【知っておきたい】10 の機械学習アルゴリズム

※本投稿は、”10 Machine Learning Algorithms You Need to Know”(https://rapidminer.com/blog/10-machine-learning-algorithms/)を翻訳したものです。

機械学習がビジネスにインパクトを与える方法を探し始めたばかりなら、おそらく最初に直面する問いは、**機械学習アルゴリズムにはどんな種類があるのか、何に向いているのか、今回のプロジェクトにはどれを選べばいいのか**、というものになるでしょう。この記事は、それらの問いに答えるのに役立ちます。

機械学習アルゴリズムを分ける方法は、いくつかあります。一つは、学習データがどのようになっているかを見ることです。データサイエンティストは学習データによって、三つのカテゴリを使い分けます。

- **教師あり学習**は、ラベル付けされた過去のデータ(ラベルは人の手で付けられることが多い)を基にアルゴリズムを訓練し、将来の結果を予測しようとします。
- **教師なし学習**は、対照的に、ラベルのないデータを使用して、アルゴリズムはデータ自身にあるルールやパターンを抽出してデータを解釈しようとします。
- **半教師あり学習**は、上記二つを混ぜたもので、ラベルがないデータが多く、教師あり(ラベルがある)データが少ない場合に用いられます。

アルゴリズムを分ける方法として、ビジネスの観点からより実践的なものがあります。それは、アルゴリズムの動作や、解く問題の種類などで分けることです。これが、この記事で述べることです。

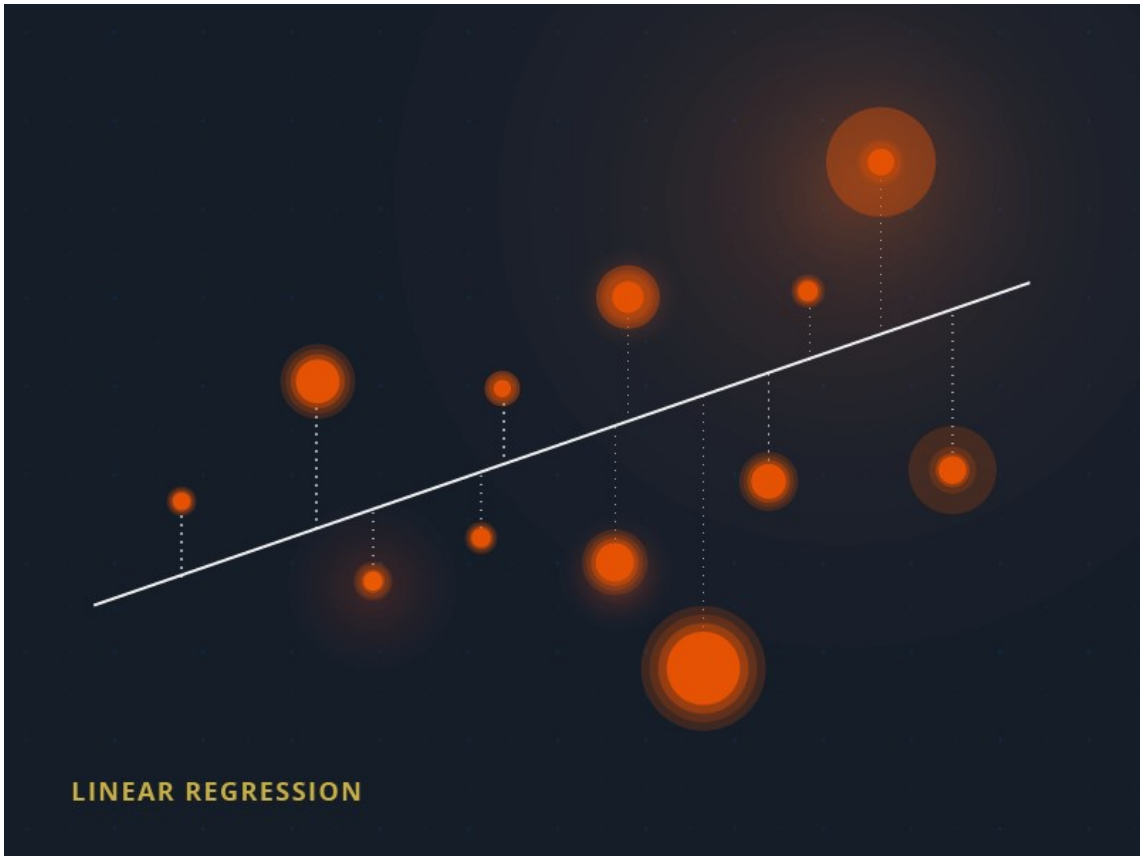
ここでは、**回帰**、**クラスタリング**、**分類**の三つの基本的なアルゴリズムについて述べます。それぞれ見ていきましょう。

