Alternative Thinking | 2019 年第 2 四半期

機械は金融を「学習」できるか？

要約

機械学習は資産運用の役に立つのでしょうか？もし役にたつとしたら、どのようにしてでしょうか。金融市場は機械学習が成功を享受してきた多くの環境とは根本的に異なっており、また、資産運用向け機械学習の研究は始まったばかりです。初期の研究結果は、機械学習ツールが投資ポートフォリオを改善できる可能性を示唆しています。機械学習技法の応用は、投資研究の自然な進化であり、今後も探求が進む分野でしょう。

PSG
ポートフォリオ・
ソリューションズ・
グループ

目次

はじめに

データの爆発的な増加と向上を続けるコンピューターの計算能力が、さまざまな分野の研究者にとてつもない機会を（そして落とし穴も）もたらしています。最も好奇心をそそる研究の一部が、「機械学習」および人工知能の幅広い領域で起きています。過去 5 年間には、これらの方法を応用することにより数々の科学分野で研究面の大幅なブレークスルーが起きました。[1] とはいえ、機械学習とは正確には何を指し、資産運用においても同様な発見をもたらすことができるのでしょうか？最初に、機械学習が問題を解く方法をどのように変えるのかを説明するわかりやすい例から始めたいと思います。

機械プログラミングから機械学習へ

自動化が必要なタスクの一例を考えてみましょう。このタスクを達成する従来のコンピュータープログラミング手法と機械学習の手法を対比してみます。

タスクは、特定のメールアドレスが、ワールドワイドウェブがメッセージのルーティングをするのに使用できると言う意味で「有効」かどうかを判定することです。[2] メールアドレスが有効であるためには、一連の基本的な基準を満たさなければなりません。たとえば、「@」記号を含んでいる必要があります。@ の後ろには、254 文字未満で、文字、数字、ハイフンとピリオドだけでできたウェブドメイン（aqr.com や yale.edu など）が続く必要があります。メールアドレスの有効性には、その他にもいくつかの明確に定義されたルールがあります。[3]

この問題に対する、従来型のコンピュータープログラミングによる解決方法は、一連の if/then 文を書くことです（例: **図表 1** の左側）。プログラムは、すべての必要条件が満たされれば「有効」を返しますが、一つの条件にでも違反すれば「無効」を返します。適切にコーディングされたプログラムは、このタスクの一例を正確に解決できます。

それでは、データ駆動型の機械学習がこの問題をどのように解くのかを考えてみます。ルールを知りインプットする人間を頼らずに、コンピューターがデータからルールを「学習」できるのでしょうか？具体的には、有効なメールアドレスと無効なメールアドレスの例、すなわちデータから、コンピューターはアドレスを分類する独自のルールに到達出来るのでしょうか？[4] そうするために、

1　例の一部には、自動運転車、リアルタイムに英語を中国語に翻訳するマイクロソフト翻訳、囲碁でイ・セドルに勝利したアルファ碁、物体の操作を学習する OpenAI のロボットハンド「Dactyl」などがあります。
2　人々と企業は毎日さまざまなプログラムを使用してこのような問題を解決しています。一般に閲覧可能なブログ投稿として以下があります。https://isemail.info/about
3　例えば次を参照。https://help.returnpath.com/hc/en-us/articles/220560587-What-are-the-rules-for-email-address-syntax-
4　この例では、データの価値がその「有効/無効」のラベルにあり、これは人間が手作業で分類するか、実験をとおして生成されます。

図表 1

メールアドレスの有効性の判定

従来のプログラミング

IF “@”を含む

AND ドメイン名が“!#$...”を含まない

AND ドメイン名が 254 文字未満

AND ...

THEN 有効

ELSE 無効

機械学習

メールアドレスの（ビッグ）データ

有効？	アドレス
0	jaime@lannister
1	hound@clegane.com
0	jon.snow@GOt.edu
1	daenerys@targaryen.org
⋮	⋮

＋

統計モデル: $Y/N = b_0 + b_1(@) + b_2(!\#\$\&\%?*) \ldots$

＝

メールアドレスが有効である確率の推定値

出所: AQR。説明目的限定。

コンピューターは統計学を利用してデータからルールを推論します。たとえば、機械学習者は有効なメールアドレスと無効なメールアドレスの数百万もの例を機械に与え、機械は「@」記号の有無が区別する重要な決め手になることを発見して独自のルールにたどり着くのです。これは、研究者がデータを与えるだけの構造化されていない機械学習の例です。十分なデータと例があれば、機械は最終的に有用なルールを見つけ出します。代替的な方法は、ある程度のルールかガイダンスを与え、機械に改良させて、さらに追加させるというものです。たとえば、「@」記号、有効なウェブドメイン名、記号のバラエティなどの重要度の高い変数を事前に指定することもできます。この第 2 のアプローチはより構造化されていて、機械がより高速に効率的に学習する手助けをしてくれます。

構造化されていないアプローチでは、機械学習者はコンピューターに有効性ルールについて何も教えません。第 2 のより構造化されたアプローチでは、機械学習者が問題となりうる変数の種類について最小限の情報を与え、それらがどのように問題となるのか、どうしたらもっともうまく利用できるのか、他にも問題となる特徴があるかなどを機械に見つけさせます。当然のことながら、ほとんどの問題は第 2 のシナリオに類似していて、研究者は機械が効率的に問題を解決できる助けとなるような、変数または関係性に関する直観をある程度持っています。いずれのケースにおいても、機械はデータを信頼し統計モデルを*推定*することによってルールや分類する能力を*学習*しようとします。

