情報Ⅰ,Ⅱでのデータサイエンスの扱い

滋賀大学データサイエンス学部長
竹村　彰通

2022年2月4日
教育家庭新聞＆レノボ・ジャパンセミナー

資料 (公的なもの)

1. 国立大学協会「２０２４年度以降の国立大学の入学者選抜制度　－国立大学協会の基本方針－」
   https://www.janu.jp/wp/wp-content/uploads/2022/01/20210128_news_001.pdf
2. 文部科学省高等学校学習指導要領（平成30年告示）解説「情報編」
   https://www.mext.go.jp/content/1407073_11_1_2.pdf
3. 文部科学省高等学校情報科「情報Ⅰ」教員研修用教材
   https://www.mext.go.jp/a_menu/shotou/zyouhou/detail/1416756.htm
4. 文部科学省高等学校情報科「情報Ⅱ」教員研修用教材
   https://www.mext.go.jp/a_menu/shotou/zyouhou/detail/mext_00742.html
5. 大学入試センター「情報サンプル問題」
   https://www.dnc.ac.jp/kyotsu/shaken_jouhou/r7ikou.html

- 以下ではこれらの資料については簡単に紹介する。
- 実際の教科書は今後検討する。

2022年1月28日　国立大学協会会長談話 (文献1関係)

２０２２年度から始まる高等学校の新学習指導要領では「情報I」が全ての生徒が学ぶ必履修科目として履修され、一方、国立大学においても既に多くの大学で、「数理・データサイエンス・AI 教育」が文理を問わず全ての学生が身に付けるべき教養科目として履修されています。このような中において「情報」に関する知識については、国立大学の教育を受ける上で必要な基礎的な能力の一つとして位置付けられていくと考えています。それらを踏まえ、今回の基本方針では、一般選抜において、これまでの「5 教科 7 科目」に「情報」を加えた、「６教科８科目」を課すことを原則としています。

指導要領解説(文献2)　情報Ⅰ

項目

（1）情報社会の問題解決
（2）コミュニケーションと情報デザイン
（3）コンピュータとプログラミング
（4）情報通信ネットワークとデータの活用

（4）情報通信ネットワークとデータの活用

情報通信ネットワークを介して流通するデータに着目し，情報通信ネットワークや情報システムにより提供されるサービスを活用し，問題を発見・解決する活動を通して，次の事項を身に付けることができるよう指導する。

ア　次のような知識及び技能を身に付けること。

(ｱ) 情報通信ネットワークの仕組みや構成要素，プロトコルの役割及び情報セキュリティを確保するための方法や技術について理解すること。

(ｲ) データを蓄積，管理，提供する方法，情報通信ネットワークを介して情報システムがサービスを提供する仕組みと特徴について理解すること。

(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。

イ　次のような思考力，判断力，表現力等を身に付けること。

(ｱ) 目的や状況に応じて，情報通信ネットワークにおける必要な構成要素を選択するとともに，情報セキュリティを確保する方法について考えること。

(ｲ) 情報システムが提供するサービスの効果的な活用について考えること。

(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。

(情報Ⅰの解説の最後の部分)

具体的に，気温や為替などの変動，匿名化したスポーツテストの結果やオリンピック・パラリンピックの記録などのデータを分析する学習活動を行う場合，グラフや表などを用いてデータを可視化して全体の傾向を読み取ったり，問題を発見したり，予測をしたりすることが考えられる。その際，データの形式や分析目的に応じた可視化の方法を選択する学習活動を通して，相関係数などの統計指標，相関関係や因果関係などのデータの関係性，調べようとするもの以外で結果に影響を与えている原因である交絡因子，データの関係性を数式の形で表す単回帰分析などについて扱うことが考えられる。

データを分析する過程については，データの分析を容易にするために必要な計算を事前に行っておくなど，データの傾向などを読むことを容易にする工夫を行う力を養うことが考えられる。更に，データを分析及び可視化するために適切なソフトウェアを活用する学習活動を通して，多くの項目のあるデータに対して，項目間の相関を見るためにデータを漏れのないように組み合わせて複数の散布図などを作成し，相関関係の見られる変数の組合せを見出し，その変数の組合せに関して回帰直線を考え，データの変化を予測する力を養うことが考えられる。

指導要領解説(文献2)　情報Ⅱ

（1）情報社会の進展と情報技術
（2）コミュニケーションとコンテンツ
（3）情報とデータサイエンス
（4）情報システムとプログラミング
（5）情報と情報技術を活用した問題発見・解決の探究

(3) 情報とデータサイエンス

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。

ア　次のような知識及び技能を身に付けること。
　(ｱ) 多様かつ大量のデータの存在やデータ活用の有用性，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。
　(ｲ) データに基づく現象のモデル化やデータの処理を行い解釈・表現する方法について理解し技能を身に付けること。
　(ｳ) 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。

