ARTICLE

応専

情報を専門としない学部・学科における情報科学教育，統計科学教育の現状と今後の展開
―2015 年度優秀教育賞における取り組みを踏まえて―

石井一夫
東京農工大学

文部科学省の特別経費による支援を受け 2011 年度から 2015 年度まで，東京農工大学において「農学系ゲノム科学領域における実践的先端研究人材育成プログラム」が実施された[1)]．筆者はこの人材育成プログラムにおいて，主に情報科学，統計科学教育に関する実践活動を担当し，その成果に関して，2016 年 6 月に本会 2015 年度優秀教育賞を授与された．本稿では，これらの人材育成活動に関する総括と，情報科学を専門としない学部，学科における情報科学，統計科学に関する今後の教育について考察する．

情報科学を専門としない学部・学科におけるデータサイエンス教育の展開

近年，ビッグデータ，人工知能，データマイニング，機械学習などデータサイエンスに関する話題がメディアに登場することも多い．書店でも統計科学を含むその関連書籍が山積みになっており，セミナ，勉強会も盛況となっている．筆者の所属する農学部，あるいは医学部，薬学部など情報科学を専門としていない学部においてもその必要性は増している．データ分析やその基礎となるプログラミング，アルゴリズム，データベース，統計科学を学習していない環境におけるデータサイエンス教育について今回の経験から考察したい．

情報処理を専門としない学部・学科では，大学院生などの教育に関しては，統計学や情報科学についての背景は白紙の状態からスタートしなければいけない．また，統計学や情報科学に関する関心や意欲が必ずしも高くない．

このため，これらの学生にとって身近な話題として情報科学，統計科学教育を捉え，自分の将来的な技能に必要であるという認識をどう持ってもらうかということから人材育成は始まる．これは，ほかの多くの学部や世間一般の人に関しても共通する課題でもある．

情報科学，統計科学の楽しさとは

情報科学や統計科学を専門としない人に理解してもらい，しっかりと学んでもらうには，その重要性をきちんと認識してもらうこと，それを実感として感じてもらうことが重要だと思う．そのためには，その内容を身近な話題に転換する．

生命科学系学部では，生物科学的な実験データの解釈になるが，それが病気の診断であったり，農産物の生育予測や，環境アセスメントであったりする．これらから何らかの定量的，定性的知識を抽出するために情報科学や統計科学が活躍するが，途中の行程をブラックボックスにしてしまい，結果の解釈のみに終始することも多い．その段階で，情報科学や統計科学と生命科学系の学生との乖離が起こる．これをブラックボックスとさせず，きちんと把握させるかがポイントになる．ブラックボックスになりがちな部分は，アルゴリズムとコーディングである．実のところ，このアルゴリズムとコーディングの部分が，データ分析で一番楽しい部分だ．しかも，その部分は，教科書では，

数式とソースコードで表現され，専門外の学生には読み飛ばされる可能性が一番高い．

ここでは，このような数式を身近な話題に転換する例として，分かりにくい理論の１つとして認識されているロジスティック回帰分析の説明を試みる．少し下品なネタであるが，アルゴリズムの楽しさを知る一例としてご容赦いただきたい．

とても可愛い女性がいて，その女性を口説き落としたいとする．実際に，口説いたところ，10 回アタックして，3 回口説くことに成功し，7 回フラれたとする．そうすると，成功した回数 3 回を失敗した回数 7 回で割った値，その値が大きければ大きいほど，口説きやすいということになる．

この値を数値的に処理しやすくするために対数をとる．これを y 軸にとり，x 軸にはデートの回数や，電話の回数，プレゼントの回数などのエフォートをとるようなグラフを作成する．これがロジスティック回帰分析だ．デートと電話とプレゼントの回数から女性を口説き落とす確率を予測することを考えよう．

