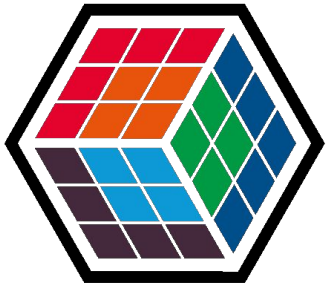


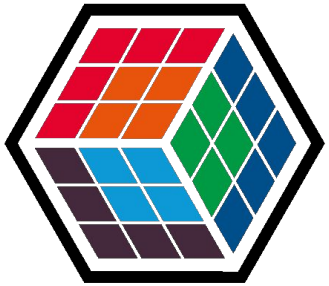
O Impacto de diferentes tipos de vieses em modelos de Aprendizado de Máquina

Trilha – Machine Learning

Agenda

THE DEVELOPER'S CONFERENCE

- Análise de Dados
- Aprendizado de Máquina
  - Viés



Análise de Dados

Análise de Dados

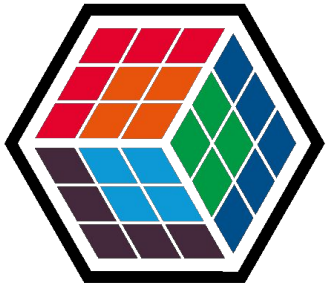
THE DEVELOPER'S CONFERENCE

"Ciência de examinar dados brutos com o objetivo de encontrar padrões e tirar conclusões sobre essa informação, aplicando um processo algorítmico ou mecânico para obter informações."

Análise de Dados

THE DEVELOPER'S CONFERENCE

- Analisar características
- Descoberta de padrões e tendências
- Compreender o processo que gerou os dados
- Extrair Informações importantes



Aprendizado de Máquina

Aplicações

THE DEVELOPER'S CONFERENCE

- Detecção de fraude
- Análise de risco de crédito
- Filtragem de phishing
- Previsão de falha de equipamentos
- Análise de currículos
- Oferta de produtos
- Auxílio em análises jurídicas
- ...

Internet of Shit Retweeted

Computer Facts
@computerfact

concerned parent: if all your friends jumped off a bridge would you follow them?
machine learning algorithm: yes.