回帰アルゴリズム

ビジネスでよく見る回帰アルゴリズムには、基本的に二種類あります。これらは、統計学とも身近な回帰を基にしています。

1. 線形回帰

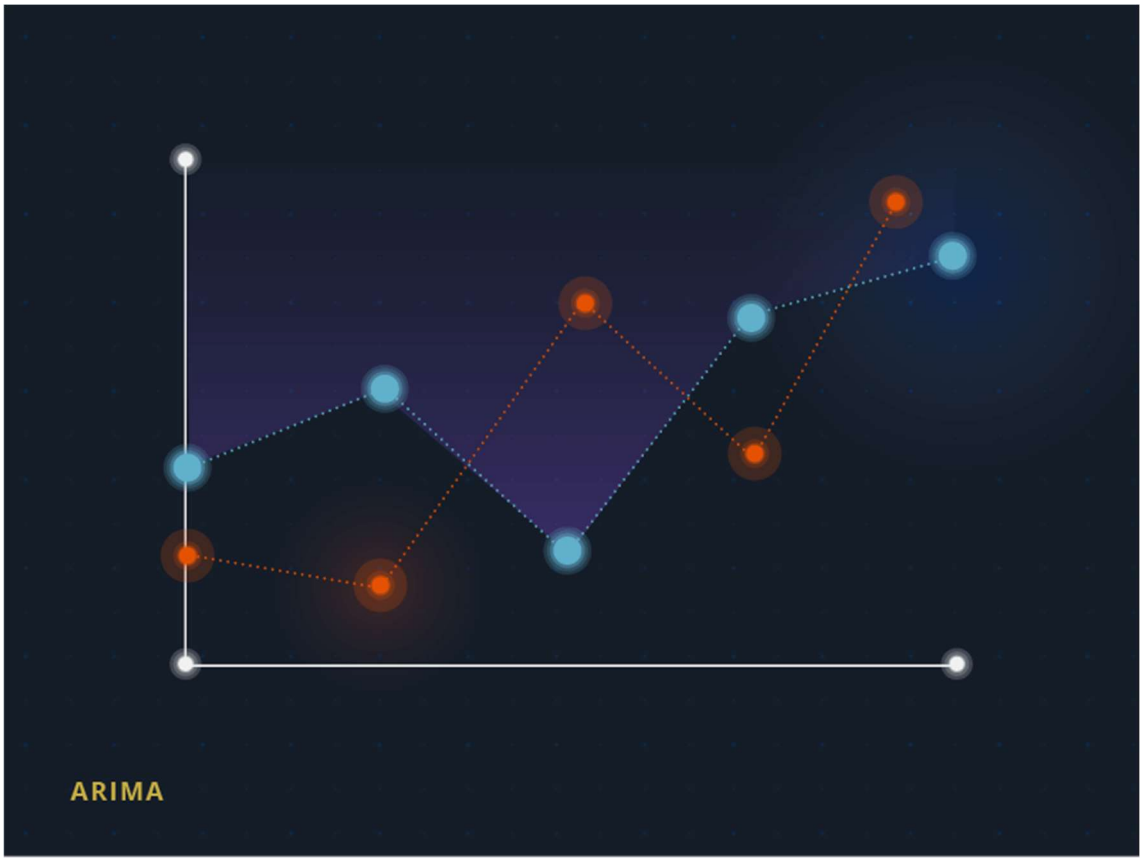
とても簡単に説明すると、**線形回帰**は**目的変数**(y 軸)と**説明変数**(x 軸)のデータ点を基に、直線を引いたものです。



線形回帰は一般的に使用されている統計モデルで、数値データを理解する万能なものと捉えられています。例えば、線形回帰は価格を決定するのに、売上に対して様々な価格の売上をマッピングすることで、商品やサービスの価格変更の影響の理解に使用することができます。各ユースケースによりますが、**リッジ回帰**、**ラッソ回帰**、**多項式回帰**なども含んだ線形回帰は、様々なものに適用できるでしょう。

