従来から行われてきた統計処理も機械学習も予測するための技術である。人は、大量にあるデータを眺めていても、そのデータが示すところを理解することは難しい。つまり、データの性質を把握するためには何らかの「説明」が必要であり、統計処理においては、そのために各種の統計値が使われていた。パーソナルコンピューターが普及し始めてから40年弱、高度な計算が可能な32ビット機が普及してから20年未満である。しかし、統計の源流は古代から始まっており、数理統計学にしても数百年以上の歴史を持つ。その大半の時間、人はデータをコンピューターなしで処理し、性質を「把握」していたのである。

これに対して機械学習は、最初からコンピューターでの利用を想定して、というよりも、コンピューターありきでつくられた技術である。大量のデータに対してひたすらに計算を続けることが可能であり、特にディープラーニングではデータに潜在する特徴の抽出を自ら行うことができる。反面、得られた予測がどうしてそうなるのかといった説明が困難という問題が発生している。

(2) 主な技術

ここでは、「予測」とは、既知のデータの属性値から未知のデータの属性値を計算することに関するものとする。回帰、分類、クラスタリングなどの技術が含まれる。また、予測に関わるそのほかの技術として、ベイズ推定による予測や、アンサンブル学習を用いた精度の高い予測、グラフ構造データによる学習と予測、ディープラーニングとシミュレーションを組み合わせた予測についても紹介する。

回帰

回帰は、統計学において、目的変数(従属変数)Yと説明変数(独立変数)Xの間の関係を明らかにすることを指す。関係が明らかになることで、例えば未知のデータ属性「家賃の高さ」の値を、既知のデータ属性である「広さ」、「築年数」、「駅からの近さ」の値から計算(=予測)できる。

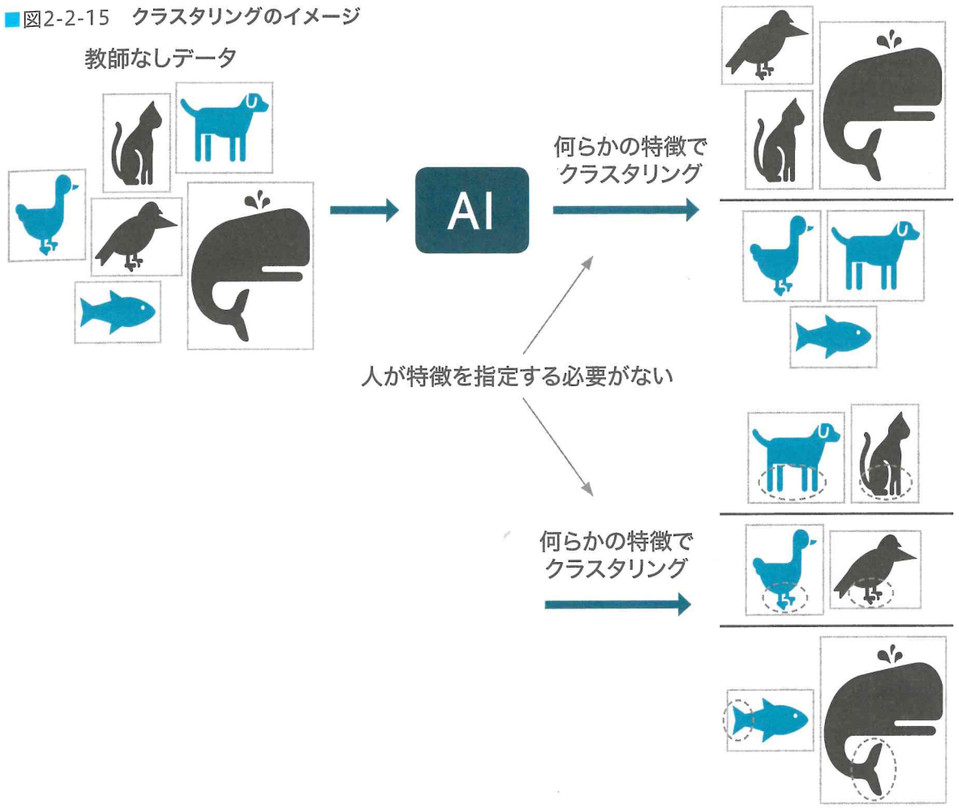
分類

分類では、事前に与えられたクラス(例えばネコやバナナ)に基づいて、与えられたデータをそれらのクラスに割り当てる。例えば質問に答えながら、動物を分類するといった場面では、動物の特徴について順に質問していき、結果その動物が属するクラスを決定(=予測)する。クラスが事前に与えられていない場合は、クラス分けそのものを既知のデータ属性の値から予測するのがクラスタリングになる。

クラスタリング

クラスタリングは、分類と違い、事前にクラスなどの分類の軸が提供されないため、与えられたデータの特徴などから自動的にクラスタを構成する(**図2-2-15**)。事前学習が不要なため、例えば、過去のデータがない場合でもクラスタ化が可能だが、クラスに基づく分類とは異なる結果が出る場合がある(色を特徴としてとらえることで、黒猫とカラスを同じクラスタに分類するなど)。クラスタリングによる予測としては、例えば、株価の変動パターンをクラスタリングすることにより頻度が高いパターンを洗い出し、現状の変動パターンと照らし合わせることで直後の変動を予測することが可能である。代表的なアルゴリズムにk平均法などがある。

■図2-2-15　クラスタリングのイメージ



なお、回帰や分類はよく似ている技術であるが、未知のデータ属性の値が、分類の場合は事前に取り得る値が決まっている離散変数=クラスであり、回帰の場合は取り得る値が事前に分かっていない離散値や連続値であるともいえる。SVM、ランダムフォレスト、決定木、ニューラルネットワーク、k近傍法は、回帰にも分類にも適用できる機械学習技術である。これ以外に回帰では、線形(多重)回帰、分類ではロジスティック回帰、ナイーブベイズなどが用いられる。**表2-2-4**は、代表的な「予測」アルゴリズムとその用途についてまとめたものである。

■表2-2-4　代表的な「予測」アルゴリズム

アルゴリズム	回帰	分類	クラスタリング
線形(多重)回帰	○	×	×
ロジスティック回帰	×	○	×
サポートベクトルマシン (Support Vector Machine)	○	○	×
ナイーブベイズ	×	○	×
決定木	○	○	×
ランダムフォレスト	○	○	×
ニューラルネットワーク	○	○	×
k近傍法	○	○	×
k平均法	×	×	○

○…利用に適している　×…利用に適していない