Gabriel Lucas Silva Machado

Métricas para *fairness* em ciência de dados

Belo Horizonte

2018/1

Gabriel Lucas Silva Machado

Métricas para *fairness* em ciência de dados

Relatório final da disciplina Projeto Orientado em Computação I do curso de Bacharelado em Ciência da Computação da UFMG

Univerisdade Federal de Minas Gerais - UFMG

Instituto de Ciências Exatas

Departamento de Ciência da Computação

Orientador: Mário Sérgio Alvim

Belo Horizonte

2018/1

Lista de tabelas

Resumo

O crescimento do uso de tomadores de decisões orientados a dados é cada vez mais perceptível em nossa sociedade. Apesar das diversas vantagens que esse modelo apresenta, ele é responsável por causar um fenômeno conhecido como discriminação algorítmica. Devido à grande popularidade desses modelos, essa discriminação acaba se tornando um problema sério e deve ser prevenida e combatida. Esse documento descreve uma abordagem matemática e computacional das principais soluções e técnicas para a detecção e correção da discriminação contida em conjuntos de dados.

Sumário

1 Introdução

A grande popularidade da *internet* e o amplo avanço tecnológico nos últimos anos causaram a viabilidade e necessidade de se ter um *smartphone* e um computador pessoal no mundo todo. A acessibilidade a essas tecnologias promoveu o crescimento do volume de dados coletados e consequentemente causou o início de uma nova era, intitulada *big data*. (WAMBA et al., 2015)

A importância do dilúvio de dados coletados mundialmente levou à ascensão da ciência de dados, que nada mais é que o estudo de como analisar, extrair e interpretar informações referentes a um conjunto de dados. Os resultados vindos dessas técnicas acabaram sendo utilizados para a criação de ferramentas de recomendação ou até mesmo de tomada de decisões. (TECHTARGET, 2017)

Graças à ciência de dados muitos problemas de difícil natureza computacional puderam ser superados utilizando técnicas de mineração de dados e aprendizado de máquina. Devido aos típicos bons resultados alcançados por essas técnicas em uma grande diversidade de problemas, estas começaram a ser usadas em aplicações que afetam diretamente a vida das pessoas, como por exemplo *credit score*, avaliação de CV, entre outros. Como grande parte dessas técnicas funcionam explorando padrões históricos nos dados, quando estas são aplicadas usando dados de pessoas é possível que o algoritmo aprenda a tomar decisões baseadas em qualquer característica delas, o que pode vir a se tornar uma mera discriminação. (ŽLIOBAITė, 2017)

Tendo esse problema em mente, o objetivo desse trabalho consiste em uma organização de conhecimento sobre como é o estado da arte referente às técnicas utilizadas para detecção e prevenção de ocorrências de discriminação nos algoritmos de classificação (*fairness*). Especificamente, o trabalho consiste no estudo individual das principais técnicas, com o propósito de explicar como funcionam, em quais situações e maneiras cada uma delas pode ser utilizada e suas vantagens e desvantagens.

2 Revisão teórica

No contexto de algoritmos de classificação, a discriminação é definida por tratamentos que desfavorecem alguém por alguma característica “protegida”, isto é, raça, gênero, deficiência, idade, entre outras. Essas discriminações podem ocorrer em duas formas distintas denominadas direta e indireta. (ŽLIOBAITė, 2017) A direta ocorre exclusivamente por causa de atributos protegidos, ou seja, o algoritmo toma decisões baseando-se diretamente nessas características. Já a indireta ocorre através de atributos não protegidos que são correlacionados com características protegidas. (TRIPATHI, 2014)

A solução para o problema da discriminação direta é trivial, pois basta retirar as características protegidas dos dados, de tal forma que o algoritmo não poderá usá-las ao tomar uma decisão. Apesar disso, a exclusão das características protegidas não é uma prática usual, pois ao fazer isso torna-se impossível detectar e corrigir a discriminação indireta.

Para detectar e quantificar a discriminação indireta é necessário o uso de métricas que avaliam comparativamente grupos favorecidos e desfavorecidos e detectam quando as decisões do algoritmo foram prejudicadas por um atributo correlacionado a uma característica protegida.

Tendo em mente o significado contextual da discriminação direta e indireta, é possível definir outros conceitos que serão pertinentes no trabalho. O primeiro desses conceitos consiste em tomadores de decisões justos, que nada mais são que algoritmos em que as escolhas feitas independem das características protegidas, isso é, toda e qualquer decisão tomada é garantidamente livre de discriminação direta e indireta. Em contrapartida, um tomador de decisão é dito como injusto quando em pelo menos uma tomada de decisão existe algum tipo de discriminação. Bons exemplos de algoritmos injustos são: um classificador de CV que rotula indivíduos extremamente semelhantes (se diferem somente na raça, por exemplo) em grupos distintos (discriminação direta), ou até mesmo um *credit scorer* cuja pontuação rejeita qualquer pessoa que more em um determinado bairro, sendo este tipicamente habitado por pessoas negras (discriminação indireta).

