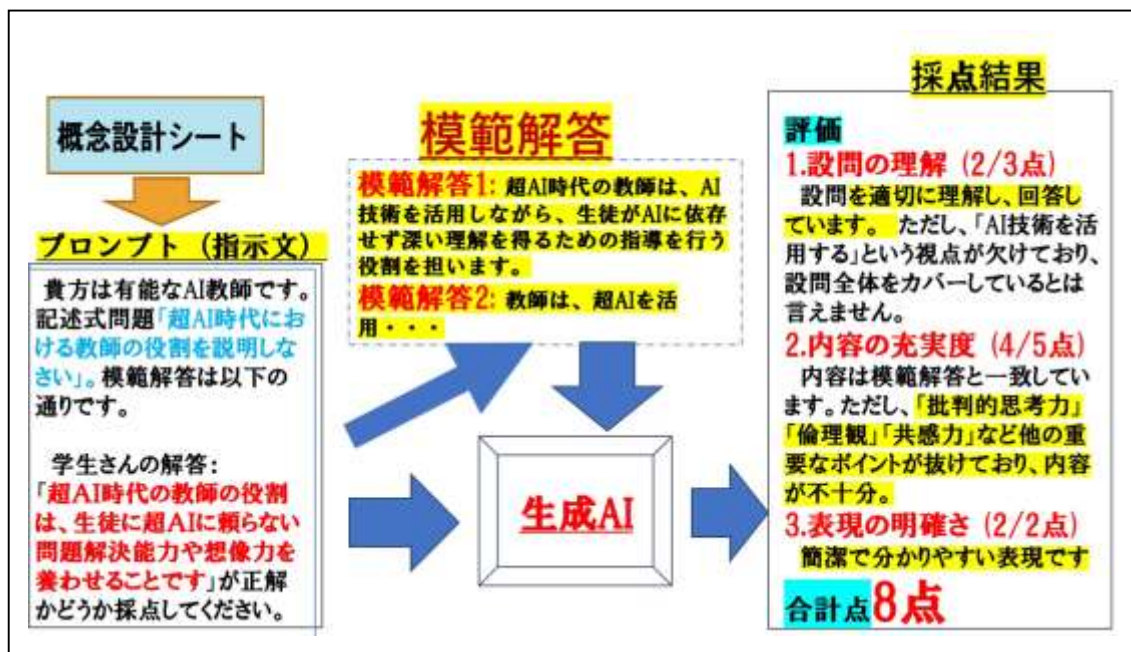


図 2 教師あり学習の例：数例の模範解答で採点

学生さんの回答が「超 AI 時代の教師の役割は、生徒の超 AI に頼らない問題解決能力や想像力を養うことです」とします。それに対して採点してくれ、総合点は 8 点となります。内訳は、「設問の理解」については 3 点中の 2 点、「内容の充実度」については 5 点中 4 点、「表現の明確さ」は 2 点中 2 点と、採点し評価理由を付けてくれます。

3. 教師なし学習

教師なし学習は、正解ラベルなしでデータの構造を抽出する手法です。例えば、学習スタイルに基づく生徒のグループ分け（クラスタリング）などが挙げられます。

教師なし学習は、大脳皮質が担う学習機能で、人間が何も教えなくても機械が自律的に学習していきます。教師なし学習では、図 1 に示すように新規の入力に対し、機械が自律的に適切な出力を出します。

教師なし学習の代表的手法は、クラスタリングの K-means 法（k 平均法）です。

k-means は、指定した k 個のクラス（集団）にデータを分類していくアルゴリズムです。重心を求め、似た特徴の画像を近くに置くことで、グループ分けする方法です。グループ分けの結果から新しいアイデアを得ることができます。

4. 強化学習

強化学習は、環境との相互作用を通じて報酬を最大化する行動を学習する手法です。学習アプリが自動的に難易度を調整したり、ロボット制御や将棋のようなゲームに使われたりします。強化学習は、大脳基底核が担う学習機能です。

