



普通教科「情報」とデータサイエンス

東京理科大学 理学部第一部 数学科 教授 清水^{しみず} 克彦^{かつひこ}

1. 新指導要領における情報科のもう一つの特徴

「情報科」の新指導要領には、新しい特徴があります。それは、共通教科「情報」においてデータサイエンスの内容が積極的に取り入れられている点です。「高等学校学習指導要領解説（平成30年告示）解説情報編」（開隆堂、2019）では、まず、「情報Ⅰ」では目標において「(1) 効果的なコミュニケーションの実現、コンピュータやデータの活用について理解を深め技能を習得するとともに、情報社会と人との関わりについて理解を深めるようにする。(p.22)（下線筆者）」が掲げられています。内容においても「(4) 情報通信ネットワークとデータの活用」のなかで、「ア、(ウ) データを表現、蓄積するための表し方と、データを収集、整理、分析する方法について理解し技能を身に付けること。(p.35)」ならびに「イ、(ウ) データの収集、整理、分析及び結果の表現の方法を適切に選択し、実行し、評価し改善すること。(p.35)」が挙げられています。「情報Ⅱ」では目標に「(1) 多様なコミュニケーションの実現、情報システムや多様なデータの活用について理解を深め技能を習得するとともに、情報技術の発展と社会の変化について理解を深めるようにする。(p.41)」が掲げられ、内容においては「(3) 情報とデータサイエンス」において、「多様かつ大量のデータを活用することの有用性に着目し、データサイエンスの手法によりデータを分析し、その結果を読み取り解釈する活動を通して、次の事項を身に付けることができるよう指導する。(p.48)」と挙げ、データサイエンスが社会に果たす役割について理解し、目的に応じた適切なデータの収集や整理、整形について理解し技能を身に付けること、データサイエンスが社会に果たす役割について理解し、目的に応じた適切なデータの収集や整理、整形について理解し技能を身に付けること、データ処理の結果を基にモデルを評価することの意義とその方法について理解し技能を身に付けること、目的に応じて、適切なデータを収集し、整理し、整形すること、将来の現象を予測したり、複数の現象間の関連を明らかにしたりするために、

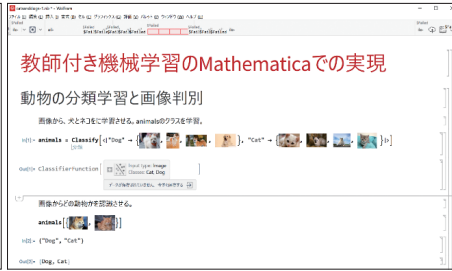
適切なモデル化や処理、解釈・表現を行うこと、モデルやデータ処理の結果を評価し、モデル化や処理、解釈・表現の方法を改善すること、などを挙げて、データサイエンスを本格的に扱っています。このようなことから、情報科の学習指導要領解説のなかには数学科との連携を記述する箇所が、17カ所もあります。

さて、実際の教科書を見ると、実教出版の「最新情報Ⅰ」、「情報Ⅱ」を参考にすると、「情報Ⅰ」では「データの活用」において尺度水準、外れ値、欠損値、最頻値、分散、標準偏差、バブルチャート、散布図、ヒストグラム、箱ひげ図、クロス集計、相関係数、移動平均、標本調査、単回帰分析、回帰直線、決定係数、交絡因子などが取り上げられています。「情報Ⅱ」ではデータサイエンスのなかで、ビッグデータ、探索的データ分析、偶然誤差と系統誤差、五数要約、標本、母集団、標本平均・分散、母平均・分散、確率変数、確率分布、点推定、不変分散、区間推定、尤度、最尤法、検定（対立仮説、帰無仮説、有意水準、両側・片側）、機械学習、教師あり学習（回帰・分類）、教師なし学習（クラスタリング、次元削減、異常検知）、単回帰モデル・多項式回帰モデル、回帰分析、ロジスティック回帰、k-近傍法、ニューラルネットワーク、クラスタリング k-means 法、過学習など非常に深い内容をどんどん紹介しています。