2.1 Conceitos estatísticos

Por se tratar de um problema de detecção de discriminação em modelos orientados por dados é natural esperar que seja necessário uma abordagem estatística do problema. Devido a essa importância, segue abaixo as principais definições de conceitos estatísticos que serão pertinentes no trabalho.

- Grau de liberdade: o grau de liberdade é um número que se refere à quantidade de observações independentes feitas menos o número de parâmetros estimados. Esse valor é utilizado para ajustar as distribuições que são usadas nos testes estatísticos;
- P-valor: o p-valor mede o quão forte é uma evidência para uma dada hipótese nula; (STATTREK, a)
- Grau de significância: o grau de significância é a a probabilidade de um teste rejeitar a hipótese nula, visto que ela é verdadeira. Por se tratar se uma probabilidade são utilizados valores entre 0 e 1, sendo que tipicamente são usados os valores 0.01, 0.05 e 0.07; (STATTREK, a)
- Valor crítico: valor x de uma distribuição que a divide em duas regiões (região de rejeição e de aceitação). Esse valor é utilizado para determinar critérios de aceitação de hipóteses em alguns testes estatísticos.

3 Referencial teórico

O estudo das métricas para detecção de discriminação indireta consiste em um dos objetivos dessa pesquisa. Por se tratar de um trabalho de organização de conhecimento, é natural encontrar alguma semelhança com outros trabalhos na literatura, e isso ocorre com a pesquisa *"Measuring discrimination in algorithmic decision making"* (ŽLIOBAITė, 2017), que faz um breve resumo sobre o assunto desse trabalho. Apesar disso, este projeto se diferencia ao entrar em detalhes não contidos no *survey*, como por exemplo as vantagens e desvantagens de cada métrica.

Um outro trabalho semelhante é o do artigo (KAMIRAN; ŽLIOBAITĖ; CALDERS, 2013), que discute sobre o mesmo assunto, porém com uma visão diferente. Nele, o objetivo é detectar e quantificar a discriminação que pode ser explicada por variáveis que realmente deveriam afetar os resultados, não importando as características protegidas.

Por fim, para a compreensão da parte estatística do problema, foram utilizados a *web page* (STATTREK, e), que contém descrições e explicações detalhadas sobre conceitos e funcionamento de várias métricas, e o *paper* (BONCHI et al., 2017), que contém uma visão probabilística do problema.

4 Análise das métricas

Nesse capítulo serão analisadas as principais métricas que compõem o estado da arte. Essas métricas são classificadas em 4 categorias, sendo estas testes estatísticos, medidas absolutas, medidas condicionais e medidas situacionais. Para facilitar o entendimento das fórmulas matemáticas serão utilizadas as notações presentes na tabela 1.

Tabela 1: Notações adaptadas de (ŽLIOBAITė, 2017)

Símbolo	**Significado**
y	Variável alvo, y_i denota a i-ésima observação.
y^i	Valor binário de uma variável alvo, y^+ denota 1 e y^- denota 0.
s	Variável protegida.
s^i	Valor de uma variável protegida. O índice 1 denota um grupo protegido, enquanto o 0 denota um grupo não-protegido. Ex: s^1 - grupo desfavorecido e s^0 - grupo favorecido.
X	Conjunto de variáveis de entrada.
z	Variável explicativa.
z^i	Valor da i-ésima variável explicativa.
N	Número de indivíduos em um conjunto de dados.
n_i	Número de indivíduos no grupo s^i. Ex: n_0 - número de indivíduos do grupo favorecido e n_1 - número de indivíduos do grupo desfavorecido.
$z = f(X)$	Variável explicativa. Ex: z^i denota um determinado nível de educação, assim todos os indivíduos com esse mesmo nível de educação serão referenciados por i.

4.1 Testes estatísticos

Testes estatísticos são ferramentas utilizadas para verificar a existência de discriminação indireta em dados. Estes testes operam através da aceitação ou rejeição de hipóteses estatísticas, que verificam o quão provável é que um resultado tenha ocorrido por acaso. Geralmente são formuladas duas hipóteses, sendo uma destas a nula, que no contexto indica que não foi detectada discriminação nos dados, e a outra delas a alternativa, que indica que foi detectada discriminação nos dados. (ŽLIOBAITė, 2017)

4.1.1 Teste t da regressão linear

O objetivo desse teste é verificar se existe uma regressão linear significativa que determina a relação entre duas variáveis aleatórias distintas. (STATTREK, d) Para isso, é necessário realizar o teste-t.

No contexto de detecção de discriminação algorítmica em tomadores de decisão devem ser utilizados como variáveis as características protegidas do indivíduo e o resultado da classificação. Também podem ser utilizadas características não protegidas no teste, porém fica a critério de quem for realizá-lo.

A estatística t

Para obter os resultados do teste t é necessário calcular uma estatística que é dada por $t = b/\alpha$, onde b é o coeficiente estimado da regressão linear e α é o erro padrão da inclinação da reta, que pode ser calculado pela a equação 4.1. Nessa equação, a função f(.) se refere ao modelo de regressão utilizado e os demais termos foram retirados das notações da Tabela 1. (ŽLIOBAITĖ, 2017)

$$\alpha = \frac{\sqrt{\sum_{i=1}^{n}(y_i - f(y_i))^2}}{\sqrt{N-2} \times \sqrt{\sum_{i=1}^{n}(s_i - \overline{s})^2}} \tag{4.1}$$

O teste t

Para realizar o teste, primeiramente é necessário definir duas hipóteses:

- Hipótese nula: o teste não detectou nenhum efeito de atributos protegidos nos resultados de classificações;
- Hipótese alternativa: o teste detectou algum efeito de atributos protegidos nos resultados de classificações.

