データサイエンスをどう捉えてどう教えるか
—早稲田大学での取り組みを交えながら—

松嶋　敏泰

データサイエンスをデータから明確で合理的な意思決定プロセスを扱う学問体系と捉え，そのプロセスを三つの部分に分け概観する．特にプロセスの後半部はほぼ最適化問題となり，例を用いて解説を試みる．三つの部分それぞれの重要性を考え，多様な受講者に合わせたデータサイエンスの教え方を考える．その具体例として早稲田大学データ科学センターの研究・教育の狙いとその教育プログラムについても触れる．

キーワード：データサイエンス，最適化，重回帰分析，意思決定，正則化

1.　はじめに

データサイエンスとは何かという問いに対して，さまざまな定義が与えられているが，どれが正しいというべきものでもなく，多様な定義があってよいのではと思っている．しかし，データサイエンスをどう教えるかを考えるには，やはりデータサイエンスの定義に立ち返らざるを得ず，結局またカオスに戻ってしまう．

本稿ではその多様な定義にはじめは踏み込まず，データサイエンスとは，データから意思決定を行う明確で合理的なプロセスを扱う学問体系としてシンプルに捉え議論を進めていく．さらに，このプロセスを数理的に問題を定式化する前半部と，定式化された問題を数理的に解く後半部に分けて議論していく．また，この意思決定プロセスの外側にある上位の概念の存在も導入し，この三つの部分について重回帰分析とその拡張法を例に説明を行う．

このようにデータサイエンスを三つの部分で捉えたもとで，教育についての議論に移る．どのような学生にどのような力をつけさせたいかで教育プログラムの基本方針が決まってくる．たとえば，上記の三つの部分をどのようなウエイトで教えていくかなどが考えられる．具体例として，早稲田大学のデータ科学センター[1] における研究・教育の狙いから，どのような視点で三つの部分のバランスを考え教育プログラムを構築しているか，その基本理念と実際のカリキュラムについても説明する．

まつしま　としやす
早稲田大学理工学術院応用数理学科
早稲田大学データ科学センター
〒 169–8555　東京都新宿区大久保 3–4–1
toshimat@waseda.jp

以下本稿の構成は，2 節では，データサイエンスの意思決定プロセスの全体像とその前半部について述べる．3 節では，プロセスの後半部について説明する．後半部はほぼ最適化の問題となっており，よくご存知の読者の皆様には冗長かもしれないが，大切な部分なので多様な例を用いながら説明していく．4 節では，前半部の定式化のために，そのデータサイエンスの体系の外の上位の概念が必要なこと，それは解決したい問題に関する専門知識と関連していることについて説明する．5 節では，データサイエンスについて一般的に論じられている必要性や有効性の議論に立ち返り，4 節までの議論も含め専門知識とデータサイエンスの組み合わせの重要性について述べる．6 節では，早稲田大学のデータサイエンス研究・教育への狙いについて述べ，それを担うデータ科学センターについて概説する．7 節では，データ科学センターにおけるデータサイエンスの教育プログラムについて説明する．

数理面に興味のある方は 2–4 節，教育面に興味のある方は 4–7 節をご覧いただければと思う．

2.　データサイエンスの捉え方

前節で述べたデータサイエンスの捉え方からはじめる．それはデータからの明確で合理的な意思決定プロセスを扱う学問体系と捉えるのであった．明確なという意味は，意思決定の問題や目的が曖昧性なく記述されるということで，そのため問題の記述言語として数学が用いられる．合理的な意思決定の部分は，論理的な演繹法を用いて判断が導出されるということで，ここにも数学や計算機による演繹が用いられる．

データを用いて決定したい，解決したい問題は何か，つまり意思決定の目的，決定したい事項を数理的に表現することが重要である．それは，データを入力とし，

意思決定の結果を出力とする写像として記述されることになる．これを意思決定写像と呼ぶことにしよう．

意思決定写像について，典型的なデータサイエンスの問題である重回帰分析で考えてみる．この意思決定問題では，p 個の説明変数ベクトル $\boldsymbol{x} = (x_1, x_2, \ldots, x_p)^{\mathrm{T}}$ と被説明変数 y の n 組のデータを用いて，y と $\boldsymbol{x}$ の関係を知ることが目的である．さらに具体的には両者を線形的に説明する関数 $y = \boldsymbol{a}^{\mathrm{T}}\boldsymbol{x}$ を求めることが決定の目的となる．$\boldsymbol{a} = (a_1, a_2, \ldots, a_p)^{\mathrm{T}}$ は回帰係数ベクトルと呼ばれる．よってこの問題の意思決定写像は，入力が n 次元のベクトル $\boldsymbol{y}$ と $n \times p$ 行列 $\boldsymbol{X}$ のデータで[1]，出力が線形関数 $y = \boldsymbol{a}^{\mathrm{T}}\boldsymbol{x}$，あるいはもっとシンプルには $\boldsymbol{a}$ ということになる．

