05 教師あり学習のしくみ

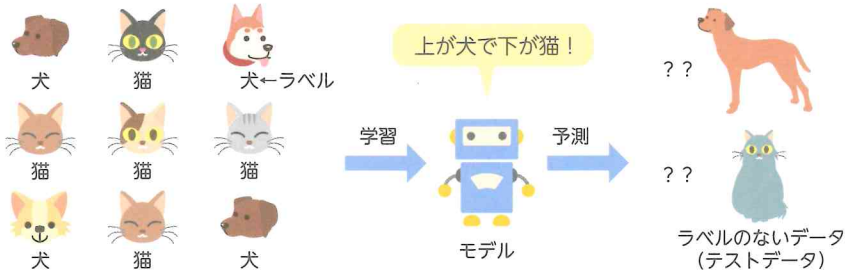
機械学習の1つである教師あり学習は、実際に行われる処理とネーミングのイメージが近いため、比較的理解しやすいといえます。その名の通り、人間がデータのラベルを通じて教師の役割を果たし、機械にお手本を教えてあげる手法のことです。

教師あり学習とは

教師あり学習(supervised learning)とは、正解となる答えが含まれたデータをモデルに学習させる方法のことです。ここでのモデルは、人工知能の脳にあたる部分と考えてかまいません。また、正解となる答えを**ラベル**といい、答えが含まれたデータを**ラベル付きデータ**（もしくは**訓練データ**）と呼びます。教師あり学習はモデルの学習にラベル付きデータを用いますが、最終的な目標はラベルのないデータ（テストデータ）を正解させることです。Section02で学んだ機械学習の例は、教師あり学習に区分されます。

例として、犬と猫の画像を分類する問題を教師あり学習によって解決することを考えましょう。犬と猫の画像一つ一つに、犬か猫かのラベルをあらかじめ人間が付けておきます。モデルは画像とラベルの対応関係を見て、どちらの画像が犬でどちらの画像が猫であるかを学習していきます。最終的に、犬や猫のラベルがなくても画像を見ただけで判断できるようになれば、うまく学習できたといえます。

■特徴量の作り方



分類と回帰

教師あり学習は分類と回帰の2タイプに分けられますが、分類は「識別」とも呼ばれることがあります。Section02では、分類を「データ全体をできるだけ分けるように線を引くこと」、回帰を「回帰はデータ全体にできるだけ重なるように線を引くこと」と解説しました。ここでは、「予測される値（答え）がどのような値をとるか」という観点で分類と回帰の違いを解説します。

まず分類は、答えが「犬/猫」、「小学生/中学生/高校生/大学生」などのカテゴリになっていることが特徴です。ここでいうカテゴリは、①連続した数値ではない（**離散値**）、②大小や順序に意味がないという条件を満たします。しかし、一見して連続した数値が答えのように見える場合でも、その数値がカテゴリ（離散値）とみなせる場合は分類識別となります。たとえば、手書きの1桁の数字が何を表すかを当てるのは分類の問題です。この場合、答えが0/1/2/3/4/5/6/7/8/9のいずれかになります。0.5や2.1といった答えは意味をなしません。また、画像を認識する時点では認識結果の数字が正しいかのみに興味があり、大小関係を意識することはありません。したがって、答えをカテゴリとみなせます。

一方の回帰は、答えが連続した数値（**連続値**）になります。株価の予測問題を考えてみると、答えが12345.6円のような中途半端な値になっても意味が通ります。そのため、株価の予測問題は回帰に分類されます。

■分類と回帰