集団化する人工知能　鈴木教授による解説シリーズ　⑨
～深層学習による株価予測 (後編)～

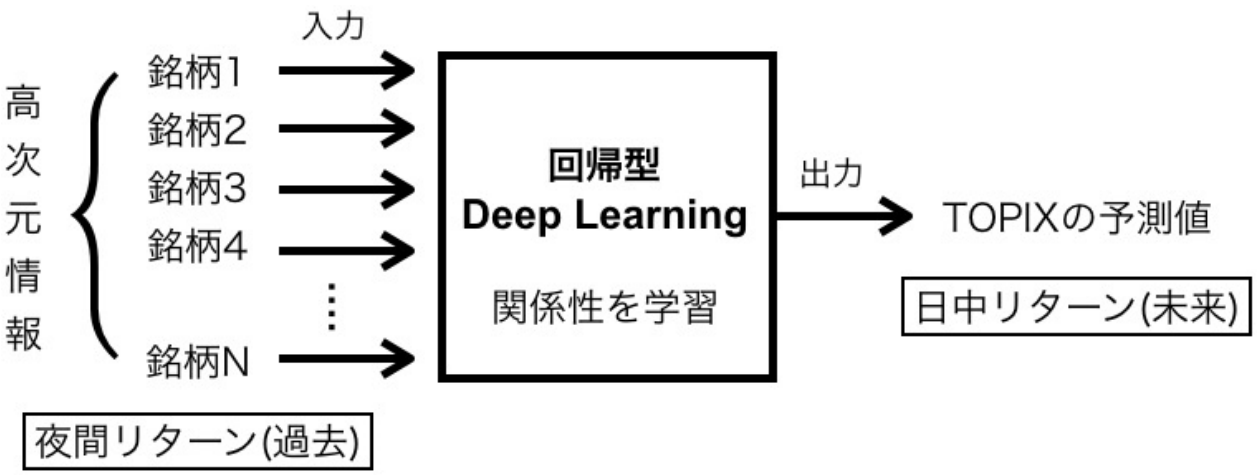
2018年10月1日

お伝えしたいポイント

- 機械学習を妨げる次元の呪い
- オートエンコーダによる次元圧縮
- 深層学習の事例紹介 (TOPIXの予測)

＜機械学習を妨げる次元の呪い＞

前回のマーケットレターの続きとして、再びTOPIXの予測を考えてみましょう。TOPIXは東証一部上場銘柄の全てから構成されるため、「入手できる全銘柄の株価を与えると、未来のTOPIXを予測する」AIを作ってみましょう。具体的には下図のように、「夜間のリターン (変化率) を入力すると、翌日の日中のリターンを出力する」箱を作ります。もし日中のリターンがプラスと予測されたならば、朝の市場開場時にTOPIX先物を買っておきます。



