ホワイトペーパー

Vertica

機械学習によるこれまでにない ビジネスのスピードとスケールの実現

Verticaのインデータベース機械学習は、コアに組み込まれているため、別パッケージをダウンロードしてインストールする必要はありません。この機能では、予測分析のプロセス全体が、超並列処理と使い慣れたSQLインターフェイスによってサポートされるので、データサイエンティストやアナリストは、ビッグデータのパワーを存分に活用し、ビジネス成果を最大化することができます。

VERTICA | MICRO FOCUS

目次

ページ

Verticaのインデータベース機械学習はビッグデータのパワーを活かして、ビジネスの成果を制約や妥協なしに最大化することを支援します。

機械学習: 競争上の優位性

データが重要な役割を果たす今日の世界では、大量のデータから意味のある情報を得られるかどうかが、競争上の優位性を左右します。高度な分析や機械学習を利用している企業は、業績が上位4分の1に入る可能性が2倍高く、効果的な意思決定を行える可能性が3倍高くなります[1] 。一方で、ビジネスリーダーの75%が、分析による価値の最大の源泉として「成長」を挙げているにもかかわらず、それらのリーダーのうち予測分析機能を導入しているのは60%に過ぎません[2]。

ビッグデータをフルに活用することで、顧客の挙動を詳細に理解して、顧客ごとに適したユーザーエクスペリエンスの提供、製品乗り換えの予防、異状検知、収益の拡大などが可能になります。しかしながら、データの速度、分量、多様性の増加によって、予測モデルの作成作業はしだいに複雑になっています。このような大量のデータセットを、ビジネスに合った速度で処理できるツールはほとんどないからです。これに対して、Verticaのインデータベース機械学習を使えば、ビッグデータのパワーを活かして、ビジネスの成果を制約や妥協なしに拡大できます。

機械学習とは何か、なぜ重要なのか

機械学習は、パターンや関係の識別だけでなく、成果の予測に関しても、基本的な手段として普及しつつあります。これによって、ビジネスの運営方法に、事後対応から事前対応へという根本的な変革が起きつつあります。機械学習を利用することで、これまで複雑すぎて手作業では不可能だったプロセスや分析が可能になります。機械学習を使用することで、データサイエンティストやアナリストは、さまざまなアルゴリズムや統計モデルによってデータを処理し、相関付けて、見かけ上ランダムな、あるいは無関係なデータセットから価値のある情報を引き出すことができます。得られる情報としては、顧客、バリューチェーンプロセス、製造工程に関するものなどがあり、ビジネスの意思決定を支援して、競争上の地位を高めるために利用できます。簡単に言えば、機械学習アルゴリズムは、情報に基づく組織の意思決定を容易にする効果があり、通常のビジネス推進、製品開発、オペレーションにとって、不可欠の情報源となります。

ただし、ビッグデータの価値をフルに引き出すには、このような機械学習アルゴリズムのモデル化、トレーニング、展開のスケールと速度を変えることが必要です。従来の分析ツールでは、今日の世界のスケールにはもはや対応できません。現在のデータ量は、従来の多くの機械学習ツールが処理できる限度を超えており、バックエンドの顧客データベースからクリックストリームの変動まで、大規模で性質の異なるデータソースの間の関係は、あまりにも複雑過ぎて、従来のツールでは隠された価値を十分に引き出すことが困難になっています。

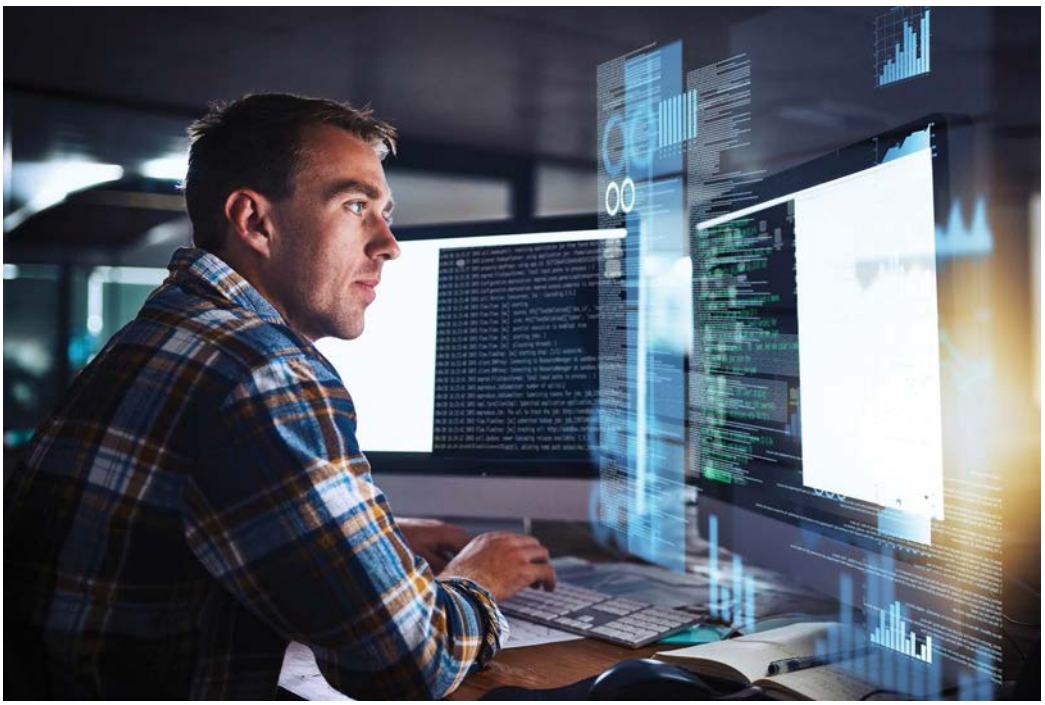
1『Creating Value through Advanced Analytics』ベイン・アンド・カンパニー、2015年2月