Após definidas as hipóteses, é necessário determinar qual o grau de significância a ser utilizado, calcular a quantidade de graus de liberdade (nesse teste é definido por $\nu = N-2$) e por fim, calcular a estatística t. Para aplicar o teste basta utilizar a distribuição t com ν graus de liberdade para calcular o p-valor, que é dado por $P(x \leq -t) + P(x \geq t)$. (STATTREK, d)

Interpretação dos resultados

A interpretação dos resultados é bastante simples, pois basta comparar o p-valor com os graus de significância definidos. Caso o p-valor seja menor, não se pode aceitar a hipótese nula, e caso contrário ela é aceita.

Vantagens e desvantagens

Vantagens:

- Métrica universal (funciona para qualquer conjunto de dados);

- Custo computacional relativamente baixo;
- Interpretação clara dos resultados.

Desvantagens:

- Caso o teste determine que a hipótese nula é a mais aceita, isso não garante a inexistência de discriminação, mas sim uma evidência de que ela não existe;
- A métrica não quantifica o quão discriminatórios são os dados, apenas detecta se houve ou não discriminação;
- É necessário estimar o coeficiente da regressão (b).

4.1.2 Teste t de duas amostras

O objetivo desse teste é verificar se as médias de valores de dois grupos são parecidas, e para isso é utilizado um teste t com uma estatística um pouco diferente da métrica anterior.

Contextualizando o teste no problema de discriminação, os grupos que serão comparados são os grupos favorecidos e desfavorecidos. (ŽLIOBAITĖ, 2017) Uma outra informação importante é que para o teste ser considerado adequado os dados devem conter as seguintes características: (STATTREK, b)

- As amostras de ambos os grupos testados devem seguir uma distribuição normal;
- As amostras de ambos os grupos testados devem ser independentes;
- As amostras de ambos os grupos testados devem ter a mesma variância.

Por fim, como esse teste t tem o objetivo (hipótese nula) de que $\mu_1 - \mu_2 = 0$ é utilizada a versão em que são analisadas as duas caudas da distribuição t. (STATTREK, b)

A estatística t

Para obter os resultados do teste t é necessário calcular uma estatística que é dada pela equação 4.2, onde δ_0^2 e δ_1^2 representam a variância dos grupos favorecidos e desfavorecidos, respectivamente. Já em relação aos demais termos, estes foram retirados das notações da Tabela 1. (ŽLIOBAITĖ, 2017)

$$t = \frac{(E(y|s^0) - E(y|s^1))}{\sqrt{\frac{(n_0-1)\times\delta_0^2+(n_1-1)\times\delta_1^2}{n_0+n_1-2}} \times \sqrt{\frac{1}{n_0}+\frac{1}{n_1}}} \tag{4.2}$$

O teste t

Para realizar o teste, primeiramente é necessário definir duas hipóteses:

- Hipótese nula: o valor esperado das classificações dos grupos protegidos e dos não-protegidos é o mesmo;
- Hipótese alternativa: o valor esperado das classificações dos grupos protegidos e dos não-protegidos não é o mesmo.

Após definidas as hipóteses, é necessário determinar qual o grau de significância a ser utilizado, calcular a quantidade de graus de liberdade (nesse teste é definido por $\nu = n_0 - n_1 - 2$) e por fim, calcular a estatística t. Para aplicar o teste basta utilizar a distribuição t com ν graus de liberdade para calcular o p-valor, que é dado por $P(x \leq -t) + P(x \geq t)$.

Interpretação dos resultados

A interpretação dos resultados é bastante simples, pois basta comparar o p-valor com os graus de significância definidos. Caso o p-valor seja menor, não se pode aceitar a hipótese nula, e caso contrário ela é aceita.

Vantagens e desvantagens

Vantagens:

- Custo computacional relativamente baixo;
- Interpretação clara dos resultados.

Desvantagens:

- Caso o teste determine que a hipótese nula é a mais aceita, isso não garante a inexistência de discriminação, mas sim uma evidência de que ela não existe;
- Para a métrica funcionar bem são necessárias determinadas propriedades nos dados;
- Essa métrica só pode ser usada para comparar duas amostras de dados (2 grupos);
- A métrica não quantifica o quão discriminatórios são os dados, apenas detecta se houve ou não discriminação.

4.1.3 Teste z para a diferença de proporção (2 grupos)

O objetivo dessa métrica é verificar se a proporção de classificações positivas é a mesma em dois grupos distintos, e para isso será utilizado um teste z.