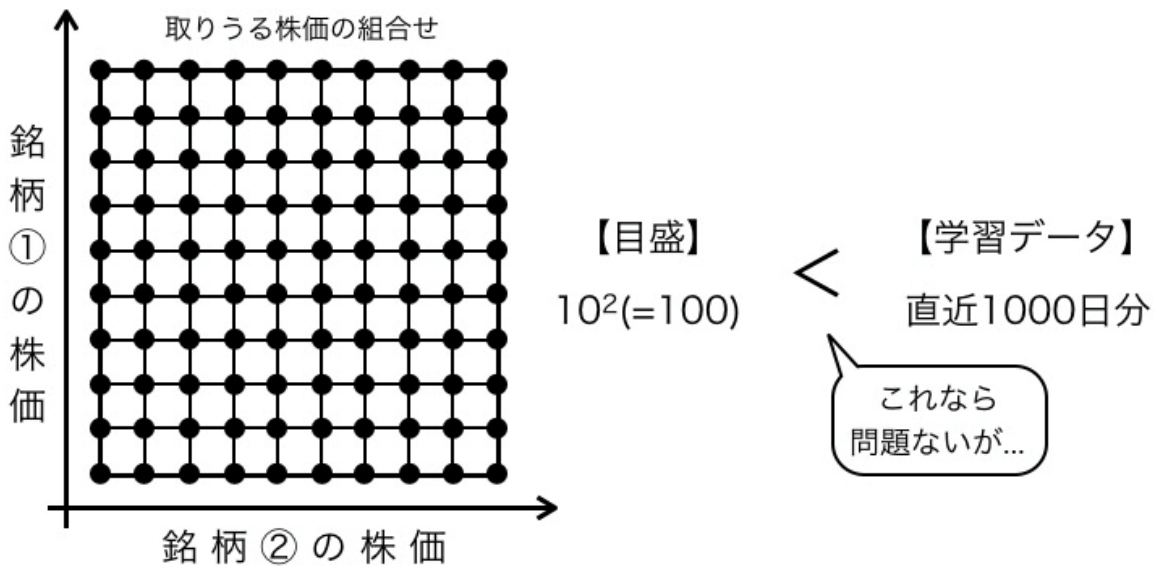
出所：大和証券投資信託委託

しかし機械学習の注意点として、入力するデータの種類が多いほど「次元の呪い」という厄介な問題が発生します。せっかく入手したデータなのだから全てAIに投入したい気持ちは分かりますが、無策で投入すると学習を悪化させる危険性があります。その理由は主に次の3点です。

1. 学習データ数が不足する

たとえば直近1000日分の株価データを用いて機械学習するとします。ここで単純に、株価は10通りの値しか持たないとします。もし3銘柄しか用いないならば、取りうる株価の組合せは10^3(＝1000)通りになり、学習データ数と同等です。

しかし300銘柄に増やすならば、取りうる株価の組合せは10^{300}通りに及ぶため、組合せ数に対して圧倒的に学習データ数が不足します。これは過学習の原因になります。

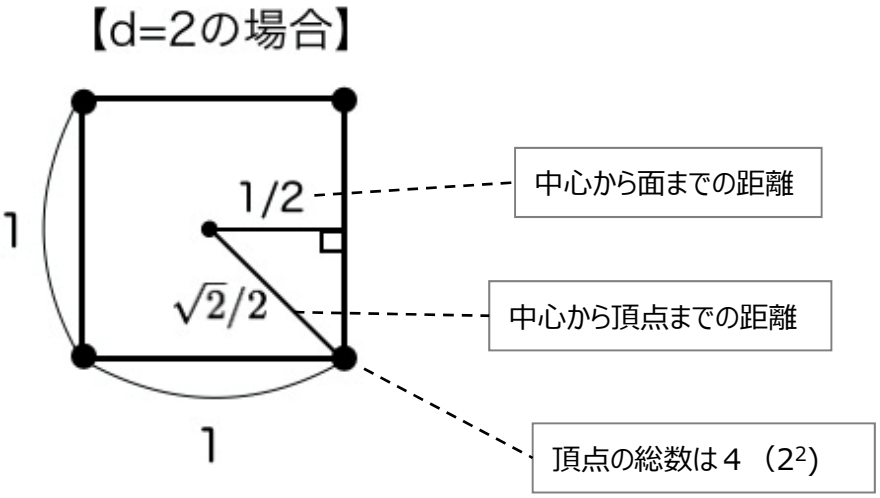


出所：大和証券投資信託委託

2. 空間内の重要度が不平等になる

前ページの図のように2銘柄なら正方形、3銘柄なら立方体のように株価が構成する空間は変化します。一般にd銘柄で構成されるd次元空間は、次の幾何学的性質を持ちます。ここで1辺の長さを1とします。