強化学習では、図 1 に示すように正しい入出力の組合せに対する報酬信号を学習することで、新規の入力に対し、適切な出力が出せるようになります。

強化学習を用いるロボット犬「A i b o」では、最終結果若しくは、途中経過に対して、程度良かったかを示す「報酬信号」に基づき、例えば「ける」ことにより報酬をなるべく大きくするように探索をします。

5. 学習過程と評価指標

学習は、訓練、検証、テストという流れで進められます。

- 1) 訓練: データでモデルを学習します。
- 2) 検証: パラメータ調整やモデルの選択を行います。
- 3) テスト: 新しいデータで性能評価を行います。

図 3 に示すように、学習過程の評価指標には、正当率 (Accuracy)、適合率 (Precision)、再現率 (Recall)、そして F 値 (バランス指標) といった 4 つの重要な指標があります。

それぞれ以下のような意味を持ちます、

- 1) 正当率: 全体のデータのうち、正しく正解を検索（分類）できた割合です。
- 2) 適合率: 正解として検索されたデータのうち、実際に正解だった割合です。
- 3) 再現率: 実際に正解のデータのうち、正しく正解と検索できた割合です。
- 4) F 値（バランス指標）: 適合率と再現率のバランスを示す指標です。

これらの評価指標は、図3のような数式で表現できます。

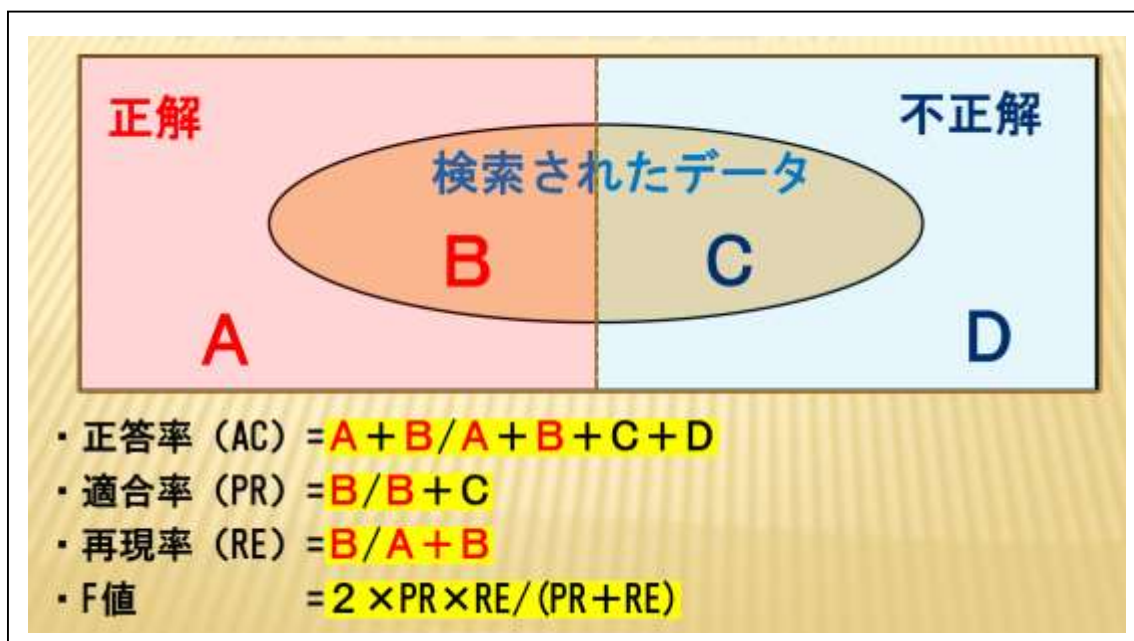


図3 学習過程の4つの評価指標

6. 過学習

過学習とは、モデルが訓練データに過剰に適合し、新しいデータに対する性能が低下する現象です。模試だけ強くて実力テストでは点が取れない生徒に例えることができます。

このような過学習を防ぐには、以下のことが挙げられます：

- 1) データを増やす：学習不足が解消します
- 2) 正則化：バリエーションが過剰に大きくならない程度に、バイアス（偏り）を抑制します
- 3) 交差検証：モデルの性能を過小適合と過剰適合の観点から測定します。