Contextualizando o teste no problema de discriminação, os grupos que serão comparados são os grupos favorecidos e desfavorecidos. (ŽLIOBAITė, 2017) Uma outra informação importante é que para o teste ser considerado adequado os dados devem conter as seguintes características: (STATTREK, c)

- A amostragem deve ser independente;
- Cada população deve conter pelo menos 10 amostras de falhas e 10 amostras de sucesso;
- A população deve ser pelo menos 20 vezes maior do que a amostragem utilizada;
- A escolha da amostragem deve ser aleatória.

Por fim, como esse teste z tem o objetivo de $P_0 = P_1$ será usada a versão em que são analisadas as duas caudas da distribuição normal. (STATTREK, c)

A estatística z

Para obter os resultados do teste z é necessário calcular uma estatística que é dada pela equação 4.3, onde os termos foram retirados da tabela 1. (ŽLIOBAITė, 2017)

$$z = \frac{p(y^+|s^0) - p(y^+|s^1)}{\sqrt{\frac{p(y^+|s^0) \times p(y^-|s^0)}{n_0} + \frac{p(y^+|s^1) \times p(y^-|s^1)}{n_1}}} \tag{4.3}$$

O teste z

Para realizar o teste, primeiramente é necessário definir duas hipóteses:

- Hipótese nula: As proporções de resultados positivos em grupos favorecidos e desfavorecidos são as mesmas ($P_0 = P_1$);
- Hipótese alternativa: As proporções de resultados positivos em grupos favorecidos e desfavorecidos são diferentes ($P_0 \neq P_1$).

Após definidas as hipóteses, é necessário determinar qual o grau de significância a ser utilizado e calcular a estatística z. Como essa estatística é normalizada utiliza-se uma distribuição normal ($N(\mu = 0, \sigma = 1)$) para calcular o p-valor, que é dado por $P(x \leq -z) + P(x \geq z)$.(STATTREK, c)

Interpretação dos resultados

A interpretação dos resultados é bastante simples, pois basta comparar o p-valor com os graus de significância definidos. Caso o p-valor seja menor, não se pode aceitar a hipótese nula, e caso contrário ela é aceita.

Vantagens e desvantagens

Vantagens:

- Sólida teoria por trás da métrica (Teorema Central do Limite);
- A métrica utiliza dados normalizados;
- Interpretação clara dos resultados;
- Custo computacional relativamente baixo.

Desvantagens:

- Para a métrica funcionar bem são necessárias determinadas propriedades nos dados;
- Caso o teste determine que a hipótese nula é a mais aceita, isso não garante a inexistência de discriminação, mas sim uma evidência de que ela não existe;
- Essa métrica só pode ser usada para comparar duas amostras de dados (2 grupos);
- A métrica não quantifica o quão discriminatórios são os dados, apenas detecta se houve ou não discriminação;

4.1.4 Teste χ^2 para a diferença de proporção (vários grupos)

O objetivo dessa métrica é verificar se as proporções de classificações positivas são iguais em N grupos distintos, e para isso será utilizado um teste χ^2.

Diferentemente da métrica anterior, que era usada para comparar grupos favorecidos e desfavorecidos, o interessante dessa nova métrica é poder comparar indivíduos de subgrupos derivados dos favorecidos e desfavorecidos. Dessa forma é possível descobrir um subgrupo específico que é vítima de discriminação algorítmica.

A estatística χ^2

Para obter os resultados do teste χ^2 é necessário calcular uma estatística que é dada pela equação 4.4, onde k indica o número de grupos comparados e os demais termos foram retirados da tabela 1. (ŽLIOBAITė, 2017)

$$\chi^2 = \sum_{i=1}^{k} \frac{(n_i - n) \times p(y^+|s^i))^2}{p(y^+|s^i)} \tag{4.4}$$

O teste χ^2

Para realizar o teste, primeiramente é necessário definir duas hipóteses:

- Hipótese nula: as proporções de resultados positivos são iguais para todo e qualquer grupo;
- Hipótese alternativa: as proporções de resultados positivos são diferentes em pelo menos um dos grupos.

Após definidas as hipóteses, é necessário determinar qual o grau de significância a ser utilizado, calcular a quantidade de graus de liberdade (nesse teste é definido por $\nu = k - 1$) e por fim, calcular a estatística χ^2. Para aplicar o teste basta utilizar a distribuição χ^2 com ν graus de liberdade para calcular o p-valor, que é dado por $P(x \geq \chi^2)$. (ŽLIOBAITė, 2017)

Interpretação dos resultados

A interpretação dos resultados é bastante simples, pois basta comparar o p-valor com os graus de significância definidos. Caso o p-valor seja menor, não se pode aceitar a hipótese nula, e caso contrário ela é aceita.

Vantagens e desvantagens

Vantagens:

- Métrica universal (funciona para qualquer conjunto de dados);
- Custo computacional relativamente baixo;
- Interpretação clara dos resultados;
- Essa métrica pode ser usada para identificar subgrupos que tendem a sofrer discriminação.

Desvantagens:

- Caso o teste determine que a hipótese nula é a mais aceita, isso não garante a inexistência de discriminação, mas sim uma evidência de que ela não existe;
- A métrica não quantifica o quão discriminatórios são os dados, apenas detecta se houve ou não discriminação;
- Resultados sensíveis, pois a comparação é feita em N grupos simultaneamente.