・中心から頂点までの距離 = √d/2
・中心から面までの距離 = 1/2
・頂点の総数 = 2^d



出所：大和証券投資信託委託

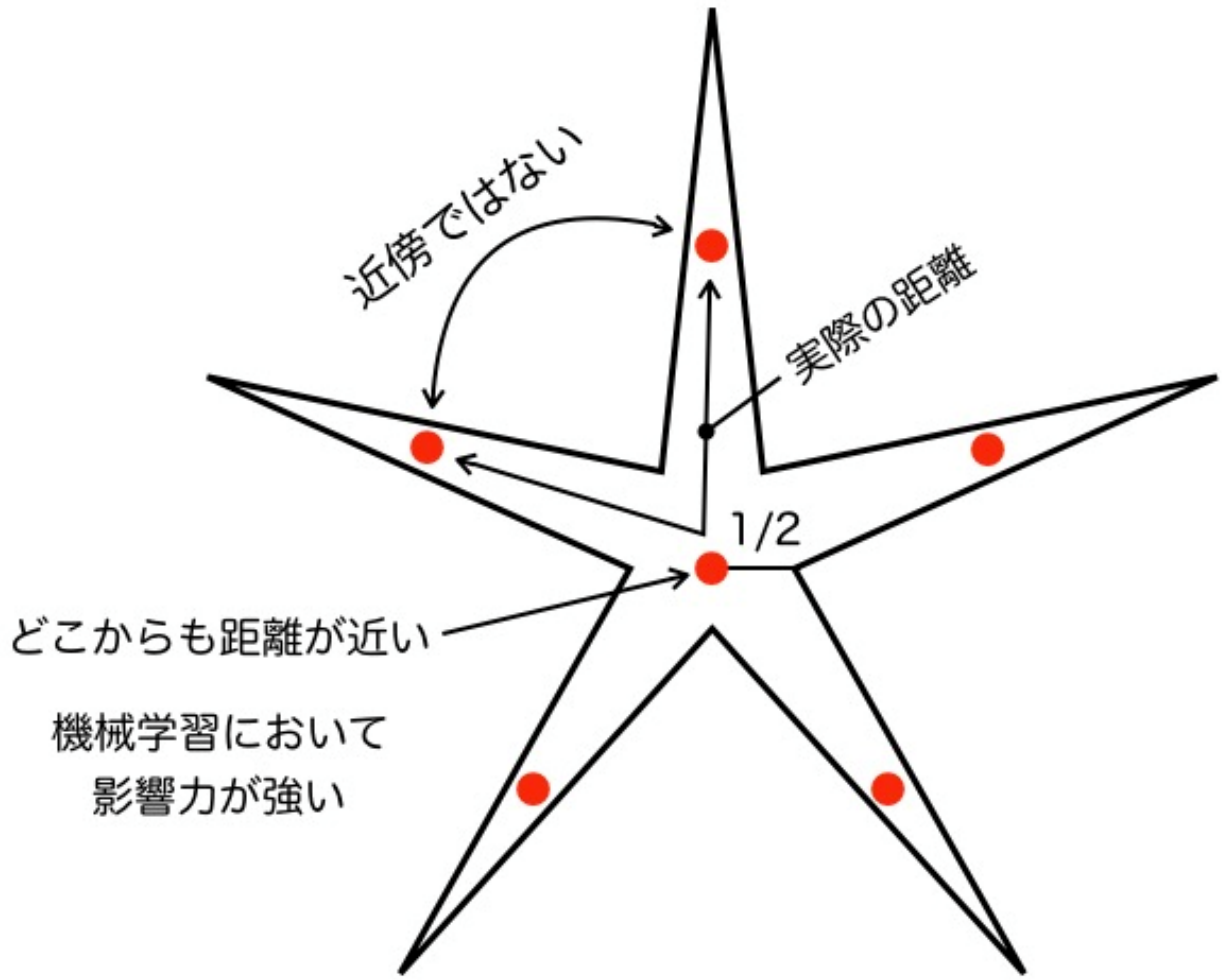
人間の直感として、次元数dを増やすほど、空間は均等に膨張していくと想像されるでしょう。しかし全く異なります。たとえば1000銘柄の場合（d=1000）を計算してみましょう。

