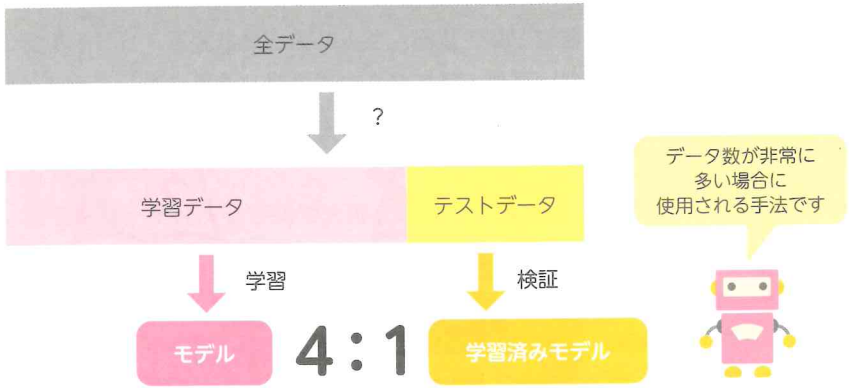
ホールドアウト検証とK-分割交差検証（K-foldクロスバリデーション）

ホールドアウト検証とは、データを学習用データとテスト用データにある割合で分割して検証する、もっとも単純な検証方法です。学習に使用するデータ数はモデルの性能に直結するためなるべく多いほうがよいですが、検証に使用するデータ数が少なすぎると未知のデータのいろいろなパターンを模倣することができなくなります。データ数が膨大な場合は、後述の交差検証をするとコンピュータの処理速度などにより学習・検証に時間がかかってしまうため、一度の学習・検証で済むホールドアウト検証が用いられます。一般に学習・テストデータの割合としては２：１や４：１、９：１などが多く使われます。

■ ホールドアウト検証

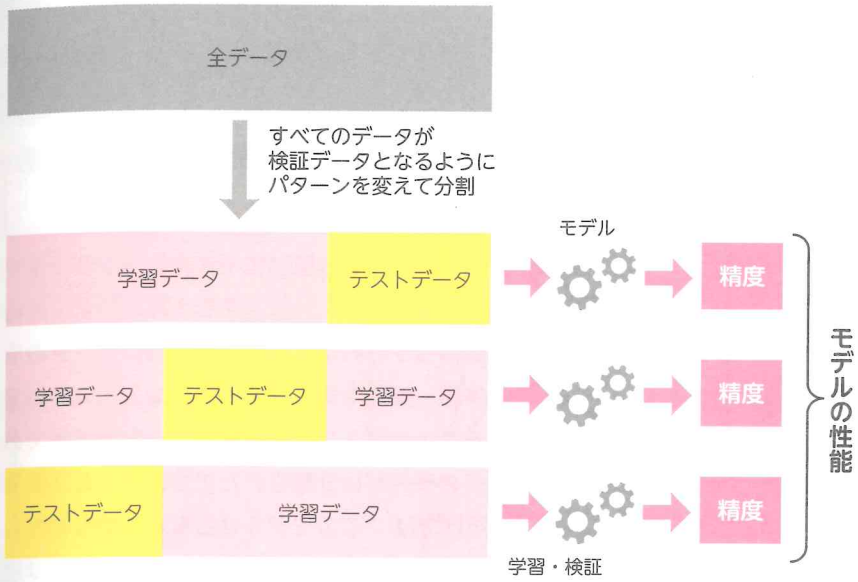


このホールドアウト検証では、全データの一部だけをテストデータに使用しますが、テストデータの選び方に偏りがある場合などに、正確に検証できないことがあります。

そこで用いられるのが、**K-分割交差検証（K-fold クロスバリデーション）**です。この手法では、すべてのデータが検証データとして利用されるよう、学習データとテストデータを入れ替えて分けるなどし、複数の組み合わせを用意します。その上で、それぞれのデータを使用して学習と検証を別々に行い、それらの検証結果から、総合的にモデルの性能を検証するのです。

K-分割交差検証は、ホールドアウト検証に比べ、パターン数によって3倍～10倍程度の計算資源が必要となりますが、現在もっとも広く使われている検証方法です。

■ K-分割交差検証



そのほかの手法としては、**Leave-one-out 交差**があります。全データから1データずつ抜き出してテストデータとし、残りすべてを学習データとする手法です。これにより得られるデータ数に応じたパターンすべてを学習・検証します。その結果から総合的にモデルの精度を検証するのです。すでに紹介した2つの手法と比較して多くの学習データを取れるため、モデルの精度向上が見込めます。しかし、データ数に比例して計算量が増大するため、近年ではデータ数が多くない場合にのみ利用される傾向にあります。

まとめ ▶ 予測結果の検証には、K-分割交差検証が良く使われる