2 **www.marutitech.com/machine-learning-predictive-analytics/** (上記のコメントを参照)。

新しいインデータベース機械学習法を使用することで、データサイエンティストは従来のツールの量的制約から解放され、これまでよりはるかに大規模なデータセットに隠されたパターンを発見して表示することができます。そして、取り込まれたデータの量が増えるに従って、学習のレベルと洗練度も高まり、さらに正確な予測を行うことで、顧客サービスや製品の改善、競争上の優位性につなげることができます。

2016年に、GoogleのAlphaGoが、囲碁プログラムとして史上初めてプロ棋士に勝ちました。囲碁は中国発祥の歴史の古いゲームで、盤上の配置の数は全宇宙の原子より多いと言われています[3]。 この勝利は、どのようにして可能になったのでしょうか。このプログラムは、機械学習アルゴリズムを利用して、オンライン囲碁対局データベースを研究したのです。これは、人間が80年間休みなしに囲碁の対局を行った場合に得られる経験と同等です。

同じことが、Googleの自動運転車についても言えます。機械学習の世界的権威であるPedro Domingos氏は、このように述べています。「自動運転車は、それ自体が運転するようにプログラムされているわけではありません。実際には、車が運転できるようにプログラムする方法は誰も知らないのです。人間は車を運転することができますが、その方法を説明することはできません。Googleの車は、何百万マイルもの運転の間に、人間の運転を観察しながら学んだのです[4]。」



「データベースのない銀行がデータベースを持つ銀行に対抗できないように、機械学習を利用しない企業は、利用している企業についていけなくなります[4]。」

3『Google just made artificial-intelligence history』、Business Insider、2016年3月
4『The Master Algorithm: How the Quest for the Ultimate Learning Machine Will Remake Our World』、Pedro Domingos著、2015年9月
www. amazon.com/The-Master- Algorithm-Ultimate-Learning/dp/0465065708

データの量、速度、多様性が増え続けるのに伴い、ビッグデータに隠された情報を見つけ出すことで、競争上の差別化を図り、利益を増加させることが不可欠になっています。

同じ機械学習アルゴリズムを使用することで、Netflixはユーザーの鑑賞履歴に基づいてお勧めの動画を表示しています。Amazonは、顧客の購入傾向に基づいておすすめの製品を表示しています。Facebookは、アップロードした写真から友達の顔を認識してタグ付けしています。このような例は、ほかにもたくさんあります。その結果として、どのような利点があるのでしょうか。個人に合わせたユーザーエクスペリエンス、ロイヤリティの向上、乗り換えの減少、ウォレットシェア（お財布シェア）の拡大などが挙げられるでしょう。これらはすべて、機械学習の力を利用した結果なのです。

当ホワイトペーパーの内容

データの量、速度、多様性が増え続けるのに伴い、ビッグデータに隠された情報を見つけ出すことで、競争上の差別化を図り、利益を増加させることが不可欠になっています。このホワイトペーパーでは、インデータベース機械学習を使用することで、小規模な分析を超大規模なデータセットに適用する際の制約を取り除き、より正確な情報を超高速で得られることを示します。また、Verticaのインデータベース機械学習アルゴリズムの独自の利点と、それによって従来のツール（C++、Java、Python、R）を補完する方法も示します。

機械学習のタイプ

機械学習アルゴリズムは、大きく2つのカテゴリに分けられます。それは、教師あり学習と教師なし学習です。教師あり学習と教師なし学習の2つは、最も広く用いられている機械学習の方法であり、Verticaにはその両方のアルゴリズムが組み込まれています。