5 当然のことながら、研究者の知識が豊富なほど、より多くのインプットを機械に与えてより効率的に学習させることができます。状況と問題の性質によって、構造化された学習も構造化されていない学習も、有益で効率的になり得ます。

モデルを「訓練」したのち（つまり、推定したのち）、新たなメールアドレスをモデルに対して提示します。新しいアドレスの属性（または「特徴量」）に基づいて、図表 1 の右側に示すとおり、良いモデルはそのメールアドレスが有効である確率の値を予測します。

従来のプログラマーも機械学習者も、同じ基本的ロジックに従います。つまり、ゲームのルールを決定するということです。両社のアプローチの違いは、積もり積もって視点の微妙な変化に結びつきます。この違いを理解するうえで役に立つ方法は、システムを 3 つの主要構成要素（インプット、連想ルール、アウトプット）に分けて考えることです。伝統的なプログラミングのアプローチは、インプットに始まり、続いて人間の理解する能力を使い、インプットをアウトプットに転換するルールである連想ルールを明示的に定義します。一方、機械学習アプローチはインプットとアウトプットがどのように関連しているかの例からスタートし、推定によって連想ルールを学習しようとします。

十分に柔軟性のあるモデルを使い、十分な数の例を与えれば、統計学の基本原則により、メールアドレスをほぼ完璧に分類する機械を訓練（つまりモデルを推定）できることが保証されます。

e メールアドレスの例では、機械学習者のアプローチはやや滑稽に見えます。無駄な労力を払わなくとも完璧に正確なプログラミングのソリューションが手の届く範囲にあるのに、どうして不完全な推定を行う必要があるのでしょうか？機械学習がこのような単純な問題に対する正しいアプローチでないのは明白です。あまりにも単純すぎるのです。しかし、現実世界の問題のほとんどはもっと複雑です。

次に、機械学習の代表的な問題であり、かつより解決が難しい問題の一例をお示しします。

これは猫の写真ですか？

従来のプログラミングのソリューションはこの種の問題に対してなすすべがありません。「これは猫ですか？」の問いに答えるには、ほぼ無限大の if/then 文のリストが必要となります。[6] この場合、プログラミングの解決法は実行不能ですが、機械学習者の解決法は実現可能です（ただし、猫と猫以外の画像を何百万も与えればの話ですが）。グーグルやその類いからわかるように（Le et al., 2011 年）、機械学習は猫の画像の特定に驚異的な正確さを示します。そのデータのすべてを使ってそうした計算のすべてをやり遂げる能力こそが、機械学習技法を便利で有益なものにしたのです。

6　たとえば、この写真の特徴を誰かに口頭で伝えようとして、それがどの動物だと思うか、あるいはその説明が猫のものかどうかをその人に尋ねてみてください。

何が新しく何がそうでないのか: データ、計算能力、および統計学

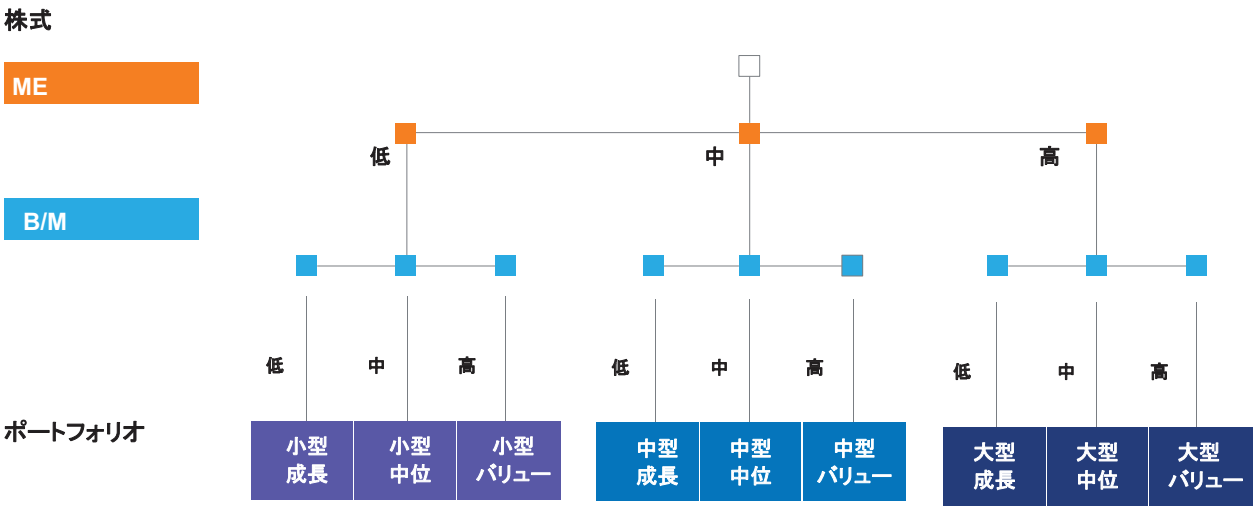
機械学習はさまざまな名前で呼ばれています（その一部は特徴を外していますが）。「深層学習」や「人工知能」と呼ばれていたとしても、実際、通常の金融への応用では、すべて統計学（の多用）の直接的拡張として理解できます。機械学習における「学習」とは、単に推定とモデル選択を意味します。実際、使われている統計学の基本原則のほとんどは、数十年前から存在するものです。何が新しいかというと、テクノロジーの発達によって膨大な量の新しいデータが生み出されるとともに、コンピューターの極めて高い計算能力によって大規模な統計モデルを実務上利用できるようになり、機械学習が実行可能になったことです。