イ　次のような思考力，判断力，表現力等を身に付けること。
　(ｱ) 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。
　(ｲ) 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。
　(ｳ) 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。

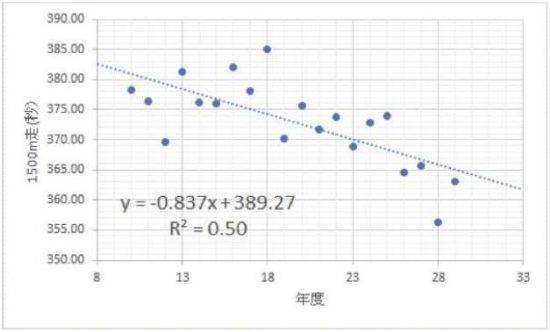
ここで行うデータの処理に関しては，回帰，分類，クラスタリング及びそれらがどのような場面で活用されているか，これらを応用して人間が判断や意思決定を行う代わりにデータを基にどのような仕組みでコンピュータが判断を行っているかを理解するようにする。回帰に関しては，重回帰分析などについて扱い，そのモデルを変更することによって結果がどのように変化するか，分類に関しては，条件付確率，近傍法，木構造などを用いた予測について扱い，これらの手法や技術がどのような場面に活用されているか，それぞれ適切なソフトウェアの活用を通して理解するようにする。全体を共通の特徴を持ったいくつかの集団に分割するクラスタリングに関しては，似たものを集団にしていく階層的方法と，集団の数を決めてから要素を所属させていく非階層的方法などについて扱い，適切なソフトウェアの活用を通して理解するようにする。その際，適切な活用場面についても考えるようにする。

ここでは，数学科における学習内容と関連する部分も含むが，数学や統計学の専門的な内容に深入りすることなく，可視化やソフトウェアによる処理の結果を基に，その概念を理解するようにする。

実際には大学でもなかなか全部はカバーしていない

「情報Ⅰ」教員研修用教材(文献3)での扱い

学習22 量的データの分析
(1)量的データと質的データ
(2)量的データの関係
(3)単回帰分析を用いた値の推測
(4)量的データの統計的仮説検定



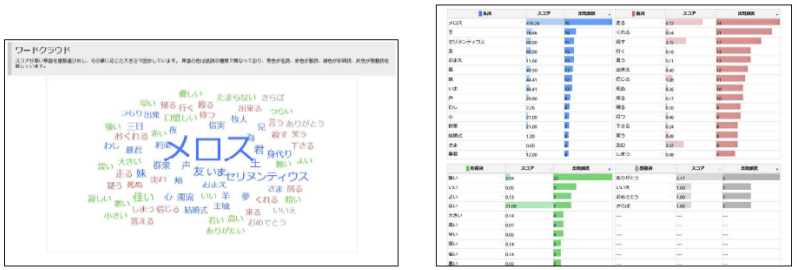
図表 9-A　散布図に 1500m 走の記録の予測モデル (回帰直線) を入れた図

学習23　質的データの分析

(1)質的データの種類とその扱い
(2)テキストデータの扱いについて
(3)テキストデータの可視化
(4)テキストの分析とその可能性

（3）テキストデータの可視化

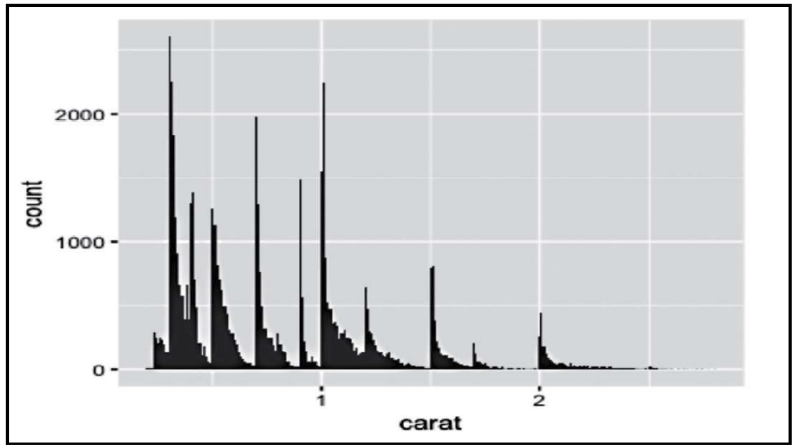
ここでは日本語のテキストマイニングの基本について考えて，実際に実習をしてみよう。株式会社ユーザーローカルでは，様々なデータ分析のツールを Web ベースで提供している。ソフトウェアのインストールなしにデータさえ用意すれば，手軽に利用できるので，授業でも活用できる。今回は，（株）ユーザーローカルのユーザーローカルテキストマイニングツール (https://textmining.userlocal.jp/) を使ってみる。以下のデータはユーザーローカルテキストマイニングツールにサンプルとして掲載されている，太宰治『走れメロス』のデータを分析した結果である。



