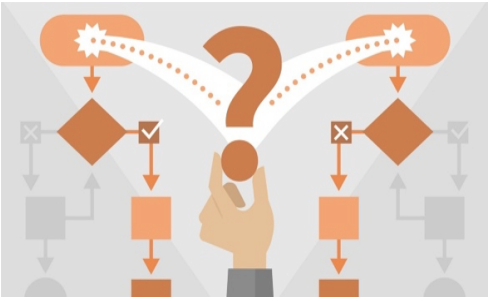
機械学習におけるバイアスについて考えてみよう

昨今、機械学習におけるバイアスへの懸念がニュースになっていますが、この話題は決して単純なものではありません。この話題に困惑している方はご安心を。この話題は複雑であるがゆえに、これほど多くの議論がなされているのです。そして、混乱の原因の一つは、「バイアス」という言葉が、必ずしも一致せず、様々な意味で使われていることにあると思われます。

今回は、機械学習における偏りについて、データサイエンスや機械学習の分野で「偏り」という言葉がどのように使われているかをご紹介します。



そもそもバイアスとは？

まずはっきりさせておかなければならないのは、データサイエンスの世界で誰かが「バイアス」と言っても、もう少し深く意味を掘り下げてみないと、その意味を正確に把握することはできないということです。というのも、「バイアス」という言葉には、データサイエンス特有の意味もあれば、一般に現実の世界で使われている意味もあります。しかも、それぞれの意味が交錯しているため、最初はわからないこともあります。

この記事では、機械学習に最も関連する「バイアス」という言葉の 3 つの異なる使い方に焦点を当てますが、世の中には、機械学習と関係するかどうかに関わらず、様々なバイアスがあることに注意してください。例えば、心理学における認知バイアスのような関連性の低いものもあります。

また、この分野ではまだ誰も気づいていないタイプのバイアスがある可能性もあり、そのバイアスはここで紹介したものよりももっと多くのレベルで作用していると思われます。科学は常に変化していますし、機械学習は（少なくともビジネスの世界での応用という点では）比較的新しい分野ですから、数年後には私たちが間違っていることになるかもしれません。

それでは、「モデルにバイアスがある」と誰かが言いそうな 3 つの主なこと、すなわち、現実のバイアス、サンプリングバイアス、モデルバイアスについて見てみましょう。

現実のバイアス

もしあなたがデータサイエンティストではなく、「バイアス」と言ったら、それはおそらく、性別、年齢、セクシュアリティ、人種、民族などのカテゴリーに基づく、社会で見られる現実世界のバイアスを意味するでしょう。このバイアスは、機械学習とは無関係に存在していますが、後述するように、明確に機械学習と関係する可能性があります。この言葉は一般的にこのように理解されているため、機械学習におけるバイアスの話を聞くと、モデルが予測によって構造的な人種差別やステレオタイプを助長している、という意味にとられる可能性が高いです。

しかし、この言葉自体は、偏ったモデルを修正しようとしている人にとっては、あまり役に立ちません。以下で説明するように、モデルが現実のバイアスを反映した予測を行っていることに気付いた場合、その理由を把握し、それを軽減したり理解したりする方法を決めるためには、もう少し深く掘り下げる必要があります。

サンプリングバイアス

データサイエンスにおける「バイアス」という言葉の一つの使い方は、「サンプリングバイアス」です。統計学の授業を受けたことがある人なら、この概念はよく知っていると思います。これは、調査しようとしている母集団（機械学習の場合は、行動を予測するモデルを構築する母集団）を正確に表さない方法でデータが収集されたことを意味します。

ここで注目すべきなのは、機械学習の世界では現実のバイアスとサンプリングバイアスは常に独立しているわけではなく、相互に影響し合うことが非常に多いということです。例えば、私が女性よりも男性を優先的に採用してきた歴史を持つ企業の人事部を担当していて、過去の採用決定に基づいてモデルを学習させた場合、そのモデルは両方の意味でバイアスを持ったものになります。つまり、学習データにはサンプリングバイアスがあり、予測には現実のバイアスが現れます。

