MO444 - Projeto Final - Primeiro resultado

Alan M Ganem
RA 178777

Novembro, 2021

1 Motivação

Modelos de machine learning em sua grande maioria exigem uma condição muito específica para que relações de causalidade sejam obtidas a partir destes. Essa hipótese é a de Independência entre variáveis, já que a maioria dos modelos de machine learning não assumem explicitamente uma estrutura de causalidade entre variáveis (com excessão de modelos como Redes Baeysianas). Desse modo, estes modelos possuem uma boa capacidade de prever, contudo, falham em definir o impacto de intervenções específicas. Isso se deve ao fato de que, uma vez que as variáveis não são independentes, a alteração de uma variável que se deseja identificar o efeito causal (recorte racial, por exemplo) não é independente das demais variáveis no dataset, e desse modo, alterar o seu valor, a fim de fazer uma análise contrafactual ("qual seria o valor de y caso apenas x fosse diferente, mantendo todas as outras variáveis constantes") exigiria que o valor de outras variáveis do dataset também fosse alteradas, uma vez que elas não são independentes.

Uma estratégia utilizada e considerada o padrão ouro para identificação de efeito causal é o ensaio randomizado controlado (Randomized Controlled Trials - **RCTs** em ingles), que consiste em aleatorizar a atribuição da variável que desejamos identificar o efeito causal (comumente chamada de **tratamento**). A aleatorização garante que a atribuição do nosso **tratamento** é independente das demais variáveis no nosso dateset, garantindo a validade da hipótese de independencia, necessária para garantir que a estimativa do efeito do nosso tratamento é não viesada.

Infelizmente, em muitos casos, aleatorização é impraticável e até mesmo impossível. No caso deste trabalho, como queremos identificar o efeito causal do recorte racial e de gênero no salário médio de indivíduos, é impossível realizar um ensaio aleatorizado, a fim de "atribuir" aleatoriamente as variáveis de raça e gênero para indivíduos. Nesse contexto, torna-se necessária a utilização de métodos de inferência causal para dados **observacionais** (quando não há aleatorização na atribuição do tratamento).

2 Descrição da técnica utlizada

Neste trabalho, será utilizado o conceito de machine learning ortogonal, que como o nome sugere, visa ortogonalizar (ou tornar independente) conjuntos de variáves, a fim de que seja satisfeita a hipótese de independência entre variáveis.

Essa estratégia consiste em utilizar o teorema de Frisch–Waugh–Lovell para um modelo genêrico (não necessariamente uma regressão linear). A intuição por trás da técnica é remover o viés na estimativa do efeito dos tratamentos fazendo regressões sobre resíduos. Em teoria, quanto fazemos a regressão de uma variável T em um conjunto de variáveis X, temos que o resíduo de T X é a parte da variância/informação de T que não é explicada por X, satisfazendo a condição de independencia entre X e T (X$\perp$T).

Para atingir esse objetivo, foi desenvolvida um estimador customizado chamado ResidualEstimator, com o intuito de implementar a estratégia de ortogonalização através da regressão de resíduos.

3 Primeiro resultado

Para inferir causalidade é primeiro necessária a definição de uma possível estrtura causal para os seus dados. Existem técnicas para determinar essas conexões utilizando dados, mas para esse trabalho, utilizaremos uma estrutura causal proposta pelo autor. Essa estrutura pode ser encontrada em 1.

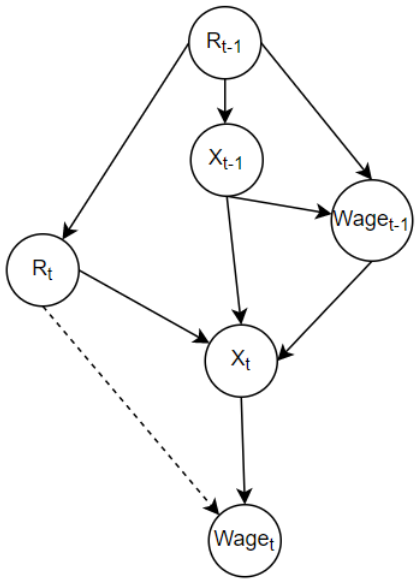

Figure 1: representação gráfica da estrutura causal proposta pelo autor. Os sub-indices t e t-1 representam a geração do indivíduo. As variáveis X são variáveis confundidoras/mediadoras, tais como maior nível de educação formal atingido, anos de educação, unidade federativa, se reside em região urbana ou rural... A variável R diz respeito ao trtamento (raça e/ou sexo) e a variável Wage é o salário. Setas pontilhadas dizem respeito ao mecanismo causal que estamos tentando inferir o efeito

o nosso intuito é identificar o efeito $R_t \rightarrow Wage_t$, isso é, o efeito direto da nossa variável sensível R_t em $Wage_t$ desconsiderando o efeito indireto, isso é, o efeito representado pelo caminho $R_t \rightarrow X_t \rightarrow Wage_t$.

