



入門講座

インフォマティクス入門ー4

教師あり学習ー分類問題ー Supervised LearningーClassificationー

小野田 崇 青山学院大学
Takashi Onoda 理工学部 経営システム工学科
教授

1 はじめに

2010年代になると、ジェフリー・ヒントンの研究者が、ニューラルネットワークの階層を4層、5層と増やし、精度の高い機械学習の実現に成功した。階層を4層、5層と増やしたニューラルネットワーク (図1参照) の学習を「ディープ・ラーニング (深層学習)」と呼ぶ¹⁾。この「深層学習」を利用したALPHAGOの出現が、第3次人工知能ブームを引き起こし、「人工知能」と「機械学習」という言葉が一般に広まった。特に、「機械学習」という言葉は初めて一般に知られるようになり、市民権を得ることになった。本稿と次稿では、この機械学習の扱う教師あり学習 (分類問題) と教師なし学習 (外れ値検知) について簡単に解説していく。

2 機械学習の2大タスク

さて、ここでは2つのクイズを考えてもらう。一つ目のクイズは表1で表される「ある、なし」クイズです。

表1には「たぬき」にはあるが「いぬ」にはない、「おちゃ」

にはあるが「コーヒー」にはない、「あかちゃん」にはあるが「こども」にはない、「おしろ」にはあるが「おうち」にはないという事例が与えられている。この事例から次のようなクイズが与えられる。「きんこ」は「ある/なし」どっち、「さいふ」は「ある/なし」どっち、「みずな」は「ある/なし」どっち、「ひまわり」は「ある/なし」どっち。みなさんの答えはいかがでしょう。答えは「きんこ」は「ある」、「さいふ」は「なし」、「みずな」は「ある」、「ひまわり」は「なし」となる。表1の「ある」の事例には「き/ちゃ/あか/しろ」といった色が隠れている。この規則が発見できれば、「きんこ」には「きん」が隠れてお

表1 「ある」、「なし」クイズ²⁾

ある	なし
たぬき	いぬ
おちゃ	コーヒー
あかちゃん	こども
おしろ	おうち

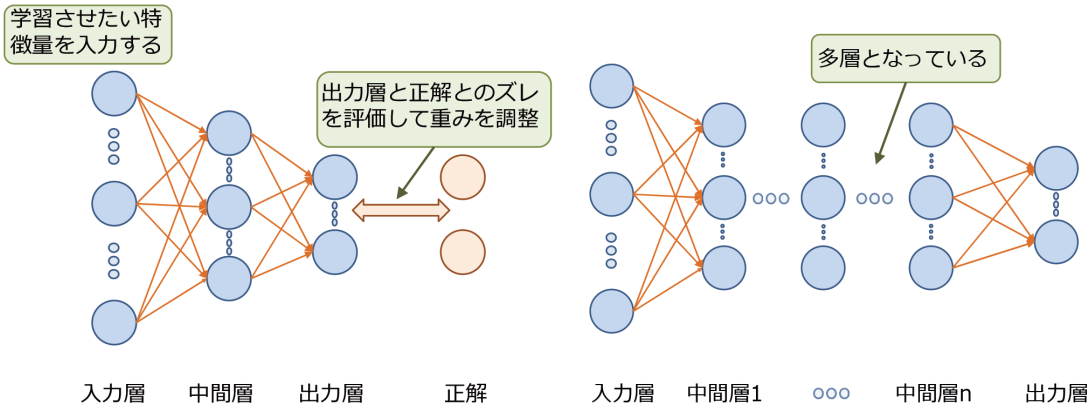


図1 従来型ニューラルネットワーク (左図) 深層学習用ニューラルネットワーク (右図)

り。「みずな」には「みず」が隠れているので「きんこ」、「みずな」は「ある」となる。

2つ目のクイズは表2に示す「仲間外れ」クイズです。表2に示した生物の中で仲間外れ生物は何でしょう。みなさんが仲間外れだと考えますか。回答例を図2、3に示す。図2は、「かたさ」（甲殻がある）と「足の本数」に基づいてグループ分けし、仲間外れの「いか」を発見した回答例です。図3は、「生息している場所」に基づいてグループ分けし、仲間外れの「くも」を発見した回答例です。