このように意思決定の目的は意思決定写像で数理的に表現できた．次は決定の良さの基準，評価基準を数理的に表現する必要がある．意思決定が合理的かどうかは，評価基準が明確でなければ議論できず，その評価基準は数理的な評価関数で記述される必要がある．

これも先の重回帰の意思決定問題の例で見てみよう．出力された線形関数 $y = \boldsymbol{a}^{\mathrm{T}}\boldsymbol{x}$ で，得られたデータ $\boldsymbol{X}$ と $\boldsymbol{y}$ の関係をよく説明していてほしいと考えているのであれば，その良さを曖昧性なく記述するために，たとえば二乗距離 $\|\boldsymbol{y} - \boldsymbol{a}^{\mathrm{T}}\boldsymbol{X}\|_2^2$ のような評価関数で数理的に記述することが考えられる．

これで，意思決定プロセスにおける目的部分を意思決定写像で，評価の基準を評価関数で数理的に記述できたことになる．データサイエンスにおける意思決定プロセスの前半部は，このように記述言語として数学を用いることで意思決定問題を明確に定式化できたことになる．

3. データサイエンスと最適化理論

データサイエンスの意思決定プロセスの後半部は，数理的に明確化された問題を合理的に解く（決定する）こととなり，数学の演繹の論理的正確性を利用することになる．もっと具体的には，前半部で定式化された評価関数を最小化（最大化）する解を求めれば良いわけで，評価関数といっていたものを目的関数とした最適化の問題そのものとなってしまう．つまり，データサイエンスの意思決定プロセスの後半部の合理的判断を担っているのは，最適化理論ということである．

あくまでイメージではオーソドックスな統計学は最適化を解析的に解く，データサイエンスではより複雑な最適化手法を用い計算機で解くような雰囲気がある．もちろんデータサイエンスは，統計学，機械学習などを含んだ総合的な体系と捉えられるので，そのような区別はそもそも存在しないのだが，あえてそのような俗っぽいイメージにも触れながら，データサイエンスの意思決定プロセスの後半部と最適化の関わりについて重回帰分析を例にいくつか述べてみたい．

まず意思決定写像の出力を回帰係数ベクトル $\boldsymbol{a}$ とした問題で，評価関数として $\|\boldsymbol{y} - \boldsymbol{a}^{\mathrm{T}}\boldsymbol{X}\|_2^2$ を考えた．意思決定プロセスの後半部では，この評価関数を目的関数として最小化する以下の最適化問題を考えればよい．

$$\min_{\boldsymbol{a}} \|\boldsymbol{y} - \boldsymbol{a}^{\mathrm{T}}\boldsymbol{X}\|_2^2 \tag{1}$$

この最適化は．$\boldsymbol{a}$ の各要素で偏微分した式を 0 とおき，p 元連立方程式を解けば最適解 $\hat{\boldsymbol{a}} = (\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{y}$ が簡単に得られる．

ご存知のように多変量解析ではこの方程式を正規方程式と呼び，この分析法を最小二乗法と呼んでいる．最適化の問題としてはプリミティブな方法で，解が解析的に解け，陽に求まる点がなんとなく統計学っぽいイメージではある．

p 個の説明変数からなる説明変数行列 $\boldsymbol{X}$ の列が従属であったり独立性が低い場合，$\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}$ が正則にならず逆行列が求まらない，解が安定しない多重共線性と呼ばれる問題が発生する．これを回避したいという評価基準がさらに加わったとしよう．それに対応するため正則化項 $\|\boldsymbol{a}\|_2^2$ を，式 (1) の目的関数に加え新たな以下の目的関数を考える方法がある．

$$\min_{\boldsymbol{a}}(\|\boldsymbol{y} - \boldsymbol{a}^{\mathrm{T}}\boldsymbol{X}\|_2^2 + \lambda\|\boldsymbol{a}\|_2^2) \tag{2}$$

制約 $\|\boldsymbol{a}\|_2^2 < C$ を考えた場合のラグランジュ関数とみなすことも可能であり，この最適解は $\hat{\boldsymbol{a}} = (\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X} + \lambda\boldsymbol{I})^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{y}$ となり，リッジ回帰 [2] と呼ばれている．目的関数の二つの項が共に l_2 ノルムなので，ほぼ最小二乗法と同様に解析的に解け，陽に表現できる点がまだ統計学らしいイメージである．

