教師あり学習 vs 教師なし学習 この違いは？

※本投稿は、”Supervised vs. Unsupervised Machine Learning: What’s the Difference?”を翻訳したものです。

機械学習は、アルゴリズムやモデルの種類が終わりなく増えていくように思えて、混乱してしまうかもしれません。しかし、スキルレベルに関係なく、誰でも機械学習を理解して使用できることを私たちは事実として知っています。

最近のリリース（RapidMiner Go）のように大きな進歩もあり、初心者も機械学習がビジネスインパクトをもたらす、パワフルなツールとして簡単に活用し始めることができます。そのようなわけで、一歩引いて機械学習のコアとなるコンセプトについて初学者向けに説明を書きたいと思いました。

そのような考えから、この投稿では機械学習の最も有名なカテゴリの二つである、教師あり学習と教師なし学習について見ていきたいと思います。

教師あり学習 vs. 教師なし学習

教師あり学習と教師なし学習の一番の違いは、モデルに何を予測したいか伝えるか、伝えないかです。

教師あり学習では、モデルの学習に使用するデータは、過去のデータ点と、そのデータ点の結果を持っています。一方、**教師なし学習**は結果を持たず、モデルの仕事はデータ自身の中にどのようなパターンが存在するのかを発見することです。

機械学習のどちらのタイプも予測分析には不可欠ですが、それらは異なる場面、異なるデータセットで活躍します。

教師あり学習

教師あり学習を学習させるためには、まずデータの結果がラベル付けされた、過去のデータセットが必要です。このデータは、運用中にモデルがアクセスできる**入力**を、既知の**結果**（入力が与えられてモデルが予測するもの）にマッピングします。

例えば、離反を予測するあるモデルでは、このデータは顧客についての様々な過去の事実（運用中の入力）で、離反するかしないか（モデルに予測してほしい結果）と対になっています。

データセットは、学習データとテストデータの二つに分割されます。

学習データは、名前から推測される通り、データ内のパターンを過去の結果にマッピングするために、モデルの学習に使用されます。モデルが作成されると、テストデータはモデルの予測と既知の結果を比較して、モデルの精度を検証します。

このシナリオは、答えが付いている問題集のように考えることができます。勉強した後は、問題に答えることができ、その後あなたの回答と問題の答えがあっているかを比較できます。

一般的にデータサイエンスの世界では、教師あり学習で訓練されたモデルは、予測が既知の結果と受容できるレベルの正確さで合っていれば成功と捉えられます。

しかしながら、ビジネスの世界では、モデルが成功かを決定する際、単純なモデルの正確さよりも、投資に対する価値とリターンを考慮するのが良いでしょう。

教師あり学習の重要性

予測モデルのゴールは、単純に学習データ内のパターンを理解するだけでなく、学んだことを未知の新しい入力データに適用して、結果がわからないデータ点を予測することです。この能力があるからこそ、予測モデルは現実のシナリオで価値あるものになります。

モデルの学習プロセス内での一般的な問題は、過学習です。**過学習**はモデルが学習データにあまりに適合している状態のことです。この状態だと、テストデータでは高い精度で予測することができますが、学習データとテストデータ以外のデータ（これが、実際にモデルに正確に予測してほしい現実世界のデータです）を予測した際には、低い精度になってしまいます。

例えば、犬と猫を見分けるモデルを訓練しているとして、犬の例にグレート・デーンとロットワイラーしか含まれていない場合、単純な大きさだけで正確に見分けられるモデルを作成することは容易です。しかしながら、このモデルが実際の世界にさらされたとき、チワワやコーギーを猫と分類してしまうでしょう。

過学習は学習データがその結果の値の中に間違いを含んでいる場合にも起こりえます。それは、当然モデルの将来の予測を歪ませてしまうでしょう。これは、モデル作成プロセス内での重要なステップとして、データの前処理と検証の重要性を強調します。

過学習を防ぐ最も良い方法は、幅広い様々なデータ点に対応できるような、単純で、一般的なモデルを使用することです。また、もちろん、モデルの学習前に学習データの整合性を検証するべきです。

