特別掲載

データ活用の基本①
―どのように統計モデルを選べばよいか―

研究員　山名　一史

目　次

１．はじめに

コンピュータの処理能力向上や利用可能なデータ量の爆発的増加、深層学習技術の飛躍的発展などを背景に、データサイエンスという言葉が人口に膾炙するようになってから数年が経つ。実証主義的な方法論、すなわち経験的に得られた仮説を実験・非実験データに基づいて検証する方法論に立脚してきた科学研究分野は言うまでもなく、政治や経営、さらにスポーツなど様々な分野においてデータサイエンスの活用が広がっている。

生命保険や損害保険といった保険商品を扱う保険業界も例外ではない。ビッグデータや機械学習・人工知能（ＡＩ）といった技術を活用し、新しい保険商品や高度な保険業務を実現しようとする取組みはインシュアテック（InsurTech）と総称される。保険会社は長年にわたって、保険の対象であるリスクや加入者に関する膨大なデータを収集・蓄積しており、データサイエンスとの親和性が高い。

分野や業種を問わず、収集・蓄積されてきたデータをどのように活用すればよいか、今回と次回の２回に分けて基本的な知識を説明する。今回はどのように統計モデルを選べばよいか、次回はニューラルネットワークに関する内容を扱う。本稿の目的は、具体的なデータ活用事例の紹介やデータ活用の是非に関する議論ではなく、データ活用の背後にある統計的な考え方を説明することである。なお、一般の読者を想定し、数式の利用は最小限にとどめた。

本稿の構成は以下の通りである。第２節ではモデル選択について概観したうえで、モデル選択の際の計算量を節約する方法として第３節で正則化、第４節では主成分分析を取り上げる。第５節ではモデル選択の際に注意する問題として高次元データと共線性に関する話題を取り上げる。

２．モデル選択

⑴　統計的予測とは

ビジネスにおいて、顧客の属性情報を用いた予測には様々な可能性がある。具体例として、保険業界の事例を取り上げると、個人の医療情報を用いた将来の入院リスクの予測、商品販売やアップセルの対象として潜在的に有望な顧客属性の予測などが挙げられる。

予測と聞くと、多くの人は「Ａが起こった（原因）のでＢが起こる（結果）と予測される」というような因果関係に基づく予測を想起するだろう。これはもちろん間違いではないが、統計学や機械学習、人工知能に基づく予測という場合、その多くはこれに該当しな

い。予測には因果関係に基づくものと、「Aが変化した時にBも変化する」という共変関係、いわゆる相関関係に基づくものとがあり、統計学的な予測の大半は後者である。

因果関係に基づく予測は分かりやすく強力なのに、なぜあまり用いられないかというと、因果関係を同定するのが非常に難しいからだ。そもそも、統計学の枠組みは因果関係との相性が悪く、因果関係を同定するためには、因果推論などの特別な手続きを経ることが要求される。とはいえ、因果関係に基づく予測でなければ実用的な結果が得られないような問題も存在しないわけではない。予測を行う際は、相関関係に基づく予測で十分なのか、一歩踏み込んだ因果関係に基づく予測まで必要なのかを最初に判断する必要がある。

それでは、顧客情報から入院リスクや保険商品の購入・加入確率を予測する問題はどちらに該当するかというと、相関関係に基づく予測で十分に実用的な結果が得られる問題と考えられる。したがって、標準的な統計学の知見を活用しやすい問題だということができるだろう[1]。

(2) 統計的予測モデル

それでは、予測について具体的に考えていこう。商品の購入・加入確率や入院リスクに関するデータを出力変数 Y、データから観測可能な p 個の異なる顧客属性や医療情報を入力変数 X と呼ぶことにする。出力変数と入力変数との間に

$$Y = f(X) + \epsilon$$

という真の関係があると仮定しよう。ここで、f は入力変数に関する未知の関数であり、ϵ は X と無関係で平均が0の撹乱項、つまり X では説明できないランダムな項である。たとえば標準的な線形モデルの場合、この関係は、

$$f = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$$

のように記述することができる。我々の分析の目標は、X が Y について持っている情報を集約した未知の関数 f を何らかの方法で特定し、特定された関数 $\hat{f}$ を予測に活用することである。これは、新しい顧客の情報が得られたとき、その顧客の購入・加入確率や入院リスクを予測することに相当し、$\hat{Y} = \hat{f}(X)$ を計算することで得られる。

(3) モデルを選ぶ方法

関数 f を何らかの方法で特定することを推定と呼ぶことにしよう。推定を行う際、複数の X のうち、どの入力変数が Y を予測するうえで重要なのかが先験的に知られている場合、重要な入力変数のみを使ってモデルの推定と予測を行えばよい。しかし、現実にはどの入力変数が重要なのか、分からないことのほうが一般的である。