2. ARIMA

ARIMA(自己回帰和分移動平均)モデルは、**回帰モデル**の特殊なタイプと考えることができます。



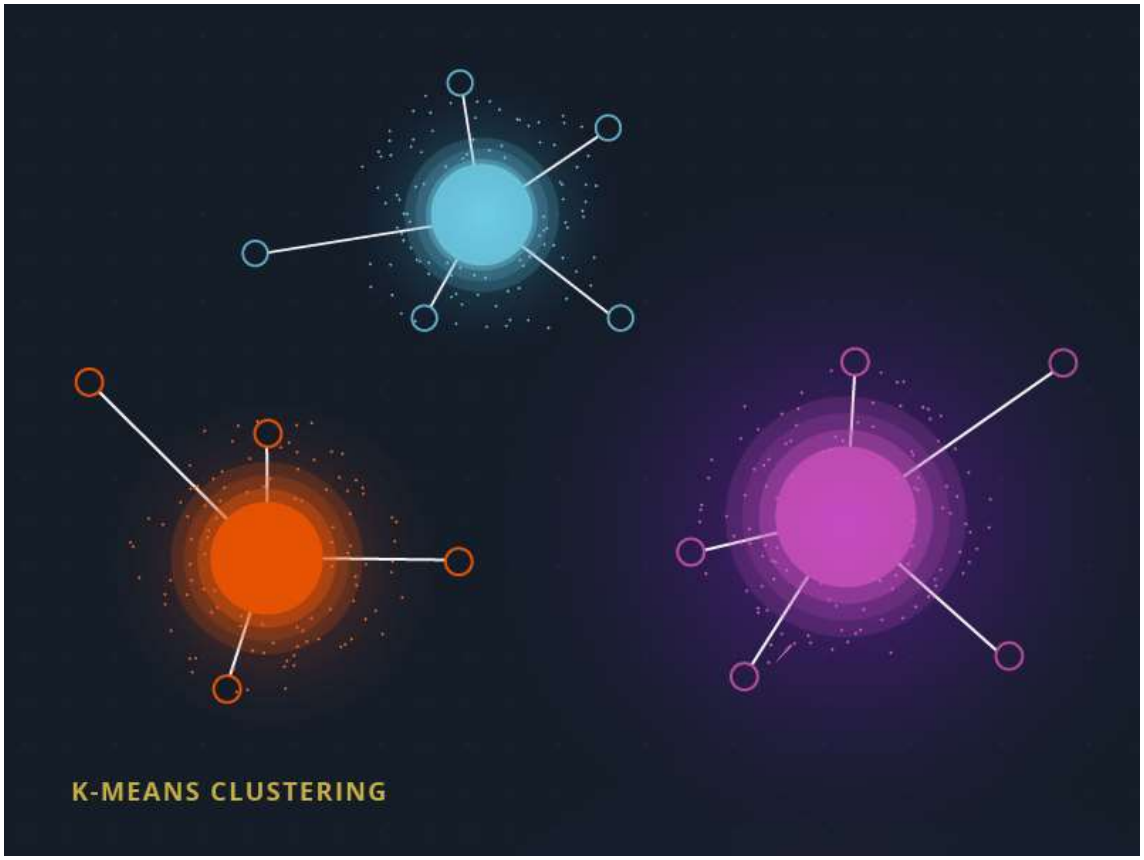
データ点を互いに独立したものと捉えるのではなく、連続したものとして解釈するため、時間依存のデータ点を探索することが可能になります。そのため、ARIMA モデルは、需要予測や価格予測などの連続した時系列分析に特に役に立ちます。

クラスタリングアルゴリズム

クラスタリングアルゴリズムは通常、データセット内のグループを見つけるのに使用されます。また、これを行うアルゴリズムにはいくつか種類があります。

3.k-means クラスタリング

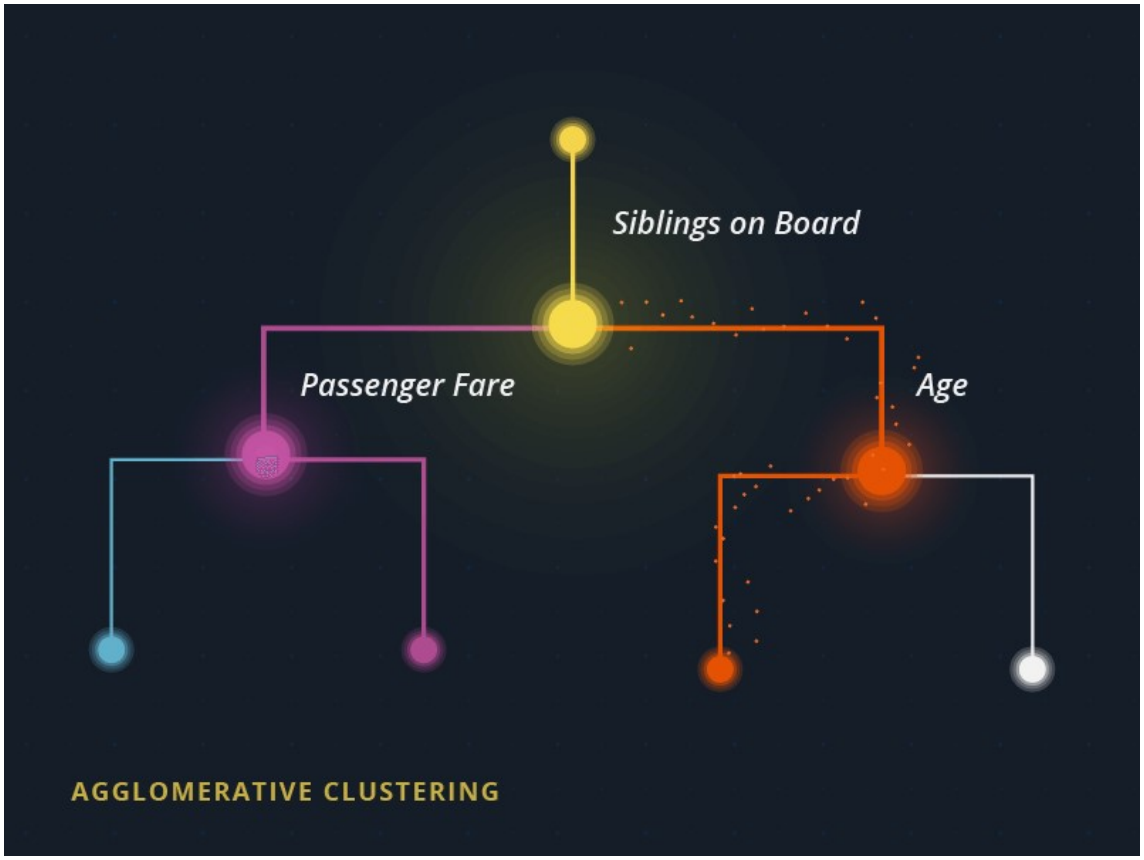
k-means クラスタリングは一般的に、関連する特徴ごとにグループに分け、グループごとにまとめます。