2 つの例: 決定木とニューラルネットワーク

機械学習の基になる基本的な統計的原理に親しんでもらうため、機械学習の手法における 2 つの柱（決定木とニューラルネットワーク）を簡単に説明します。

順次ソートとも呼ばれるツリーモデルは、ポートフォリオを組む金融研究者にはおなじみの統計的概念です。というのも、そのような手法を用いてポートフォリオを組むことが多いからです。たとえば、**図表 2** では、株式リターンの観測値とともに、2 つの「ソート」変数として企業の時価純資産（Market Equity, “ME”）と簿価時価比率（Book to market, “B/M”）を想定しています。ツリーはまず企業サイズに基づいて株式をソートし、サイズの面で類似したいくつかのグループに分けます。[7] グループの数は、サイズの面でグループがどのように「異なっているか」や、サイズとリターンとの関係性によって決定されます。次に、各サイズグループの中で、株式が B/M に基づいてさらにソートされます。ツリーの最終的な「葉」は、これらの特性の面で互いに類似した株式のグループで構成され、それが株式のポートフォリオになります。

図表 2
ポートフォリオツリーの例



出所: AQR。説明目的限定。

7　ツリーモデルは、各枝で観測値を 2 つのグループの分ける二分木であることが多いのですが、（上記の例の三分構造のように）その他の選択肢もあります。

一例として「大型バリュー」銘柄のモデルツリーのリターン予想とは、単に大型バリュー・ポートフォリオの平均リターンのことです。金融の学術研究では、何十年もの間、ポートフォリオソートを使用してきましたが（Fama and French, 1992）、これこそ決定木が行っていることなのです。

同様に、ニューラルネットワークモデルの考え方も古くからあり、元来は 1940 年代と 1950 年代の神経科学者らによって考え出されたものです（McCulloch and Pitts, 1943; Rosenblatt, 1958）。ニューラルネットワークをめぐるミステリーの一部は、その神経科学用語に由来しています（「ニューロン」、「活性化関数」、「結合」）。しかし、ニューラルネットワークの基となる統計学の基本原則は単純明快でなじみやすいものです。

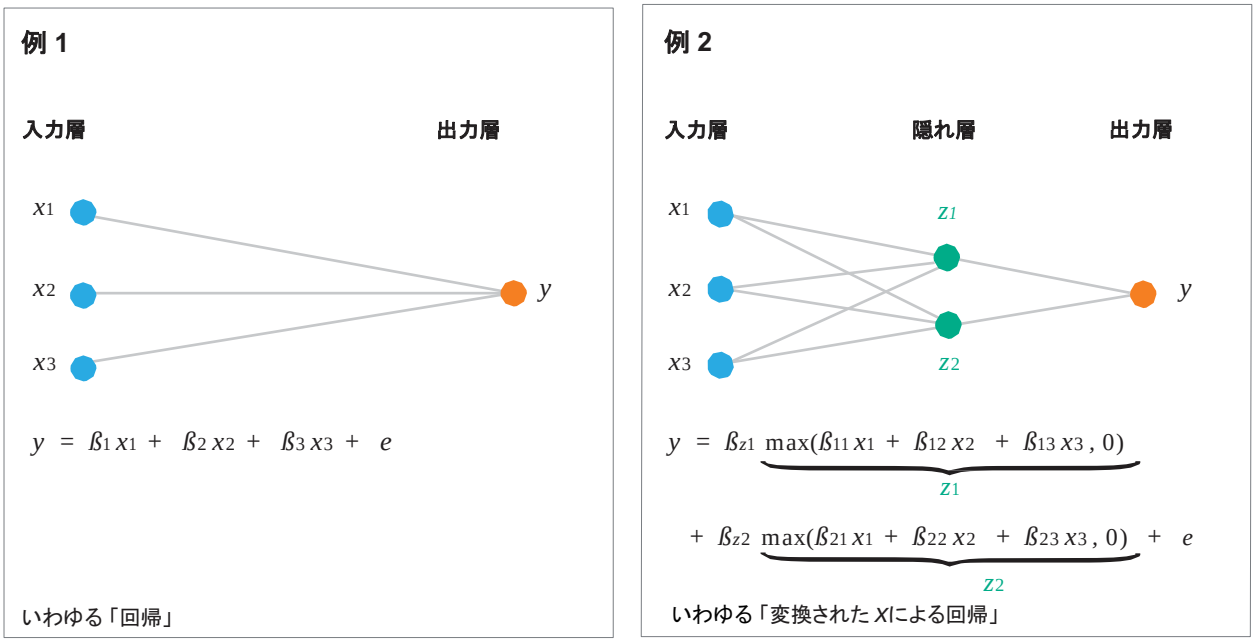
図表 3 では、2 つの単純なニューラルネットワークの例を示しています。最初の例は、単一の「入力層」と「出力層」を持つ、ニューラルネットワークとしては最も単純な「アーキテクチャ」を示しています。入力から出力への線は

ネットワークの中の「結合加重」、いわば影響の状態を表します。この例では影響の流れは左から右への一方向になっていて、この単純な「順伝播型（フィードフォワード）」ネットワークを構成しています。入力は単純に予測子/リグレッサー/独立変数（x）で、出力は従属変数もしくは結果（y）です。目的は、入力が出力にどのように影響を及ぼすのかを理解して、予測に役立てることにあります。

たくさんの用語が出てきましたが、このモデルの推定は単純にできます。例の中で β と表記された「結合加重」は、最小二乗法（OLS）回帰分析によって求められます。もちろん、手が込んでいるだけの統計関数を使って、より複雑な「結合加重」を求めることもできます。