- **教師あり学習**が用いられるのは、すべてのデータがラベル付けされており、アルゴリズムが入力データから出力を予測することを学習する場合です。学習アルゴリズムは、トレーニング入力のセットと、それに対応する正しい出力のセットを取り込みます。その後、アルゴリズムは、実際の出力と正しい出力を比較し、誤りを見つけ、それに基づいてモデルを変更することで、学習を行います。このプロセスは、トレーニングデータに基づいて、必要なレベルの精度をモデルが実現できるまで反復されます。

  教師あり学習では、過去のデータを用いて未来の事象を予測することが多く行われます。たとえば、過去の購入行動に基づいてプロモーションに最適な対象顧客を見つけたり、過去の財務行動に基づいて信用度を予測したりすることです。一般的なアルゴリズムとしては、決定木、ナイーブベイズ分類器、ランダムフォレスト、線形回帰、ロジスティック回帰などがあります。

- **教師なし学習**が用いられるのは、データがすべてラベル付けされておらず、アルゴリズムが入力データから内在的構造を学習する場合です。教師あり学習と異なり、正解も「教師」も存在しません。

  アルゴリズムは、自分でデータを解析してパターンを識別します。たとえば、マーケティングキャンペーンの対象とする、類似した特性を持つ顧客集団の発見などです。教師なし学習で用いられる一般的な分類技法としては、相関ルール学習やクラスタリング技法（階層的クラスタリング、K-meansなど）があります。

機械学習の大規模適用への課題

組織が収集して保存している大量のデータに機械学習を実際に適用しようとすると、いくつかの課題が生じます。予測分析は、特にビッグデータが関わると、非常に複雑になることがあります。データセットが大きい方が正確な結果が得られるので、ビジネスの進捗と同じ速度で情報を得るためには、ハイパフォーマンスの分散型並列処理が必要です。さらに、最新の分散型並列エンジンを利用できるように、機械学習アルゴリズムを書き直す必要もあります。

従来の機械学習アプリケーションやツールを使用する場合は、データの小さいサブセットだけを使用して（ダウンサンプリング）モデルの構築とチューニングを行う必要があるため、不正確な結果、遅れ、コストの増加が生じ、重要な情報を得るために時間がかかることになります。

- **開発時間の増加:** 大量のデータをシステム間で移動する際の遅れにより、予測分析モデルの作成に時間がかかり、価値を引き出すまでの時間が長引きます。
- **不正確な予測:** 大規模なデータセットは、従来の方法ではメモリと処理能力の制限のために処理できないため、データのサブセットだけが分析されます。これにより、得られる情報の精度が下がり、それに基づく意思決定のリスクが増えることになります。
- **展開の遅れ:** 予測モデルの実環境への展開は複雑なプロセスなので、時間と手間がかかり、ビッグデータの活用への障害になります。
- **コストの増加:** データを移動し、予測モデルの複製を構築し、複数のプラットフォームでそれらを実行して必要な結果を得るために、追加のハードウェア、ソフトウェアツール、管理と開発のリソースが必要となります。

Verticaは、大量のデータ処理を前提として一から設計されており、バランスのとれた分散型のカラム型方式の採用により、ビッグデータ分析の課題に有効に対処できます。

「Verticaの新しいインデータベース機械学習機能は本当に便利です。当社の機械学習モデルをVerticaで当社のデータを使ってトレーニングして、プラットフォームと一緒に出荷してお客様のクラスターで動作させることができます。このようなことを他のツールでやろうとすれば、もっとリソースが必要になります。」

– Fidelis Cybersecurityデータサイエンティスト、Abhishek Sharma氏

Vertica Analytics Platformの高速・大規模な予測分析

Vertica Analytics Platformは、大量の多様なデータを保存することができ、主要な機械学習アルゴリズムを組み込んでいるため、上記の障害の多くを取り除くか最小化することができます。Verticaは、大量のデータ処理を前提として一から設計されており、バランスのとれた分散型のカラム型方式の採用により、ビッグデータ分析の課題に有効に対処できます。

超並列処理により、ペタバイト単位のデータ処理が可能で、最も厳しい条件のユースケースにも対応できます。カラムストア機能によりデータの圧縮が可能なので、ビッグデータ分析のクエリ時間を、従来のテクノロジーに比べて、時間単位から分単位、あるいは分単位から秒単位に短縮できます。さらに、Verticaはフル機能の分析システムとして、パターンマッチング、地理空間分析、モンテカルロシミュレーションなどの高度なSQLベースの分析機能を備えています。