重回帰の決定問題として重要なものに，p 個の説明変数 $\boldsymbol{x}$ を y の説明に役立っているものだけに絞り込みたい，いわゆる変数選択問題がある．意思決定として，説明に使う必要のない説明変数 x_i の回帰係数 $a_i = 0$ とすると考えても良いので，ここまでと同様に回帰係数ベクトル $\boldsymbol{a}$ を意思決定写像の出力として考えること

---

[1] データ $\boldsymbol{y}$, $\boldsymbol{X}$ は次のように正規化されているとする．

$$\sum_{i=1}^{n} y_i = 0,\ \sum_{i=1}^{n} x_{ij} = 0,\ \sum_{i=1}^{n} x_{ij}^2 = 1 \quad j \in \{1, 2, \ldots, p\}$$

ができる．有名な AIC[3] や BIC[4] などにおける目的関数は以下になる．

$$\min_{\boldsymbol{a}}(\|\boldsymbol{y}-\boldsymbol{a}^{\mathrm{T}}\boldsymbol{X}\|_2^2+\lambda\|\boldsymbol{a}\|_0) \tag{3}$$

リッジ回帰の正規化項の l_2 ノルムに対して，l_0 ノルムを考えていることになる．正則化項が有効な説明変数の数に対応するので，これがペナルティになって説明変数が絞り込まれることは直感的にも明らかであろう．なぜこの正則化項をつけるかについては後で述べることとする．

この問題の最適解を求めることは，基本的には 2^p の指数オーダの組み合わせを探索しなければならず，最適化問題としては非常に厄介になってしまう[2]．

さらに，l_2 ノルムと，l_0 ノルムの間の l_1 ノルムを正則化項として用いると以下のような目的関数になる．

$$\min_{\boldsymbol{a}}(\|\boldsymbol{y}-\boldsymbol{a}^{\mathrm{T}}\boldsymbol{X}\|_2^2+\lambda\|\boldsymbol{a}\|_1) \tag{4}$$

lasso (least absolute shrinkage and selection operater) [5] と呼ばれる方法で，名前のとおり l_2 ノルムと l_0 ノルムの両者の良いとこ取りを狙ったものであるが，過学習やスパース性など色々理由づけがされている．

この目的関数の最適解を求めるにはどのような方法があるのだろうか．目的関数に絶対値が入り微分ができないため，劣微分で考える必要があるなど，最適化問題としては難しくなっている．l_0 ノルム正則化項の最適化よりはまだましなようだが，最適化の専門家でない著者からするともう相当難しい問題に感じる．

歴史的には LARS (Least Angle Regression)[6] という方法が有名で，貪欲法の変数増加法と似ているが l_1 ノルムの性質と KKT 条件をうまく使い，変数増加法と同様な基準で説明変数を取り入れながら，回帰係数を控えめな値に決めていくことで，説明変数の数と同じ p ステップで lasso 問題をほぼ解くことができる．各ステップが解析的ではあるので，ぎりぎりまだ統計学っぽいと思えるが，p が大きくなると逆行列計算の負荷が大きく，大規模データの処理には向かない．

そこでこの問題を解く方法としていくつかの最適化法が登場してくることになる．その一つが ISTA (Iterative Shrinkage Thresholding Algorithm)[7] と呼ばれる近接勾配法の一種で，l_2 ノルムと l_1 ノルムが混在した目的関数の計算困難性を，近接項を加えリプシッツ定数を利用することで目的関数を変形して平方完成すること，x の座標ごとに目的関数を分離すること，l_1 ノルムで場合分けして二つのノルムが混在していても最適値が求められる軟判定しきい値関数を用いることなどで計算を簡素化している．これ以外にも勾配法のさまざまなアルゴリズム[3]が提案されている．このように，勾配降下法，座標降下法などのアルゴリズムが登場すると，なんとなくデータサイエンスっぽいイメージになるが，問題自体は変わっていない．

この ISTA は，目的関数を上界式で上から抑えて，その上界式における最小値を逐次的に求めていきながら，最適解に近づいていく上界最小化法のアルゴリズムとしても解釈される．実はそのようなアルゴリズムは統計分野でも従来から使われている．潜在変数を含む確率モデルの最尤推定量を求める EM アルゴリズム (Expectation Maximization Algorithm)[10] がまさに，期待値ステップ (E-step) で上界式をもとめ，最大化ステップ (M-step) で最大化を交互に行っていることになる．先程，勾配法を使うのがデータサイエンスっぽいといったが，実は従来から統計学でも勾配法が用いられていた．

この EM アルゴリズムは，少し変形することでベイズ決定理論で重要となるパラメータの事後確率の近似計算[4]に利用できる．このアルゴリズムは変分ベイズアルゴリズム [11] と呼ばれ，さまざまなベイズ決定理論