4.1.5 Teste U de Mann-Whitney (Rank)

O objetivo dessa métrica é verificar se a proporção de classificações positivas é a mesma em dois grupos distintos, e para isso será utilizado um teste U.

Contextualizando o teste no problema de discriminação, os grupos que serão comparados são os grupos favorecidos e desfavorecidos. (ŽLIOBAITė, 2017) Uma outra informação importante é que para o teste ser considerado adequado, os grupos comparados devem conter uma quantidade pequena de indivíduos.(SPBHWEB,)

A estatística U

Antes de obter a estatística é necessário ordenar os dados utilizados (de acordo com algum critério). Os índices em que cada instância permaneceu após a ordenação são definidos como o seu *ranking*.

Para obter os resultados do teste U é necessário calcular uma estatística que é dada pela equação 4.5, onde R_i indica a soma das posições dos rankings de todos elementos do grupo i, já os demais termos foram retirados da tabela 1. (SPBHWEB,) (MANN; WHITNEY, 1947)

$$U_i = n_0 \times n_1 + \frac{n_i \times (n_i + 1)}{2} - R_i \tag{4.5}$$

O teste U

Para realizar o teste, primeiramente é necessário definir duas hipóteses: (ŽLIOBAITė, 2017)

- Hipótese nula: As distribuições do grupos favorecidos e desfavorecidos são as mesmas ($P_0 = P_1$);
- Hipótese alternativa: As distribuições dos grupos favorecidos e desfavorecidos são diferentes ($P_0 \neq P_1$).

Com as hipóteses definidas, o próximo passo é definir o grau de significância a ser utilizado. Feito isso, basta calcular as estatísticas U dos dois grupos e compará-las, escolhendo a com menor valor ($U_{stat} = \min(U_0, U_1)$). O último passo do teste é verificar o valor crítico do U (U_{crit}). Para isso, basta checar uma tabela de valores críticos de U com os valores de n_0, n_1 e o grau de significância.

Interpretação dos resultados

A interpretação dos resultados é feita baseando-se nos valores de U_{stat} e U_{crit}. Se $U_{stat} > U_{crit}$, a hipótese nula é aceita. Caso contrário, a hipótese alternativa é a mais adequada.(SPBHWEB,)

Vantagens e desvantagens

Vantagens:

- A métrica funciona bem para um conjunto de dados pequeno (grupos com poucos indivíduos);
- Interpretação clara dos resultados.

Desvantagens:

- Caso o teste determine que a hipótese nula é a mais aceita, isso não garante a inexistência de discriminação, mas sim uma evidência de que ela não existe;
- A métrica não quantifica o quão discriminatórios são os dados, apenas detecta se houve ou não discriminação;
- A métrica não funciona para um conjunto de dados com muitos indivíduos.

4.2 Medidas absolutas

Medidas absolutas medem de forma quantitativa a discriminação indireta. Geralmente comparam grupo de indivíduos (divididos pelas características protegidas) e verificam o quão diferentes eles são, ou o quanto de discriminação um grupo sofre se comparado a outro. Nelas, é assumido que todos os indivíduos são semelhantes, de tal forma que as medidas absolutas consideram qualquer diferença no tratamento entre grupos protegidos e não-protegidos discriminatória. Por causa disso, geralmente não são usadas sozinhas em um *dataset*, mas sim para reunir evidências essenciais para se aplicar testes estatísticos ou medidas condicionais. (ŽLIOBAITė, 2017)

4.2.1 Diferença entre médias

Funcionamento da métrica

A intuição para essa métrica é comparar a média dos valores esperados de dois grupos distintos, como mostra a equação 4.6, e usar esse resultado para determinar o quão discriminatório é o algoritmo. (ŽLIOBAITĖ, 2017)

$$d = E(y^+|s^0) - E(y^+|s^1) \tag{4.6}$$

Apesar disso, no contexto de algoritmos de classificação binária, é mais interessante usar a diferença da taxa de aceitação entre um grupo protegido e um não-protegido, como mostra a equação 4.7. Esse resultado é chamado de *discrimination score*, e funciona como um indício de que houve discriminação (valor próximo de 1) ou não (valor próximo de 0). (CALDERS; VERWER, 2010)

$$d = P(y^+|s^0) - P(y^+|s^1) \tag{4.7}$$

Vantagens e desvantagens

Vantagens:

- Custo computacional baixo;
- Medida objetiva e facilmente implementável.

Desvantagens:

- Não existe uma interpretação clara de resultados, isto é, o quão próximo de 0 um valor deve ser para se considerar que não foi detectada discriminação;
- Uso restrito em classificadores binários;
- A métrica sempre indicará que não existe discriminação caso não tenha indivíduos no grupo protegido analisado. (SILVA, 2017)

4.2.2 Diferença normalizada entre médias

Funcionamento da métrica

Assim como a métrica anterior, esta é utilizada para algoritmos de classificação binária, e o cálculo realizado nessa técnica é representado pela equação 4.8. Apesar da semelhança nos métodos, essa métrica se diferencia ao utilizar técnicas para transformar os valores resultantes em valores de Kappa (COHEN, 1960), que podem ser facilmente interpretados.