データを複製して別のプラットフォームで処理する場合、通常は複数のベンダーの製品が必要なので、複雑さとコストが増加しますが、Verticaは、インデータベースで大規模なデータセットに対して高度な予測モデリングを実行するために最適化されたプラットフォームであるため、その必要がありません。SQLベースの分析と同じ速度、スケール、パフォーマンスが、機械学習アルゴリズムに対して利用可能になります。しかも、その両方が1つのシステムで動作するため、さらに簡素化とコストの削減を実現できます。

Vertica Analytics Platformによる機械学習の実装

機械学習は非常に大規模なデータセットに対して使用したときに最も効果を発揮するので、ビッグデータの高速処理向けに設計されたVerticaにはまさに最適です。機械学習機能をVerticaで展開するには、主に2つの方法があります。Verticaのインデータベース機械学習と、ユーザー定義拡張（UDx）です。

インデータベース機械学習

Verticaのインデータベース応用機械学習アルゴリズムでは、最も広く用いられているいくつかの機械学習モデルをネイティブに作成して展開できるので、高い精度で意思決定を高速化できます。インデータベース機械学習アルゴリズムは、Verticaのコアに組み込まれているため、別パッケージをダウンロードしてインストールする必要はありません。このアルゴリズムには、次の特長があります。

- **スケーラビリティ：** RやPythonといった外部ツールのほとんどは、処理できるデータセットのサイズに制限があるため、分析のためにダウンサンプリングが必要となり、大量のデータを解析する利点が薄れてしまいます。これに対して、Verticaのインデータベース機械学習は、大規模なデータセットをサポートしているため、より正確な情報をより多く得ることができます。

- **シンプルさ:** Verticaのネイティブ取り込み、データ準備、モデル管理機能は、データマイニングのライフサイクル全体をカバーするので、データをエクスポートして解析のために別のツールにロードし、結果をまたエクスポートしてVerticaに返す必要がありません。さらに、機械学習モデルのトレーニング、テスト、展開には、使い慣れたSQL形式のインターフェイスが使用できるので、新しい技術を学んだり、特殊なスキルを持つ専門家を新たに雇用する必要はありません。
- **速度:** Verticaのインデータベース機械学習では、Verticaの超並列処理 (MPP) アーキテクチャーを利用して、情報取得までの時間を短縮できます。必要な場合は、クラスター内の複数のノードを使用することで、計算を高速化できます。

Verticaでは、大規模なデータセットに対するインデータベース予測分析を実行するためのいくつかの機械学習機能が備わっており、予測の精度を高めて、隠された情報へのアクセス速度を加速するために役立ちます。

ユーザー定義拡張 (UDx)

Verticaは、何百ものアプリケーション、データソース、ETL、ビジュアリゼーションモジュールと接続できます。また、標準で接続できないものでも、UDxを使用して容易に統合できます。UDxは、Vertica Analytics Platform用のユーザー独自の分析ツールやデータロードツールを開発するための仕組みであり、新しいタイプのデータ解析や、新しいデータタイプの解析とロードなどのために使用できます。UDxの開発には、C++、Java、Python、Rプログラミング言語とVertica SDKを組み合わせて使用します。このため、標準のSQLでは困難あるいは実行速度が不十分な分析処理の実現に適しています。

さまざまな種類のユーザー定義拡張機能 (関数、変換、集計、分析、ロードなど) がVerticaのMPP能力を利用して実現されており、分析対象のデータ (構造化、半構造化、または非構造化) に合わせた手続き型プログラミング言語を利用することでその能力と柔軟性を高めています。Verticaのユーザーインターフェイスを使えば、プログラミングの展開が容易になり、運用手順がシンプルになるとともに、コードの再利用が促進されます。ただし、UDxはVerticaのデータ分析機能を拡張することはできますが、速度やスケールに関してはVerticaのインデータベース機械学習機能の速度やスケールに関して該当しません。

インデータベース機械学習機能

Verticaでは、大規模なデータセットに対するインデータベース予測分析を実行するためのいくつかの機械学習機能が備わっており、予測の精度を高めて、隠された情報へのアクセス速度を加速するために役立ちます。機械学習アプリケーションの主要なユースケースの一部は、分類、クラスタリング、予測に関連しています。Verticaのビルトイン機械学習アルゴリズムでは、これらすべての領域を、K-means、線形回帰、ロジスティック回帰、ナイーブベイズによって扱うことができます。

K-meansアルゴリズムは、教師なし学習アルゴリズムの一種です。すなわち、入力データはラベル付けされていません。K-meansの目的は、n個の観測値をk個のクラスターに分割することです。