[2] 実務的な変数増加法などでは，そこまで取り入れた説明変数を用いた回帰式と $\boldsymbol{y}$ との残差ベクトルを求め，そのベクトルとまだ説明に使っていない各説明変数 $\boldsymbol{x}_i$ との内積を求め，その値の大きいものを逐次取り入れていく貪欲法が用いられる．

[3] この他にもたとえば，この双対問題のラグランジュ関数に罰金項を加えた拡張ラグランジュ関数法 [8] を用いたものもある．さらにその交互版で ADMM (Alternating Direction Method of Multipliers)[9] なども用いられている．ADMM は拡張ラグランジュ関数法の目的関数の変数 $\boldsymbol{a}$ を形式的に二つの変数 $\boldsymbol{a}$ と $\boldsymbol{b}$ として目的関数を $\boldsymbol{a}$ を用いた l_2 ノルムの項と $\boldsymbol{b}$ を用いた l_1 ノルムの項で別々に表現し，二つの変数は等しい $\boldsymbol{a}=\boldsymbol{b}$ という制約を加えた新たな目的関数をつくり，$\boldsymbol{a}$ を含む目的関数と $\boldsymbol{b}$ を含む目的関数を交互に最適化していくアルゴリズムである．

[4] 事後確率計算の困難性の主原因は積分計算にあり，それを回避するためにパラメータの事前分布として，データを発生させる分布の共役事前分布を仮定することで解析的に事後分布を求めることが行われてきた．これも統計学っぽいといえるかもしれない．しかし，指数分布族でないようなさまざまなデータ発生の確率モデルが用いられるようになり，そのような都合の良い事前分布は仮定できなくなってきた．そのために用いられたのがマルコフ連鎖モンテカルロ法 (MCMC: Markov chain Monte Carlo methods) に代表される乱数を用いた数値積分法である．この数値積分法に対してもう一つの主流となるのが，変分ベイズ法のような，確率分布を強引に小さな独立なブロックに分割して，各ブロックを逐次的にあるいは同時並行的に最適化（この場合は疑似部分事後確率を計算）していく方法がある．このような方法は，統計力学では平均場近似，符号理論では sum-product アルゴリズムに代表されるメッセージ伝搬アルゴリズムなどと同等であり，さまざまな分野で用いられている．

の推論に用いられている．

また，大規模データの解析という意味では，確率的最適化の恩恵をあげておかなければならないだろう．大規模なデータを全部使って一気に推定などの決定を行うことは計算量的に困難なため，確率勾配降下法では逐次的にデータを与えて勾配降下法を行っていく．データサイエンスのさまざまな問題で多く見受けられるアルゴリズムである．実は統計学でも最尤推定量を求めるため，データを逐次的に与えて推定値を更新していくほぼ同様なアルゴリズム [12] が使われていた．

統計学っぽいとかデータサイエンスっぽいとあえて書いてみたものの，最初に申し上げたようにそんな区別はもともとなく，上記のように古くからさまざまな最適化法が使われており，最適化理論がデータサイエンスの意思決定プロセスの後半部を渾然一体となって支えてきたことが，この重回帰の一連の例からだけでも明らかなのではと思う．この節のまとめとして，意思決定写像とその出力の評価関数が決まった，つまりデータサイエンスの意思決定プロセスの前半部が首尾よく終了した後の後半部は，全く数理的な議論のみで合理的な決定が行われる構造になっていることをあらためて確認しておく．

4. データサイエンスの意思決定プロセスとしての限界

ここまで，データサイエンスの意思決定プロセスを前半部と後半部に分けて説明したが，データサイエンスの体系の上位の階層の概念についても考えてみたい．2 節では，プロセスの前半部で意思決定問題における目的部分を意思決定写像で，評価の基準を評価関数で数理的に記述できたことになると書いたが，その目的がなぜ出てきたのか，その評価基準もなぜその評価でいいのかを考えると，その上の概念である上位の目的や評価基準が浮かび上がってくる．

数理的には同じ意思決定写像と評価関数で表現される意思決定問題も，その上位の目的や評価が異なっている場合もある．まずはそれを例を通して説明していく．

重回帰問題の意思決定写像は，入力がベクトル $\boldsymbol{y}$ と行列 $\boldsymbol{X}$ のデータで，出力が関数 $y=\boldsymbol{a}^{\mathrm{T}}\boldsymbol{x}$ であった．また，写像の出力の評価関数として二乗距離 $\|\boldsymbol{y}-\boldsymbol{a}^{\mathrm{T}}\boldsymbol{X}\|_2^2$ を考えた．これらの関数がでてくるための上位の概念として，たとえば次の二つの立場があると考えられる．