1つ目の「ある／なし」クイズの場合、与えられた事例から「ある」、「なし」を分類するルールを発見して、発見したルールを未知の対象に適用して分類する。2つ目の「仲間外れ」探しクイズの場合、ある視点から対象をグループ分けし、それぞれの対象がどのグループに属するのかを評価する。「仲間外れ」探しの場合は、グループに属さない対象を見つけることになる。

「ある／なし」クイズ、「仲間外れ」探しクイズは、それぞれ機械学習の2大タスクである「教師あり学習」と「教師なし学習」に対応する。厳密には、「ある／なし」クイズは教師あり学習の中の分類問題というタスクを表し、「仲間外れ」探しクイズは、教師なし学習の中のクラスタリングというタスクを表す。

表2 「仲間外れ」クイズ：仲間外れは誰？

くも
やどかり
いか
たこ
毛ガニ
花咲ガニ
ざりがに

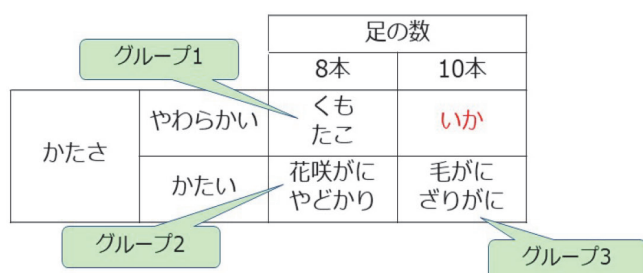


図2 「かたさ」と「足の本数」でグループ分類

3 教師あり学習（分類問題）

教師あり学習とは、試行錯誤するうちに入力に対する出力の関係を、どういう入力のときにどういう出力をすればよいかわかってくることを意味する。この教師あり学習は、入出力の推定問題と考えることができる。入出力の推定問題の目的は、入力 $\mathbf{x}$ が与えられたとき、対応する出力 y を予測することである。例えば、前述の「ある／なし」クイズの場合、入力 $\mathbf{x}$ は「きつね」、「たぬき」、「おもちゃ」などであり、出力 y は「ある」か「なし」である。つまり入出力の推定問題とは、 $y=f(\mathbf{x})$ となる関数を生成する問題となる。

教師あり学習（入出力の推定）を計算機上に実現するためには、この入出力の推定問題を数理的にとらえる必要がある。出力の「ある」、「なし」については、その数理的な（計算機可読な形式の）表現が比較的簡単である。例えば、「ある」を「+1」、「なし」を「-1」に割り当てればよい。前述した「ある／なし」クイズの場合の入力である「コーヒー」、「あかちゃん」、「こども」などは計算機上でどのように扱えばいいのであろうか？つまり、教師あり学習における入力をどのように数理的（計算機可読な形式）に表現すればいいのであろうか？一般に、教師あり学習における入力は、入力対象の特徴量を列挙した D 次元の実数値ベクトル $\mathbf{x}$ で表現される。通常、この実数値ベクトル $\mathbf{x}$ は特徴ベクトルと呼ばれ、特徴ベクトルが構成する領域は特徴空間と呼ばれる。この特徴ベクトルは以下の式で一般に表される。

$$\mathbf{x} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_D \end{pmatrix} \in \mathcal{R}^D \dots\dots\dots (1)$$

上式の $x_1, x_2, \dots, x_D$ には、特徴量の値が入る。例えば、 $\mathbf{x}$ が「りんご」であったとする。この場合、「りんご」を表すために、 x_1 には特徴量として赤みの度合いの値が入り、 x_2 には特徴量としてビタミンCの含有量が入り、 x_D には特徴量として形状の値が入る。

特徴ベクトル $\mathbf{x}$ の設計は完全にドメイン（解きたい問題）



図3 生息場所でのグループ分類

依存であり、一般的な解はない。特徴ベクトル $\mathbf{x}$ は、ユーザの目的に合わせて、ユーザが設計する必要がある。教師あり学習の学習データ（「ある／なし」クイズの場合は表1に示した「ある」の事例と「なし」の事例）は、 N 個の入力ベクトルと出力値とのペアで以下の式のように与えられる。

$$\left\{ \left(\mathbf{x}^{(1)}, y^{(1)} \right), \left(\mathbf{x}^{(2)}, y^{(2)} \right), \dots, \left(\mathbf{x}^{(N)}, y^{(N)} \right) \right\} \dots\dots\dots (2)$$