顧客セグメンテーション戦略を進めようとしている企業は、k-means クラスタリングを使用すると、ある顧客グループが反応するようなマーケティングキャンペーンをより効率的に立てられるでしょう。別の k-means クラスタリングのユースケースでは、過去に保険会社を騙す傾向があったデータを使用して、現在のケースを調査することで保険金詐欺を発見することが考えられます。

4. 凝集型&分割型クラスタリング

凝集型クラスタリングはデータやクラスタの階層関係を見つけるのに使用される方法です。



ボトムアップのアプローチを使用し、各データ点をそれ自身のクラスタに分け、似たクラスタ同士を結合させます。対照的に、**分割型クラスタリング**は反対のアプローチをとります。すべてのデータ点は同じクラスタにあると仮定し、そこから似たクラスタに分割していきます。

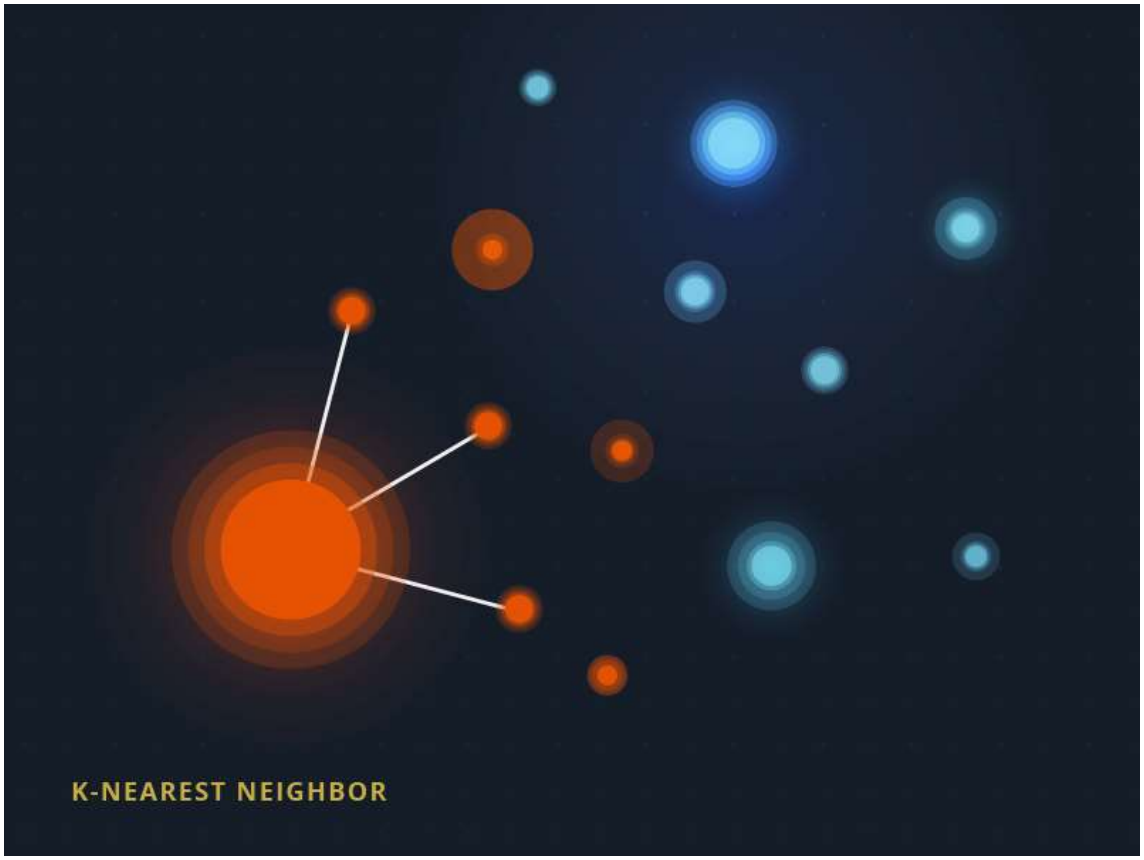
これらのクラスタリングアルゴリズムのタイムリーな使用例は、ウィルスの追跡です。DNA 解析を行うと、研究者は変異率や感染パターンをより理解することができるようになります。

分類アルゴリズム

分類アルゴリズムはクラスタリングアルゴリズムと似ていますが、クラスタリングがデータ内のカテゴリの発見とカテゴリへのデータ点の整理の両方に利用されるのに対し、分類は事前に定義されたカテゴリへの整理に使用されます。