A normalização consiste na divisão pela maior discriminação possível. Dessa forma, δ terá valor 1 para algoritmos que classificam todos os candidatos de um grupo não-protegido (favorecido) acima de todos os indivíduos de grupos protegidos, e δ terá valor 0 para um classificador não discriminante.(ZLIOBAITE, 2015)

$$\delta = \frac{p(y^+|s^0) - p(y^+|s^1)}{\min(\frac{p(y^+)}{p(s^0)}, \frac{p(y^-)}{p(s^1)})} \tag{4.8}$$

Vantagens e desvantagens

Vantagens:

- É possível ter uma interpretação clara e objetiva dos resultados;
- Custo computacional baixo;
- Medida objetiva e facilmente implementável.

Desvantagens:

- Métrica restrita a classificadores binários;
- Valores retornados pela métrica que oscilam em torno de 0,5 não expressam muita informação.

4.2.3 Área embaixo da curva ROC

Funcionamento da métrica

A curva ROC consiste em um gráfico gerado a partir da plotagem $sensibilidade$ x $1-especificidade$. Essa curva tem propriedades estatísticas interessantes que podem ser exploradas. Em relação ao assunto desse trabalho, a propriedade mais relevante é a área embaixo dessa curva, que pode ser calculada segundo a equação 4.9, onde I(.) é uma função indicadora que é 1 quando a condição é verdadeira e 0 quando é falsa. (ŽLIOBAITĖ, 2017)

$$AUC = \frac{\sum_{\forall(s^i,y^i)\in D^0} \sum_{\forall(s^j,y^j)\in D^1} I(y_i < y_j)}{n_0 \times n_1} \tag{4.9}$$

No contexto de detecção de discriminação algorítmica, essa área indica o quanto uma determinada variável afetou na pontuação dada por um classificador. Para utilizar essa propriedade com o propósito de detectar discriminação, basta comparar atributos protegidos de indivíduos pertencentes a grupos favorecidos e desfavorecidos. Dessa forma, caso essa área seja um número próximo de 1 é uma indicação de que o atributo afeta muito o resultado, e caso o valor dessa área seja próximo de 0,5, indica que o atributo tem pouca influência na classificação. (CALDERS et al., 2013)

Vantagens e desvantagens

Vantagens:

- Essa métrica consegue medir separadamente o efeito de cada característica protegida nos resultados.

Desvantagens:

- Em relação às outras medidas, esta tem uma complexidade computacional relativamente alta ($O(n \times m)$, onde n e m indicam o número de observações feitas em indivíduos de grupos favorecidos e desfavorecidos, respectivamente);
- A área é calculada sem normalizar de acordo com a máxima discriminação possível. (SILVA, 2017)

4.2.4 Taxa de impacto

Funcionamento da métrica

A taxa de impacto é definida pela razão entre a proporção de classificações positivas de indivíduos de um grupo protegido sobre a proporção de classificações positivas de indivíduos de um grupo não-protegido, como pode ser visto pela equação 4.10. A inversa da taxa de impacto ($\frac{1}{r}$) foi baseada no conceito de razão de verossimilhança, que mede o quão favorável uma hipótese A é em relação a uma hipótese B. (ŽLIOBAITė, 2017)

$$r = \frac{p(y^+|s^1)}{p(y^+|s^0)} \tag{4.10}$$

Em relação à interpretação de valores, aqueles menores que 1 indicam existência de discriminação, enquanto os maiores ou iguais a 1 indicam a inexistência de discriminação.

Vantagens e desvantagens

Vantagens:

- Essa métrica funciona para comparar grupos disjuntos. (SILVA, 2017)

Desvantagens:

- Não há uma interpretação clara para valores próximos de 1;
- A métrica pode não funcionar bem em grupos de tamanhos muito diferentes.

4.2.5 Razão Elift

Funcionamento da métrica

A intuição para essa métrica é determinar o quanto a probabilidade de um indivíduo receber uma classificação positiva aumenta ao saber que ele pertence a um grupo não protegido. (SILVA, 2017) Para calcular isso, basta seguir a equação 4.11. (ŽLIOBAITĖ, 2017)

$$r = \frac{p(y^+|s^0)}{p(y^+)} \tag{4.11}$$

Idealmente o valor dessa razão deve ser próximo de 1, pois isso indica que os eventos S (variável protegida) e Y (resultado do algoritmo) são independentes. Na prática, o valor dessa razão não será 1, porém é possível definir o quão próximo de 1 esse valor deve ser para que as decisões do algoritmo possam ser consideradas neutras. Esse limite é conhecido como η-*neutrality*, e é definido pela a equação 4.12. (FUKUCHI; SAKUMA; KAMISHIMA, 2013)

$$\frac{p(y, s)}{p(y) \times p(s)} \leq 1 + \eta \tag{4.12}$$

Vantagens e desvantagens

Vantagens:

- Interpretação de valores bem definida;
- A métrica possui bom comportamento para grupos pequenos, visto que considera em seu cálculo a probabilidade de um indivíduo ter determinada característica.

