Análise de Redes Sociais

Justiça no Aprendizado de Máquina

Prof. Luciano Antonio Digiampietri

Justiça Algorítmica (*Algorithmic Fairness*)

Justiça Algorítmica (*Algorithmic Fairness*)

A justiça algorítmica tipicamente refere-se às tentativas de corrigir o viés algorítmico em processos de decisão automatizados. (MITCHELL et al., 2021)

Justiça Algorítmica (*Algorithmic Fairness*)

Não existe uma definição objetiva e compartilhada por toda a comunidade da computação sobre o que seria um algoritmo ético, justo e imparcial.

Justiça Algorítmica (*Algorithmic Fairness*)

Por exemplo:

Justiça Algorítmica (*Algorithmic Fairness*)

Por exemplo:

Alguém pode dizer que o classificador "mais justo" é aquele que possui maior *acurácia*.

Justiça Algorítmica (*Algorithmic Fairness*)

Por exemplo:

Alguém pode dizer que o classificador "mais justo" é aquele que possui maior *acurácia*.

Ou considerar que é aquele em que a *distribuição das classes* no conjunto de testes é mais semelhante com a distribuição do conjunto de treinamento.

Justiça Algorítmica (*Algorithmic Fairness*)

Por exemplo:

Alguém pode dizer que o classificador "mais justo" é aquele que possui maior *acurácia*.

Ou considerar que é aquele em que a *distribuição das classes* no conjunto de testes é mais semelhante com a distribuição do conjunto de treinamento.

Ou considerar que é aquele que não usa nenhum *atributo sensível* em seu modelo.

Justiça Algorítmica (*Algorithmic Fairness*)

Alguns dirão que o algoritmo mais justo é o que usa a maior quantidade de dados/atributos.

Justiça Algorítmica (*Algorithmic Fairness*)

Alguns dirão que o algoritmo mais justo é o que usa a maior quantidade de dados/atributos.
Outros dirão que os modelos não deveriam considerar sexo, etnia, orientação sexual, deficiência

Justiça Algorítmica (*Algorithmic Fairness*)

Alguns dirão que o algoritmo mais justo é o que usa a maior quantidade de dados/atributos.
Outros dirão que os modelos não deveriam considerar sexo, etnia, orientação sexual, deficiência, endereço, sobrenome ...

Justiça Algorítmica (*Algorithmic Fairness*)

Alguns dirão que o algoritmo mais justo é o que usa a maior quantidade de dados/atributos.
Outros dirão que os modelos não deveriam considerar sexo, etnia, orientação sexual, deficiência, endereço, sobrenome ...
Mas um algoritmo de auxílio ao diagnóstico médico provavelmente precisará saber o sexo, talvez a etnia e outras informações sensíveis.

Justiça Algorítmica (*Algorithmic Fairness*)

Existe uma questão fundamental ao tentarmos projetar algoritmos mais justos: há ao menos três "*mundos diferentes*" (MITCHELL et al., 2021):

Justiça Algorítmica (*Algorithmic Fairness*)

Existe uma questão fundamental ao tentarmos projetar algoritmos mais justos: há ao menos três "*mundos diferentes*" (MITCHELL et al., 2021):

[1] O mundo de acordo com o dados.

Justiça Algorítmica (*Algorithmic Fairness*)

Existe uma questão fundamental ao tentarmos projetar algoritmos mais justos: há ao menos três "*mundos diferentes*" (MITCHELL et al., 2021):

[1] O mundo de acordo com o dados.

[2] O mundo como ele é.

Justiça Algorítmica (*Algorithmic Fairness*)

Existe uma questão fundamental ao tentarmos projetar algoritmos mais justos: há ao menos três "*mundos diferentes*" (MITCHELL et al., 2021):

[1] O mundo de acordo com o dados.

[2] O mundo como ele é.

[3] O mundo como ele "deveria/poderia" ser.

Justiça Algorítmica (*Algorithmic Fairness*)

Os modelos costumam ser construídos de acordo com os dados (por razões óbvias) ...

Justiça Algorítmica (*Algorithmic Fairness*)

Os modelos costumam ser construídos de acordo com os dados (por razões óbvias) ...

gerados/otimizados de acordo com alguma medida específica de 'desempenho' ...

Justiça Algorítmica (*Algorithmic Fairness*)

Os modelos costumam ser construídos de acordo com os dados (por razões óbvias) ...

gerados/otimizados de acordo com alguma medida específica de 'desempenho' ...

processo que, muitas vezes, insere novos vieses.