図表 2　ワードクラウド
出典：ユーザーローカルテキストマイニングツール (https://textmining.userlocal.jp/)

学習24　データの形式と可視化

(1)質的データとその種類
(2)データの分析と可視化
(3)データの可視化と問題発見



図表 6　グラフの例（ダイアモンドのカラット数）
転載：「R ではじめるデータサイエンス」P78 グラフ（(株）オライリー・ジャパン）

「情報 II 」教員研修用教材(文献4)での扱い

第3章 **情報とデータサイエンス** ····· 105

大学のデータサイエンス学部で学ぶような内容になっている。

学習17の一部

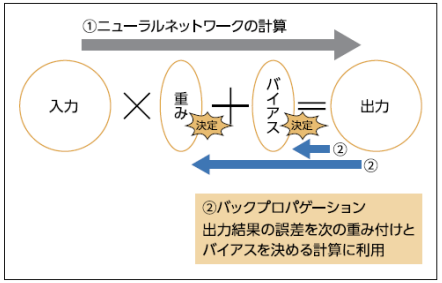
(3)バックプロパゲーションによる学習

バックプロパゲーション(誤差逆伝播法)は前項(2)勾配降下法をもとにしたニューラルネットワークの学習で最適な重みとバイアスを探す代表的な手法である図表14。

(4) オーバーフィッティングとその防止

オーバーフィッティング(過学習)とは学習の際に特定のデータにだけ過剰に対応し，学習に用いていない他のデータでは正しくならない状態のことである。

これらの技術をPython等のプログラミング言語で実装することもできるが，ここでは，ニューラルネットワークを容易に構築するためにNeural Network Console（SONY）を利用する。Neural Network Consoleには，Webブラウザで動作するクラウド版とWindowsアプリ版がある。このサイトを活用して，学習15でも扱った手書き数字データセットMNISTをニューラルネットワークで学習させてみよう。



図表 14　学習とバックプロパゲーション

「情報サンプル問題」(文献5) 第3問がデータサイエンス
(大問3問のうちの一つ)

第３問 次の文章を読み，後の問い(**問１～４**)に答えよ。

Ｓ高等学校サッカー部のマネージャーをしている鈴木さんは，「強いサッカーチームと弱いサッカーチームの違いはどこにあるのか」というテーマについて研究している。鈴木さんは，ある年のサッカーのワールドカップにおいて，予選で敗退したチーム（予選敗退チーム）と，予選を通過し，決勝トーナメントに進出したチーム（決勝進出チーム）との違いを，データに基づいて分析することにした。このデータで各国の代表の 32 チームの中で，決勝進出チームは 16 チーム，予選敗退チームは 16 チームであった。

分析対象となるデータは，各チームについて，以下のとおりである。

- 試合数…大会期間中に行った試合数
- 総得点…大会で行った試合すべてで獲得した得点の合計
- ショートパス本数…全試合で行った短い距離のパスのうち成功した本数の合計
- ロングパス本数…全試合で行った長い距離のパスのうち成功した本数の合計
- 反則回数…全試合において審判から取られた反則回数の合計

鈴木さんは，決勝進出チームと予選敗退チームの違いについて，このデータを基に，各項目間の関係を調べることにした。データの加工には，表計算ソフトウェアを活用し，表１のデータシートを作成した。

決勝進出チームと予選敗退チームの違いを調べるために，決勝進出の有無は，決勝進出であれば 1，予選敗退であれば 0 とした。また，チームごとに試合数が異なるので，各項目を 1 試合当たりの数値に変換した。

表 1　ある年のサッカーのワールドカップのデータの一部（データシート）

	A	B	C	D	E	F	G	H	I	J	K
1	チーム ID	試合数	総得点	ショートパス本数	ロングパス本数	反則回数	決勝進出の有無	1 試合当たりの得点	1 試合当たりのショートパス本数	1 試合当たりのロングパス本数	1 試合当たりの反則回数
2	T01	3	1	834	328	5	0	0.33	278.00	109.33	1.67
3	T02	5	11	1923	510	12	1	2.20	384.60	102.00	2.40
4	T03	3	1	650	269	11	0	0.33	216.67	89.67	3.67
5	T04	7	12	2257	711	11	1	1.71	322.43	101.57	1.57
6	T05	3	2	741	234	8	0	0.67	247.00	78.00	2.67
7	T06	5	5	1600	555	9	1	1.00	320.00	111.00	1.80