口説き落とした回数を，フラれた回数で割った値を「オッズ」という．競馬場でウロウロしているオヤジからすると，「オッズ」といえば当たり馬券の枚数を，ハズレ馬券の枚数で割った値のことだ．この値が小さければ賭け金の戻りが大きくなる．この値の対数をとった値を y とし，これに払ったエフォートを x とする．この口説き落とした回数をフラれた回数で割り，その対数をとることを「ロジット変換」という．これを数式で書くと以下のようになる．

$$\log\left(\frac{p(x)}{1-p(x)}\right)=\beta_0+\beta_1 x$$

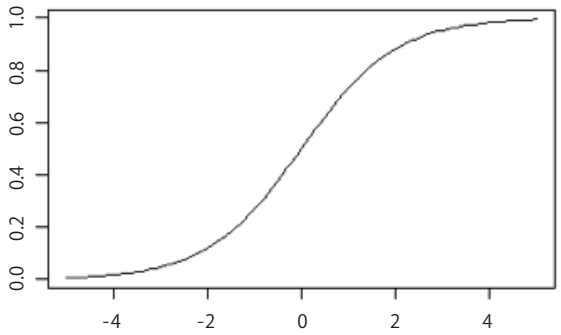
左辺のカッコの中の分子 $p(x)$ が口説き落とした確率，分母 $1-p(x)$ はフラれた確率であるが，両方の確率とも，それぞれの回数を試行回数で割っているので，この値を求める場合には口説き落とした回数を，フラれた回数で割るだけでよい．

左辺を y とし，右辺の回帰式 $y=\beta_0+\beta_1 x$ の係数 β_0 と β_1 を求めることが，ロジスティック回帰分析だ．これを，以下のように対数をはずし，$p(x)$ を求める．

$$\frac{p(x)}{1-p(x)}=e^{\beta_0+\beta_1 x}$$

$$p(x)=\frac{e^{\beta_0+\beta_1 x}}{1+e^{\beta_0+\beta_1 x}}$$

y 軸を $p(x)$ に，x 軸を x に，それぞれプロットすると以下のようなシグモイド曲線が得られる．



なお，計算上は口説き落とした回数をフラれた回数で割り，その対数を y 軸にすれば，単なる回帰分析になる．あとは $y=\beta_0+\beta_1 x$ の係数 β_0 と β_1 を求めればよい．x が，デートの回数や，電話の回数，プレゼントの回数などの複数の変数（多変量）になる場合は，これをベクトル $X=(x_1,\cdots,x_n)^T$ に変換して計算すればよい．数式で表現すると以下のようになる．ここで $\beta=(\beta_1,\cdots,\beta_n)^T$ は係数ベクトルである．

$$\log\left(\frac{p(X)}{1-p(X)}\right)=\beta_0+\beta_1 x_1+\beta_2 x_2+\ldots+\beta_n x_n$$

$$=\beta_0+\sum_{i=1}^{n}\beta_i x_i=\beta_0+\beta^T X$$

$$\frac{p(X)}{1-p(X)}=e^{\beta_0+\beta^T X}$$

$$\mathrm{p}(X)=\frac{e^{\beta_0+\beta^T X}}{1+e^{\beta_0+\beta^T X}}$$

このように，数式を身近な話題に置き換えれば，

やっていることはとてもやさしいことが理解できる．その内容を読み飛ばしたのではもったいない．機械学習やデータマイニングで表現されているアルゴリズムの難度は，大抵はその程度だ．少しの努力で世界が開けてくる．もしかすると，女性を口説き落とすための法則（数式モデル）を発見できるチャンスを逃してしまうかもしれない．ブラックボックスにすることで失われるものはあまりに大きい．

生命科学分野の学部・学科の課題

生命科学分野では，次世代シーケンサの普及により，ゲノムレベルの大量データが実験現場に持ち込まれる機会が増え，インターネットやICTを駆使した機器やそれらからの出力データが蓄積していることから，これらのデータを処理したり，分析したりするという要請が増えている．これらの変化があまりに急であるために，関連学部での人材育成に関する理解は必ずしも進んでいない．

数式やコードを駆使することに対するアレルギーのようなものは根強くあり，実際に行われている教育も表面的なものになりやすい．数式やコードを駆使し，自分でプログラミングを行ってデータ分析を行うようにならないと，なかなか実感としても，実践的に有意義なものにならない．

データ分析に必要な知識・技能としては，次の項目が挙げられる．

(1) 微積分，線形代数など統計学に登場する基本的数学の理解
(2) 古典的な統計学，ベイズ統計学，機械学習，人工知能などの概要理解
(3) SQLなどデータベースの理解と操作
(4) Perl, Python, Ruby，シェルスクリプトなどの基本的なスクリプト言語やRやMatlab，SASなどドメイン固有言語（DSL）や専用ソフトの精通と駆使
(5) ゲノム科学や医療統計，農業ITなど専門分野へ展開できる知識