Fontes de Viés

Suresh e Guttag (2021) listam sete fontes de danos:

Fontes de Viés

Suresh e Guttag (2021) listam sete fontes de danos:

1. **Viés Histórico**

Fontes de Viés

Suresh e Guttag (2021) listam sete fontes de danos:

1. **Viés Histórico**
2. **Viés da Representação**

Fontes de Viés

Suresh e Guttag (2021) listam sete fontes de danos:

1. **Viés Histórico**
2. **Viés da Representação**
3. **Viés de Medição**

Fontes de Viés

Suresh e Guttag (2021) listam sete fontes de danos:

1. **Viés Histórico**
2. **Viés da Representação**
3. **Viés de Medição**
4. **Viés da Agregação**

Fontes de Viés

Suresh e Guttag (2021) listam sete fontes de danos:

1. **Viés Histórico**
2. **Viés da Representação**
3. **Viés de Medição**
4. **Viés da Agregação**
5. **Viés do Aprendizado**

Fontes de Viés

Suresh e Guttag (2021) listam sete fontes de danos:

1. **Viés Histórico**
2. **Viés da Representação**
3. **Viés de Medição**
4. **Viés da Agregação**
5. **Viés do Aprendizado**
6. **Viés da Avaliação**

Fontes de Viés

Suresh e Guttag (2021) listam sete fontes de danos:

1. **Viés Histórico**
2. **Viés da Representação**
3. **Viés de Medição**
4. **Viés da Agregação**
5. **Viés do Aprendizado**
6. **Viés da Avaliação**
7. **Viés de Implantação**

1. Viés Histórico

1. Viés Histórico

“Surge quando o mundo como é ou foi leva a um modelo que produz resultados prejudiciais. Tal sistema, mesmo que reflita o mundo com precisão, ainda pode causar danos a uma população.”

“Pesquisas recentes mostraram que *embeddings* de palavras, que são aprendidos a partir de grandes corpora de texto refletem preconceitos humanos.”

1. Viés Histórico

“Surge quando o mundo como é ou foi leva a um modelo que produz resultados prejudiciais. Tal sistema, mesmo que reflita o mundo com precisão, ainda pode causar danos a uma população.”

“Pesquisas recentes mostraram que *embeddings* de palavras, que são aprendidos a partir de grandes corpora de texto refletem preconceitos humanos.” Por exemplo, associando palavras relacionadas a certas etnias com palavras de baixo calão, ou ‘aprendendo’ que algumas profissões são femininas.

2. Viés da Representação

2. Viés da Representação

“A amostra de desenvolvimento sub-representa alguma parte da população e não consegue generalizar bem para um subconjunto da população.

Se a seleção da população alvo: (i) não refletir a população de uso; (ii) contém grupos sub-representados; (iii) método de amostragem for limitado ou irregular.”

2. Viés da Representação

“A amostra de desenvolvimento sub-representa alguma parte da população e não consegue generalizar bem para um subconjunto da população. Se a seleção da população alvo: (i) não refletir a população de uso; (ii) contém grupos sub-representados; (iii) método de amostragem for limitado ou irregular.”

Um classificador para reconhecimento facial que foi treinado apenas com imagens de pessoas brancas e tem um desempenho muito inferior no reconhecimento de outras pessoas.

3. Viés de Medição

3. Viés de Medição

“O viés de medição ocorre ao escolher, coletar ou computar atributos e rótulos a serem usados em um problema de previsão.”

“Diferentes origens: (i) O *proxy* é uma simplificação exagerada de algo mais complexo; (ii) O método de medição varia entre os grupos; (iii) A precisão da medição varia entre os grupos.”

3. Viés de Medição

“O viés de medição ocorre ao escolher, coletar ou computar atributos e rótulos a serem usados em um problema de previsão.”

“Diferentes origens: (i) O *proxy* é uma simplificação exagerada de algo mais complexo; (ii) O método de medição varia entre os grupos; (iii) A precisão da medição varia entre os grupos.”

(i) Aceita-se um aluno pela média ponderada estimada; (ii) Avalia-se programadores pela média de *erros*, medida de forma diferente; (iii) Um diagnóstico é baseado na sensação de dor do paciente (varia de acordo com gênero e etnia).

