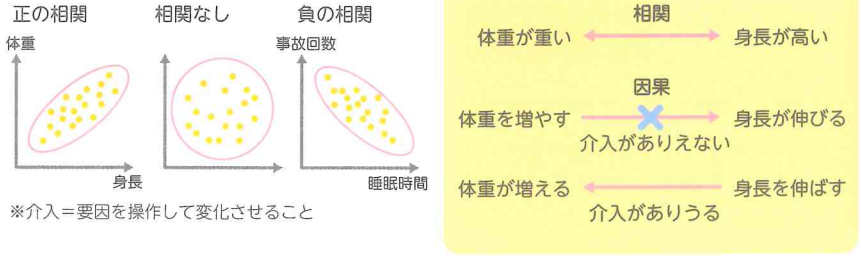
21 相関と因果

データから相関関係を導くことは比較的かんたんですが、因果関係を導くのは難しいことです。データを利用する機械学習においては、両者の区別は非常に重要です。ここでは両者の違いのほか、データから因果関係を分析する手法も解説します。

相関関係と因果関係

まずは相関関係について知っておきましょう。相関関係とは、「ある変数が大きいときに、他の変数も大きい」「ある変数が大きいときに、他の変数は小さい」といった関係のことです。前者を正の相関関係、後者を負の相関関係といいます。たとえば、身長が高い人ほど体重も重い傾向にあるため、身長と体重は正の相関があると言えます。次に因果関係とは、「ある変数を変化させたときに、他の変数も変化する」関係のことです。相関関係とは別物であることを理解しましょう。「身長が高い人ほど体重が重い」という相関関係においては、「身長を伸ばせば体重が重くなる」因果関係は正しいかもしれません。しかし、「体重を増やせば身長が伸びる」因果関係は正しくありません。体重が増えても身長は伸びず、ただ太っていくだけになってしまいます。統計学（や機械学習）では相関関係を分析することはできますが、相関関係だけで因果関係を結論づけるには無理があります。

■ 相関関係と因果関係



疑似相関

疑似相関とは、「本当は因果関係がない要素の間に、見えない外部の要因の影響で因果関係ができているように見えること」です。この見えない外部の要因のことを「交絡変数」「交絡因子」「共変数」と呼ぶこともあります。たとえば、「スキーをする人が多い時期は暖房器具を買う人が多い」という関係は疑似相関と言えます。ここでの交絡変数としてはたとえば気温が考えられるでしょう。もし気温が低いなら、暖房器具を買う人は多くなるはずです。また、気温が低いと雪が多くなり、スキーをする人も多くなります。そのほか、「小学生の計算テストの成績と50m走のタイム」の関係も疑似相関の可能性が高いと言えます。この場合であれば、交絡変数は学年です。学年が上がれば計算の能力は向上し、50m走のタイムも上がるのは自然であると言えるでしょう。

ただし「スキー人口を増やせば、暖房器具の需要が高まる」「計算テストの成績を上げれば、50m走のタイムも向上する」のように因果関係に無理やり結びつけてしまうと、かえって本質を見誤ってしまいます。

■ 疑似相関