2:20 PM · Mar 15, 2018

Garbage in, Garbage out

Garbage in, Garbage out

THE DEVELOPER'S CONFERENCE

Dados

Garbage in, Garbage out

THE DEVELOPER'S CONFERENCE

Dados

Modelo

Garbage in, Garbage out

THE DEVELOPER'S CONFERENCE

Dados

Modelo

Resultado

Garbage in, Garbage out

THE DEVELOPER'S CONFERENCE

Dados

Modelo

Resultado

Dados

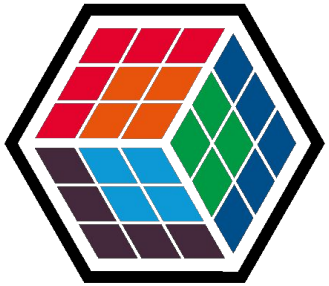
Modelo

Resultado

“

THE DEVELOPER'S CONFERENCE

Sua análise só é tão boa quanto forem os seus dados



Viés em Aprendizado de Máquina

Viés Indutivo

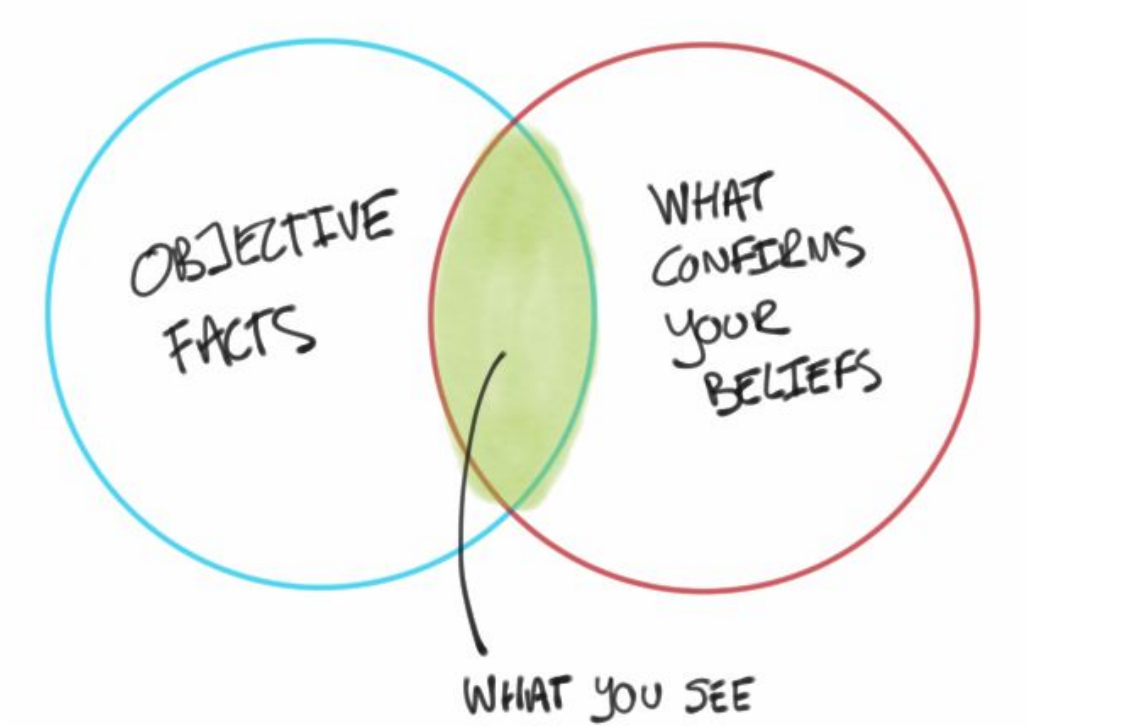
THE DEVELOPER'S CONFERENCE

- Hipótese

Viés Indutivo

THE DEVELOPER'S CONFERENCE

- Hipótese
- Representação utilizada define a preferência (viés) de representação do algoritmo


OBJECTIVE FACTS
WHAT CONFIRMS YOUR BELIEFS
WHAT YOU SEE

Exemplo

THE DEVELOPER'S CONFERENCE

Extreme *she*	**Extreme *he***
1. homemaker	1. maestro
2. nurse	2. skipper
3. receptionist	3. protege
4. librarian	4. philosopher
5. socialite	5. captain
6. hairdresser	6. architect
7. nanny	7. financier
8. bookkeeper	8. warrior
9. stylist	9. broadcaster
10. housekeeper	10. magician

Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings (2016)

Exemplo

THE DEVELOPER'S CONFERENCE

Extreme 'he'

1. escritor → poeta
2. cantor→ músico
3. pintor → escultor
4. secretario → secretário
5. ator → actor
6. historiador → poeta
7. arquiteto→ arquitect
8. fotógrafo → cineasta
9. advogado → empresário
10. juiz → juíz

Extreme 'she'

1. atriz → actriz
2. escritora → poetisa
3. pesquisadora → bióloga
4. sindica → medicamen
5. diretora → coordenadora
6. matemática → astronomia
7. historiadora → pesquisadora
8. garçonete → stripper
9. secretaria → secretária
10. enfermeira → psicóloga

Is there Gender bias and stereotype in Portuguese Word Embeddings? (2018)

Exemplo

THE DEVELOPER'S CONFERENCE

Extreme 'he'

1. escritor → poeta
2. cantor→ músico
3. pintor → escultor
4. secretario → secretário
5. ator → actor
6. historiador → poeta
7. arquiteto→ arquitect
8. fotógrafo → cineasta
9. advogado → empresário
10. juiz → juíz

Extreme 'she'

1. atriz → actriz
2. escritora → poetisa
3. pesquisadora → bióloga
4. sindica → medicamen
5. diretora → coordenadora
6. matemática → astronomia
7. historiadora → pesquisadora
8. garçonete → stripper
9. secretaria → secretária
10. enfermeira → psicóloga

Is there Gender bias and stereotype in Portuguese Word Embeddings? (2018)