4. Viés da Agregação

4. Viés da Agregação

“Surge quando um modelo *one-size-fits-all* é usado para dados nos quais existem grupos subjacentes que devem ser considerados de maneira diferente.”

“O viés de agregação pode levar a um modelo que não é ideal para nenhum grupo ou a um modelo adequado à população dominante (por exemplo, se também houver viés de representação)”

4. Viés da Agregação

“Surge quando um modelo *one-size-fits-all* é usado para dados nos quais existem grupos subjacentes que devem ser considerados de maneira diferente.”

“O viés de agregação pode levar a um modelo que não é ideal para nenhum grupo ou a um modelo adequado à população dominante (por exemplo, se também houver viés de representação)”

“Gírias ou emojis têm diferentes significados em diferentes grupos e palavras ou frases que podem transmitir agressão em um contexto, mas são letras de um rapper local.”

5. Viés do Aprendizado

5. Viés do Aprendizado

“O viés de aprendizado surge quando as escolhas de modelagem amplificam as disparidades de desempenho em diferentes exemplos nos dados”(HOOKER, 2021)

“Ao se priorizar a acurácia geral pode-se levar a uma diminuição da revocação da classe minoritária.”

5. Viés do Aprendizado

“O viés de aprendizado surge quando as escolhas de modelagem amplificam as disparidades de desempenho em diferentes exemplos nos dados”(HOOKER, 2021)

“Ao se priorizar a acurácia geral pode-se levar a uma diminuição da revocação da classe minoritária.”

Ao se escolher, por exemplo, uma árvore de decisão menor (um modelo mais ‘inteligível’), há mais chance do modelo focar nas características discriminantes mais frequentes, por exemplo, dos elementos da classe majoritária.

6. Viés da Avaliação

6. Viés da Avaliação

“Ocorre quando os dados de referência/teste usados não representam a população de uso.”

“O viés de avaliação surge devido ao desejo de comparar quantitativamente os modelos entre si, podendo levar ao *overfitting* a um *benchmark*. Isso é especialmente problemático se o *benchmark* sofre de viés histórico, de representação ou de medição.”

6. Viés da Avaliação

“Ocorre quando os dados de referência/teste usados não representam a população de uso.”

“O viés de avaliação surge devido ao desejo de comparar quantitativamente os modelos entre si, podendo levar ao *overfitting* a um *benchmark*. Isso é especialmente problemático se o *benchmark* sofre de viés histórico, de representação ou de medição.”

“Na literatura relatou-se um desempenho bastante inferior de uma ferramenta comercial de análise facial em imagens de mulheres de pele escura.”

7. Viés de Implantação

7. Viés de Implantação

“Ocorre quando há uma incompatibilidade entre o problema que um modelo pretende resolver e a maneira como ele é realmente usado.”

“Um sistema que é construído e avaliado como se fosse totalmente autônomo, enquanto na realidade ele opera em um sistema sociotécnico complexo moderado por estruturas institucionais e tomadores de decisão humanos”

7. Viés de Implantação

“Ocorre quando há uma incompatibilidade entre o problema que um modelo pretende resolver e a maneira como ele é realmente usado.”

“Um sistema que é construído e avaliado como se fosse totalmente autônomo, enquanto na realidade ele opera em um sistema sociotécnico complexo moderado por estruturas institucionais e tomadores de decisão humanos”

O uso de reconhecimento facial, planejado para ajudar na identificação de potenciais foragidos, sendo usado como única ferramenta para prisão.

Referências

1 **Shira Mitchell, Eric Potash, Solon Barocas, Alexander D'Amour, Kristian Lum. Algorithmic Fairness: Choices, Assumptions, and Definitions. Annual Review of Statistics and Its Application, 8:1, 141-163, 2021.**
https://doi.org/10.1146/annurev-statistics-042720-125902

2 **Harini Suresh and John Guttag. A Framework for Understanding Sources of Harm throughout the Machine Learning Life Cycle. In Equity and Access in Algorithms, Mechanisms, and Optimization (EAAMO '21), October 5–9, 2021.**
https://dl.acm.org/doi/fullHtml/10.1145/3465416.3483305

3 **Sara Hooker. 2021. Moving beyond 'algorithmic bias is a data problem'. Patterns 2, 4 (2021), 100241.**

Análise de Redes Sociais

Justiça no Aprendizado de Máquina

Prof. Luciano Antonio Digiampietri