エンドツーエンドの機械学習管理

データの準備からモデルのスコア付けと展開に至るまで、Verticaは機械学習プロセス全体をサポートします。データの準備のために、正規化、異常値検出、サンプリング、不均衡データ処理、欠損値補完といった機能が利用できます。大規模なデータセットを使用して機械学習モデルの作成、トレーニング、テストを行った後、ROCテーブルや混同行列といったモデルレベルの統計値を使ってモデルを評価できます。

データの準備から展開まで、Verticaは機械学習のプロセス全体をサポートします。

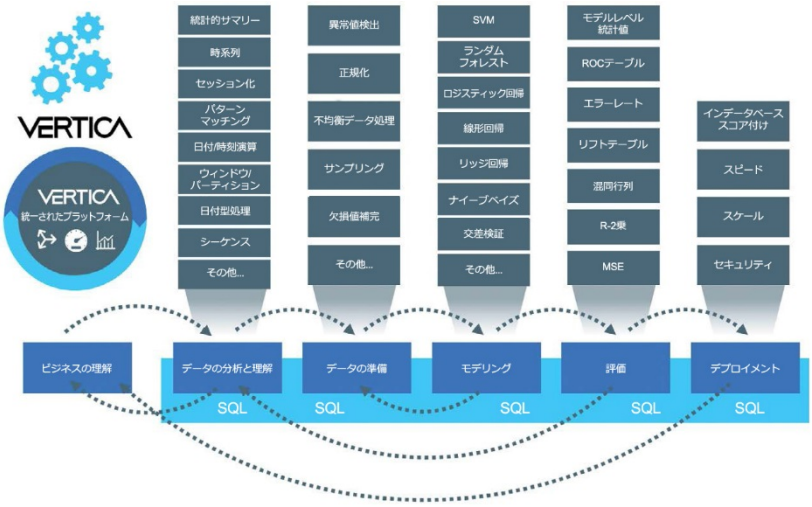


図1

エンドツーエンドの機械学習管理

K-Meansクラスタリング

クラスタリングは、データ内の自然なグループ化を検出します。同じクラスター内の項目同士は、クラスター外の項目との間に比べて、共通点が多くなります。

K-meansアルゴリズムは、教師なし学習アルゴリズムの一種です。すなわち、入力データはラベル付けされていません。K-meansの目的は、n個の観測値をk個のクラスターに分割することです。

このアルゴリズムは、ラベル付けされていないデータを受け取り、データポイント間の類似性に基づいて、データポイントをk個の異なるクラスターに分類します。この分割では、各観測値が、最も近い平均値を持つクラスターに割り当てられます。この最も近い平均値は、クラスター中心と呼ばれます。結果のモデルを使用して、後で既存のクラスターに新しいデータを追加することができます。データは、平均値すなわちクラスター中心が最も近いクラスターに割り当てられます。

K-meansには、次のようなさまざまなアプリケーションがあります。

- **顧客のセグメント化:** 顧客や購入者を、年齢、収入、製品の好みといった類似属性に基づいて、いくつかのグループ（クラスター）にセグメント化します。これは、プロモーションのターゲット設定、サポートの提供、クロスセル (商品の組み合わせ販売) 機会の発見などに利用できます。
- **詐欺行為の検出:** 特定のグループ（クラスター）に一致しない個別の観測値を見つけたり、詐欺行為の可能性が高いクラスターのタイプを特定したりすることができます。

K-meansアルゴリズムを適用するための基本的なSQL構文を次に示します。

KMEANS ('model_name', 'input_table', 'input_columns', 'num_clusters')

この関数のオプションのパラメーターとしては、除外する列、ε値、最大反復回数、クラスター中心初期値決定方法、クラスター中心初期値テーブル、距離決定方法、出力ビュー、キー列があります。

関数名	機能
kmeans	入力テーブル/ビューに対するk個のクラスター中心を判定し、オプションで各行とそれに割り当てられたクラスターを出力します。
summarize_model	クラスター中心といくつかの評価指標を表示します。
apply_kmeans	既存のK-meansモデルを受け取って、入力テーブルの行を適切なクラスターに割り当てます。

ロジスティック回帰

ロジスティック回帰アルゴリズムは、独立変数 (特徴) と従属変数 (結果) の間の論理的関係に基づいてデータをグループに分類するために使用されます。ロジスティック回帰の結果は、真/偽、合格/不合格、はい/いいえ、1/0といった結果を表す2進値です。ロジスティック回帰モデルは、次のような用途に使用できます。

- ローン申請者の信用履歴、収入、ローン条件 (予測変数) から、申請者がローンを返済できない確率 (応答) を判定します。その結果は、ローンの承認、不承認、あるいはローン条件の変更のために利用できます。
- システムの動作条件と診断測定 (予測変数) に基づいて、特定の機械部品が故障するかメンテナンスが必要になる確率 (応答) を予測します。
- 患者の年齢、血圧、喫煙習慣、飲酒習慣などの因子 (予測変数) に基づいて、特定の薬剤や治療が患者に対して有効である確率 (応答) を判定します。