ここで、 $\mathbf{x}^{(i)}$ は i 番目の事例の入力ベクトルを表し、 $y^{(i)}$ は i 番目の事例の入力ベクトルに対する正しい出力値（例えば、答えが「りんご」なら+1、「りんご」でないなら-1）を表す。

教師あり学習では、上述した与えられた入力ベクトルに対する、あるべき出力値を訓練データとして、入出力の関係を学習する。この学習には、ニューラルネットワーク³⁾、RadialBasis Function ネットワーク⁴⁾、畳み込みニューラルネットワーク⁵⁾、深層学習¹⁾、サポートベクトルマシン⁶⁾、Random Forest⁷⁾ など様々な機械学習手法が用いられる。機械学習手法にはそれぞれ長所、短所があるので、ユーザの目的に合った機械学習手法を選択することが重要となる。

4 分類問題の事例

ここでは、教師あり学習の中の分類問題の事例を紹介する。

4.1 信用リスク評価

信用リスク評価は、ある顧客に融資を行ってよいか否かを判断する問題である。信用リスク評価における顧客を様々な特徴量で表現する顧客ベクトル $\mathbf{x}$ を図4に示す。顧客ベクトル $\mathbf{x}$ によって信用リスクを評価できるように、そのベクトル $\mathbf{x}$ は顧客の信用に関係する「過去に延滞したことがあるか否か」(x_1)、「リボ払い使用率」(x_2)、「カードの使用限度額」(x_D)などで構成される。この顧客ベクトル $\mathbf{x}$ が入力ベクトルとなる。

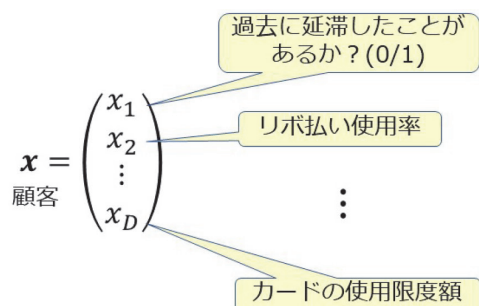


図4 信用リスク評価における顧客ベクトル

この入力ベクトルに対する正しい出力値 y は、融資してよい（返済してくれる）場合には $y=+1$ となり、融資してはいけない（貸し倒れる）場合には $y=-1$ となる。

信用リスク評価問題では、過去の顧客ベクトル $\mathbf{x}$ とその顧客ベクトル $\mathbf{x}$ に対する出力値（返済してくれた $y=+1$ 、あるいは、貸し倒れた $y=-1$ ）のペアの事例を訓練データ（教師）として N ($\gg D$) 個用意する。そして、与えられた顧客ベクトル $\mathbf{x}$ に対する、あるべき出力値 $y=+1$ or -1 を訓練データとして、顧客ベクトル $\mathbf{x}$ に対する融資する／融資しないの関係を学習する。この学習には、ニューラルネットワーク、サポートベクトルマシン、Random Forestなど様々な機械学習手法が用いられる。ある機械学習手法によって、訓練データを正しく学習できた計算機を分類器と呼ぶ。この分類器に顧客ベクトルを入力すると、融資してよいか否かの出力値が得られる。

4.2 テキスト分類

テキスト分類は、ある自然言語で表現された文書がどのカテゴリーに入るかを判断する問題である。ここではネットニュースの記事（文書）が「スポーツ」、「政治」、「経済」、「エンターテインメント」などのどのカテゴリーに入るかを判断する問題を考える。ネットニュースのカテゴリー分類における記事（文書）を単語の出現回数で表現する文書ベクトル $\mathbf{x}$ を図5に示す。計算機で対象を扱う場合、特徴量は数値で表す必要がある。自然言語を扱う場合も、文書を何らかの方法で数値化しなければならない。そのための代表的な手法の一つが、Bag of Wordsである。Bag of Wordsとは文書中に出現する単語を数え、その数を特徴量とする手法である。文書における出現順序は考慮しないシンプルな表現となっている。図5に示した文書ベクトル $\mathbf{x}$ はBag of Wordsで表現されている。 x_1 は「野球」という単語がその文書に含まれている回数を表し、 x_2 は「AKB」という単語がその文書に含まれている回数を表し、 x_D は「グループ」という単語がその文書に含まれている回数を表す。この文書ベクトル $\mathbf{x}$ が入力ベクトルと