このような場合、もっとも単純な方法は入手可能なすべての入力変数を使って推定を行う方法である。これは誰でも思いつく単純な

1　経済・経営学系の分野において、標準的な統計学との親和性が高い問題はそれほど多くない。なぜなら、経済・経営学系のデータの多くは経済主体の非連続的な行動変容を潜在的に含んでいるにもかかわらず、標準的な統計モデルは経済主体の意思決定を明示的に組み込んでいない（誘導型という）ためだ。経済主体の意思決定が変化する以前の行動に基づいた相関関係をいくら正確に予測できたとしても、意思決定が少し変化した途端に予測能力が低下するのであれば、あまり実用的な予測とは言えないだろう。この意味で、医療情報から将来の入院リスクを予測するような問題は、入力変数と出力変数との間に経済主体の意思決定が介在しないため、統計学の知見を非常に活用しやすい問題といえる。また、顧客情報から商品の購入・加入確率を予測するような問題も、意思決定の変化による影響を受けにくいと考えられるため、統計学との親和性が比較的高い。しかし、社会的情勢の変化などで顧客の価値観が大きく変化するような事態が生じた場合、予測の妥当性が低下する可能性はあるだろう。このような場合、経済主体の意思決定を明示的に組み込んだモデル（構造型という）を用い、因果関係に基づく予測を行う必要がある。

方法であるが、入力変数の数が増えれば増えるほどモデルが複雑化して解釈が困難になる。そのため、入力変数の数はできる限り絞り込んで、必要最小限にとどめおくほうが好ましい。それでは、入力変数をどのように絞り込めばよいだろうか。

入力変数を絞り込むもっとも自然な方法は総当り、つまり、p 個の入力変数について考えられるすべての組み合わせを試し、何らかの「当てはまりの基準」に基づいて最良の変数の集合に絞り込む方法である。たしかに入力変数の数が少なければ、この方法は有効である。しかし、入力変数の数が増加するにしたがって、組み合わせの数が 2^p で幾何級数的に増加するため、計算が困難になり、場合によっては不可能になることが予想される。そこで、予測精度を保ちつつ、計算量を節約する方法を考えたい。

計算量を節約する方法としては、入力変数を段階的に増やしていくか、または減らしていくことで、入力変数の少ないモデルを選ぶ方法が分かりやすいだろう。入力変数を増やしていく場合を例にとると、入力変数がまったくないモデルを出発点とし、入力変数を１つ加えたとき、p 個すべてについて、推定された予測モデルとデータとの「当てはまりの基準」を計算する。当てはまりの改善が見られた場合、もっとも当てはまりがよかったモデルを新たな出発点として、さらに入力変数を１つ追加する。これを、当てはまりが改善しなくなるまで繰り返せばよい。このように変数を増加または減少させる計算方法は、必ずしも大域的に最適なモデルを導くことが保証されるものではないものの、予測精度と計算量とのより良いトレードオフを与えてくれる、という観点から有用な方法である。

⑷ 当てはまりの基準

ところで、「当てはまりの基準」として、我々はどのような統計量を採用するべきだろう。当てはまりの基準としてもっとも基礎的なものは平均２乗誤差（MSE）

$$MSE = \frac{1}{N}\sum_{i=1}^{N}\left(Y_i - \hat{f}(X_i)\right)^2$$

で、これが小さければ小さいほど、予測モデルの当てはまりがよいということができる。ただ、ここで注意したいのは、用いている Y_i が既存顧客の情報に過ぎない点である。というのも、我々の目的は既知の、既存顧客の属性と商品や既往歴との関係を正確に説明することではなく、顧客の属性から、将来の商品の購入・加入確率や入院リスクという未知の情報を予測することである。推定した $\hat{f}$ がどれだけうまく既存顧客のデータを説明できていたとしても、それは予測性能の高さを必ずしも意味しない。もちろん、この二つには重複する部分があるかもしれないが一致することはない。そこで、このギャップを埋めるため、モデルを選択する際には、マローズの C_p や自由度調整済み決定係数、赤池情報量基準（AIC）やベイズ情報量基準（BIC）といった手法や交差検証を用いる方法が提案されている。本稿は一般の読者を想定しているため、個別の方法について詳しい説明はしないが、これらの手法を用いると、重要でない入力変数もモデルに加えてしまう予測モデルの複雑化や、既存のデータでたまたま観測された例外的な性質を規則性として誤って認識し、予測モデルに反映してしまう、いわゆる過学習[2]

2　たとえば、生年月日や血液型など、加入する保険商品の選択に何の影響も及ぼさないはずの顧客情報について考えてみよう。たまたま、ある保険商品Ａに12月生まれで、血液型がＯ型の契約者が異常に多かった場合、予測モデルはこれらの要素を考慮し、予測に反映させる可能性がある。こうした要素の反映は既知のデータの当てはまりを改善するかもしれないが、将来の予測精度の低下に繋がるため、適切な処置とはいえない。