Além disso, como a base de dados da PNADc possuem muitas variáveis (mais de 200), as variáveis de controle X_t foram escolhidas através de conhecimento prévio do autor e foram posteriormente filtradas utilizando uma técnica

de análise multivariada (RandomForest + PermutationImportance) para seleção de variáveis importantes. É possível ver um exemplar dos gráficos gerados nesa análise em 2.

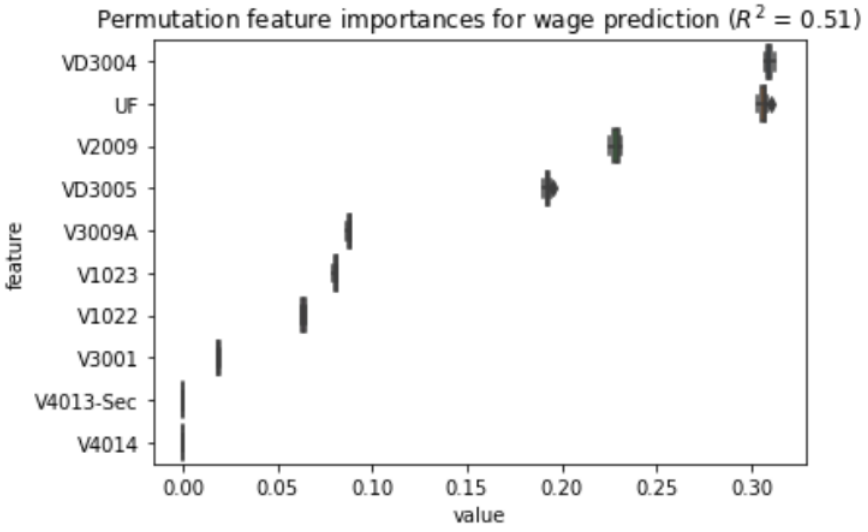

Figure 2: A importância de permutação de features é realizada fazendo permutações aleatórias de cada feature e observando o impacto desssa permutação na métrica de validação

Na seção "Finding Confounders" do notebook associado a esse trabalho, estão as variáveis utilizadas na análises de associação, bem como os seus valores de associação tanto com o salário dos indivíduos quanto com as variáveis sensíveis "sexo" e "raça".

4 Validação de resultados em inferência contrafactual

Um dos maiores problemas na validação de modelos causais e análise contrafactual ("qual seria o salário dessa pessoa caso mantivessemos todas as demais variáveis constantes e mudassemos apenas o seu sexo/raça ?") é a não possibilidade de observar esse resultado contrafactual (exceto em dados simulados) nos datasets, o que torna inviável tecnicas de validação cruzada comumente utilizadas na validação de modelos de Machine Learning.

Por esse motivo, existe uma grande preocupação com o formalismo das técnicas propostas recentemente na área, como por exemplo a de double/debiased machine learning, que tenta garantir a convergência assintótica do estimador em questão, quando exposto a muitos dados. Essas garantias se tornam importantes

em um cenário em que o efeito real de um tratamento não pode ser medido diretamente através dos dados.

Entretanto, com os resultados obtidos (3), mostra-se necessário uma revisão na tecnica utilizada.

```
V2007-Bin  V2010-Bin
Feminino   Branca ou Amarela           8.673617e-19
           Preta, Parda ou Indígena   -5.261189e-03
Masculino  Branca ou Amarela           0.000000e+00
           Preta, Parda ou Indígena    1.866967e-03
Name: race_effect, dtype: float64
```

Figure 3: O impacto direto estimado do recorte racial no salário. O caso base é o valor "Branco ou Amarelo", e o valor contrafactual é obtindo fazendo uma "intervenção" nesse valor para "Preta, Parda ou Indígena". Pode-se observar que o impacto é praticamente insignificante, o que mostra a necessidade de uma análise mais aprofundada

5 Próximos passos

Na próxima entrega, será testada uma outra técnica de ortogonalização, utilizando o matching de unidades semelhantes em relação às covariantes mais significativas. Essa técnica permite encontrar unidades do experimento semelhantes entre sí, exceto pelas variáveis sensíveis. Essa técnica utilizará, além das técnicas já utilizadas, o algoritmo de KNeighbors para encontar essas unidades no espaço de representação aprendido pelo algoritmo supervisionado utilizado (nesse caso, RandomForests).

Mais informações sobre o código podem ser encontradas neste repositório: `https://github.com/AlanGanem/MO444/tree/master/trabalho_final`