モデルバイアス

データサイエンスで耳にするバイアスの 3 つ目のタイプは、あるカテゴリーの確率を過剰に予測したり、過小に予測したりすることで、モデルが偏った予測をしていると言われる場合です。これは偏ったデータの結果かもしれませんが、サンプリングバイアスとは異なり、サンプリングバイアスの場合はデータに偏りがあります。

ちょっとした余談ですが、アルゴリズム・バイアスの考え方について説明します。機械学習は、非常に高度で抽象的なレベルで、データセット、モデルを作成するアルゴリズム、そして予測を行うモデルの 3 つの部分から構成されていると考えることができます。ディシジョンツリーが自動的に女性よりも男性を優先して採用を決定することはありませんし、畳み込みニューラルネットワークが大学入試で白人を優先することもありません。
今後、このようなモデルの使用が増えたり、新しいタイプのモデルが開発されたりすると、モデルを作成するアルゴリズム自体にも何らかのバイアスが発見される可能性があります。しかし今のところ、モデルのバイアスという言葉は、モデルを作成したアルゴリズムではなく、学習したモデルに存在するバイアスを意味していると考えてよいでしょう。

モデルの予測について「バイアス」という言葉を使うことは、最も理解しにくいことです。なぜなら、予測の内容（例えば、製造工程を早く止めてしまうなど）に関わらず、モデルが歪んだ予測をしている場合、それが現実のバイアスとは無関係であっても、バイアスがかかっていると呼ぶ人がいるからです。一方、データサイエンスの用語でモデルが偏っているかどうかに関わらず、モデルの予測が私たちの定義や現実世界のバイアス（例えば、大学入試で特定の人種を優遇するなど）と一致する場合には、「モデルのバイアス」という言葉を使う人もいます。

機械学習のバイアスの例

それでは、これまで説明してきたバイアスの種類ごとに、実際にどのようなバイアスなのかをイメージできるような例を見てみましょう。

◆ サンプリングバイアスの一例

サンプリングバイアスは、モデルの学習や予測よりもデータが先行するため、また、統計学など他の分野でもよく取り上げられるトピックであるため、まずはサンプリングバイアスについて説明します。

例えば、私がアメリカの全国規模の小売店で働いていて、よく一緒に買われる商品に基づいて、チェックアウト時にクーポンを提供するモデルを作りたいとします。データウェアハウスからデータを取り出して見てみると、データの約 80%がボストンかニューヨークで購入されたものであることがわかりました。ボストンやニューヨークのような都市部とそうではない地方では、人々の購買習慣が異なることは直感的に理解できるでしょう。ですので、このデータをもとにしたモデルを全米に展開するとなると、問題が発生します。

今までで説明した他のタイプのバイアスとは異なり、この例のように、モデルのトレーニングを始める前にデータの問題が見つかることがよくあります。これらデータクリーニングやデータ前処理の作業で、データサイエンティストは忙殺されているように思われがちですが、実際にモデル作成のプロセスに欠かせないものなのです。ヒストグラムのような簡単なツールを使うだけで、データが一方向に偏っていて、モデルのトレーニング中に問題を起こす可能性があることがわかります。これが、次の例にあるバイアス（モデルのバイアス）の原因の 1 つと考えられます。

◆ モデルバイアスの例

純粋なモデルバイアスについては、現実のバイアスとは関係のない例を見てみましょう。例えば、あなたが化学メーカーで、ある桶が沸騰するタイミングを予測し、熱源を遮断して製品を保存できるように設計されたモデルがあるとします。トレーニングやテストではうまくいっていたのですが、実際に運用を開始してみると、火曜日に沸騰するイベントを予測する確率が 400%も高いことに気がつきました。

これは怪しいと思いますが、なぜモデルはこのような挙動を示すのでしょうか？サンプリングバイアスのせいかもしれません。モデルを学習させた沸騰イベントの 95%がたまたま火曜日に発生し、モデルは正確な予測を行うための非常に効果的な方法を学習したのかもしれません（少なくとも利用可能なデータに関しては）
しかし、他にも理由があるかもしれません。もしかしたら、火曜日に働いている人が、勘違いして他の人と違うデータを処理・入力しているのかもしれません。実際にはタンクの温度は変わっていませんが、モデルが何かおかしいと判断して、沸騰イベントを知らせるのに十分な差があるのです。