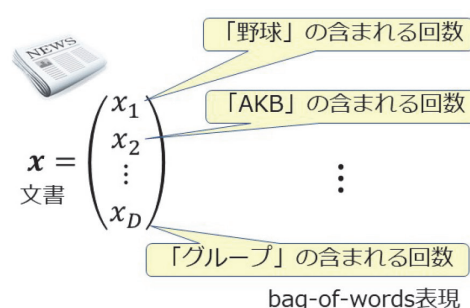


図5 テキスト分類における文書ベクトル

なる。この入力ベクトルに対する正しい出力値 y は、ニュースのカテゴリー「スポーツ」、「政治」、「経済」、「エンターテインメント」などとなる。

テキスト分類問題では、過去の文書ベクトル $\mathbf{x}$ とその文書ベクトル $\mathbf{x}$ に対する出力値（その文書のカテゴリーが「スポーツ」、「政治」、「経済」、「エンターテインメント」などのどのカテゴリーに入るかの情報）のペアの事例を訓練データ（教師）として $N (\gg D)$ 個用意する。そして、与えられた文書ベクトル $\mathbf{x}$ に対する、あるべき出力値（その文書の正しいニュースのカテゴリー）を訓練データとして、文書ベクトル $\mathbf{x}$ に対するニュースカテゴリーの関係を学習する。この学習には、ニューラルネットワーク、サポートベクトルマシン、Random Forest など様々な機械学習手法が用いられる。ある機械学習手法によって、訓練データを正しく学習できた計算機を分類器と呼ぶ。この分類器に文書ベクトルを入力すると、どのニュースカテゴリーに分類されるのかの出力値が得られる。

4.3 手書き文字認識

手書き文字認識は、手書き文字画像がある文字である ($y=+1$) か否 ($y=-1$) かを識別する問題、または、手書き文字画像が「0」、「1」、「2」、 $\dots$ 、「9」のどの数字なのかを識別する問題である。ここでは、手書き文字認識として、書かれた文字が「崇」である ($y=+1$) か「崇」である ($y=-1$) かを識別する問題を考える。手書き文字認識に用いられる手書き文字ベクトルの一例を図6に示す。計算機で対象を扱う場合、特徴量は数値で表す必要がある。手書き文字を扱う場合も、手書き文字を何らかの方法で数値化しなければならない。図6は手書き文字をベクトル化する一つの例を示している。図6に示した「崇」は縦16画素、横16画素の合計256画素で表現された画像である。図6の場合、手書き文字ベクトルは256次元 ($D=256$) のベクトルで表現される。 x_1 には1番目の画素に含まれる黒の割合が入り、 x_2 には2番目の画素に含まれる黒の割合が入り、 x_D には256番目の画素に含まれる黒の割合が入ることになる。この手書き文字ベクトル $\mathbf{x}$ が入力ベク

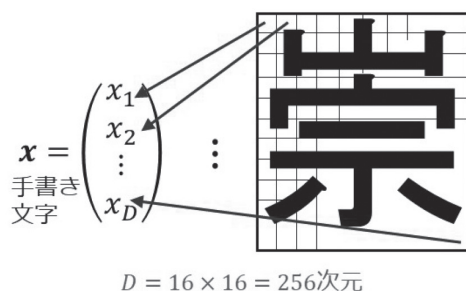


図6 手書き文字認識における手書き文字ベクトル

トルとなる。この入力ベクトルに対する正しい出力値 y は、入力ベクトルが「崇」である場合には $y=+1$ となり、入力ベクトルが「崇」である場合には $y=-1$ となる。

手書き文字認識では、過去の手書き文字ベクトル $\mathbf{x}$ とその手書き文字ベクトル $\mathbf{x}$ に対する出力値（「崇」の場合 $y=+1$ 、あるいは、「崇」の場合 $y=-1$ ）のペアの事例を訓練データ（教師）として $N (\gg D)$ 個用意する。そして、与えられた手書き文字ベクトル $\mathbf{x}$ に対する、あるべき出力値 $y=+1$ or -1 を訓練データとして、手書き文字ベクトル $\mathbf{x}$ に対する「崇」/「崇」の関係を学習する。この学習には、ニューラルネットワーク、サポートベクトルマシン、Random Forest など様々な機械学習手法が用いられる。ある機械学習手法によって、訓練データを正しく学習できた計算機を分類器と呼ぶ。この分類器に手書き文字ベクトルを入力すると、「崇」であるか「崇」であるかの出力値が得られる。