例 2 では、「隠れ層」と呼ばれる複雑性の層を加えています。隠れ層は入力層と出力層の中間に位置し、ひねりを加える結合です。ニューラルネットワークの用語を使えば、隠れ層は

図表 3
ニューラルネットワークの簡単な図解



出所: AQR。説明目的限定。

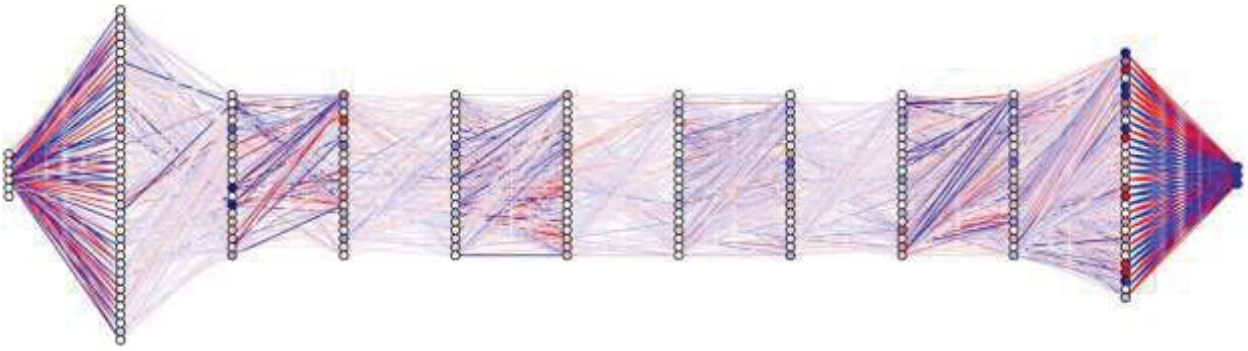
z1 と z2 で表された 2 つの中間「ニューロン」からなり、それが入力層から情報を受け取り、何らかの形で加工して結果を出力層へと送ります。統計学の言葉を使えば、変数 x の変換を施していることになります。
例2では、xの正の値を取り出し、負の値はゼロにして出力します。こうした隠れ層の使用がニューラルネットワークを非常に強力にしているのです。

隠れ層はモデルに複雑性を加えますが、基本的なアイデアは同じです。y を x1、x2、x3 に対して回帰せずに、モデルはまず z1 と z2 に加工して、次に y が z1 と z2 にどのように関係しているのかを計測します。回帰分析でレグレッサーを使用する前に変換する（例: ボラティリティによるスケーリング）ことは、資産運用の研究でよく行われています。ニューラルネットワークは、事前にデータ変換を特定せずに、この段階をモデルの*内部*に持ち込んでいるのです。統計学を使って、y の最良の予測子を学習するため多くの可能な変換を調べます。これは、信頼できる推定を行うのにとてつもない処理能力と膨大なデータを要する力強いイノベーションなのです。

ビッグデータ、高速プロセッサー

図表 3 の例は単純です。しかし、ニューラルネットワークはどのようにしてより難しいタスクをこなせるのでしょうか？上記のような単純なニューラルネットワークがたくさんあるとして、それらをあらゆる興味深く複雑な方法で積み重ねれば、極めて複雑なモデルが得られます。**図表 4** は「深層」ニューラルネットワークのタイプの一例を示しています（「深層」とはネットワーク内に隠れ層とネットワークが多数あることを意味します）。この例は、地震モデリング、コンピュータービジョンや自動運転車など多様なアプリケーションで成功が実証されています。数千個もの多数の小さなネットワークを積み重ねることによって、非常に柔軟性があって、幅広い結果と現実世界の現象の複雑性を記述するのに必要な変数間の相互作用を捉えることができるモデルが得られます。こうしたことは、膨大なデータとネットワーク経路のすべてを計算するだけのとてつもない計算能力があって初めて可能になります。歴史的には、統計モデルが猫の画像を認識するのを妨げていたのは、

図表 4
地球物理学からの深層ニューラルネットワークの例



出所: DeVries et al.（2017 年）。注記: 図は、地震の挙動を説明するのに有用なことが実証されたニューラルネットワークのアーキテクチャを示しています。

統計モデルがわずかなパラメータしかない「小さな」ものだったからです。なぜでしょうか？それは計算能力と記憶装置が不足していたためです。参考までに、1983 年当時の Apple IIe は、64KB の RAM と 1.02MHz のプロセッサーを誇っていました。筆者が本稿の執筆に使用している 2015 年発売の MacBook Pro は 16GB の RAM と 2.8GHz のプロセッサーを積んでおり、Apple IIe を RAM の容量で 26 万倍、プロセッサー速度で 2,800 倍も上回っています。[8] さらに、今日の統計分析を走らせる機械に比べたら MacBook Pro も形無しです。そのため、機械学習における技術革新の巨大な飛躍は、技法というよりはテクノロジーの進歩によって成し遂げられたものだと言えます。

さらに、計算能力の向上に伴って、膨大な情報源が自由に使える状況になっています。毎年、私たちは計算能力とデータの飛躍が勢いづいている様を目にしています。[9] 私たちがデータを捉えて蓄積する能力は、それを分析して理解する能力をはるかに超えており、そのため機械学習がそのギャップを埋めるうえで有益になり得るのです。先の長い話ですが。