教師あり学習の使用例

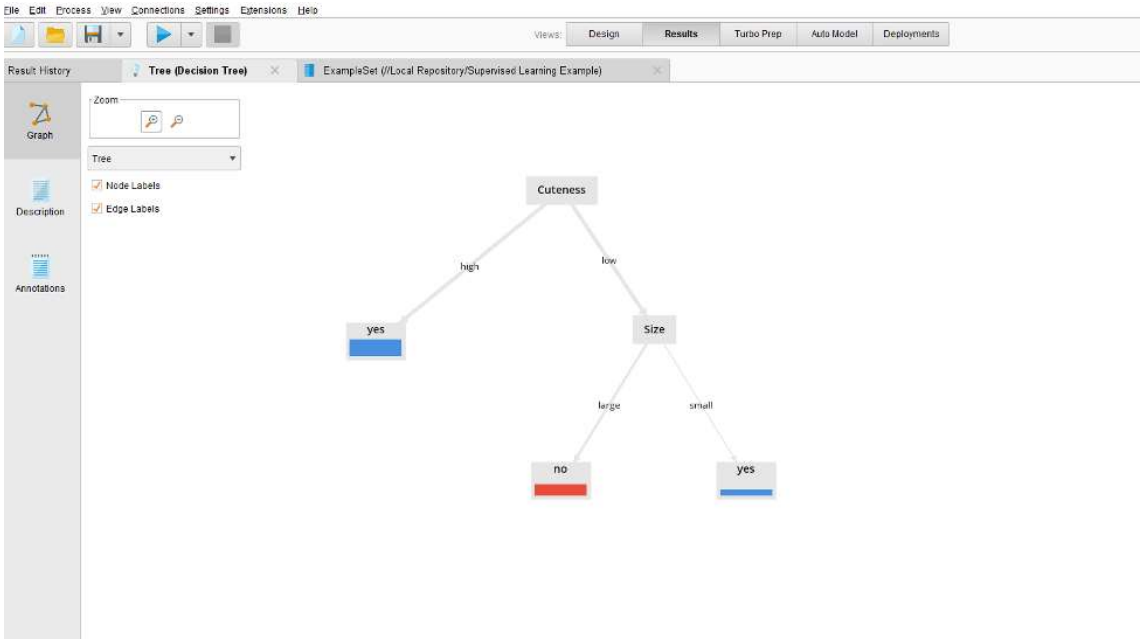
教師あり学習の使用例は、**分類**と**回帰**に分けることができます。

分類のケースでは、モデルはデータがどのグループに属するかを予測します。例えば、継続顧客と離反しそうな顧客です。**回帰**の場合は、モデルは数値を予測します。例えば、工場の部品がどれくらいで修理が必要かを予測します。

教師あり学習の例

教師あり学習の良い例は、分類木です。決定木は使いやすく、視覚化しやすいです。名前から推測されるように、これは複数の条件を使用して最終的な結論にたどり着きます。決定木はとても理解しやすく説明しやすいため、よく選択されます。これらは、ビジネスでの機械学習の実装においてキーとなる要素です。

決定木は再帰的なトップダウン戦略を使用します。データセットから木のトップとなる最初の属性（スプレッドシートでは列）が選択され、二つのカテゴリに分割します。以下の例は、私たちの創立者である Ingo Mierswa から提供されたものです。犬を様々な属性で分け、条件を受け入れられるか受け入れられないかで分けて予測を行います。



Cuteness（可愛らしさ）からスタートして、データは可愛らしさが高いか低いかで分けられます。もし可愛らしさが高ければ、犬は常に受け入れられます。これは、純粋なカテゴリをもち、枝はここで終了することを意味しています。

可愛らしさが低ければ、犬のサイズが条件になり、サイズで新しいカテゴリに分類します。ここより、あまり可愛らしさがなくサイズが大きな犬はずっと選ばれませんが、あまり可愛らしさがな

くサイズが小さな犬はここで選ばれることがわかります。このようにして、完全な決定木ができました。

教師なし学習

教師あり学習とは違い、教師なし学習は「正しい答え」を含まないデータを使用します。代わりに、これらのモデルはデータ自身が持つデータ内の構造を見分けるように作成されます。例えば、異なるデータ点がどのように一緒のカテゴリにグループ分けされるのかがわかります。

教師なし学習の重要性

これらの特徴により、効率的なマーケティングや顧客をグループに分ける要素の特定のために、共通の属性をもとに潜在的な顧客をカテゴリにわけるトランザクションデータにとって、教師なし学習は特に有用です。