図-1　人材育成プログラムによるパソコン実習風景

残念ながら，教育現場では，特に，情報科学を専門としない学部，学科においては，長期にわたってこれらの教育はなされておらず，カリキュラム的にも，教育人材的にも課題は大きいと考える．

農学系ゲノム科学人材育成プログラムの概要

東京農工大学の「農学系ゲノム科学人材育成プログラム」では，ゲノム科学をテーマとする大学院生から研究課題を募集し，採択された課題について，その研究に関する個別指導を行い，採択者による成果報告会などを実施した[1)]．ゲノム解析において，高度なプログラミングや統計解析についての指導を行った（**図-1**参照）．ゲノム情報としては，次世代シーケンサからは数千万エントリのゲノム配列データが産生され，これを処理するために，自然言語処理を含むテキスト処理，データベース，集計，数値計算などを行った．

2011～2015年度までに延べ289名の応募者から245名の研究課題を採択して個別指導と50件を超すセミナ，講習会を実施した．その結果，120件を超す学会発表，9件の学会賞などの受賞，13件の論文，6件の特許出願などの成果があった（**表-1**参照）．

データサイエンス教育とその実践に関する課題

実施したゲノム科学人材育成プログラムは，

	2011	2012	2013	2014	2015	合計
セミナ，講習会など	13	12	17	9	5	56
学会発表	7	25	32	41	17	122
受賞	1	0	5	2	1	9
原著論文	1	0	4	5	3	13
書籍，総説，報告書	1	2	13	11	3	30
外部での講演（招待講演など）	2	4	8	12	4	30
海外国際学会での招待講演	0	0	0	3	0	3
新聞，雑誌，Web そのほかの記事	8	10	16	14	2	50
特許出願	0	1	1	2	2	6

表 -1　ゲノム科学人材育成プログラム（2011 ～ 2015 年度）の成果一覧

数値的には成功だと思われるが，課題も残った．3 カ月単位の個別指導やセミナであるため，学生にじっくり基礎からプログラミングや統計学などの演習や指導を実施する系統的な教育は行いがたい．残念ながら，プログラミングやアルゴリズムを深く学ぶ時間はとても取れず，得られた結果の解釈に終始しがちであった．その結果，採択された学生で，自分でコーディングをし，アルゴリズムを実装してデータ分析を行えるレベルまで達した学生は本当に少ない．

データ分析や，情報科学，統計科学の重要性を認識するまでには至るものの，実際にデータ分析を行う研究者を育てるというレベルまでは到達しにくい．やはり，個別指導やセミナではなくきちんとしたカリキュラムを組み，数学やプログラミングをしっかり学び 1 ～ 2 年じっくりとトレーニングを積まないとなかなか人材は育ちにくい．テキストや自習書も最近はいろいろ出てきているが，まだまだ不足している．

今後の在り方

農学系ゲノム科学人材育成プログラムは，一定の成果を上げ 2015 年度で終了したが，その後の資金的な支援などはなく後継カリキュラムなどを設置するメドは残念ながら立っていない．その意味では，データ分析教育がこの分野で根付いていくにはいまだに道は険しい．個人的には，本会 IT フォーラム「ビッグデータ活用実務フォーラム」などの協力もあり，勉強会「マシンラーニングのら猫勉強会」を開始して努力を継続している [2]．月に 1 回有志で勉強会を実施しており，30 名近くの参加者を得て，最新の機械学習や人工知能の情報交換を行っている．

参考文献

1) 石井一夫：農学系ゲノム科学領域における情報科学・統計科学教育の取り組み，情報処理，Vol.55, No.5, pp.500-503 (May 2014).
2) マシンラーニングのら猫勉強会，https://machinelearning.doorkeeper.jp/

（2016 年 8 月 29 日受付）

石井一夫（正会員） kishii@cc.tuat.ac.jp

東京農工大学特任教授．数理モデリング，予測分析，データマイニング，機械学習，計算機統計学，ビッグデータなどを専門とする．徳島大学大学院医学研究科博士課程修了．フランス国立遺伝子多型解析センター，ノースウエスタン大学 Feinberg 医学部などを経て現職．日本技術士会フェロー，APEC エンジニア，IPEA 国際エンジニア．