また，データシートを基に，統計処理ソフトウェアを用いて，図 1 を作成した。

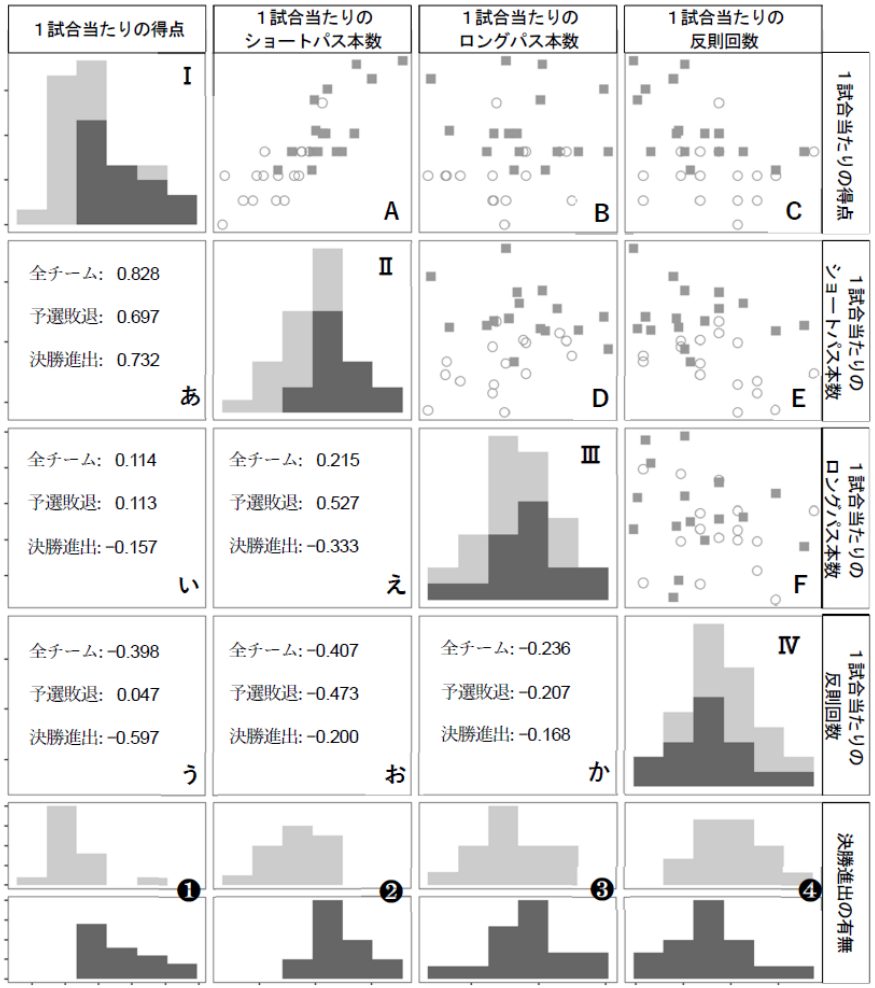


図１　各項目間の関係

図１の**Ⅰ**～**Ⅳ**は，それぞれの項目の全参加チームのヒストグラムを決勝進出チームと予選敗退チームとで色分けしたものであり，❶～❹は決勝進出チームと予選敗退チームに分けて作成したヒストグラムである。**あ**～**か**は，それぞれの二つの項目の全参加チームと決勝進出チーム，予選敗退チームのそれぞれに限定した相関係数である。また **A**～**F** は，それぞれの二つの項目の散布図を決勝進出チームと予選敗退チームをマークで区別して描いている。例えば，図１の **A** は縦軸を「１試合当たりの得点」，横軸を「１試合当たりのショートパス本数」とした散布図であり，それに対応した相関係数は**あ**で表されている。

問１　次の問い（ａ・ｂ）に答えよ。

ａ　次の文章を読み，空欄［ア］～［ウ］に入れる最も適当なものをそれぞれの解答群のうちから一つずつ選べ。ただし，空欄［ア］・［イ］の順序は問わない。

図１を見ると，予選敗退チームにおいてはほとんど相関がないが，決勝進出チームについて負の相関がある項目の組合せは，１試合当たりの［ア］と［イ］である。また，決勝進出チームと予選敗退チームとで，相関係数の符号が逆符号であり，その差が最も大きくなっている関係を表している散布図は［ウ］である。したがって，散布図の二つの記号のどちらが決勝進出チームを表しているかが分かった。

［ア］・［イ］の解答群

⓪ 得点　　① ショートパス本数　　② ロングパス本数　　③ 反則回数

［ウ］の解答群

⓪ A　　① B　　② C　　③ D　　④ E　　⑤ F

まとめ

- 大学入学共通テストの情報Ⅰは多くの国立大学で使われる。
- 情報Ⅰでは相関や回帰が扱われる。テキストデータも扱われる。
- 情報Ⅱでは大学レベルの機械学習やAI手法が扱われる。
- 国際的に見ても高度な内容を含んでいる。
- 現実には、大学でも教える教員が不足している。高等学会で教えるのはかなり困難に思われる。