ロジスティック回帰アルゴリズムは、独立変数（特徴）と従属変数（結果）の間の論理的関係に基づいてデータをグループに分類するために使用されます。線形回帰は、主に連続体上で線形関係にある連続的な数値結果を予測するために用いられます。

線形回帰は、主に連続体上で線形関係にある連続的な数値結果を予測するために用いられます。

ロジスティック回帰アルゴリズムを適用するための基本的なSQL構文を次に示します。

LOGISTIC_REG ('model_name', 'input_table', 'response_column', 'predictor_columns')

この関数のオプションのパラメーターとしては、除外する列、オプティマイザーのタイプ、ε値、最大反復回数、必要な出力のタイプ (2進値、確率) があります。

関数名	機能
logistic_reg	トレーニング (2種類の最適化方法: BFGSとNewton)
summarize_model	モデルの概要の読み取り
predict_logistic_reg	スコア判定

線形回帰

線形回帰は、主に連続体上で線形関係にある連続的な数値結果を予測するために用いられます。

線形回帰を使用すると、独立変数 (特徴) と従属変数 (結果) の間の線形関係をモデル化できます。線形回帰モデルの応用例を以下に示します。

- 住宅の価格 (応答) を、家屋の居住面積、寝室の数、浴室の数といった特徴 (予測変数) の関数としてモデル化します。
- サービスやグッズに対する需要 (応答) を、その特徴 (予測因子) に基づいてモデル化します。たとえば、ノートパソコンのさまざまなモデルの需要を、モニターサイズ、重量、価格、オペレーティングシステムなどに基づいてモデル化します。
- コンクリートの圧縮強度 (応答) と、セメント、スラグ、フライアッシュ、水、高性能減水剤、粗骨材といった成分の量 (予測因子) の間の線形関係を判定します。

線形回帰アルゴリズムを適用するための基本的なSQL構文を次に示します。

LINEAR_REG ('model_name', 'input_table', 'response_column', 'predictor_columns')

この関数のオプションのパラメーターとしては、除外する列、オプティマイザーのタイプ、ε値、最大反復回数があります。

関数名	機能
linear_reg	トレーニング (2種類の最適化方法: BFGSとNewton)
summarize_model	モデルの概要の読み取り
predict_linear_reg	スコア判定
mse	評価 (平均2乗誤差)
rSquared	評価 (R 2乗値)

ナイーブベイズ

ナイーブベイズアルゴリズムは、データを分類するために使用されます。このアルゴリズムは、いくつかの特徴を予測変数として使用して、特定のクラスまたは複数のクラスの確率を計算します。たとえば、電子メールが迷惑メールである確率を予測するには、一般的に迷惑メールと関連付けられる単語を使用します。同じ方法で、文書を内容に基づいて、たとえばニュース、金融、スポーツなどに分類することもできます。

この教師あり機械学習アルゴリズムには、次のようなアプリケーションがあります。

- 迷惑メールフィルター
- 文書の分類
- 画像の分類
- 消費者の習慣

ナイーブベイズアルゴリズムを適用するための基本的なSQL構文を次に示します。

NAIVE_BAYES ('model_name', 'input_table', 'response_column', 'predictor_columns')

この関数のオプションのパラメーターとしては、除外する列、α値があります。上記の機械学習アルゴリズムに加えて、Verticaでは、サポートベクターマシン（SVM）およびランダムフォレスト向けのインデータベース機械学習機能も提供されています。詳細については、**www.vertica.com/machinelearning**をご覧ください。

データの準備と正規化

データの準備は、データ分析のための重要な前処理段階であり、データ分析プロジェクト中にデータサイエンティストやアナリストが費やす時間のほとんどを占める場合もあります。Verticaには、補間、パターンマッチング、イベント系列の結合、高度な集計、異常値検出、シーケンスといった豊富な分析機能が組み込まれています。これらの機能は、データ準備プロセスをサポートし、データ分析作業の効率を高めることで、価値実現までの時間を短縮するために役立ちます。

また、機械学習モデルのトレーニング前のデータ準備には、データの正規化などのその他の機能も使用できます。正規化の目的は、主に、スケールが異なる数値データを同等のスケールに揃えることです。このようなデータ正規化機能を使用しないと、異なる特徴の値の間に大きな差異がある場合に、機械学習アルゴリズムのパフォーマンスが低下することがあります。予測分析のための機械学習には、2種類のデータ準備方法があります。正規化、MinMax、Zスコアを使用するものです。