・中心から頂点までの距離 = √1000/2 = 15.8

・中心から面までの距離 = 1/2

・頂点の総数 = 2^{1000}

中心から面までの距離に対して、頂点までの距離は31倍強となり、その頂点数は膨大になります。空間はいわば「ウニ」のような形になります。

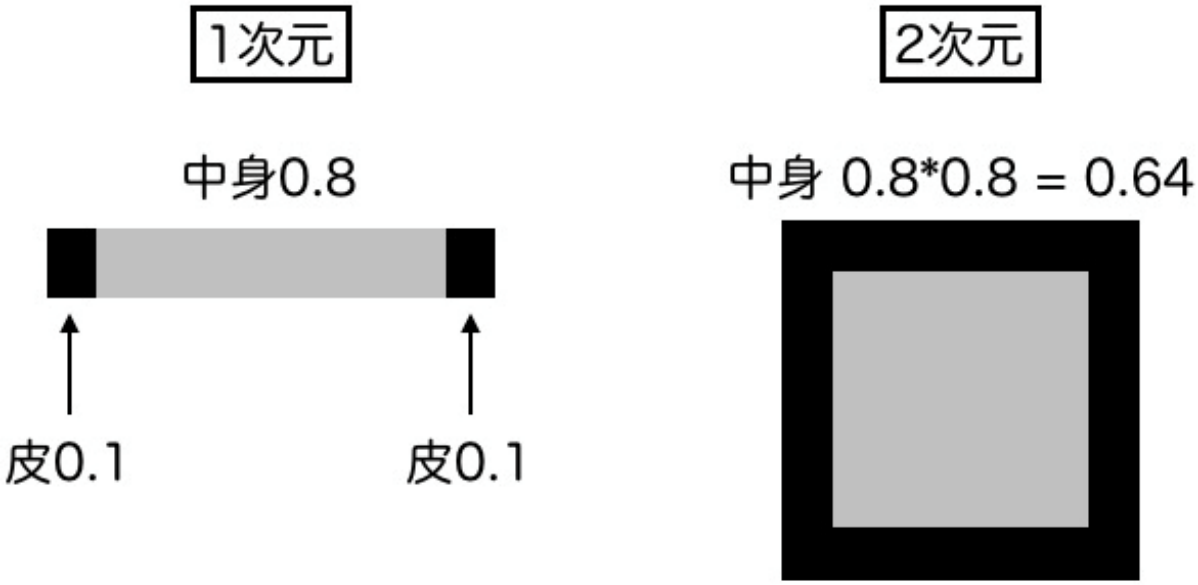


出所：大和証券投資信託委託

空間が不均一に広がる弊害として、空間内のパワーバランスが乱れてしまいます。機械学習では、距離が近い学習データほど状況が類似しているため重要視します。しかしトゲ内のデータは互いに離れてしまうため、いかなる場所からも中心部ほど距離が近くなります。結果として中心部の学習データが機械学習を支配し、常に同じような出力しか返さなくなります。

3. 内側と外側が入れ替わる（球面集中現象）

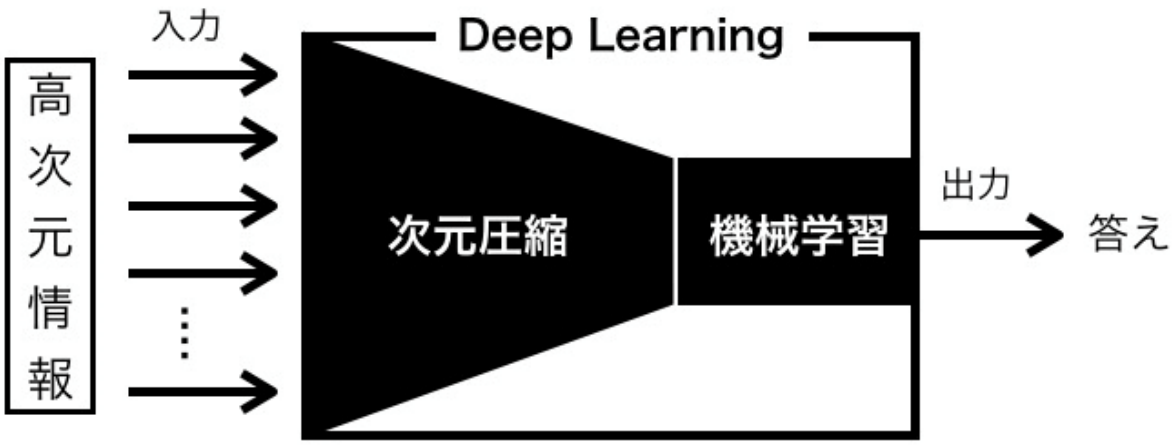
さらに高次元空間には直感に反する面白い性質があります。空間の面に包み紙がついたお菓子をイメージして頂き、全体の8割が食べれる部分 (内側)、その他2割が包み紙 (外側) とします。これを2次元空間で考えると内側は$(8割)^2=6.4$割になり、10次元空間では$(8割)^{10}=1.1$割に減少します。つまりほとんどが外側になり、食べれなくなってしまいます。大福、または、まんじゅうに例えるなら、あんこが減って皮ばかりになります。これを球面集中現象と呼びます。これにより距離が近いと考えていた学習データが全て遠い外側になってしまい、それぞれの重要度が均一化します。その結果、何を頼りに学習すべきか分からなくなってしまいます。



出所：大和証券投資信託委託

＜オートエンコーダによる次元圧縮＞

そこで、高次元情報を機械学習する場合は、あらかじめ次元圧縮しておく必要があります。その方法の一つに、これまで何度かご紹介したオートエンコーダを用います。次元圧縮によりその後の機械学習を単純化できるため、過学習やデータ不足の問題を緩和できます。