5.k 近傍法

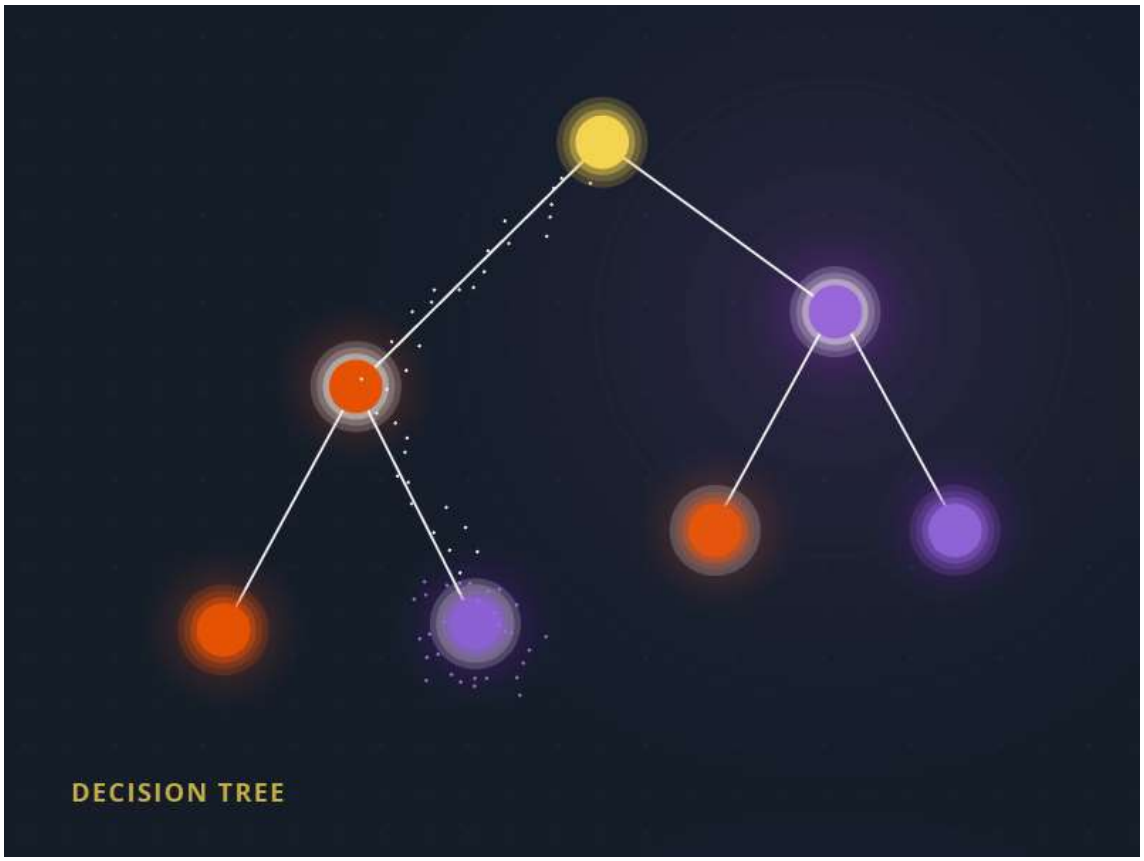
k-means クラスタリングと混同しないようにしましょう。**k 近傍法**はパターン分類の手法で、提示されたデータを見て、すべての過去のデータを確認し、最も似たものを特定します。



k 近傍法は、よくクレジットカード取引の行動分析に用いられ、過去の取引と比較します。クレジットカードを使用した海外での取引のような、異常な行動はカード発行会社の不正検知部署からの電話に繋がるかもしれません。また、この機械学習アルゴリズムは視覚パターン認識にも使用され、現在は小売業者の損失防止策の一部としてよく用いられています。

6. ツリーベースアルゴリズム

決定木やランダムフォレスト、勾配ブースティング決定木などを含む**ツリーベースアルゴリズム**は、分類問題を解くのに使用されます。決定木は、多くのカテゴリ値を持つデータセットの理解を得意としており、データの一部が欠損している場合でも効果的です。