Verticaには、補間、パターンマッチング、イベント系列の結合、高度な集計、異常値検出、シーケンスといった豊富な分析機能が組み込まれています。これらの機能は、データ準備プロセスをサポートし、データ分析作業の効率を高めることで、価値実現までの時間を短縮するために役立ちます。

Verticaで線形回帰モデルのトレーニングにカテゴリ予測変数を考慮する必要がある場合には、あらかじめ変数を数値に変換しておきます。カテゴリ変数を数値に変換するには、いくつかの方法があります。たとえば、ワンホットエンコーディングを使用する方法があります。

Verticaのインデータベース機械学習機能の実装

次に示すのは、Prestigeデータセットを使用した線形回帰モデルの例です[5]。このデータセットは、収入とその他の変数の間の関係を判定するために使用されます。この例では、データセットをロードし、Verticaのlinear_reg関数を使用することで、予測変数の応答への影響を理解する方法を示します。

データセットには次の情報が含まれます。

- 職種名
- 教育 (年数)
- 収入 – 現職者の1971年の平均収入 (ドル)
- 女性 – 現職者の女性の割合
- プレステージ – この職種のPineo-Porterプレステージ (職業威信) スコア (1960年代中頃に実施された社会調査に基づく)

5 「Prestige Data Set」、カナダ統計局、 カナダ国勢調査 (1971年)、 第3巻、パート6

- 国勢調査 – カナダ国勢調査の職種コード
- タイプ – 職種のタイプ。bcはブルーカラー、wcはホワイトカラー、profは専門職、管理職、技術職を表します。

ここでの目標は、このデータセットを使用して、データセット内の他の値から収入を予測する線形回帰モデルを構築し、モデルの適合度を評価することです。

初めに、このモデルに使用する変数をどのように選択すればよいでしょうか。タイプ列は除外できます。Verticaは現時点ではカテゴリ予測変数をサポートしていないからです。職種名列と国勢調査列には固有の値が多数含まれるため、このユースケースでは収入の予測に役立つとは考えられません。したがって、残るのは、教育、プレステージ、女性の各列となります。

メモ: 実際には、Verticaで線形回帰モデルのトレーニングにカテゴリ予測変数を考慮する必要がある場合には、あらかじめ変数を数値に変換しておきます。カテゴリ変数を数値に変換するには、いくつかの方法があります。たとえば、ワンホットエンコーディングを使用する方法があります。

カテゴリ変数を数値に変換するには、いくつかの方法があります。たとえば、ワンホットエンコーディングを使用する方法があります。

ステップ1: データのロード

次に示すのは、Prestigeデータセットを格納するテーブルの定義です。

```
=> DROP TABLE IF EXISTS public.prestige CASCADE;
=> CREATE TABLE public.prestige
        ( occupation VARCHAR(25), education NUMERIC(5,2), -- avg years of education
        income INTEGER, -- avg income
        women NUMERIC(5,2), -- % of woman
        prestige NUMERIC(5,2), -- avg prestige rating
        census INTEGER, -- occupation code
        type CHAR(4) -- Professional & managerial (prof)
        )               -- White collar (wc)
                        -- Blue collar (bc)
                        -- Not Available (na)
```

Verticaでは、PREDICT_LINEAR_REG関数によって線形回帰モデルが入力テーブルに適用されます。この関数の詳細情報は、Verticaドキュメントに記載されています。

PrestigeデータセットからVerticaテーブルにデータをロードするには、次のSQLステートメントを使用します。

```
=> COPY public.prestige
FROM stdin
        DELIMITER ','
        SKIP 1
        ABORT ON ERROR
        DIRECT
;
```

ステップ2: 線形回帰モデルの作成

次に、Verticaの機械学習関数LINEAR_REGを使用して、線形回帰モデルを作成します。

モデルを作成するには、LINEAR_REG関数をpublic .prestigeテーブルに次のように適用します。
ステートメントでは、incomeが応答であり、予測変数はeducation、women、prestigeです。
=> SELECT LINEAR_REG('prestige', 'public .prestige', 'income', 'education,women,prestige');
このステートメントでは、次の式の係数を求めようとしています。

$$income = \alpha + \beta_1 education + \beta_2 women + \beta_3 prestige \tag{4}$$

モデルを作成したら、SUMMARIZE_MODEL関数を使用して、モデルの特性を観察します。

```
=> SELECT SUMMARIZE_MODEL('prestige');
SUMMARIZE_MODELは、次の情報を返します。
SUMMARIZE_MODEL| coeff names : {Intercept, education, women, prestige}
coefficients: {−253.8390442, 177.1907572, −50.95063456, 141.463157}
p_value: {0.83275, 0.37062, 4.1569e−08, 8.84315e−06}
```

