08 統計と機械学習の違い

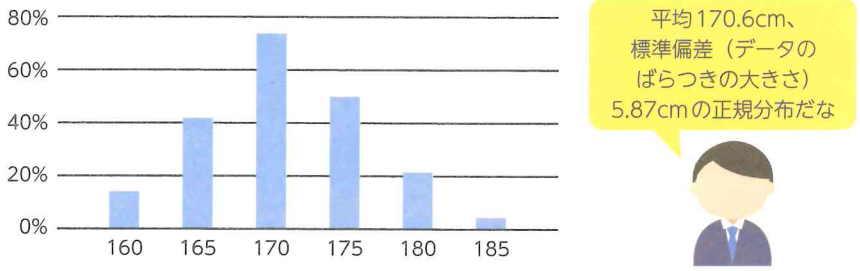
機械学習と同様にたくさんのデータを扱う分野として、統計があります。両者は理論に共通する部分が多い一方で、応用に対する考え方の違いから、その線引きが難しいといえます。ここでは「ツール」としての見方を通して、違いを整理していきます。

統計と機械学習では、導く情報が違う

世の中には、ある都市の気温や企業の株価、個人の1年間の体重の増減に至るまで、多種多様なデータがあります。これらに対して、統計は「**なぜこのようなデータが出るのか**」を教えてくれます。一方、機械学習は「**これからデータがどう変わっていくのか**」を、それぞれ教えてくれるのです。もっとも、厳密には両者を線引きすることは難しく、このような整理はあくまでイメージの違いをわかりやすくするためであることを留意してください。

その上で、統計の理解をより深めるため、身長分布（身長のばらつき）を例に考えてみましょう。文部科学省のHPでは毎年、就学中の児童および生徒の身長が学校健診で集められ、公開されています。その中の1つである、17歳（高校3年生）の身長データをヒストグラム（柱状グラフ）で表してみると下図のようになります。

■身長をヒストグラムで表した場合にわかること



参照：「学校保健統計調査 学校保健統計調査-結果の概要（平成30年度）」（http://www.mext.go.jp/component/b_menu/other/__icsFiles/afieldfile/2019/03/25/1411703_03.pdf）

さて、ここでもし「日本の高校3年生の身長について説明してください」と言われたとして、「160cmの人が14%で……」などとくだくだしく述べたら、どうでしょう。伝えるのに時間がかかる上に、不正確です。こういった場合に、統計のモデルを使って説明すると、簡潔かつ正確に伝えることができるのです。

ここで詳しくは触れませんが、先ほどの身長の分布も含め、自然界の多くの数値の分布（ばらつき）は「正規分布」とよばれる統計モデルに当てはめることができます。正規分布は、平均付近が一番高く、平均から離れるにつれ低くなっていく、左右対称の形状を取ることが特徴です。冒頭の説明に正規分布のモデルを使うと、「平均170.6cm、標準偏差（データのばらつきの大きさ）5.87cmの正規分布です」と述べることができます。このように、統計は今あるモデルを使ってうまく「**データを説明する」ための分野**と言い換えることもできるでしょう。

一方、機械学習は「**データを予測する**」ことにフォーカスした分野です。同じく身長の例を挙げると「2050年の日本の高校3年生の平均身長を推測してください」と言われたとき、平均身長の推移を推測できるモデルを即座に思い付く人はいないでしょう。このような場合に、身長推移のデータを入力として、Section05で扱った回帰を利用すると推測が可能になるのです。

■機械学習はデータを予測できる