教師なし学習の使い道

教師なし学習には三つのタイプがあります。

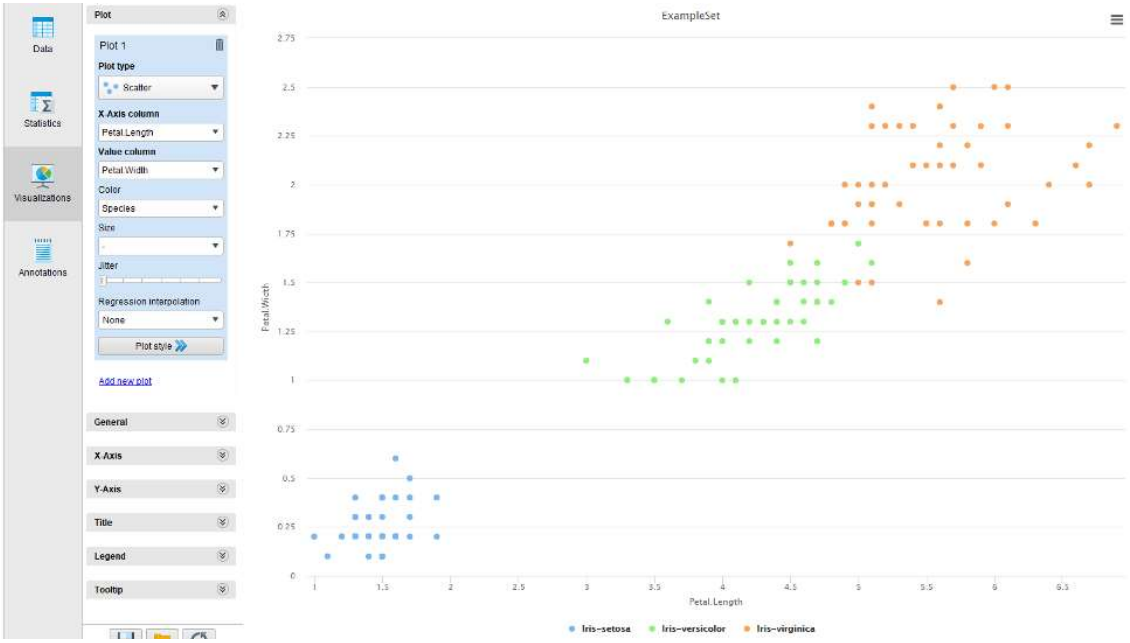
1. **クラスタリング**（グルーピング）は各データと似ているデータ内のアイテムを見つけようとします。例えば、互いに似ている顧客の選別などです。
2. **トピック検出**はクラスタリングの一種で、テキストからトピックを判別します。
    - **異常検知**（外れ値検出）は、データの中から他のデータとは異なるアイテムを見つけます。
3. **頻出パターンマイニング**は Y を購入した人が X を購入する可能性が高いかを判断します。

教師なし学習の例

k-means クラスタリングは理解しやすい教師なし学習アルゴリズムの一つです。このアルゴリズムは k 個の中心点を定め、その周囲のデータ点をクラスタ化します。そして、最も近い中心点ごとにデータ点をグループに分け、全てのデータ点がいずれかのカテゴリに属するまで続けます。

（RapidMiner Studio で利用可能な）Iris データセットを用いると、この動きを見ることができます。このデータは、アヤメの花の種類ごとにがく片の長さ、がく片の幅、花びらの長さ、花びらの幅を測定したものです。

私たちはデータセット内に三種類のアヤメの花が含まれていることを知っていますが、ラベルを無視して、教師なし学習モデルがこれらの測定値を基に、データセットの花の種類を正確に分ける様子を見ることができます。



今回は、周囲にクラスタを作成する中心点を 3 にしました。結果は、一つ目のクラスタに 50 個の花を、二つ目のクラスタに 39 個、三つ目のクラスタに 61 個の花を分けました。上記の散布図を見ると、データがはっきりとクラスタリングされたことがわかり、各花の最も可能性の高い種にラベル付けすることができました。

半教師あり学習

すべてのユースケースが教師あり学習や教師なし学習に該当するわけではありません。特にデータのラベルが一部しかない、またはデータの結果が全くない場合などは、半教師あり学習が使用されることがあります。

例えば、メールがスパムかそうでないかをカテゴリ分けするのに、教師なし学習を使用することもできます。そこから、二つのグループの単語の頻度を分析し、届いたメールがスパムかそうでないかの分類に教師あり学習の手法を使用できます。

そのため、教師あり学習と教師なし学習の相違点と類似点を考慮すると、どちらが良いのか不思議に思うかもしれません。

各手法はそれぞれ状況に合う強みを持ちますが、データサイエンスサービスの長である Martin Schmitz はきっぱりと教師あり学習と述べています。

彼は”A Human’s Guide to Machine Learning”で、「教師あり学習が適用できるなら、教師あり学習を使用しましょう」と述べています。

これは、教師なし問題の中では、どのクラスタリングが良いか測りにくいことが原因です。教師なし問題を教師あり問題にすることは、たとえ初期のデータへラベルを追加する作業に多大な労力を必要とする場合でも、最適なモデルを開発するキーとなることがよくあります。

さいごに

教師あり学習と教師なし学習の手法はデータサイエンティストにとってパワフルなツールで、一つの記事では説明できないくらいたくさんの使用法と例があります。しかし、両方をしっかりと理解することは、自分にとって何がベストなのかを見極める最初の第一歩です。

機械学習からどのようにビジネスが利益を上げるかを見たい場合は、KSK アナリティクスにお問い合わせください。そして、潜在的なユースケースを探り、ビジネスにどのようにインパクトをもたらせるか探索していきましょう。