出所：大和証券投資信託委託

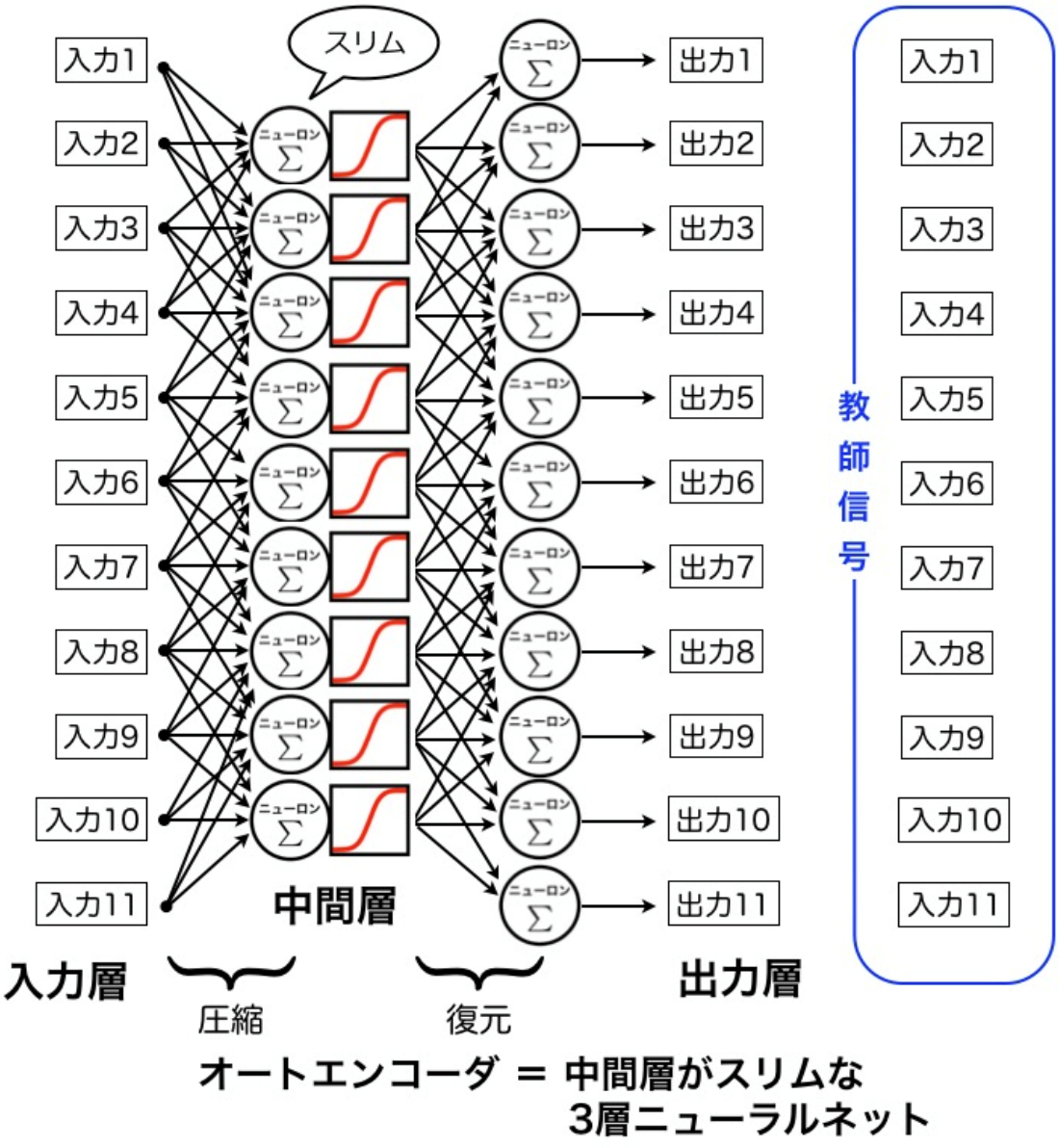
オートエンコーダは、通常の3層ニューラルネットワークと動作原理は同じです。学習アルゴリズムも代表的な逆誤差伝搬法を用います。ただし以下2点において独特な学習ルールがあります。