金融は別物である

メディアが盛んに取り上げているように、機械学習はかつてなら考えられなかったことを実現しています。機械は猫を認識できるばかりか、人の話を認識し、車を運転し、複雑な戦略ゲームの世界チャンピオンを負かすことができるまでになっています。機械は何でもできるかのように思えます。通常はこの辺で興奮や誇大広告や根拠の薄い推測が入り込んできます。機械学習はあまりにも多くの驚異的なことを成し遂げてきたので、株式銘柄選択のような金融のタスクまで席巻しまうのが当然の結末のように思えるかもしれません。しかし、この結論は決して明白ではなく、まだ研究の裏付けも得ていません。

何が金融を別物にしているのでしょうか？本セクションでは、機械学習が証明済の良好な実績を持つ分野と金融とを分け隔てているいくつかの特徴を取り上げます。

低いシグナル対ノイズ比

おそらく最も重要な違いは、あるシステムの中にどれだけの予測可能性があるのかを要約したシグナル対ノイズ比に関連して理解できます。たとえば、猫の画像認識を考えてみましょう。千枚のインスタグラムの画像を渡されれば、人間ならばその中で猫を含む画像をほぼ確実にみつけることができるでしょう。ピンボケ写真の中の猫や、おかしな姿の犬を見間違えたりすることも 1 枚や 2 枚はあるでしょう。そうした高い成功率は、この設定が高いシグナル対ノイズ比の環境であることを示しています。信号（猫の画像）が、写真の中のノイズ源（ピンボケ、背景画像など）を圧倒しています。機械学習は、そうした環境で最も力を発揮します。

これを金融、とりわけリターン予測と対比してみてください。シグナル対ノイズ比は弱いどころか、絶えずゼロに近いところまで引き下げられることになります。まず、シグナル対ノイズ比が弱い一つの理由は、

8　http://applemuseum.bott.org/sections/computers/IIe.html. を参照。
9　ソフトウェア会社の Domo は、2018 年の平均的な 1 日の平均的な 1 分間に、人類は 1,300 万ものテキストメッセージを送り、390 万回のグーグル検索を走らせ、1,400 回ウーバーを利用し、49,800 枚の写真をインスタグラムに投稿し、68,500 ドルの Venmo 取引を精算したと推定しています（https://www.domo.com/learn/data-never-sleeps-6).こうしたデータのすべてが記録され蓄積されています。

金融市場にはノイズが極端に多いためです。世界最高の株式銘柄や投資ポートフォリオであれ、どの 1 日、四半期、あるいは 1 年を取っても、予想もしていなかったニュースによってパフォーマンスの大きな振れを経験しています。[10] 次に、金融市場のシグナルは低いと予想されており、今後も低いままで維持されるでしょう。低いシグナル対ノイズ比は、市場の不幸な偶然の一致というわけではないのです。むしろ、それは利益の最大化と競争という単純なわかりやすい経済の動きによって保証され、絶えず強化されている特徴なのです。仮にトレーダーが将来の価格上昇を確実に予測できる情報（強いシグナル）を持っているとしたら、その情報に対して受け身で座っていることはあり得ません。必ずトレーディングを開始します。彼らの予測情報を活用する行為は価格を押し上げるため、市場から予測可能性の一部を吸い上げます。そして価格がわずかに上昇しても、彼らは買いの手を緩めません。彼らは持っている情報を使い果たすまで、その情報が予測した水準まで価格が完全に調整されるまで買い続けます。利益指向のトレーディングに情報を活用することによって、投資家には最小限の予測可能性しか残らなくなります。予測可能性がすでに価格に織り込まれている場合には、市場を動かす唯一のことは、予期していないニュースかショック、すなわちノイズということになります。市場における競争がリターンの予測可能性を失わせるという発想は新しいものではありません。ノーベル賞を受賞した効率的市場仮説（Fama、1970 年）に関する研究の根底にある発想そのものなのです。

進化する市場

低いシグナル対ノイズ比が提起する機械学習の課題は、市場が持つ適応していく性質とダイナミックな特性によってさらに複雑になっています。研究者が資産価格を予測するのに役立つミスプライシングを捉える新たなシグナルを見つけたとしたら、そのシグナルは広く知られていくにつれ、さらに多くのトレーダーがシグナルに基づいて行動するため、価格はより迅速に修正されます。最終的に市場がその情報を吸収し、市場にいるエージェントの行動そのものによってデータ発生プロセスは変化します。同様に、技術的イノベーションが経済の構造変化を引き起こす可能性があり、人間の市場との相互作用を再形成します。機械学習の最前線では、そうした適応的現象に役立つ可能性のあるツールを開発していますが（Arora et al., 2012、Li and Hoi, 2014 のオンライン学習アルゴリズムなど）、それは金融が他の多くの機械学習研究よりも複雑だという事実を浮き彫りにしています（猫の画像の例では、アルゴリズムが猫の認識ができるようになってから猫が犬に変身し始めるようなことはありません）。

短いサンプルと構造化されていないデータ

金融（および経済学全般）が持つもう一つの大きな違いとして、研究分野が実際には「ビッグデータ」の環境にないことが挙げられます（それでもビッグデータの分析手法は有用ですが）。金融の統計的分析、より広くはマクロ経済学などは基本的に時系列の分野です。リターン予測の例では、

10 市場のボラティリティの意味を伝えるならば、平均的な個別株式には年間 5%の現金を上回る期待リターンがあり、年間 50%のリターンのボラティリティがあるのです。ボラティリティは期待リターンの 10 倍あるというわけです。これは仮想的シナリオで、説明のみを目的としています。仮想的シナリオには、予想結果に関して本来的な限界があります。