図4は「最適解」がバイアス・バリエーション曲線の交わる点であること、バイアス・バリエーションのトレードオフがあることを示しています。学習不足では偏り誤差が多くバイアスが高くなり、学習量が多すぎると「過学習」となり、ばらつき誤差が増大し、バリエーションも高くなります。

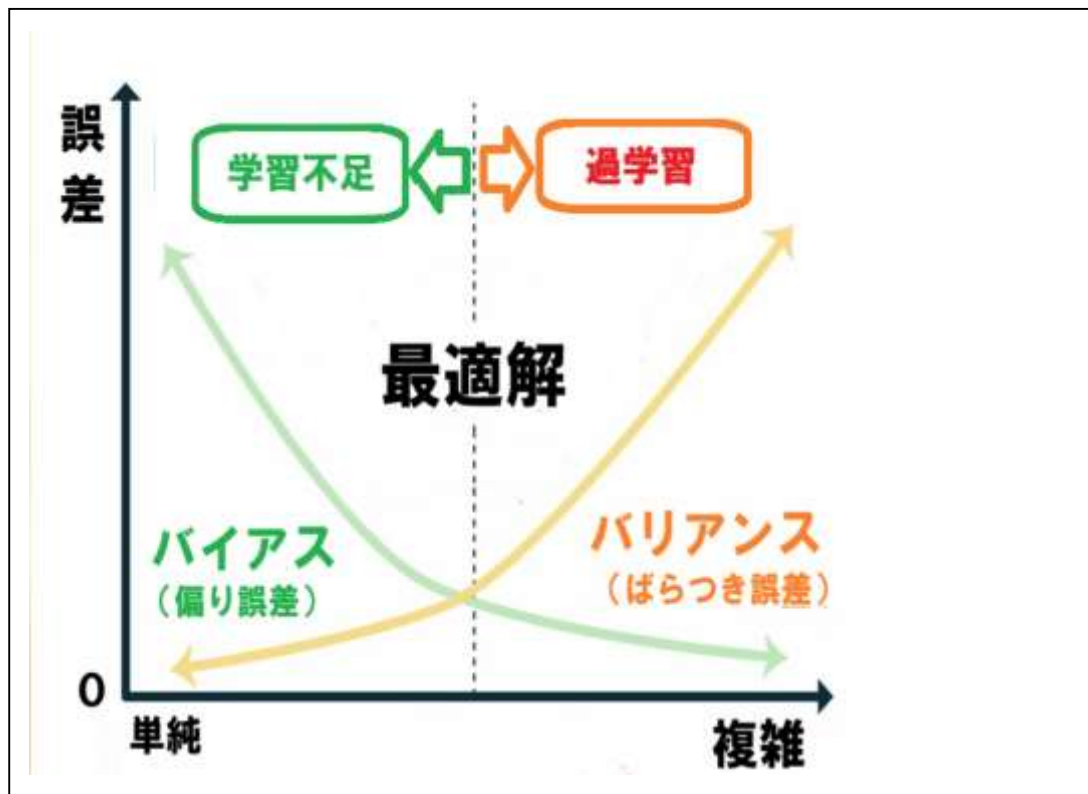


図4 バイアス・バリエーションのトレードオフ

これらのことから、

- 1) 単純すぎるモデルは、バイアスが高く、偏り誤差がでます。
- 2) 複雑すぎるモデルは、バリエーションが高く、ばらつきが多く、誤差や過学習リスクが出ます。
- 3) 適切なバランスが最適解となります。

6. まとめ

まとめますと、今後

- 1) データサイエンスは社会・教育の両面で重要性が増すので、
- 2) 教育者は未来を見据えた教育を実践する必要があります。
- 3) 未来の教育は、単なる知識の伝達だけでなく、データを活用した問題解決能力や倫理観を育むことが求められるだろうと言えます。

【課題】

- ・データサイエンスの未来において、社会や産業界で期待される役割と、その教育的意義について述べなさい。
- ・今後の教育現場において、データリテラシー教育を推進するために必要な取り組みや内容について具体的に述べなさい。
- ・データサイエンスの発展に伴う倫理的・社会的課題を挙げ、それに対して教育現場でどのような対策や教育内容を取り入れるべきか、あなたの考えを述べなさい。