4.4 油中ガス分析データに基づく電力用変圧器の内部不具合診断

ここでは、電気協同研究会第65巻 第1号「電力用変圧器改修ガイドライン」⁸⁾に記載されている油中ガス分析データに基づく電力用変圧器の内部不具合の判定基準について紹介する。

この油中ガス分析データに基づく電力用変圧器の内部不具合診断は、油中のガスの発生量から電力用変圧器の内部に不具合が発生している ($y=+1$) か否 ($y=-1$) かを識別する問題である。この問題の場合、油中ガス分析データベクトル $\mathbf{x}$ は表3に示す7種類のガスの発生量で構成される。つまり、油中ガス分析データベクトル $\mathbf{x}$ は7つの特徴量 ($x_1, x_2, \dots, x_7$) から構成され、 x_1 は水素の発生量、 x_2 はメタンの発生量、 x_3 はエタンの発生量、 x_4 はエチレンの発生量、 x_5 はアセチレンの発生量、 x_6 は一酸化炭素の発生量、 x_7 は二酸化炭素の発生量となる。この油中ガス分析データベクトル $\mathbf{x}$ が入力ベクトルとなる。この入力ベクトルに対する正しい出力値 y は、電力用変圧器に内部不具合がある場合には $y=+1$ となり、内部

表3 7種類の油中ガス

No.	ガスの種類
1	H ₂ (水素)
2	CH ₄ (メタン)
3	C ₂ H ₆ (エタン)
4	C ₂ H ₄ (エチレン)
5	C ₂ H ₂ (アセチレン)
6	CO (一酸化炭素)
7	CO ₂ (二酸化炭素)

不具合がない場合には $y = -1$ となる。

油中ガス分析データに基づく電力用変圧器の内部不具合診断では、過去の油中ガス分析データベクトル $\mathbf{x}$ とその油中ガス分析データベクトル $\mathbf{x}$ に対する出力値（電力用変圧器に内部不具合がある場合には $y = +1$ 、内部不具合がない場合には $y = -1$ ）のペアの事例を訓練データ（教師）として $N (\gg D)$ 個用意する。そして、与えられた油中ガス分析データベクトル $\mathbf{x}$ に対する、あるべき出力値 $y = +1$ or -1 を訓練データとして、油中ガス分析データベクトル $\mathbf{x}$ に対する内部不具合「あり」/内部不具合「なし」の関係を学習する。

電気協同研究第65巻第1号に記載されている油中ガス分析データに基づく電力用変圧器の内部不具合の判定基準は、同報告書 (1) の付録に記載されているデータ、内部不具合の変圧器102件と正常変圧器1030件、合計1132件を使用し（ただし、改造時に不具合を発見した事例等、付録で様相診断検討の対象外と記された内部不具合変圧器15件、重複データ、欠損値のあるデータの正常変圧器3件を除外）、線形サポートベクトルマシンによって生成されたものである。

参考文献

- 1) Y.LeCun, Y.Bengio and G.E.Hinton : Nature, 521 (2015), 436.
- 2) <https://asoppa.com/hiramekiasoppa/2838326>
- 3) 甘利俊一：神経回路網の数理—脳の情報処理様式（システム・サイエンス・シリーズ），産業図書，(1978)
- 4) <https://www.cc.gatech.edu/~isbell/tutorials/rbf-intro.pdf>
- 5) A.Krizhevsky, I.Sutskever and G.E.Hinton : Proc. of the 25th International Conference on Neural Information Processing Systems, Vol.1, (2012), 1097.
- 6) 小野田崇：サポートベクターマシン（知の科学シリーズ），オーム社，(2007)
- 7) L.Breiman : Machine Learning, 45 (2001), 5.
- 8) 電気協同研究会第65巻 第1号，電力用変圧器改修ガイドライン，(2009)

(2020年8月31日受付)

5 おわりに

本稿では教師あり学習、厳密には教師あり学習の分類問題について、事例を中心に紹介した。読者が本稿を読んで、教師あり学習を利用してみたいと考えてもらえれば幸いである。