Métricas para *fairness* em ciência de dados

DCC
DEPARTAMENTO DE CIÊNCIA DA COMPUTAÇÃO

Gabriel Lucas Silva Machado

gabriel.lucas@dcc.ufmg.br

Orientador: Mário Sérgio Alvim

UFMG

INTRODUÇÃO

- Discriminações algorítmicas são tratamentos feitos por algoritmos que desfavorecem alguém devido a uma característica “protegida” (raça, gênero, deficiência, etc).
    - Discriminação direta: a característica “protegida” afeta diretamente o resultado.
    - Discriminação indireta: resultados são afetados por atributos correlacionadas com características “protegidas”.
- A discriminação algorítimica ocorre devido à existência de discriminação em *datasets,* que são utilizados nas fases de treinamento. Dessa forma o algoritmo aprende a ser discriminatório pelos próprios dados.

MOTIVAÇÃO

- *Data mining* e *machine learning* já estão sendo usados em ferramentas que afetam diretamente a vida das pessoas.
    - Exemplo: análise de CV, *credit scoring.*

OBJETIVOS

- Análise individual das principais métricas que compõem o estado da arte.
    - Fundamentos por trás da métrica.
    - Intuições sobre o funcionamento.
    - Contextualização dos usos nas áreas de *data mining* e *machine learning.*
    - Principais vantagens e desvantagens.

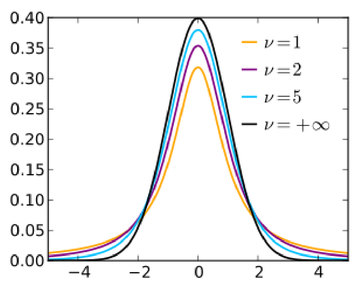
TIPOS DE MÉTRICAS

Tipo de Métrica	O que mede?	Tipo de discriminação
Testes estatísticos	Presença ou não de discriminação	Discriminação indireta
Medidas absolutas	Gravidade da discriminação	Discriminação indireta
Medidas condicionais	Gravidade da discriminação	Discriminação indireta
Medidas situacionais	Propagação da discriminação	Discriminação direta e indireta

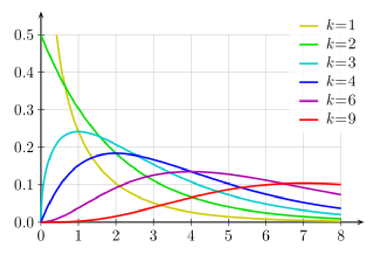
Testes estatísticos

- Verificam a existência de discriminação indireta em modelos empíricos, comparando-os com um modelo teórico em que não existe discriminação.

DISTRIBUIÇÃO t



DISTRIBUIÇÃO χ^2



DISTRIBUIÇÃO NORMAL (0,1)

Medidas absolutas

- Quantificam a discriminação existente partindo da premissa de que qualquer tratamento diferente entre grupos favorecidos e desfavorecidos é discriminação.

Medidas condicionais

- Quantificam a discriminação causada somente por atributos “não protegidos” correlacionados com as características “protegidas”.

Medidas situacionais

- Detectam o quanto uma determinada discriminação se espalha por um *dataset*.

RESULTADOS

- Um documento que contém uma análise de 17 métricas, como descrita na seção “objetivos”. Dessas métricas, 5 são classificadas como testes estatísticos, 8 como medidas absolutas, 2 como medidas condicionais e 2 como medidas situacionais.