立場 1) 目的は 2 節でも書いた直接的なもので，データ $\boldsymbol{y}$, $\boldsymbol{X}$ の関係をよく説明する線形関数 $y=\boldsymbol{a}^{\mathrm{T}}\boldsymbol{x}$ を求めたい．そのよく説明するという評価基準は二乗誤差で測ることにしたいという考え方である．

立場 2) 意思決定したい対象が観測データそのものではなく，そのデータが発生してきている母集団（分布）について知りたいという立場である．そのための数理的モデルとして，データを確率変数として捉え，ある確率モデルから発生していると仮定を置くことになる．たとえば確率変数 Y がパラメータ $\boldsymbol{a}$ と σ でパラメタライズされた確率分布 $P(y|x,\boldsymbol{a},\sigma)$ から発生していると仮定する．さらに具体的には以下のよう正規分布を仮定することが多い．

$$Y\sim\mathcal{N}(\boldsymbol{a}^{\mathrm{T}}\boldsymbol{x},\sigma^2) \tag{5}$$

この仮定のもと意思決定の目的は，得られた観測データにおける y と $\boldsymbol{x}$ の関係 $y=\boldsymbol{a}^{\mathrm{T}}\boldsymbol{x}$ ではなく，母集団として y と $\boldsymbol{x}$ がどのような構造をもっているかということにある．この例では，パラメータ σ を既知として確率分布 $P(y|x,\boldsymbol{a},\sigma)$ を仮定すると，母集団の構造はパラメータ $\boldsymbol{a}$ さえわかればよいことになる．よって，意思決定写像としては立場 1 と同じであるが，これはいわゆる統計的決定理論におけるパラメータ推定の決定関数として解釈され，出力の $\boldsymbol{a}$ もパラメータの推定値として解釈されることになる．

この立場の場合，上位の評価基準として，意思決定写像の出力のパラメータ推定量の良さを考える必要がある．推定量の評価基準は統計学においてさまざまであるが，たとえば尤度を最大化する評価基準 $\max_{\boldsymbol{a}} P(y|x,\boldsymbol{a},\sigma)$ を考えると，簡単な式変形で $\min_{\boldsymbol{a}}\|\boldsymbol{y}-\boldsymbol{a}^{\mathrm{T}}\boldsymbol{X}\|_2^2$ と同じ評価関数が導かれる．この評価関数を最小化する出力はもちろん最尤推定量となる．

以上の例で見たように，数理的な表現上は意思決定写像，評価関数が同じであっても，その上位の目的，評価基準，背景の仮定，が異なっていることがある．この例の場合はいわゆる記述統計的立場と数理統計的立場の違いである．

もう少し違う例も見ておこう．リッジ回帰も，直接的な目的では既に説明した解の安定性のために，正則化項をつけた評価関数 $\min_{\boldsymbol{a}}(\|\boldsymbol{y}-\boldsymbol{a}^{\mathrm{T}}\boldsymbol{X}\|_2^2+\lambda\|\boldsymbol{a}\|_2^2)$ とすることを述べた．

この問題についても，上記の立場 2) のように母集団について意思決定したい上位の目的も考えられる．仮定する数理モデルとしても，データ発生に同様な分布 $P(y|x,\boldsymbol{a},\sigma)$ を仮定したもとで，さらにパラメータ $\boldsymbol{a}$ 自体も確率変数と考えた，いわゆるベイズ決定理論の枠組みを考えることも可能である．この立場の場合も意思決定写像の出力はやはりパラメータ $\boldsymbol{a}$ の推定量とし

て解釈されることになり，その推定量の良さの評価基準としてたとえば事後確率の最大化 $\max_{\boldsymbol{a}} P(\boldsymbol{a}|\boldsymbol{y},\boldsymbol{X},\sigma)$ を考えると出力は事後確率最大推定量となる．たとえば確率分布モデルとして式 (5) を仮定したうえに $\boldsymbol{a}$ の事前分布として多次元正規分布 $\mathcal{N}(\boldsymbol{0},\frac{1}{\lambda}\boldsymbol{I})$ を仮定して，事後確率最大を評価基準にするとやはり評価関数 $\min_{\boldsymbol{a}}(\|\boldsymbol{y}-\boldsymbol{a}^{\mathrm{T}}\boldsymbol{X}\|_2^2+\lambda\|\boldsymbol{a}\|_2^2)$ が導出される．

l_0 ノルムを正規化項とした評価関数を用いた変数選択問題の場合，本当にさまざまな上位の目的と評価基準，背景の仮定，から同様な意思決定写像や評価基準が導出されている．たとえば，AIC の場合は意思決定写像で出力された線形関数で予測を行った場合の良さを Kullback-Leibler 情報量で評価している[5]．その評価値の漸近不偏推定量を最小化する評価基準を考えると，式 (3) と同じ評価関数が導出される．l_0 ノルムを正規化項とした類似[6] の変数選択またはモデル選択基準 [13] も，さまざまな上位の概念から導出されてきており，それぞれの違いを理解して使い分ける必要がある．

l_1 ノルムも同様で，過学習だからスパースにという定性的な理由から，安易にこの評価関数で意思決定をすれば良いというわけではないことをご理解いただけると思う．もし予測に用いたときに有効な回帰係数 $\boldsymbol{a}$ を求めたいという上位の目的と評価基準があるのなら，適切な仮定のもとそれに適合した評価関数が決められるべきであろう．

