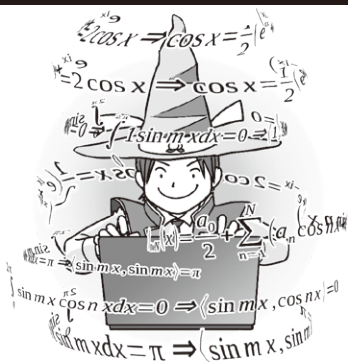




CPUに関するデータ・セットから コンピュータの性能を決める要因を分析する **線形代数 × データ・サイエンス!** **Pythonで回帰分析にトライ**

宮下 修人

本稿では、データ・サイエンスの世界で使われている線形代数の応用例として、一般化逆行列を使った線形回帰モデルのパラメータ最小2乗推定について解説します。

回帰モデルは、複数のデータの間にある関係を説明したり、あるデータの値が分かっているときに別のデータがどのような値になるのか予測したりするための統計モデルです。このようなデータに基づく予測はアマゾンなどのECサイトで使われている、商品のレコメンデーション・システムなどにも応用できます。

Python + NumPyを使用することで、スクリプトを数式に近い形式で記述できるので、本稿ではその方法を解説します。

また、データ分析の実施例として、メモリ・サイズなどのCPUに関するデータを使って、コンピュータの性能を決める要因を調べる分析の手順についても解説します。コンピュータの性能を決める要因について説明する線形回帰モデルを構築し、一般化逆行列を使ってこのモデルの未知パラメータを求めてみます。

これからの組み込みエンジニアには データ・サイエンスの知識が必須？

● ハードウェアの進化とデータ・サイエンス

半導体技術の進歩によるハードウェア性能の向上によりもたらされた大規模かつ複雑なデータを利用するために、データ・サイエンスが注目されています。

データ・サイエンスの知識は、データ計測のための機器設計や開発でも必要になります。そのため、エレクトロニクス分野のエンジニアもデータ・サイエンスに関する知識が要求される機会が増えていくと思われます。そのときに武器になるのが、基本的な数学の知識です。

● データ・サイエンスに要求される知識

データ・サイエンスの世界では高度な数学が使われることもあります。データ分析のためのソフトウェ

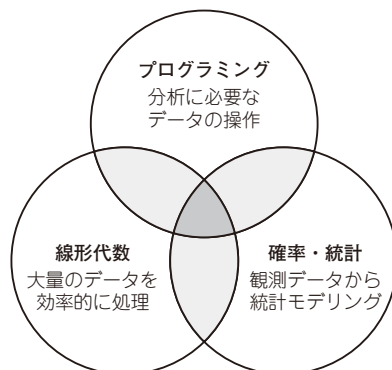


図1 データ分析のためのソフトウェア・ライブラリを正しく利用するために必要なスキル
ただし、この図には描かれていないが、データ・サイエンスの理論を理解するためには微分、積分の知識も必要になる

アやライブラリを正しく利用するときに重要なのは、基本的な確率・統計と線形代数、そしてプログラミングの能力です(図1)。データ・サイエンスでは、これらの能力を組み合わせ、観測データから有益な洞察を引き出すことが求められます。

● 確率・統計と線形代数の間の考え方の差

確率・統計と線形代数の間には、対象とするデータへの考え方に隔たりがあります。

例えば、統計学はその発展の歴史的経緯から、母集団から抽出された小標本を対象とした分析を中心に説明される傾向があります。一方で、線形代数が威力を発揮するのは、大規模な計算を効率的に行う場合なので、確率・統計とは想定するデータの規模が大きく異なります。

このような確率・統計と線形代数の間の考え方の違いは、初学者を悩ませる原因になります。そのため、まずはデータ・サイエンスにおける線形代数の利用についての具体的なイメージを持つことが重要です。本稿ではその一例として、一般化逆行列を使って線形回帰モデルのパラメータの最小2乗推定について解説します。