(1) 入力データと同じデータを出力するように学習する。つまり教師信号は入力データそのものとする。

(2) 入力層と出力層のニューロン数を等しく、中間層のニューロン数をそれ以下とする。

つまりオートエンコーダは、中間層がタイトな3層ニューラルネットワークによって入力データの次元を圧縮し、出力層において再び元の高次元情報に復元する役割を持ちます。

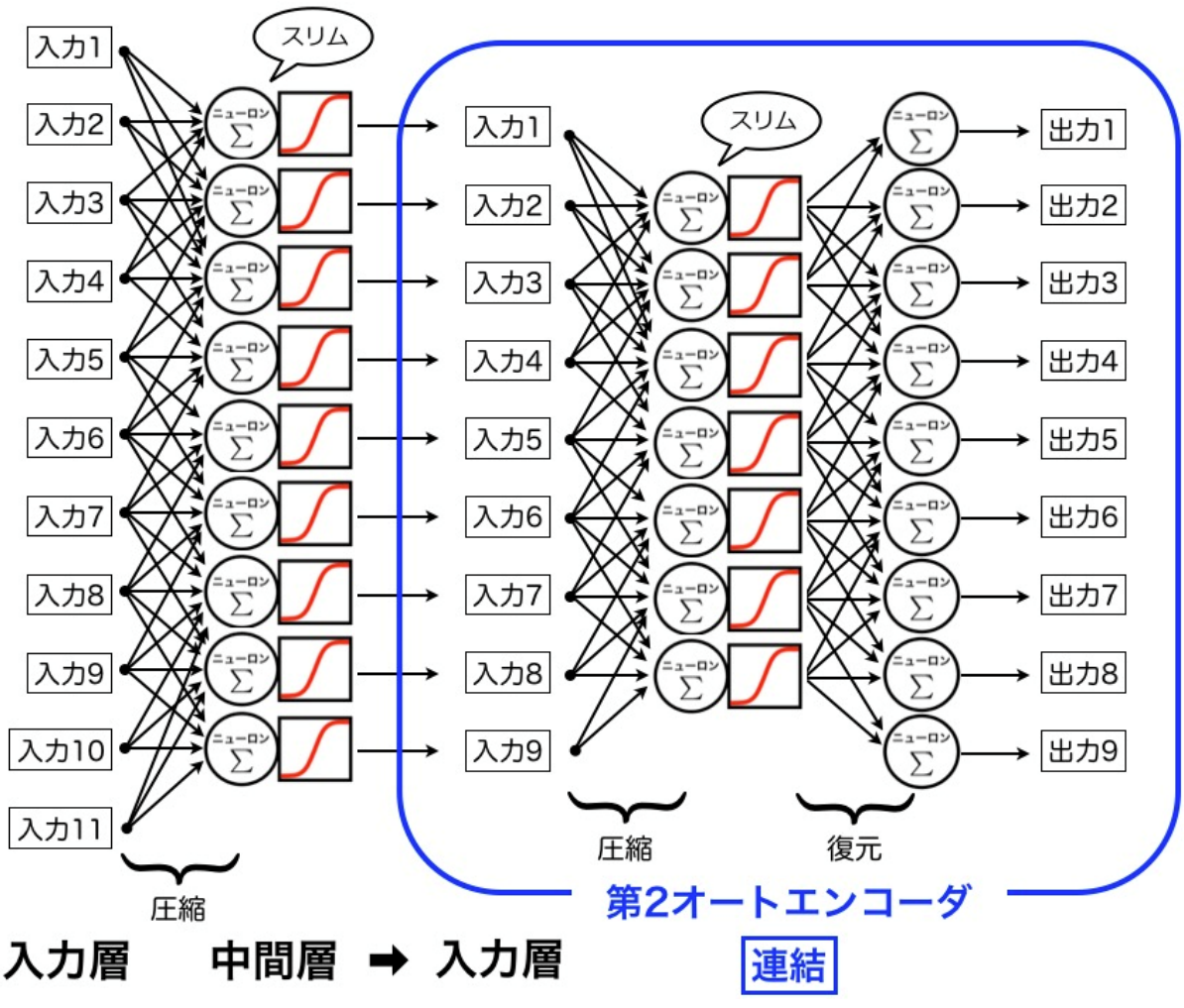
シンプルなオートエンコーダは下の図のような形になります。



出所：大和証券投資信託委託

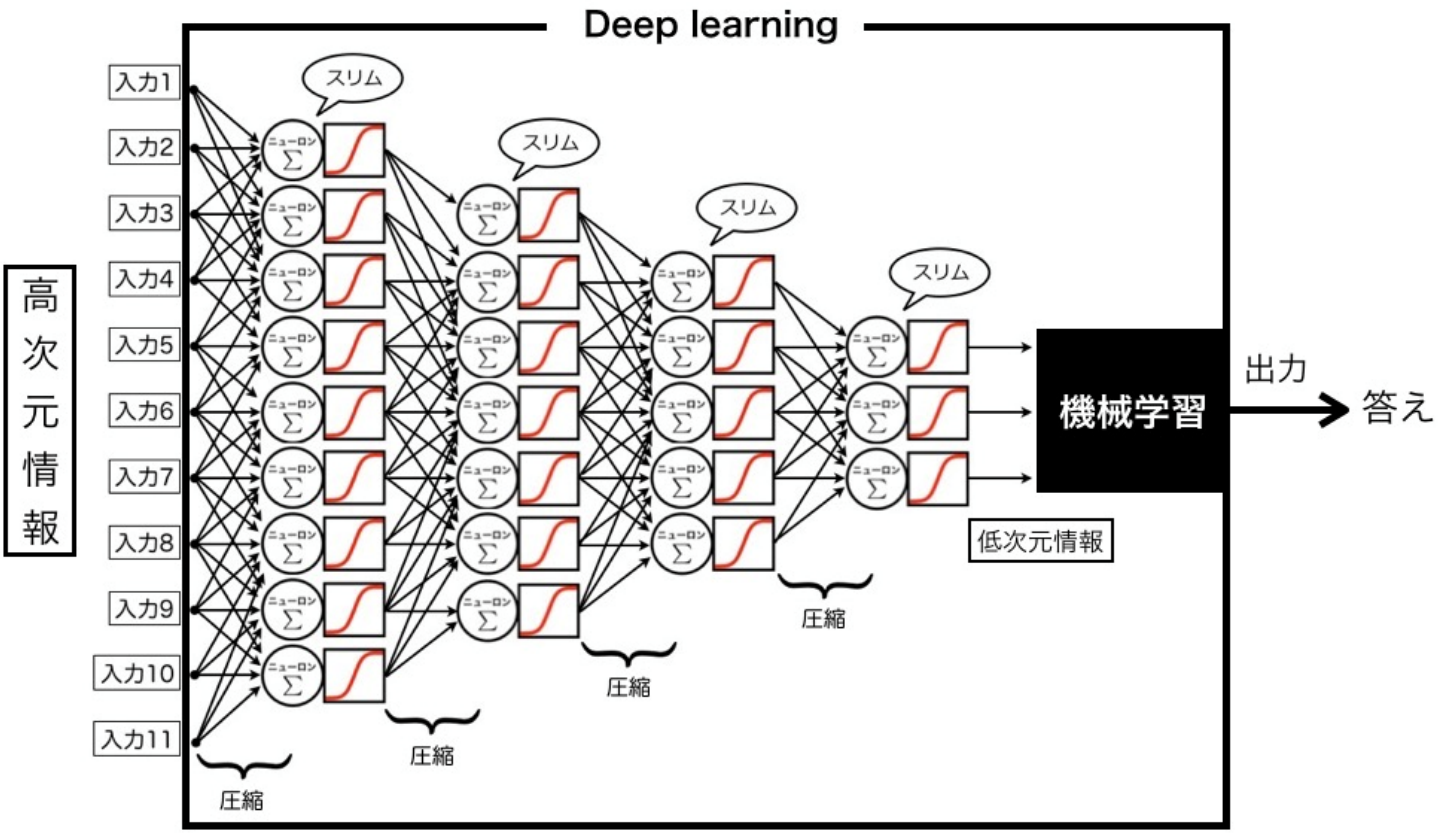
中間層において情報は圧縮されるものの再び復元できるならば、元の入力情報の特徴は失われていません。そこで出力層を除去し、中間層の出力を新しいオートエンコーダに入力します。

下の図では、前ページのオートエンコーダに第２のオートエンコーダを追加しています。



出所：大和証券投資信託委託

図のように同様の圧縮および連結を繰り返すことで、元の入力情報の特徴を保持しつつ、次元圧縮していくことができます。これを事前学習 (Pre-training) といい、次元圧縮後の低次元情報を機械学習することで次元の呪いを緩和することができます。この一連のプロセスを統合して深層学習 (Deep learning) と呼びます（イメージ図は次のページ）。