以上の例で見たように，数理的に記述された意思決定写像，評価関数は，より上の概念から決まる上位の目的，評価基準，背景の仮定から導出されてきているわけである．同じ意思決定写像で記述できたとしても，その上位の目的についてたとえば立場 1 と 2 のどちらをとりたいのかは，その決定を行う主体の意図によるものであるし，その意図に関連して上位の評価基準も決まってくることになる．背景の仮定についても，たとえば母集団や分布を仮定して良いのか，さらにパラメータの事前分布まで仮定してよいのかは，意図や決定問題の背景に関する専門知識からしか決定できない．

ここがデータサイエンスの限界で，明確で合理的な意思決定プロセスを扱う体系の拠り所が，極めて曖昧なところから出発しているという矛盾である．プロセスの後半部はまさに数理的演繹で明確で合理的であるが，前半部の意思決定問題の明確な定式化や合理的評価のための評価基準は，データサイエンスの体系の外の概念によって決めざるを得ないということである．

誤解していただきたくないのは，データサイエンスはちっとも明確で合理的でないではないか，ということを説明しようとしているわけではなく，決定主体の意図やデータの背景の専門知識をうまく前半部に取り入れることができれば，あとはデータサイエンスの体系により極めて明確で合理的な意思決定プロセスが展開できるということである．

5. データサイエンスと各専門分野

これまで述べてきたようにデータサイエンスの合理的意思決定プロセスの体系はそれ単体だけでは機能せず，上位の概念としての意思決定主体の意図や，対象とする決定問題に関する専門分野の知識があってはじめて機能するのであった．ここで，視点を変え，一般にデータサイエンスとは何か，その必要性や有効性について論じられている事柄について振り返ってみたい．

データからの意思決定の重要性については従来から論じられてきたはずであるが，近年「データサイエンス」という言葉が登場し注目を集めブームにまでなった大きな要因としては，数値のみならず，文字，音声，画像などの多様なデータが，情報・通信インフラの発展により容易に世界中から収集可能になった点と，データを分析する理論や技術の進歩があげられる．可能であればデータや事実からの意思決定をすることが望ましいが，今まではそれが困難であったさまざまな対象でそれが可能となり，それらの分野でもデータサイエンスが注目され活用されるようになってきたという流れである．データサイエンスの発展により，さまざまな分野において今まで想像もされていなかった新仮説や知見がデータから導出される可能性が高まり，これを新しい知の創造プロセスと捉えようとする考え方が出てきている[7]．

以上をまとめると，どのような専門分野でも明確で

---

[5] 具体的には，予測の評価として，説明変数を絞り込んだ，つまり回帰係数の一部が 0 であることも許した回帰係数ベクトル $\boldsymbol{a}$ を用いた予測分布 $P(y|x,\boldsymbol{a},\sigma)$ と真の分布 $P(y|x,\boldsymbol{a}^*,\sigma)$ の違いを以下の Kullback-Leibler 情報量で評価する上位の評価基準を考えている．

$$\sum_{y\in Y} P(y|x,\boldsymbol{a}^*,\sigma)\log\frac{P(y|x,\boldsymbol{a}^*,\sigma)}{P(y|x,\boldsymbol{a},\sigma)}$$

この評価基準は真の分布が含まれていて実際には計算できないため，この評価基準のテーラー展開の 2 次までの近似で漸近不偏推定量求めたものが AIC に対応する．

[6] λ にいろいろなバリエーションがあり，たとえば $\lambda=\frac{1}{2}\log n$ などがある．

[7] これは Jim Gray の “Data science as a fourth paradigm of science” という言葉に象徴されるかもしれない．いわゆるをデータ駆動型の知の創造プロセスとも呼ばれる理念である．

合理的意思決定のためにデータサイエンスが必要な時代となってきた．また，前節まで述べてきたことのまとめは，データサイエンスは決定問題に関する専門分野の知識がないと機能しないということであった．まとめのまとめは，データサイエンスと各専門分野はセットにしないと機能しない，さらに両者をセットにすると強力な知的活動の武器になるということになろう．