11 効率的市場において、リターンの予測可能性は必ずしも完全に失われるわけではありません。たとえば、情報の利用が過度なリスクテークを要求したり、投資家が取引コストに直面したり、あるいはインサイダー取引の場合のように法律上の制約に従わねばならないとすれば、投資家がすべての情報を利用するに至らない可能性があります。それでも、容易に得られる利益があれば、競合するトレーダーによって捉えられてしまうので、残る予測可能性は小さく、捉えるのが難しいはずです。予想値は説明目的限定で、パフォーマンスを保証するものではなく、変更される場合があります。配分はいつでも変更される可能性があります。

予測のためにより大きくより良いデータセットを常に使うことができます。しかし、私たちが目標とする結果の変数（例えば株式リターン）の観測値の数には限界があります。株式リターンの新たなデータは、時間の経過によってのみ生まれるのです。

資産運用における従来のインプットは、エクセルのスプレッドシートに収まるうまく構造化されたデータの類いです。列が予測変数で、行が反復観測値であって、こうしたデータは容易に統計分析に使えるタイプのデータです。 対照的に、興味深い新たなデータ情報源の多くには、「構造化されていない」データという特徴があります。そうしたデータには、ニュース記事やツイートのようなテキストデータ、インスタグラムへの投稿や YouTube の動画といった画像データ、そして詳細な指値注文の履歴のような形式の市場データまでもが含まれます。

ほとんどの構造化されていないデータセットについて、データの履歴は短くなっています。たとえば、ソーシャルメディアの投稿にはせいぜい 10 年分のデータがあるだけです。限られた期間の時系列しかなければ、意味のあるバックテストを行うことが難しくなります。履歴が短ければ、戦略のパフォーマンスを正確に推定することがさらに難しくなり、最終的には非常に強いシグナルさえもポートフォリオの中でわずかなウェイトしか与えられなくなる場合があります。

解釈できることが必要

一部の機械学習モデルはいわゆるブラックボックスです。ですが、モデル内部の仕組みが理解可能であることが、資産運用では使い勝手の良さになります。アセットマネージャーには、クライアントのポートフォリオが負うリスクを理解して伝える受託者責任がありますので、モデルが解釈可能であることを特に重視することになります。

金融だけが解釈可能なモデルを必要としているわけではありません。医師は、機械学習の医療診断の決定要因を理解して、アルゴリズムに頼ったために生じる意図せざる悪い結果を避けようとしますし（Cabitza et al., 2017）、政府と規制当局は政策の中に暗示的または明示的な偏りがないか常に気を配っています（金融機関の融資判断など（Hardt et al., 2016））。このように広範な需要があるため、解釈可能性は機械学習の研究における一つの優先事項になっています（Doshi-Velez and Kim, 2017; Vellido et al., 2012）。機械学習が不透明なブラックボックスである必要はありません。構造的アプローチによって、発見を強化するためにデータを効率的に使用すると同時に解釈と直観的理解も与えることができます。より意味があって直観的に理解できる結論を金融の機械学習モデルから引き出せる興味深い研究の方向性がたくさんあります。

研究の最前線

金融と、機械学習が発展するそれ以外の分野との間には決定的な違いがあるため、「機械は金融を学習できるか？」という問いかけに対して簡単には答えられません。業界として、資産運用における機械学習の有益性について、私たちには取り急ぎの理解しかありません。だからこそ、この分野の研究が貴重なのです。問題の大きさは巨大であり、今後進むべき道は掘り下げていって研究に精を出すことなのです。

逸話ではなく、分析が必要

人々が金融における機械学習を議論する場合、「自分はマネージャーXYZ がどのようにそれを行ったかの話を聞いた」などという「逸話」がよく話題になります。資産運用分野での機械学習の利点に関する方法論研究はまだ未成熟な段階にあります。しかし、初期の研究には希望の持てる話があります。Gu, Kelly, and Xiu (2018)は、機械学習法には株式銘柄選択戦略のパフォーマンスをアウトオブサンプルで有意に改善する可能性があることを示唆しています。彼らは、そのアウトパフォームの性質に対する新たな洞察も与えています。たとえば、改善はより高度なモデル（ツリー構造やニューラルネットワーク）の間で最も顕著に現われており、それは単純な手法にはない非線形予測子の相互作用を取り入れいることに負うところが大きいというものです。成果は漸進的で、経済的にも統計的にも有意ですが、決して革命的ではありません。

経済理論と機械学習の組み合わせ

統計分析の基本原理は、理論とモデルパラメータが代替し合うということです。モデルの中により多くの構造を組み入れれば、より少ないパラメータの推定で済むようになり、モデルはより効果的に観測値を使ってノイズを遮断できるようになります。つまり、モデルはノイズを遮断できるから役に立つのです（ニューヨーク市の地図は細かい点を省略しているからこそ市内を動き回るのに役に立つのです）。しかし、過度に単純化されたモデルは一部のシグナルを遮断してしまうことがありますから、データが豊富でシグナル対ノイズ比が高い環境においては、不必要に小さいモデルを使わない方が賢明です。しかしながら、シグナル対ノイズ比が低く、ノイズを遮断するメリットがシグナルの一部を失うコストを上回る場合には、単純さが美点になります。資産運用においては、シグナル対ノイズ比が低い問題にまず経済理論を持ち込んでデータのある側面を説明させることによって取り組むことから始め、データのうち理論が何も語らない側面を捉えるために機械学習ツールを使って補完するという可能性があります。それが当てはまる学術研究の例には、Kelly et al. (2017)、Kelly et al. (2018)、および Gu, Kelly, and Xiu (2019)があり、基本的経済構造から始めて、モデルの一つの側面にのみ機械学習を導入しています。