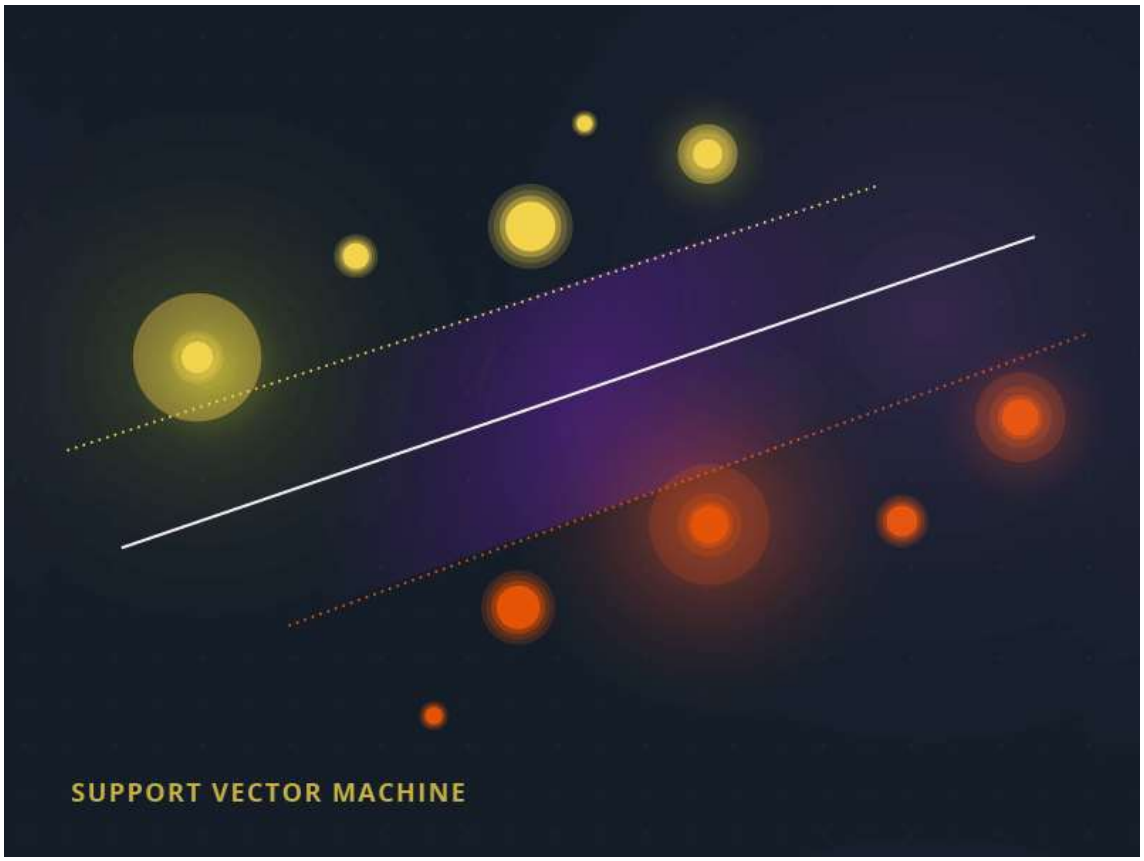
これらは、予測モデルにまず使われるもので、「どの戦略をもっと行うべきか？」という問いに答えてくれるため、マーケティングで役立ちます。決定木は、メールマーケティングの担当者が、あるキャンペーンに対して、どの顧客が注文する可能性が高いかを決定するのに役立つかもしれません。

ランダムフォレストは複数の決定木を使用して、より完全な分析を行います。ランダムフォレストでは、複数の木が作成され、複数の木の結果の平均を取って予測を行います。

勾配ブースティング決定木(GBTs)も決定木を用いますが、反復的なアプローチをとり、各決定木モデル内の間違いをなくそうとます。GBTs はデータサイエンティストにとって、最も強力な予測手法の一つと広く捉えられています。また他のケースでは、製造業者が利益を最大化させるために、製品やサービスの価格の最適化に使用することができます。

7. サポートベクターマシン

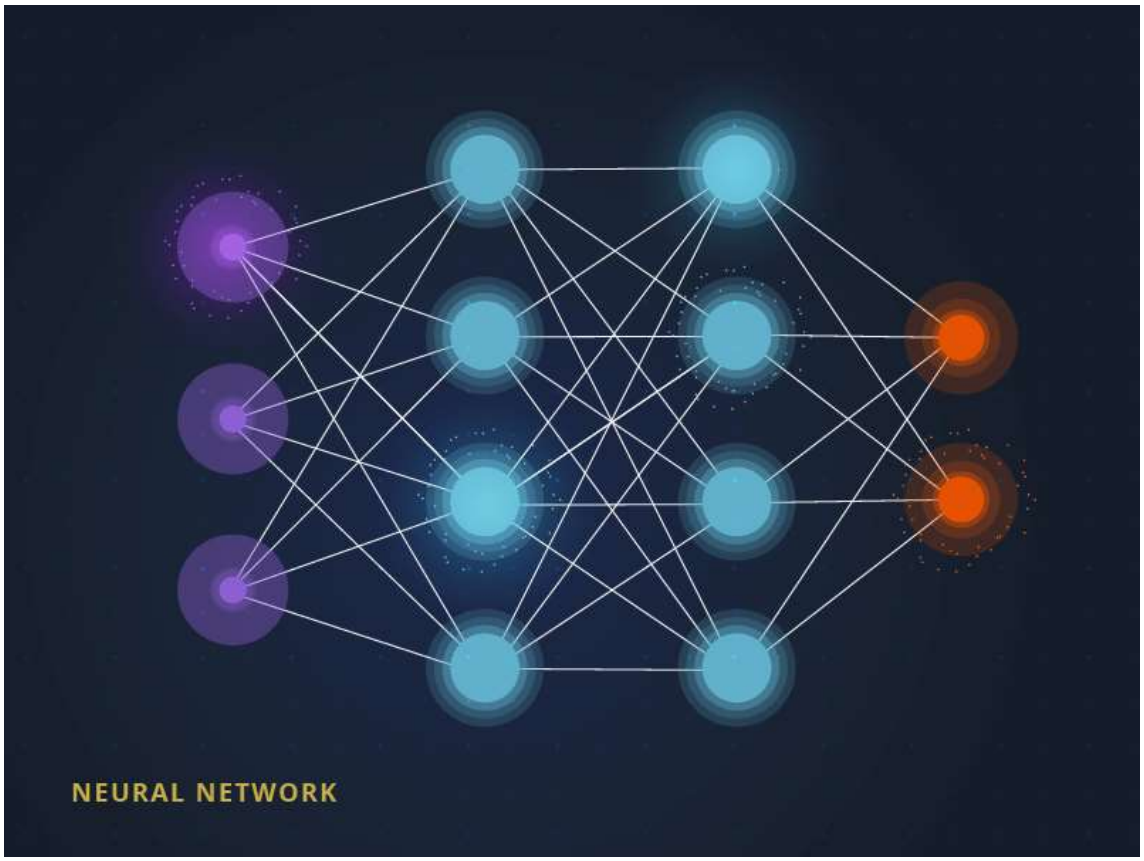
サポートベクターマシン(SVM)は、一部の専門家によると、最も人気のある機械学習アルゴリズムです。SVM は分類(時には回帰)アルゴリズムで、データセットをクラスに分けるのに使用されます。例えば、クラス間に境界線を引いて二つのクラスに分割します。