6. データサイエンスにおける早稲田大学の狙い

ここまでに述べてきたようにデータサイエンスとさまざまな専門分野が融合することで，人類の意思決定や知の創造のプロセスが大きく変わる大変革期が訪れていると考えられる．早稲田大学では，この大変革期に最も変わらなければならないのは，実は知の創造と継承の拠点である大学の研究・教育ではないかという視点から，早稲田大学データ科学センター [1] を 2017 年 12 月に全学的な組織として設置し，2018 年 4 月から本格的に稼働させている．

データ科学センターでは，総合大学の強みである人文社会系や理工系のさまざまな分野の専門知識とデータサイエンスの融合を軸として，研究面と教育面の両面での効果を目指している．研究面ではデータを基にした知の創造プロセスの進化・深化と，それを発展させたさまざまな分野との融合研究による革新的な研究，たとえば複雑でグローバルな社会問題の解決などを目指している．教育面ではそれぞれの専門性の上にデータを活用する能力をもった社会で有用な人材の育成を目指している．

早稲田大学のデータ科学センターは上記の目標を具現化するため，その組織の体制や運営法にいくつかの工夫をこらしている．早稲田大学の各学術院（学部と大学院を含む組織）を縦糸にたとえると，データ科学センターは大学本部直轄の全学横断的な横糸の組織であり，センター自体の専任教員と各学術院を本属とする兼任センター員で構成されている．これはここまで述べてきた専門的知識とデータサイエンスが組み合わされることにより機能するという視点からは，縦糸の政治，経済，文，法，経営学などさまざまな専門分野と，データ科学センターのデータサイエンスが横糸となって組み合わされることで大学全体として機能することを目指していることになる．

たとえば，各学術院で行われているさまざまな専門分野の研究，特にデータを活用した研究に対して，センターが横のパイプ役となり各専門研究とデータサイエンスを融合させることにより，革新的な研究成果が生み出されることを狙っている．分野融合型研究の必要性が叫ばれるものの，学術院間の壁に阻まれ成果がなかなか生まれないのが実情であったが，センターの横糸機能でその壁が取り除かれ大きな進展の可能性がでてきている．

教育面でもセンターの横糸機能を最大限に生かそうとしている．すべての学術院においてその専門分野とデータサイエンスの融合を目指した教育の展開が求められているが，各学術院にデータサイエンスの専任教員を置くことの困難性，各学術院の独自のカリキュラムの中に新たにデータサイエンスの教育カリキュラムを組み込むことの困難性など課題が山積している．これらを解決する意味でもセンターの分野横断的な横糸の教育体制は有効で，この後の節でその具体的内容については説明する．

7. データ科学センターにおける教育の展開

総合大学でさまざまな専門分野がある多様性は，データサイエンスとの組み合わせを考えた場合に強みであると述べていたが，教育プログラムを展開しようとすると逆にさまざまな問題が発生する．私立大学ではセンター試験を通過しない学生の数理系知識のレベルに差がある．さまざまな学部がある総合大学では学生が身につけている専門知識や興味のある対象も多様である．数理系に不得意意識のある文系の学生でも理解でき，全学共通であり，また個別の専門分野にもある程度対応可能なデータサイエンスの教育のしくみやカリキュラムはどうあるべきか苦悩した．

これらの問題点を乗り越えるための基本戦略も，2–4 節を通じて述べてきたデータサイエンスをどのように捉えるかという考え方であった．プロセスの前半部は数理を用い，決定問題の目的についてはデータを入力として決定結果を出力とする意思決定写像で，決定結果の良さについては評価関数でそれぞれ記述した．後半部は合理的判断のため数理的演繹を用い，評価関数を目的関数とした最適化法などでの数理的なアルゴリズムで最適な解を求めた．

文系学生の数理の知識のレベル差を考えた場合，前半部では論理的考え方や，記述言語としての数学の利用が主であるので，ここまでなら高校の数学程度で理解可能と思われる．後半部の数理的に最適解を求める部分を本当に理解させることは文系の学生には正直ハードルが高いかもしれない．理系の学生にはこの部分をしっかり教えなければと思っている．センターが重視している専門分野とデータサイエンスの関連の部分は，

4 節で述べた決定主体の意図と専門知識から前半部の数理的記述に落とし込んでいく部分に対応する．そこで，教育プログラムの中心を意思決定プロセスの前半部に置き，そのさらに上位の専門知識から数理に落とし込んでいく部分と，後半部の数理的に解を求める部分の両方向につなげていくような構成とした．

各学術院独自のカリキュラムへの影響を少なくした全学横断的教育プログラムとするため，オンデマンド授業（e ラーニング）を中心とすることで学習時間やカリキュラムの柔軟性を担保している．このような対面でない授業を補完するため，常設の相談窓口も設置し，授業に関する質問に限らずデータサイエンスに関してある程度高度な質問にも答えられる LA (Learning Assistant) が対応している[8]．さらに，学生の専門分野や興味の多様性に対応するため，コンテンツをモジュール化し，部分的なカリキュラム内容の組み換えや，独立した利用なども可能としている．