理論の活用に加えて、賢明な方法論的イノベーションが、金融の機械学習アプリケーションの中のノイズに対処するうえで役に立ちます。猫の画像認識は、機械学習の初期の成功例でしたが、シグナル対ノイズ比が高かったため、この作業は手の届くところにぶら下がっている果実のようなものでした。

現在の研究の最前線は、水中で撮影された写真の画像を認識するとか（Jin and Liang, 2017）、大勢の人がいる部屋での音声認識（Serdyuk et al., 2016）などといった、ノイズが多く含まれたより難しい問題へと移っています。

リターン予測だけではない

リターン予測が（低いシグナル対ノイズ比と非定常的性質がゆえに）機械学習にとりわけ難しい課題を突きつけていることを強調しましたが、その他の重要な金融分野の問題も機械学習からもっと恩恵を受ける可能性がある点を認識することも重要です。主な例は、リスク管理、取引コスト管理やファクター構築などを含むポートフォリオの実行です。

Engle (1982) のノーベル賞受賞研究を一例とする長い文献リストは、金融市場のリスクには高水準の予測可能性があることを示しています。

つまり、リスク予測は比較的高いシグナル対ノイズ比という恩恵を受けるということです。それに関連して、価格インパクトから生じる取引コストはかなり高水準の確度で予測可能です（Frazzini et al., 2018）。リターン予測とは異なり、投資家行動がこの予測可能性を消し去る明白な傾向はなく、おそらく機械学習はリスクと取引コストモデリングにより適していると言えるでしょう。

つまり、アルファだけがすべてではないのです！金融に応用された機械学習のほとんどの議論と、ならびに「逸話」のほとんどは、アルファの創出に特化しています。新しいデータと機械学習をアルファの構築に使うことは（つまり、新しい独特なリターン予測可能性の源泉を見つけること）、金融市場の中で最も競争が激しい分野に向かうことになります。より多くの投資家が同様なデータとツールを用いて市場に参入すれば、ミスプライシングが修正され、アルファはゼロへと圧縮されます。反対に、資産運用研究の有望な分野では機械学習を投資の別の側面を改善するために使っています。

まとめ：進化だが革命ではない

金融機械学習は、定量的運用の分野における次なる大きな飛躍となる可能性があります。資産運用実務における機械学習の現状を理解するには、2 つのキーポイントがあります。第一に、研究はまだ始まったばかりで、多くの重要な疑問にまだ答えられていません。第二に、初期の研究成果によると、機械学習ツールを活用してポートフォリオのパフォーマンスを経済的にも統計的にも有意に改善できる可能性があります（Gu, Kelly, and Xiu 2018）。しかし、成果は進化的ではありますが、革命的ではありません。

機械学習の背後にあるアイデア、すなわち新しいデータセットを活用して頑健で加法的なポートフォリオのパフォーマンスを見つけ出し、定量的方法を用いて体系的に情報を抽出することは、定量的運用プロセスの常套手段です。何十年間にもわたり、アセットマネージャーは人力に頼った分権的なやり方で統計的学習を行ってきました。機械学習はそのプロセスを機械化するシステマティックなアプローチを投資プロセスに提供します。そして機械学習は、マネージャーが以前は活用されていなかった構造化されていないデータを含むより新しい情報源からの情報を高速に処理することを可能にします。また、金融市場の複雑な現実をうまく捉えようとして次第に柔軟性を高める経済モデルによって検索するツールを提供します。金融分野での機械学習の進化は、まだ始まったばかりです。

ポートフォリオ・ソリューションズ・グループについて

ポートフォリオ・ソリューションズ・グループは、AQR のクライアントがより優れたポートフォリオの成果を達成するのを助け、幅広い投資コミュニティにユニークな洞察をお届けすることを目指しています。

本稿の研究について、私たちは Bryan Kelly, Ronen Israel, Tobias Moskowitz の各氏の貢献に感謝いたします。また、有益なコメントをくれた Gregor Andrade, Pete Hecht, Antti Ilmanen, Michael Katz, Lasse Pedersen, Dan Villalon の各氏にも感謝いたします。

References

Arora, Sanjeev, Elad Hazan, and Satyen Kale, 2012, The multiplicative weights update method: a meta-algorithm and applications, *Theory of Computing* 8, 121−164.

Cabitza, Federico, Raffaele Rasoini, and Gian Franco Gensini, 2017, Unintended consequences of machine learning in medicine, *JAMA* 318, 517−518.

DeVries, Phoebe MR, T Ben Thompson, and Brendan J Meade, 2017, Enabling large-scale viscoelastic calculations via neural network acceleration, *Geophysical Research Letters* 44, 2662−2669.

Doshi-Velez, Finale, and Been Kim, 2017, Towards a rigorous science of interpretable machine learning, *arXiv preprint arXiv:1702.08608*.

Engle, Robert F, 1982, Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation, *Econometrica: Journal of the Econometric Society* 987−1007.

Fama, Eugene F, 1970, Efficient capital markets: A review of theory and empirical work, *The Journal of Finance* 25, 383−417.

Fama, Eugene F, and Kenneth R French, 1992, The cross-section of expected stock returns, *The Journal of Finance* 47, 427−465.

Frazzini, Andrea, Ronen Israel, and Tobias J Moskowitz, 2018, Trading costs.

Gu, Shihao, Bryan Kelly, and Dacheng Xiu, 2018, Empirical asset pricing via machine learning, Technical report, National Bureau of Economic Research.