これには無数の線を引くことができますが、SVM は最適な線を見つけるのに役立ちます。データサイエンティストは SVMs を画像分類、顔認識、手書き文字認識、バイオインフォマティクスなど、ビジネスの幅広い分野で使用しています。

8. ニューラルネットワーク

ニューラルネットワークはパターンを認識し、可能な限り人間の脳を模倣するように設計されたアルゴリズムの集合体です。ニューラルネットは、脳のように、たとえ本来意図されていなかったものでも、状況の変化に適応することができます。



ニューラルネットは、学習データに犬の画像を与えると、犬を認識するようになります。一度アルゴリズムが学習データを処理すると、新しい画像を「犬」か「犬でない」かに分類することができます。ニューラルネットワークは画像の分野で最も活躍していますが、テキストや音声、時系列データなどにも使用することができます。ニューラルネットワークには様々な種類があり、それぞれのタスクで働くように最適化されています。

ニューラルネットワークをビジネスに適用したものには、気象予報、顔検出と顔認識、音声のテキストへの書き写し、株価市場予測などがあります。マーケティング担当者は、ニューラルネットワークを使用して、あるコンテンツに最も反応しそうな顧客へのキャンペーンを考えられます。

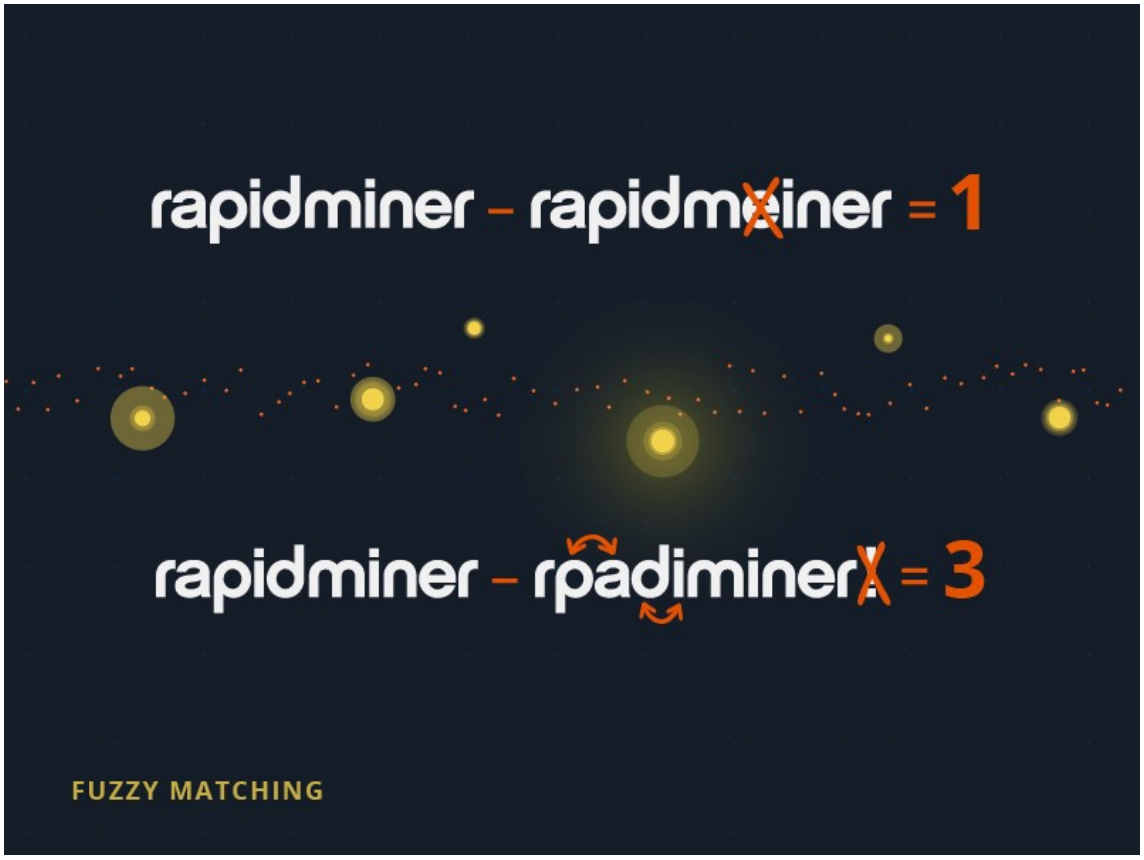
ディープラーニングは、実はニューラルネットワークの一部分で、大規模なデータセットを分析することで「学習」します。ディープラーニングにはビジネスケースが無数にあり、多くの場合、一般的な機械学習アルゴリズムを凌駕します。一般的に、ディープラーニングは特徴量の生成に人の入力を必要としないため、例えば、テキストや音声、画像認識、自動運転や他のたくさんのものを解釈するのが得意です。