このようなバイアスに遭遇した場合、それを特定して軽減するためには、データサイエンスの手法にとらわれない考え方が必要になることがあります。

◆ 現実のバイアスの例

現実世界のバイアスの例は、ここで紹介する他の例よりも見つけやすく、多くの点で直感的に理解できます。これは、バイアスという言葉の実際の使い方が、私たちの日常的な理解と一致しているからです。言うまでもなく、この種のバイアスの例は、人事や大学入試などの分野で現実のバイアスを持つモデルが実用化された際に、時折ニュースになることがあります。

例えば、ある人事部が、オンラインシステムで応募してきた候補者を採点するモデルをトレーニングしたとします。数か月後、このシステムでは、アフリカ系アメリカ人の候補者が白人の候補者と同等の経験を持っていても、低いスコアを与えていることが判明しました。このような予測が実行されているとしたら、現実のバイアスがかかっていることになります。

前述したように、これは単なる説明的なラベルであり、モデルが現実のバイアスを生み出しているという問題を特定するだけで、その理由や解決方法を示すものではありません。モデルからこのような予測が出た場合、さまざまな潜在的な要因を調べる必要があります。サンプリングバイアスなのか？モデルが本番で取得するデータに問題があるのではないか？他の関係する因子があるのではないか？

機械学習モデルの偏りへの対応

この時点で、偏ったモデルを見つけた場合、どうすればいいのか気になっていることでしょう。前述の通り、RapidMiner を使ってモデルが何をしているかを調べる方法など、このトピックについては今後改めて説明しますが、もしモデルに偏りがあることがわかった場合、その原因を探るために今できることがいくつかあります。

1 つ目は、すでにお話ししたように、データを見ることです。サンプリングバイアスは、モデルが偏ってしまう最も一般的な理由の 1 つです。モデルが偏っている、という情報を持ってデータを見直すと、元のデータに気づかなかったパターンが見つかるかもしれません。ヒストグラムを見るだけでなく、特徴量の重みを計算して、どの列がモデルの予測に最も影響を与えているかを確認することもできます。

データを調べることは、モデルの説明可能性に向けた一般的な取り組みのほんの一部と考えることができます。これは、モデルが何をしているのか、なぜしているのかを理解するのに役立ちます。
モデルがどのような意思決定を行っているかによっては、外部要因も考慮する必要があります。外部要因とは、モデルがおかしな予測をする原因となっているもので、前述の火曜日の沸騰現象に関する要因のように、モデルが予測を行う前にデータがどこから来ているのか、どのように処理されているのかを含みます。

バイアスの相互作用

すでに述べたように、これらの異なる種類のバイアスは相互に影響し合うことが多く、これがモデルバイアスをどのように見れば理解できるのか、混乱の原因の一つとなっています。

これは重要なポイントです。なぜなら、問題が発生したときに、どのような種類のモデルバイアスを考えるのかという定義や議論に惑わされないようにするためです。問題があることを特定することが最初のステップですが、偏った予測をしているモデルがある場合、その問題を何と呼べばいいのかを理解するために多くの時間を費やす必要はありません。

まとめと今後の展望

この時点では、機械学習のバイアスに最も関連性のあるタイプ、それらの相互作用、そして偏った予測をしているモデルがある場合に問題を発見する方法について、理解を深めてください。

最終的には、バイアスを軽減するための取り組みは、データとモデルの動きを見ている人間が解決策を見出すことになります。なぜなら、バイアスは最終的には社会的な問題であり、モデルはトレーニング中であれ本番運用中であれ、見たデータを反映しているだけだからです。そのため、データリテラシーが高く、バイアスを認識している人間が、モデルの動作を見て、今後の改善方法を考える必要があるのです。

つまり、バイアスはサンプリングやモデリングの技術的な問題だけではなく、純粋な技術的解決策だけでは対処できない現実の問題なのです。意識を高め、説明可能なモデルを構築してこそ、何が起きているのかを理解し、発見したバイアスに対処するための計画を立てることができるのです。