Gu, Shihao, Bryan T Kelly, and Dacheng Xiu, 2019, Autoencoder asset pricing models, *Available at SSRN*.

Hardt, Moritz, Eric Price, Nati Srebro, et al., 2016, Equality of opportunity in supervised learning, *Advances in neural information processing systems*, 3315−3323.

Jin, Leilei, and Hong Liang, 2017, Deep learning for underwater image recognition in small sample size situations, *OCEANS 2017-Aberdeen*, 1−4, IEEE.

Kelly, Bryan, Seth Pruitt, and Yinan Su, 2018, Characteristics are covariances: A unified model of risk and return, Technical report, National Bureau of Economic Research.

Kelly, Bryan T, Seth Pruitt, and Yinan Su, 2017, Instrumented principal component analysis.

Le, Quoc V, Marc'Aurelio Ranzato, Rajat Monga, Matthieu Devin, Kai Chen, Greg S Corrado, Jeff Dean, and Andrew Y Ng, 2011, Building high-level features using large scale unsupervised learning, *arXiv preprint arXiv:1112.6209*.

Li, Bin, and Steven CH Hoi, 2014, Online portfolio selection: A survey 46, 35.

McCulloch, Warren S, and Walter Pitts, 1943, A logical calculus of the ideas immanent in nervous activity, *The bulletin of mathematical biophysics* 5, 115−133.

Morgan, James N, and John A Sonquist, 1963, Problems in the analysis of survey data, and a proposal, *Journal of the American statistical association* 58, 415−434.

Rosenblatt, Frank, 1958, The perceptron: a probabilistic model for information storage and organization in the brain, *Psychological review* 65, 386.

Serdyuk, Dmitriy, Kartik Audhkhasi, Philemon Brakel, Bhuvana Ramabhadran, Samuel Thomas, and Yoshua Bengio, 2016, Invariant representations for noisy speech recognition, *arXiv preprint arXiv:1612.01928*.

Vellido, Alfredo, Jose David Martın-Guerrero, and Paulo JG Lisboa, 2012, Making machine learning models interpretable, *ESANN*, volume 12, 163−172, Citeseer.

Disclosures

This document has been provided to you solely for information purposes and does not constitute an offer or solicitation of an offer or any advice or recommendation to purchase any securities or other financial instruments and may not be construed as such.The factual information set forth herein has been obtained or derived from sources believed by the author and AQR Capital Management, LLC ("AQR") to be reliable but it is not necessarily all-inclusive and is not guaranteed as to its accuracy and is not to be regarded as a representation or warranty, express or implied, as to the information's accuracy or completeness, nor should the attached information serve as the basis of any investment decision.This document is not to be reproduced or redistributed to any other person.The information set forth herein has been provided to you as secondary information and should not be the primary source for any investment or allocation decision.

Past performance is not a guarantee of future performance.

This presentation is not research and should not be treated as research.This presentation does not represent valuation judgments with respect to any financial instrument, issuer, security or sector that may be described or referenced herein and does not represent a formal or official view of AQR.

The views expressed reflect the current views as of the date hereof and neither the author nor AQR undertakes to advise you of any changes in the views expressed herein.It should not be assumed that the author or AQR will make investment recommendations in the future that are consistent with the views expressed herein or use any or all of the techniques or methods of analysis described herein in managing client accounts.AQR and its affiliates may have positions (long or short) or engage in securities transactions that are not consistent with the information and views expressed in this presentation.

The information contained herein is only as current as of the date indicated and may be superseded by subsequent market events or for other reasons.Charts and graphs provided herein are for illustrative purposes only.The information in this presentation has been developed internally and/or obtained from sources believed to be reliable; however, neither AQR nor the author guarantees the accuracy, adequacy or completeness of such information.Nothing contained herein constitutes investment, legal, tax or other advice nor is it to be relied on in making an investment or other decision.

There can be no assurance that an investment strategy will be successful.Historic market trends are not reliable indicators of actual future market behavior or future performance of any particular investment which may differ materially and should not be relied upon as such.Target allocations contained herein are subject to change.There is no assurance that the target allocations will be achieved, and actual allocations may be significantly different than that shown here.This presentation should not be viewed as a current or past recommendation or a solicitation of an offer to buy or sell any securities or to adopt any investment strategy.

The information in this presentation may contain projections or other forward-looking statements regarding future events, targets, forecasts or expectations regarding the strategies described herein, and is only current as of the date indicated.There is no assurance that such events or targets will be achieved and may be significantly different from that shown here.The information in this presentation, including statements concerning financial market trends, is based on current market conditions, which will fluctuate and may be superseded by subsequent market events or for other reasons.Performance of all cited indices is calculated on a total return basis with dividends reinvested.

Diversification does not eliminate the risk of experiencing investment losses.Broad-based securities indices are unmanaged and are not subject to fees and expenses typically associated with managed accounts or investment funds.Investments cannot be made directly in an index.

Neither AQR nor the author assumes any duty to, nor undertakes to update forward looking statements.No representation or warranty, express or implied, is made or given by or on behalf of AQR, the author or any other person as to the accuracy and completeness or fairness of the information contained in this presentation, and no responsibility or liability is accepted for any such information.By accepting this presentation in its entirety, the recipient acknowledges its understanding and acceptance of the foregoing statement.

263856

www.aqr.com

AQR Capital Management, LLC Two Greenwich Plaza, Greenwich, CT 06830 **P** +1.203.742.3600 **F** +1.203.742.3100