このような方針でつくられた具体的教育プログラムは次のようなものとなった．まず全学を対象としたデータ科学の入門シリーズは，データ科学入門 α，β，γ，δ の四つのクォータのフルオンデマンド授業で構成され，この四つで最低限のデータサイエンスの考え方が身につくことを狙っている．このシリーズでは先程述べた意思決定プロセスの前半部を中心にして，データから専門知識を活かしながら意思決定を行う考え方について学べるよう，体系的であるが数理的知識を必要最低限におさえた構成を心がけ，どの学部の学生，特に文系の学生でも興味がもてるように工夫している．データ科学入門シリーズでも考え方の基礎となるのは統計学であるが，機械学習の考え方も組み込み融合再整理したカリキュラムとなっている．

一方，統計学を中心とした全学を対象とした入門シリーズである統計リテラシー α，β，γ，δ はやはり 4 クォータのフルオンデマンド授業で既に数年前から運用されていて，この統計シリーズを開発したノウハウがデータ科学入門シリーズでも生かされている．その他にもこれらのシリーズをコアとしていくつかの科目が設置されており詳しくはセンターの HP[1] をご覧いただきたい．

8. おわりに

データサイエンスをデータからの明確で合理的な意思決定プロセスを扱う体系と捉えると，それ単体では機能しない体系で，各分野の専門知識と融合してはじめて機能することがより明らかになった．さらに早稲田大学データ科学センターでは，このデータサイエンスと各分野の専門知識の融合を軸に教育と研究を展開していることを概説した．本拙稿の結びのかわりに，早稲田大学の創設者である大隈重信が 130 年以上前に残した，統計（データ）からの意思決定の重要性について述べている言葉をあげさせていただく．「現在の国勢を詳明せざれば，政府すなわち施政の便を失う．過去施政の結果を鑑照せざれば，政府その政策の利弊を知るに由なし」[14].

[8] さらに高度な研究活動のサポートとしては，全学の大学院生，ポスドク，研究員，教員向けにデータ科学研究相談窓口も設置されている

参考文献

[1] 早稲田大学データ科学センター，HP，https://www.waseda.jp/inst/cds/（2020 年 7 月 31 日閲覧）

[2] A. E. Horel and R. W. Kennard, “Ridge regression: Biased estimation for nonorthogonal problems,” *Technometrics*, **12**, pp. 55–67, 1970.

[3] H. Akaike, “A new look at the statistical model identification,” *IEEE Transactions on Automatic Control*, AC-19, pp. 716–723, 1974.

[4] G. E. Schwarz, “Estimating the dimension of a model,” *Annals of Statistics*, **6**, pp. 461–464, 1978.

[5] R. Tibshirani, “Regression shrinkage and selection via the lasso,” *Journal of the Royal Statistical Society: Series B*, **58**, pp. 267–288, 1996.

[6] B. Efron, T. Hastie, I. Johnstone and R. Tibshirani, “Least angle regression,” *Annals of Statistics*, **32**, pp. 407–499, 2004.

[7] I. Daubechies, M. Defrise and C. De Mol, “An iterative thresholding algorithm for linear inverse problems with a sparsity constraint,” *Communications on Pure and Applied Mathematics*, LVII, pp. 1413–1457, 2004.

[8] M. R. Hestenes, “Multiplier and gradient methods,” *Journal of Optimization Theory and Applications*, **4**, pp. 303–320, 1969.

[9] D. Gabay and B. Mercier, “A dual algorithm for the solution of nonlinear variational problems via finite element approximation,” *Computers and Mathematics with Applications*, **2**, pp. 117–140, 1976.

[10] A. P. Dempster, N. M. Laird and D. B. Rubin, “Maximam likelihood from incomplete data via the EM algorithm,” *Journal of the Royal Statistical Society: Series B*, **39**, pp. 1–38, 1977.

[11] H. Attias, “Inferring parameters and structure of latent variable models by variational Bayes,” *Uncertainty in Artificial Intelligence*, pp. 21–30, 1999. *Springer*, 2006.

[12] H. Robbins and S. Monro, “A stochastic approximation method,” *Annals of Mathematical Statistics*, **22**, pp. 400–407, 1951.

[13] 松嶋敏泰，“統計モデル選択の概要，” オペレーションズ・リサーチ：経営の科学，**41**, pp. 369–374, 1996.

[14] 総務省統計局，統計の偉人たち，https://www.stat.go.jp/library/meiji150/ijin/（2020 年 7 月 31 日閲覧）