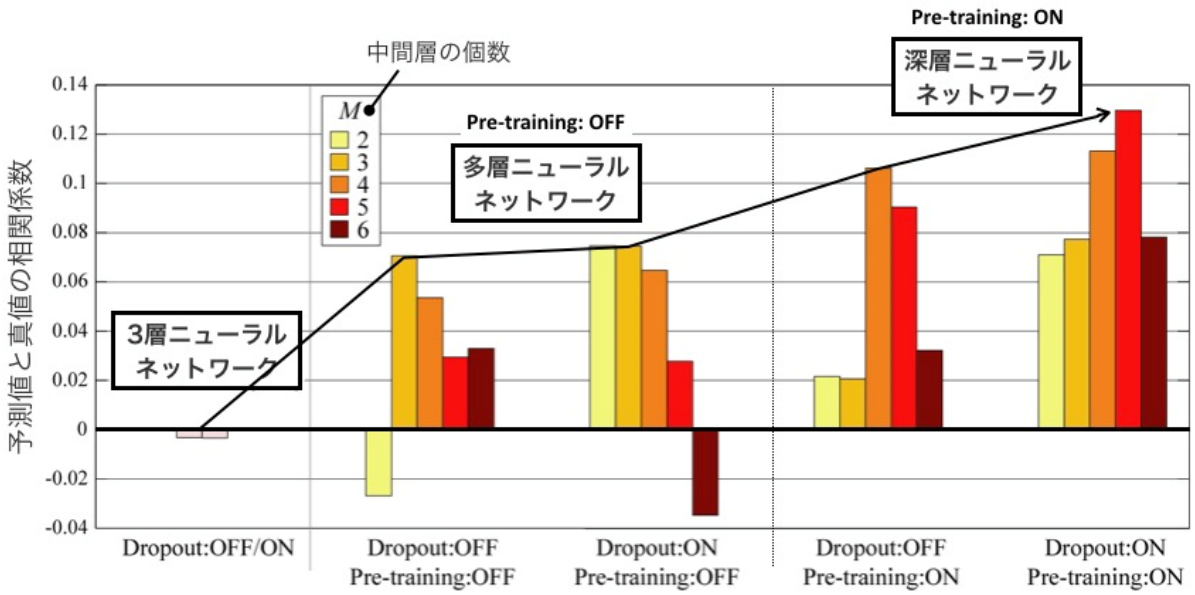


出所：大和証券投資信託委託

<深層学習の事例紹介 (TOPIXの予測)>

最後に、冒頭のTOPIXの日中リターンの予測結果を示します。2008年～2012年を学習期間とし、2013年を予測対象としました。詳細は割愛しますが、過学習を抑制するDropout※や、オートエンコーダによるPre-trainingを使う場合 (ON) と使わない場合 (OFF) を比較しました。なお分析結果は、茨城大学鈴木研究室より国際会議にて発表したものです(※)。

※Dropoutとは、ニューラルネットワークの学習時に、一定割合のニューロンをランダムに休止させることで過学習を抑制するテクニック。



出所：茨城大学

(※) Tomoya Onizawa, Takehiro Suzuki, Tomoya Suzuki: "Predictability of Financial Market Indexes by Deep Neural Network," Proc. of Nonlinear Theory and its Applications (NOLTA), 4 pages, 2017.

TOPIXがランダムウォークのように予測不可能であれば、いかなるテクニックを駆使しても予測結果は改善しないはずです。しかしDropoutやPre-trainingをONにするほど、予測精度は改善する傾向にあります。つまりTOPIXの短期変動にはわずかながら法則性を秘めており、深層学習によってその法則性を自動抽出できる可能性を示唆しています。

【鈴木教授　プロフィール】

鈴木智也（すずきともや）
新潟県新潟市生まれ．IFTA国際検定テクニカルアナリスト(MFTA)．平成17年東京理科大学大学院理学研究科物理学専攻博士課程修了．理学博士．同年東京電機大学工学部電子工学科助手，平成18年より同志社大学工学部情報システムデザイン学科専任講師，平成21年より茨城大学工学部知能システム工学科准教授を経て，平成28年より同大学教授，さらに平成29年より大和証券投資信託委託(株)クウォンツ運用部特任主席研究員を兼務．平成30年より茨城大学大学院理工学研究科機械システム工学専攻長および領域長，CollabWiz株式会社代表取締役．
研究分野は，時系列解析，テキスト解析，機械学習，人工知能，金融工学など実践的なデータサイエンスに従事．電子情報通信学会，情報処理学会，人工知能学会，日本テクニカルアナリスト協会，日本証券アナリスト協会各会員．

【Market Letter　鈴木教授による解説シリーズ　バックナンバー】

第1回 AI運用に挑む	http://www.daiwa-am.co.jp/market/html_ml/ML20171207_1.html
第2回 集団化する人工知能	http://www.daiwa-am.co.jp/market/html_ml/ML20180125_1.html
第3回「２年目のジンクス」を集合知AIで緩和	http://www.daiwa-am.co.jp/market/html_ml/ML20180301_1.html
第4回 時系列データの見えない法則をつかむ	http://www.daiwa-am.co.jp/market/html_ml/ML20180409_1.html
第5回 愚かな人間心理・カモにするAI	http://www.daiwa-am.co.jp/market/html_ml/ML20180501_2.html
第6回 ナイトメア★アノマリーを狙え	http://www.daiwa-am.co.jp/market/html_ml/ML20180601_1.html
第7回 ブルーオーシャンAI戦略	http://www.daiwa-am.co.jp/market/html_ml/ML20180703_1.html
第8回 深層学習による株価予測（前編）	http://www.daiwa-am.co.jp/market/html_ml/ML20180802_2.html