第1章

決定木の拡張版「ランダムフォレスト」

牧野 浩二

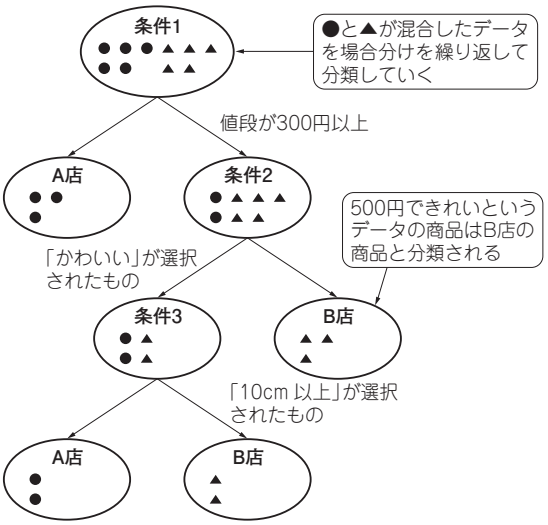

図1　ベースになる場合分けを繰り返して分類する手法「決定木」

決定木を拡張した方法である「ランダムフォレスト」を紹介します．決定木は，名前に「木」が入っているように，木構造を持つ分類方法です．ランダムフォレストは，日本語に無理やり訳すと，「でたらめな森」となります．森を構成する木には決定木を使い，決定木を作るときはデータをランダムに選びます．

ランダムフォレストは，普通の決定木に比べて多くのデータが必要となります．そこで本章では，IRIS（アイリス）データという，データの分類でよく使われる「あやめ」データと，Boston（ボストン）データという「ボストンの家の価格」データを使うことにしました．この2つのデータや統計分析ソフト「R」を使って決定木とランダムフォレストで分類問題と回帰問題を解いてみます．

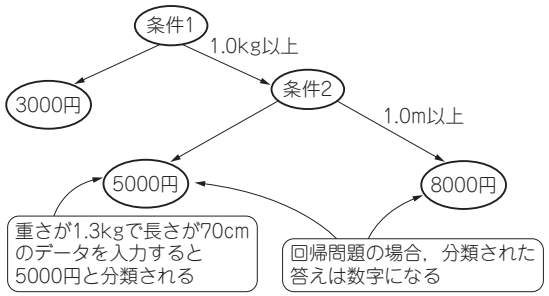

図2　回帰問題は分類された答えが数字

決定木がベース

● できること

▶その1：分類

分類問題とは，例えるとデータを入れるとどの店の商品かを当てるような問題です．決定木は，**図1**のように，多くのデータから分類するためのルールを作ります．分類のルールは，例えば**図1**では値段が「300円以下」ならば左に分類，「かわいい」ならば右に分類とした場合，A店で買った商品かB店で買った商品かを分類するためのルールができたことになります．決定木は，新しいデータを用意してこのルールで分類し，どれになるかで答えを出す方法です．実際に分類してみると，例えば「500円」で「きれい」というデータの商品は，B店の商品であると分類されます．

▶その2：回帰

回帰問題とは，例えるとあるデータを入れたらそのものの価値を当てるような問題です．**図2**に示すように回帰問題も分類問題と同じように木構造を作りますが，分類された答えが，数字になっています．例えば「重さ1.3kg，長さ70cm」のデータを入力した場合，5000円として分類されます．