Desvantagens:

- Essa métrica assume que os grupos avaliados consistem em conjuntos disjuntos de indivíduos. (SILVA, 2017)

4.2.6 Razão de chances

Funcionamento da métrica

A razão de chances indica o quão mais provável é que um grupo tenha classificações positivas em comparação a outro grupo. (ŽLIOBAITĖ, 2017) No contexto de discriminação em algoritmos de classificação, comparam-se grupos protegidos com não-protegidos. Assim a razão é determinada conforme a equação 4.13.

$$r = \frac{p(y^+|s^0) \times p(y^-|s^1)}{p(y^+|s^1) \times p(y^-|s^0)} \tag{4.13}$$

Nessa razão, compara-se a chance de indivíduos de grupos protegidos terem vantagem sobre indivíduos de grupos não protegidos (numerador) com o oposto (denominador). Logo, valores em que $r < 1$ indicam a existência de discriminação, enquanto valores próximos de 1 indicam equilíbrio.

Vantagens e desvantagens

Vantagens:

- Custo computacional baixo;
- Medida objetiva e facilmente implementável.

Desvantagens:

- Não existe uma interpretação clara de resultados, isto é, o quão próximo de 1 um valor deve ser para se considerar que não foi detectada discriminação.

4.2.7 Informação mútua normalizada

Funcionamento da métrica

A informação mútua consiste em uma medida de quanta informação dois conjuntos compartilham um sobre o outro. A intuição por trás da métrica é que a informação mútua entre a classificação e os atributos protegidos deve ser nula, ou seja, não existe dependência entre eles. (SILVA, 2017)

Para se calcular a informação mútua entre um atributo protegido e uma dada classificação utiliza-se a fórmula 4.14. Em relação a equação 4.15, esta representa a entropia, que é uma medida de incerteza que um determinado conjunto carrega. A entropia é utilizada para normalizar a informação mútua, como pode ser visto na equação 4.16.

$$I(s, y) = \sum_{(s,y)} p(s, y) \times \log \frac{p(s, y)}{p(s) \times p(y)} \tag{4.14}$$

$$H(y) = -\sum_{y} p(y) \times \log p(y) \tag{4.15}$$

$$NMI = \frac{I(y,s)}{\sqrt{H(s) \times H(y)}} \tag{4.16}$$

Vantagens e desvantagens

Vantagens:

- Trabalha diretamente em cima da definição de *fairness* (os atributos protegidos não devem influenciar os resultados);
- Essa métrica funciona com variáveis categóricas;
- A métrica também funciona com variáveis numéricas, basta substituir os somatórios por integrais. (ŽLIOBAITė, 2017)

Desvantagens:

- A interpretação dos resultados pode não ser trivial, pois na prática dificilmente duas variáveis são completamente independentes, logo é necessário saber o quão próximo de 0 um valor deve ser para se considerar que não foi detectada discriminação.

4.2.8 Resíduos balanceados

Funcionamento da métrica

Essa métrica funciona comparando valores preditos pelo classificador com os contidos no *dataset* (ŽLIOBAITė, 2017). A intuição usada para fazer essa comparação é verificar se o erro médio cometido pelo classificador em diferentes grupos é o mesmo (SILVA, 2017), e isso é feito na equação 4.17, onde y representa o valor contido no *dataset* e $\hat{y}$ representa o valor previsto pelo classificador.

$$d = \frac{\sum_{i \in D^1} y_i - \hat{y}_i}{n_1} - \frac{\sum_{j \in D^0} y_j - \hat{y}_j}{n_0} \tag{4.17}$$

Tendo isso em mente, os valores de d retornados indicam discriminação quando forem positivos, e o equilíbrio quando forem próximos de 0.

Vantagens e desvantagens

Vantagens:

- Essa métrica funciona em cima da acurácia das precisões, o que a torna bem peculiar;
- Possui bom comportamento com grupos de tamanhos bem distintos, pois a medida é normalizada.

Desvantagens:

- Essa métrica assume que não existe discriminação no *dataset*;
- É necessário conter as informações referentes às classificações no *dataset* de treinamento.
- Custo computacional relativamente alto em relação às demais métricas ($O(n_0 + n_1)$).

4.3 Medidas condicionais

As medidas condicionais tentam calcular o quanto da diferença de tratamento de grupos (protegidos e não-protegidos) pode ser explicada somente por características não protegidas. No geral são usadas em conjunto com as medidas absolutas com o objetivo de encontrar a porcentagem da quantificação de discriminação encontrada que vem somente de atributos não-protegidos (correlacionados com protegidos). (ŽLIOBAITė, 2017)

4.3.1 Diferença inexplicável

Funcionamento da métrica

A intuição usada nessa métrica é que parte da diferença de classificações entre grupos protegidos e não protegidos pode ser explicada por variáveis que realmente deveriam afetar os resultados, e a verdadeira discriminação está na parte não explicada. (SILVA, 2017)

A equação 4.18 denota como é calculada a diferença inexplicável, e nessa equação o termo d_e indica a quantidade dessa diferença que pode ser explicada, e é calculado pela fórmula 4.19. (ŽLIOBAITė, 2017)

$$d_u = p(y^+|s^0) - p(y^+|s^1) - d_e \tag{4.18}$$

$$d_e = \sum_{i=1}^{m} \frac{p(y^+|s^0, z^i) + p(y^+|s^1, z^i)}{2} \times (p(z^i|s^0) - p(z^i|s^1)) \tag{4.19}$$

Não existe uma interpretação clara dos resultados, mas intuitivamente valores próximos de 0 podem indicar pouca discriminação ou até mesmo inexistência. Já os demais valores podem indicar que existe discriminação. É válido ressaltar que pode ser necessário comparar o resultado relativamente ao valor da métrica de diferença entre médias.