これらの係数を使用して、式 (4) を次のように書き直せます。

$$\begin{aligned} income = & -253.8390442 \\ & + 177.1907572 * education \\ & - 50.950663456 * women \\ & + 141.463157 * prestige \end{aligned} \tag{5}$$

最後に、線形回帰モデルがデータにどの程度適合するか、すなわち適合度を測定する方法を調べてみましょう。Verticaでは、PREDICT_LINEAR_REG関数によって線形回帰モデルが入力テーブルに適用されます。この関数の詳細情報は、Verticaドキュメントに記載されています。

ステップ3: 適合度の評価

線形回帰モデルが観測されたデータにどの程度適合するかを判定する一般的な方法の1つは、決定係数です。この係数は、次の式で定義されます。

$$R^2 = 1 - \frac{\sum_i (\hat{y}_i - y_i)^2}{\sum_i (y_i - \bar{y})^2} \qquad (6)$$

決定係数R2の範囲は、0（不適合）と1（完全適合）の間です。決定係数を計算するには、VerticaのRSQUARED関数を使用します。

```
=> SELECT RSQUARED(income, predicted) OVER()
FROM ( SELECT
        income,
        PREDICT_LINEAR_REG (
                prestige, women
                USING PARAMETERS OWNER='dbadmin',
                MODEL_NAME='prestige')
        AS predicted
        FROM public.prestige
        ) x ;
```

rsq	コメント
0.63995924449805	102行中102を使用

メモ: OWNERパラメーターは、Vertica 8.1で廃止される予定です。

決定係数の評価は、多くの場合、調査する領域に依存します。社会科学では、0.6という係数はかなりよい値と見なされます[6]。

モデルを評価する際には、複数の指標を考慮することが重要です。1つの指標でよい値が得られたとしても、モデル自体がニーズを満たすとは限りません。適合度の評価に使用されるR 2乗値やその他の指標について理解しておくことが重要です。

さまざまな産業での機械学習の使用例

Vertica Analytics Platformのインデータベース機械学習機能は、さまざまな業種でビジネス上の利益をもたらすために導入されています。

Vertica Analytics Platformのインデータベース機械学習機能は、さまざまな業種でビジネス上の利益をもたらすために導入されています。

6『Linear Models of R, second edition』、CRC Press、Julian J. Faraway著、2014年

Verticaのインデータベース機械学習機能を使用すれば、ビッグデータを活用しながら、予測分析のプロセスをシンプル化、高速化することで、意思決定を改善し、競争上の立場を強化し、情報を入手するまでの時間を短縮できます。

- **金融サービス**企業では、詐欺行為の検出、投資機会の発見、ハイリスクプロファイルの顧客の特定、ローン申請者の債務不履行確率の判定などに利用できます。
- **政府**機関、たとえば公共安全や公益事業などの分野では、機械学習を利用して、詐欺行為の検出、アイデンティティ窃盗の防止、スマートメーターからのデータの分析による効率向上および費用節約手段の発見などが可能です。
- **通信サービスプロバイダー**は、ネットワークプローブやセンサーからのさまざまなデータを利用して、ネットワークパフォーマンスの分析、容量の制約の予測、エンドユーザーへの高品質なサービス提供の維持などが可能です。
- **マーケティング/セールス**分野では、機械学習を利用することで、購入パターンの分析、顧客のセグメント化、ショッピングエクスペリエンスの個別化、ターゲットを絞ったマーケティングキャンペーンの実施などが可能です。
- **石油/ガス**業界では、機械学習を利用することで、鉱物分析による新エネルギー源の発見、石油配送の合理化による効率向上とコスト削減、機械またはセンサーの故障の予測を通じた予防的メンテナンスなどを実現します。
- **運輸**業界では、トレンドを分析してパターンを発見することで、顧客サービスの改善、ルートの最適化、利益率の向上などを実現できます。

まとめ

Verticaのインデータベース機械学習機能を使用すれば、ビッグデータを活用しながら、予測分析のプロセスをシンプル化、高速化することで、意思決定を改善し、競争上の立場を強化し、情報を入手するまでの時間を短縮できます。同時に、C++、Java、Python、Rでプログラムされたユーザー定義拡張がサポートされているので、選択の自由も得られます。

Verticaを体験できる技術ブログ、評価版のダウンロードについては、以下のサイトをご覧ください。

- ブログポスト:「Verticaによる機械学習」
- 無料トライアル (評価版) のダウンロード: Vertica Analytics Platform

詳細情報

www.vertica.com

Vertica
150 Cambridgepark Drive
Cambridge, MA 02140

詳細情報: **www.vertica.com**

マイクロフォーカスエンタープライズ株式会社
jp-info-enterprise@microfocus.com
www.microfocus-enterprise.co.jp

www.vertica.com