VII. 多変量解析から機械学習へ

VII-1. 機械学習とは何か

VII-1-1. コンピューターの発達と学習

学習とは、正しい判断をするために、過去の事実を集約して未来を予測し、正しい判断をするためのシステム（モデル）を作ることです。その中には、モデルに含まれるいくつかの関数を適切に選択することや、関数に含まれる係数を最適化することが含まれます。つまり、回帰分析、判別分析、クラスタリング、主成分分析等々の統計解析は、すべて学習という知的作業そのものなのです。学習には、教科書を使った勉強のように、結果が出ていて予測すべき答え（被説明変数）が与えられてやるそれに至るためのシステムを作る学習と、与えられたデータセットの中に予測すべき正解が与えられないままに、データがどんな構造になっているのかをデータ同士の関係から導き出すという学習があります。また、赤ちゃんが泣いたり声を出したり動いたりして、得られる結果（報酬）から何かを学んでいくように、何かの行為をしてその結果（報酬）からシステムを変更していくという能動的な学習もあります。これらは、それぞれ、機械学習でいうところの教師あり学習（supervised leaning)、教師なし学習（unsupervised leaning)、強化学習（reinforcement leaning）にあたります。重回帰分析では被説明変数が量的尺度（間隔尺度または比例尺度、ただしダミー変数を含む）で与えられ、判別分析では何に属するかという離散的な名義尺度で表わされているという違いがありますが、重回帰分析や判別分析は教師あり学習、クラスター分析、主成分分析、対応分析などは教師なし学習です。強化学習に相当する多変量分析は何かと考えてみましたが、MCMC とか勾配降下法とかコンピュータを使った計算方法の中にそれに近い考え方がありますが、多変量解析でぴったりとそれにあたるものは思いつきません。我々のやっていることの中で探すと実験科学の手法が似ています。

もともと、そういうものがあったにもかかわらず、近年になって、教師あり学習、教師なし学習などという概念が出てきたのは、コンピュータとその周辺の技術・情報が発達してきたためでしょう。人間にできないコンピュータの機能は、データの蓄積の大きさとそのデータを用いた計算の速さです。コンピュータの発達と並行して、最近ではデータ化された様々な膨大な情報(Big data)が共有されるようになってきました。共有されていなくても、お金を出せばデータが手に入ることがあります。データの集積・整理は極めて重要な知的作業ですが学習ではありません。データから正しい判断をするためのシステム（モデル）をつくることが学習です。人の記憶力や計算速度には限界があるので、Big data の分析は、コンピュータの計算力に頼らざるを得ません。コンピュータの発達がもたらしたものは、計算速度や除法の蓄積量のような量的拡大だけではありません。解析的方法で微分方程式が解けなくても、コンピュータの計算速度の速さを使えば、数値積分で実質的に使える近似解を求めることが可能です。コンピューターを使って、従来出来なかった解法が可能になっているのです。そういう計算のためのプログラムがたくさん開発されています。今後も新しいものが次々と出てくるでしょう。そうした背景から、コンピューターによる解法のプログラムを書くこ

とを意識して、教師あり学習や教師なし学習という概念が出てきたのだと思います。機械学習の入門の部分で、教師あり学習の例としてしば紹介されるのが判別分析です。線形判別分析はすでに VI-1-3 で説明しました。線形判別分析には等分散という制約があります。ここでは、機械学習が従来の多変量解析が持っている弱点（制約）をどのように克服するのかを理解するために、線形判別分析から出発して、それを機械学習的な教師あり学習に発展させるという形で、段階的に教師あり学習としての判別分析（多クラス分類システム）を作っていきます。

VII-1-2.線形判別分析の限界

線形判別分析（VI-1-3）では、線形判別分析は、A であるか B であるかというような一対のどちらであるかという判別に使われることが普通と解説しました。実は、3 つ以上のクラスに分類するときにも使えないことはないのです。ただ、3 つ以上あった場合に、判別分析の前提になっている正規性と等分散性が怪しくなってきます。例えば、2 次元平面で表せるデータで、5 つのグループがあり、それぞれのグループが 2 次元平面上で図 97 のように分布していたとします。それぞれのグループをクラス青、クラス赤、クラス緑、クラス黄色、クラス白と呼ぶことにします。それぞれのクラスの分布範囲は同じ色で示した楕円の範囲に分布しているとなんとなく思ってください。この楕円の大きさ、形、傾きはそれぞれ違っています。このデータセットは解説のために筆者が作ったのですが、それぞれの楕円で、長軸方向、単軸方向にはデータは正規分布しています。しかし、この中の 2 クラスを選び出して、その中間に判別境界線を引いたときに、判別境界線からの距離は正規分布しているとは限らないし、等分散でもないでしょう。一対ずつ個別に考えれば、問題を解決する方法がないわけではないでしょうが、コンピュータが得意とする網羅的な分析では、そんな面倒なことはしたくないというのが分析者の気分ですし、判別平面という直線的な境界で仕分けるということも、仕分けるグループが多次元空間中にたくさんあった場合には不自然です。コンピュータの計算速度の速さを使えば、この問題は解決して、不自然な前提条件のない人の感覚に合った無理のない学習になります。

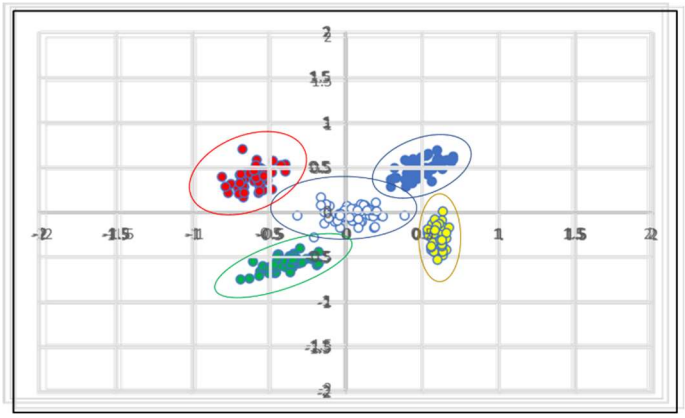

図 97. 2 次元データの分布の事例