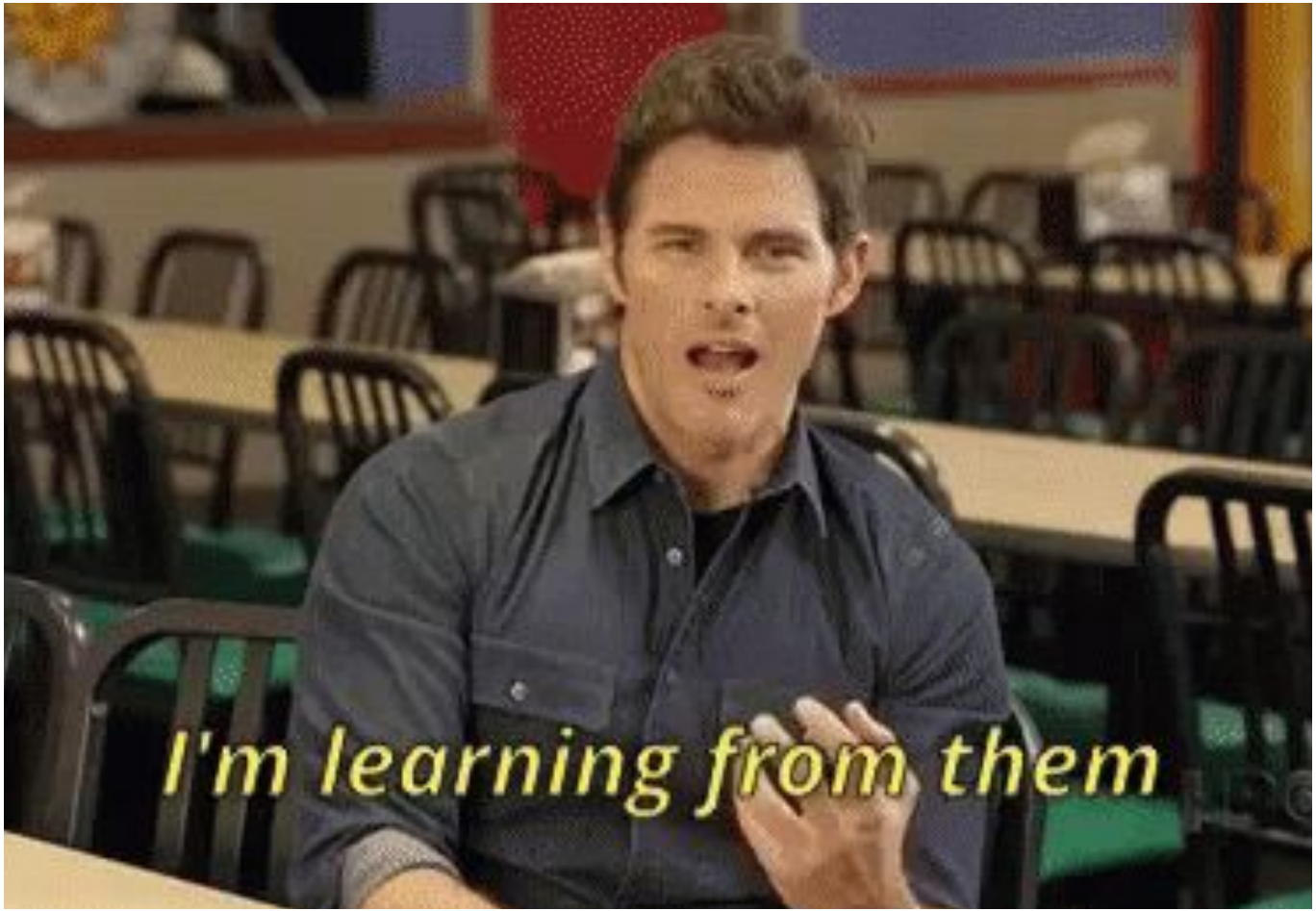
Exemplo

THE DEVELOPER'S CONFERENCE

"Google Photos confundia pessoas negras com macacos"

"Ferramenta utilizada pela justiça americana tende a ser mais rigoroso com negros do que com brancos"

"Algoritmos do Uber pagam mais para motoristas do sexo masculino"



THE DEVELOPER'S CONFERENCE

“

THE DEVELOPER'S CONFERENCE

Aquilo que não se pode medir, não se pode melhorar

William Thompson

“

THE DEVELOPER'S CONFERENCE

Remoção de bias produz um impacto negativo na qualidade do modelo.

Kamishima, T. et al. (2012)
Zliobaite, I. (2015)

THE DEVELOPER'S CONFERENCE

O que escolher?

Checklist

THE DEVELOPER'S CONFERENCE

- Já listamos como essa tecnologia pode ser atacada ou abusada?
- Testamos nossos dados de treinamento para garantir que sejam justos e representativos?
- Estudamos e compreendemos possíveis fontes de viés em nossos dados?
- Nossa equipe reflete a diversidade de opiniões, origens e tipos de pensamento?
- Que tipo de consentimento do usuário precisamos coletar para usar os dados?
- Temos um mecanismo para coletar o consentimento dos usuários?

Ethics and Data Science (2018)

Checklist

- Já explicamos claramente o que os usuários estão consentindo?
- Temos um mecanismo de reparação se as pessoas forem prejudicadas pelos resultados?
- Podemos desligar este software em produção se ele está se comportando mal?
- Testamos a imparcialidade em relação a diferentes grupos de usuários?
- Nós testamos taxas de erro diferentes entre diferentes grupos de usuários?
- Testamos e monitoramos o desvio do modelo para garantir que nosso software permaneça justo ao longo do tempo?

Ethics and Data Science (2018)

5C's

THE DEVELOPER'S CONFERENCE

Clareza

Clareza na obtenção do consentimento

Consentimento

Fonte de Dados e distribuição

Consistência e Confiabilidade

Manter o que foi acordado

Controle e Transparência

Transparência sobre a manipulação dos dados

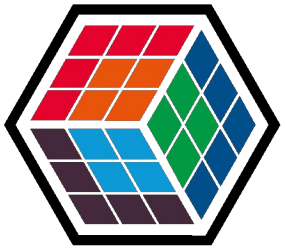
Consequências

Resultados indesejados gerados

Ethics and Data Science (2018)

THE DEVELOPER'S CONFERENCE

- Algoritmos não são a prova de falhas
- Não confiar cegamente
- Não utilizar algoritmos como desculpa



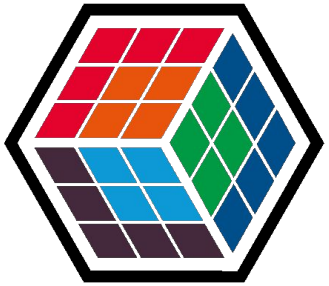
THE DEVELOPER'S CONFERENCE

Trilha – Machine Learning

Brenda Salenave Santana

bssantana@inf.ufrgs.br

@brendasalenave



THE DEVELOPER'S CONFERENCE