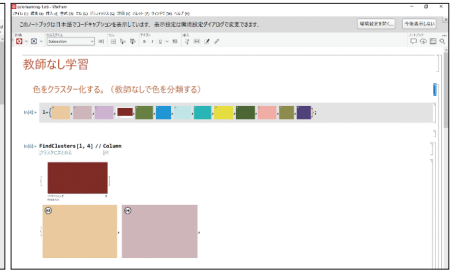
文科省も今までにないデータサイエンスの内容、また、情報科の教員もあまり習っていない内容があるため、「高等学校情報科のための特設ページ (https://www.mext.go.jp/a_menu/shotou/zyouhou/detail/1416746.htm)」を設けて、指導ビデオや教材を提供しています。また、統計局も「統計指導のために」(<https://www.stat.go.jp/teacher/comp-learn-04.html>)を用意しています。このサイトでは「高等学校における「情報Ⅱ」のためのデータサイエンス・データ解析入門」を設けて、pythonのプログラムやデータを提供しています。先生方もご活用ください。



【図1】 MDASHのロゴ



【図2】 教師つき学習のMathematicaによる実行



【図3】 教師なし学習のMathematicaによる実行

2. データサイエンス強調の背景

このような積極的な推進の背景の一つにMDASHがあります。MDASHとはMathematics, Data Science and AI Smart Higher Education (スマート高等教育における数学・データサイエンス・AI)の略で文科省が進める「数理・データサイエンス・AI教育プログラム認定制度」で示されるAI時代のリテラシーを指しています【図1】。これは全国の高等教育機関(大学・高等専門学校)で数理・データサイエンス・AI教育強化拠点コンソーシアムが定めたモデルカリキュラムの示すリテラシーを指します。ほぼすべての高等教育機関で必修科目として履修されています。理科大では「AI・データサイエンス」という科目で全学生が履修しています。これはリテラシーレベルで年間50万人(大学・高専は100%、高校)、応用基礎レベルで年間25万人(大学・高専は50%、高校の一部)の習得を目指すもので、1. 社会におけるデータ・AI活用、2. データリテラシー、3. データ・AI活用における留意事項について、学ぶこととなります。この中核の一つがデータサイエンスであり、情報Ⅰと情報Ⅱにおけるデータサイエンスの内容は、このMDASHの示すリテラシーの高大接続のための内容と言えます。

しかし、先に挙げた様々なデータサイエンスの内容は当然、コンピュータの使用によって支えられ、pythonのプログラミングなどが前提となっています。pythonには様々なモジュールが用意されていて便利ではありますが、やはり、高校生にとっては困難であることは否めません。そこで今回はローコード(少ないコード)でのデータサイエンスの実行を紹介したいと思います。

3. ローコードでデータサイエンス

今回、紹介させていただくのはMathematicaを使って、ローコードで機械学習を演示するものです。機

械学習とは実教出版の「情報Ⅱ」では「データに潜むパターンや構造をデータから自動的に学習する方法(p.40 [0010 1000])」(この教科書ではページ数を2進数でも振っています。 $2^5+2^3=32+8=40$)です。教師あり学習(supervised learning)とは「入力とラベルの対の集合の訓練データによって、入力となるデータから出力となるデータを予測するモデルを学習すること」で、回帰もしくは分類の問題となります。

今回の例は、ごくごく少ない例の犬と猫の画像にDogとCatのタグをつけた訓練データからanimalというモデルを作る教師あり学習を行い、新しい犬と猫の画像がDogかCatかを分類させるものです。その実行結果は上ようになります【図2】。

実際に書いているコードは、Classifyという命令とanimalsに新しい画像を入力し、DogかCatのタグのどちらかを分類させているだけです。たった、4つの画像の2つの集合の訓練データですが、上手く分類してくれています。このような演示を見て、高校生は教師つき機械学習のイメージが湧くのではないのでしょうか。高校生にpythonのモジュールを実行させるよりはイメージが湧くような気がします。

次は、教師なし学習の演示を紹介したいと思います。教師なし学習(unsupervised learning)とは「入力のデータに対するラベルが存在せず、入力のデータ集合のみをもとにそれらのパターンや構造をよく表すようなモデルを学習すること」です。今回は、色のクラスタリングを行う例を紹介します。色見本の集合を与え、色見本はRGBの値のみで定義できるので、それを4つのクラスタにするように求めたものが上の図です【図3】。

1に色見本の集合を定義し、FindCluster[1,4]//Columnと4つのクラスタに分けるように求めた結果です。最初の2つのクラスタまでの図を示しました。これも直観的に分かりやすいと思います。先生方の役にたつ事例の紹介になりましたら幸いです。