その他の機械学習アルゴリズム

上記のカテゴリに加えて、ファジーマッチングや特徴量選択アルゴリズムのような、モデルの作成時や学習時に使用されるアルゴリズムがあります。

9. ファジーマッチング

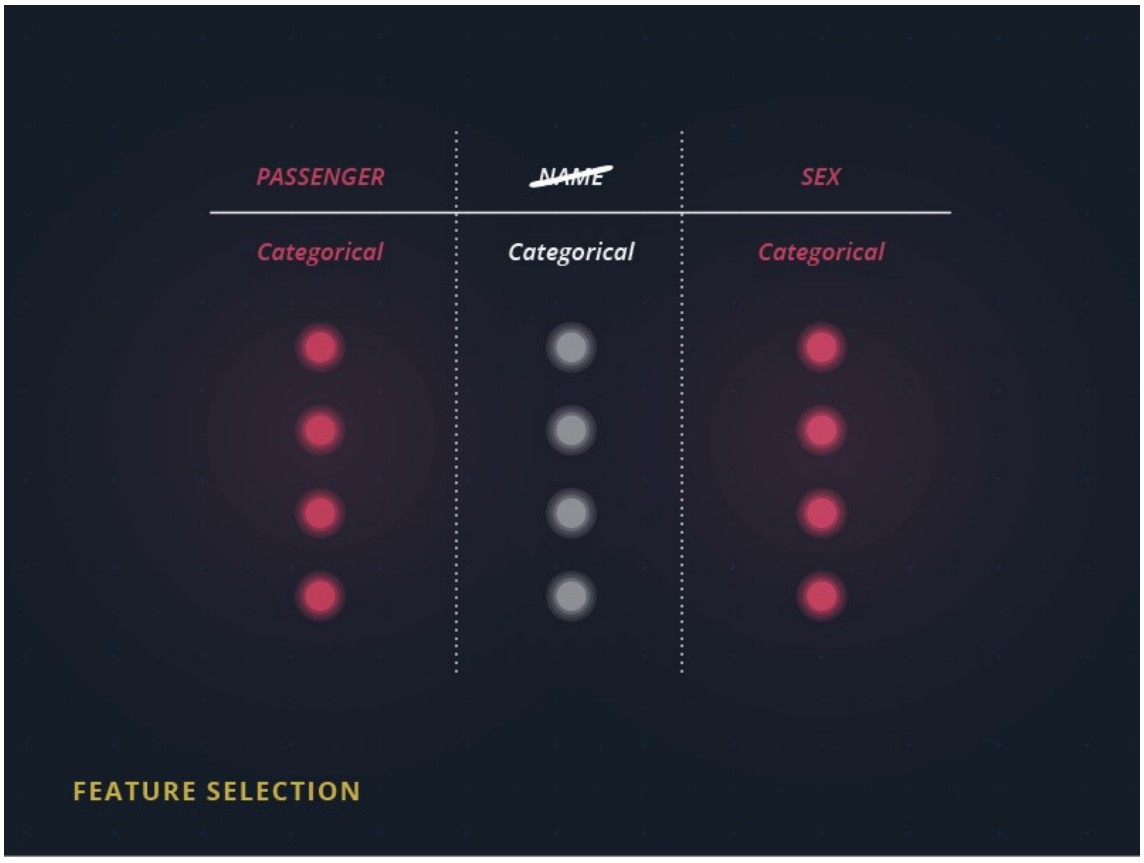
ファジーマッチングはクラスタリングアルゴリズムの一種で、タイプミスのような、データの問題によって語が完全一致していないときでも一致させることができます。一部の自然言語処理タスクでは、前処理にファジーマッチングを用いると、3～5%結果を向上させることができます。



典型的なケースは、顧客のプロフィール管理です。ファジーマッチングは非常に似た住所を同じものと判定でき、二つのよく似た住所にユニークなレコード ID とソースファイルを使用できるようになります。

10. 特徴量選択アルゴリズム

特徴量選択アルゴリズムはモデルから入力パラメータの数を減らすのに使用されます。入力変数が少なくなれば、モデルのパフォーマンスを向上させるだけでなく、モデルの実行にかかる計算コストも減らせるかもしれません。



PCA や **MRMR** のような一般的に使用されている手法は、特徴量の減ったセットから可能な限り多くの情報を取得するのに有用です。特徴量のサブセットを使用すると、モデルがノイズに惑わされることも少なくなり、アルゴリズムの計算時間も減らせるため、有益な場合が多いです。例えば、特徴量選択はビジネスの競合関係を表示するのに使用されてきました。

最初のプロジェクトを軌道に乗せる方法も含めて、機械学習をもっと深く知りたい場合は、RapidMiner の Human's Guide to Machine Learning Projects を確認してください。