
Statistic
Visualisation
Pattern recognit.
DATA science
Data Mining
Machine Learning
Neurocomputing
Database & Data procss

Tema: 20 lições sobre preconceito contra sistemas de aprendizado de máquina

Aluno: Thiago Aquino(tas4)

Objetivo

Apresentar um pouco dos problemas recorrentes em aprendizagem de máquinas devidos ao enviesamento da base da dados.

Introdução

O uso de sistemas com aprendizagem de máquinas está gerando um novo ecossistema repleto de técnicas e infraestrutura, porém o aprendizado sobre sua plena capacidade, ainda está no começo. Esse novo ecossistema é muito interessante, se observarmos o que é possível ser feito com isso. No entanto, existem problemas realmente preocupantes, como formas de preconceito, estereotipagem e determinação injusta, entre outros.

Lições

→ Estudos sobre “bias” cresce exponencialmente nos últimos 3 anos.

→ “Bias” são diferentes coisas de diferentes grupos.

→ Nos termos simples, podemos dizer que “bias” é como um desvio que produz danos.

→ Existem técnicas para resolver o problema de “bias”.

→ Não existe bala de prata para resolver “bias” em ML.

Lições

→ “Bias” causa dois tipos de danos

→ Danos de Alocação, quando o sistema aloca ou retém determinados grupos como uma oportunidade de recurso.

→ Danos de Representação, quando os sistemas reforçam a subordinação de certos grupos.

→ Cada decisão de design tem consequências e implicações sociais poderosas.

→ O conjunto de dados não reflete apenas na cultura, mas também na hierarquia do mundo em que foram criados

O que é Bias?

→ Uma linha diagonal;

→ Preconceito indevido;

→ Diferença sistemática entre amostra e população

→ Underfitting VS Overfitting

→ Julgamento baseado em noções de preconceitos em oposição a avaliação imparcial.

→ É um desvio que produz algum tipo de dano

De onde vem o Bias?

Do conjunto de treinamento!!!

Tipos de Danos

→ Estereótipos: Se relacionarmos com a nossa vida, o estereótipo e nosso viés também são inevitáveis, não porque somos ruins ou julgadores, mas porque o fazemos para classificar as coisas mais rapidamente, mas deveríamos evitá-lo e isso é difícil porque já fomos treinados a ser assim muito parecido com o modo como o aprendizado de máquina funciona.

→ Reconhecimento: Quando se tem um conjunto de treinamento com apenas um tipo de dado, ex: só mulheres, após feito todo o processo de treinamento, ao chegar na validação e continuar apresentando dados semelhantes, o sistema irá validar corretamente, porém quando chegar na fase de testes e apresentar dados diferentes, ex: homens, aos que foram usados para treinar e validação, os mesmos não serão reconhecidos pelo sistema.

→ Difamação: É quando as pessoas atribuem rótulos culturalmente ofensivos e inadequados a algo ou a alguém. Devido ao conjunto de treinamento está com essa tendência, o sistema também se comportará dessa mesma maneira.

→ Subrepresentação: Quando não se é possível representar completamente, ou se representa parcialmente algo, mesmo após ter passado pelo processo de treinamento. Uma pesquisa de imagens de "CEOs" resultou em apenas uma mulher como CEO na parte mais inferior da página. A maioria era do sexo masculino branco.

Exemplos


jews should
jews should be wiped out
jews should leave israel
jews should
jews should get over the holocaust
jews should go back to poland
jews should apologize for killing jesus
jews should all die
jews should be perfected
jews should not have a state

Como lidar com esses problemas?

→ Testar o sistema com diferentes populações;

→ Rastrear o ciclo de vida do conjunto de dados de treinamento, para checar os possíveis desvios demográficos;

→ Trabalhar com pessoas interdisciplinares;

→ Pensar mais na ética da classificação

Perguntas

Quem vai se beneficiar do sistema que estamos construindo?

E quem pode ser prejudicado?


Statistic
Visualisation
Pattern recognit.
Data Mining
Obrigado!
Machine Learning
Neurocomputing
Database & Data procss