数値データの整形

・離散化

離散化は連続した値をある区分に分けることです。遊園地の来場者数を予測するとき、データの値に来場者の年齢があるとして、この年齢を年代別に分けるのが離散化です。来場者の年齢をそのまま使うと、1歳差のデータであっても別の値とみなされるため、データの特徴を表す量としてふさわしくない場合があります。年代別に大きく区分すれば、年齢のわずかな違いを吸収できます。

・対数変換

対数変換は値のlog（対数）を取る（logに変換する）ことです。正の値を持つ数値データにおいて、長い裾を短く圧縮し、小さい値を拡大することができます。機械学習では正規分布（きれいな山型）に近いデータが効果を発揮しやすいため、対数変換は有効な手段の一つです。

・スケーリング

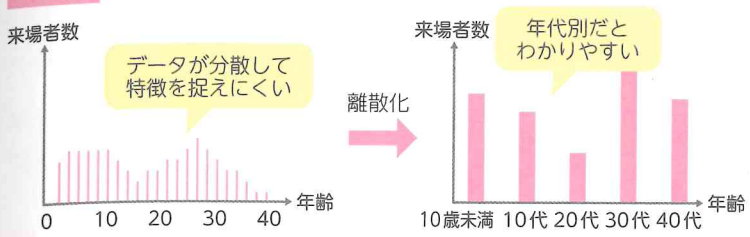
スケーリングは値の範囲を変換することです。データによっては値の取りうる範囲が決まっていない場合があります。たとえば、遊園地の来場者数は取りうる上限が決まっていません。しかし、線形回帰やロジスティック回帰などのアルゴリズムは値の大きさに影響されやすいため、値の範囲を変換する必要があります。代表的なスケーリングの方法にMin-Maxスケーリングと標準化があります。Min-Maxスケーリングは最小値を0、最大値を1にし、データの範囲を0〜1にすることです。標準化は値の平均を0、分散を1にすることです。なお、上で紹介した対数変換の後に標準化を行う場合もあります。

まとめ

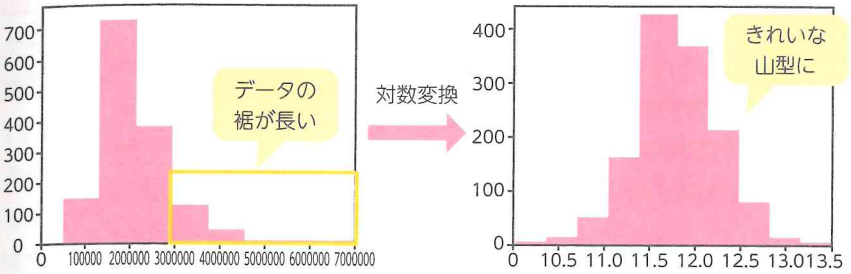
- データや使用するアルゴリズムを考慮して、整形方法を選ぶ

■数値データの整形

離散化



対数変換



スケーリング