といった問題に陥る危険性を小さくすることができる。

3．正則化

前節では、すべての入力変数の部分集合を用いたモデルを推定し、当てはまりの度合いを比較することで、予測精度の高いモデルを選択する二段階の方法について議論した。二段階の方法は理解しやすいが、入力変数の数が大きくなると、変数増加や減少によって計算量を節約したとしても、あまり効率的な方法とはいえない。そこで、関数 f のパラメータの大きさ[3]に制約を設けたもとで推定を行い、パラメータの推定値を0に近づける[4]、または0にすることで、推定とモデル選択を同時に行う方法を次に考えよう。このように、元々あったパラメータ推定のための最適化問題に新しい制約を導入し、制約付き最適化問題として推定プロセスを再定義する手法を正則化（Regularization）という[5]。正則化法を使う場合、推定とモデル選択は同時に1回だけ行えばよいので、これまでに学んできた二段階の方法に比べて、計算量を大きく節約できるという利点がある。

正則化法を使う際は、元の最適化問題と新たに導入した制約との相対的な重み付けの度合いを決定する必要がある。この重みの値は正則化パラメータとよばれ、元のパラメータの推定値に影響を与えるため、値を適切に調整する必要がある。一般に、この値が大きければ大きいほど、モデルの柔軟性が低下するため、当てはまり性能が低下し、値が小さければ小さいほど、過学習の危険性が高まる。

正則化パラメータの決定には交差検証法が有効とされている。たとえば、k分割交差検証を例にとると、既存顧客のデータを同じサイズのk個のブロックに分割し、（k－1）個のグループを使ってモデルを推定、残りの1個のグループでMSEを計算する。これをk回繰り返し、その平均をとることで得られる交差検証誤差が最小になるように正則化パラメータを選べばよい。

パラメータの推定値を0に近づける代表的な手法がリッジ回帰である。リッジ回帰は、これまでに見てきたような入力変数を絞り込む方法とは異なり、常に p 個すべての入力変数を含んだ予測モデルを扱う[6]ため、とくに入力変数の数が多いとき、モデルの解釈が困難になるという欠点がある。

そこで、リッジ回帰とは異なる制約条件を用い、上記の欠点を克服した手法がLassoである。Lassoはリッジ回帰と異なり、正則化パラメータを十分に大きくすると、パラメータの推定値を0にすることができる。これは、モデルから一部の入力変数が除外されることを意味し、結果的に解釈が容易な予測モデルを作ることができる。

4．主成分分析

これまでの議論では、計算量を節約する方法として、予測に適した入力変数の部分集合を選ぶ二段階の方法、さらに制約付き最適化を行うことで、パラメータ推定値を0に近づける方法について学んできた。他方、入力変数を組み合わせて、入力変数の特徴をうまく再現できるような成分を抽出し、その成分を使って推定や予測を行うことで計算量を節約する二段階の方法もある。このように次元を

3　厳密にはノルムというべきである。
4　つまり、パラメータの次元は高次元に保ちつつ、関数の表現能力を抑える。
5　たとえば、L_1 ノルムを用いた正則化は L_1 正則化、L_2 ノルムを用いた正則化は L_2 正則化とよばれる。最小二乗法の誤差に L_1 ノルムを加えたものをリッジ回帰、L_2 ノルムを加えたものをLassoという。
6　正則化パラメータをどれだけ大きくしても、入力変数は除外されない。

削減するための方法を主成分分析と呼ぶ。

古典的な主成分分析の手法が主成分回帰である。主成分回帰は、入力変数から主成分を作成し、主成分を新たな入力変数とした予測モデルのパラメータを推定する。主成分回帰は、第一段階の主成分を作成する過程で、第二段階の予測がまったく考慮されていないため、モデルの予測性能を最大化するという観点からは、必ずしも適切な方法とはいえない。得られた主成分が、予測性能の観点から最適だとは保証されていないからである。

主成分回帰の欠点を改善する方法としては、たとえば部分最小二乗法が用いられることがある。部分最小二乗法は、主成分を作成する際、入力変数だけでなく出力変数の情報を利用することで、主成分回帰の問題に対処している。

5．高次元データと共線性

ここまで、変数を増加または減少させて入力変数を絞り込む方法、またリッジ回帰やLassoといった正則化の手法について学んできた。こうした手法がとくに有効なのは、入力変数の数 p が十分に多く、すべての組み合わせを試す計算が物理的に困難か不可能な状況である。

ただ、p が多ければそれだけ共線性、すなわち入力変数の間に強い相関が存在する可能性が高まる点には注意が必要である。共線性が存在する場合、各入力変数が出力変数に及ぼす影響を正確に識別することは原理的に不可能となる。そのため、正則化や変数の絞り込みを通じて得られたパラメータ推定値や予測モデルの妥当性は大幅に低下する。これはどういうことかというと、同じ母集団から独立に得られたデータを用いて同様の推定を行ったとしても、まったく異なった予測モデルが得られる可能性があることを意味する。予測の信頼性が大幅に低下するのは言うまでもない。残念ながら、こうした技術的困難を克服することは容易ではない。一種の過学習とみなし、得られた予測モデルを過度に信頼するのではなく、交差検証法などを用いてモデルの頑健性を確認するなどの作業が重要になるだろう。