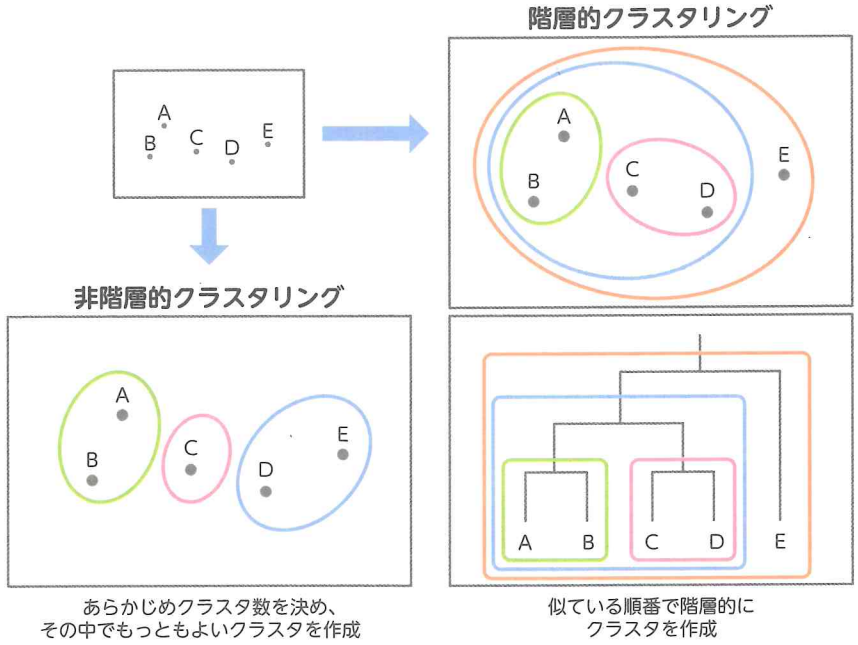
教師なし学習は「クラスタリング」ができる

教師なし学習で実現できるタスクとして代表的なのは「**クラスタリング**」です。クラスタリングはデータの中から特徴の似ているデータをグループ（クラスタ）ごとに分けるタスクです。先ほどの例で見ると、クラスタリングは野菜や果物をどの観点で見るとうまく分けられるのかを考えることにあたります。そんなクラスタリングの手法には大きく分けて「**階層的クラスタリング**」と「**非階層的クラスタリング**」の2種類があります。階層的クラスタリングとは、特徴の似ているクラスタ同士を1つずつ結合させていき、最終的に1つの大きなクラスタになるまでくり返すことでクラスタリングを行う手法です。対して非階層的クラスタリングとは、初めにクラスタ数（下図では3つ）を設定し、そのクラスタ数でもっともよくデータを分けることができるようクラスタリングを行う手法です。なお本書では、Section31にて非階層クラスタリングの代表的なアルゴリズムである「k平均(k-means)法」を紹介しています。

■階層的クラスタリングと非階層的クラスタリング

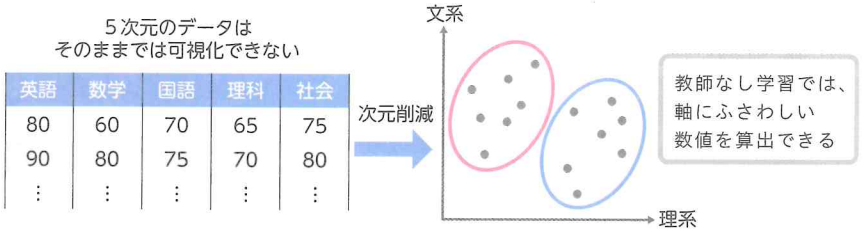


教師なし学習は「次元削減」ができる

教師なし学習において、クラスタリングの次に代表的なタスクといえば「**次元削減**」でしょう。次元削減は、データから重要な情報だけを抜き出し、あまり重要でない情報を削減するタスクです。ここでの次元とは、データの項目の数です。たとえば、ある中学生1人のデータとして英語・数学・国語・理科・社会の成績という5つの項目があるならば、5次元のデータとなります。

次元削減の一例としては、データの可視化が挙げられます。私たちが多次元データを直感的に理解するためには、データの次元を人間が視認することができる3次元以下に落とした上で可視化しなくてはなりません。たとえば、中学生の5教科の成績のデータがたくさん集まったとします。このとき横軸に「数学の点数」、縦軸に「国語の点数」として2次元グラフを書くことで、グラフの形から、このデータが「文系」と「理系」の2クラスタから構成されていると推測することができるでしょう。しかし、縦軸として一番ふさわしいのは「英語と国語の合計点」や「英語:国語:社会を2:2:1の割合で加えた点数」かもしれません。教師なし学習で次元削減を行えば、データの特徴がわかりやすい軸を求めることができ、有効なデータの可視化を行うことができます。次元削減については、Section32にてより丁寧に取り上げています。

■次元削減



まとめ

- **教師なし学習の最終目標は「データの特徴をとらえる」**
- **教師なし学習は「クラスタリング」と「次元削減」ができる**