Vantagens e desvantagens

Vantagens:

- Utiliza uma nova abordagem para o conceito de discriminação;
- Aprimora a ideia da métrica diferença entre médias.

Desvantagens:

- É necessário escolher as variáveis explicativas a serem usadas, o que pode não ser trivial;
- A escolha errada de uma variável explicativa afeta gradativamente os resultados (SILVA, 2017);
- Para verificar se um algoritmo é justo, pode ser necessário checar os valores da métrica em um número exponencial de grupos (ROTH, 2017);
- Medida não normalizada em relação à discriminação máxima possível;
- Não existe uma interpretação clara de resultados, isto é, o quão próximo de 0 um valor deve ser para se considerar que não foi detectada discriminação.

4.3.2 Razão Belift

Funcionamento da métrica

A intuição dessa métrica se assemelha com a da razão Elift, porém ao invés de determinar o quanto a probabilidade de um indivíduo receber uma classificação positiva aumenta ao saber que ele pertence a um grupo não protegido, ela verifica o quanto atributos correlacionados com as características protegidas de um indivíduo afetam os resultados em classificações de grupos desfavorecidos. (ŽLIOBAITė, 2017)

A equação 4.20 denota o cálculo da medida, nessa fórmula X^r representa o conjunto de atributos correlacionados com características protegidas e X^a o conjunto de todos os atributos não protegidos. (MANCUHAN; CLIFTON, 2014)

$$belift = \frac{p(y^+|s^1, X^r, X^a)}{p(y^+|X^a)} \tag{4.20}$$

Em relação à interpretação dos resultados, valores próximos de 1 indicam a inexistência de discriminação, enquanto os demais valores indicam o contrário.

Vantagens e desvantagens

Vantagens:

- É possível estimar as probabilidades usando redes bayesianas (MANCUHAN; CLIFTON, 2014).

Desvantagens:

- Muitos atuantes das áreas de aprendizado de máquina e mineração de dados não estão familiarizados com redes bayesianas, e suas construções podem não ser triviais (ŽLIOBAITė, 2017);
- Não existe uma interpretação clara de resultados, isto é, o quão próximo de 1 um valor deve ser para se considerar que não foi detectada discriminação.

4.4 Medidas situacionais

As medidas situacionais se diferem das demais por tratarem a discriminação direta. O objetivo delas é detectar o quanto a discriminação se espalha no *dataset*, ou seja, a quantidade de indivíduos do conjunto de dados que são vítimas da discriminação direta. (ŽLIOBAITė, 2017)

4.4.1 Teste situacional

Funcionamento da métrica

A intuição para essa métrica é que indivíduos semelhantes devem receber classificações semelhantes. A forma em que isso é quantificado é bastante interessante, pois para cada indivíduo são analisados os κ vizinhos mais próximos (em características não protegidas), filtrados por grupos (favorecidos e desfavorecidos). A ideia da métrica é comparar a diferença das médias desses dois grupos, como mostra a equação 4.21. (ŽLIOBAITė, 2017)

$$diff(u_i) = \frac{\sum_{u_j \in D(s^0, \kappa | u_i)} y_j}{\kappa} - \frac{\sum_{u_j \in D(s^1, \kappa | u_i)} y_j}{\kappa} \tag{4.21}$$

Em relação à interpretação dos resultados, basta comparar o valor calculado na equação 4.21 utilizando a fórmula 4.22, onde $D(s^1)$ representa o conjunto de todos os indivíduos de um grupo protegido, t representa um limite máximo de aceitação (definido pelo usuário), u_i denota o indivíduo i e $I(.)$ uma função indicadora que vale 1 caso a condição seja verdadeira e 0 caso contrário.

$$f = \frac{\sum_{u_i \in D(s^1)} I(diff(u_i) \geq t)}{|D(s^1)|} \tag{4.22}$$

Vantagens e desvantagens

Vantagens:

- Boa interpretação de resultados, basta utilizar a equação 4.22;

- Boa métrica para se ter uma noção inicial do quão discriminatório é um determinado classificador;
- Cumpre bem os objetivos propostos pelas medidas situacionais.

Desvantagens:

- É necessário definir o valor de t;
- É necessário definir alguma métrica para quantificar a semelhança entre indivíduos;
- É necessário estimar um valor para κ, o que pode não ser trivial;

4.4.2 Consistência

Funcionamento da métrica

A intuição por trás dessa métrica é que indivíduos semelhantes devem receber classificações semelhantes, logo basta comparar cada indivíduo com os seus κ vizinhos mais próximos. A fórmula 4.23 quantifica essa informação, onde $D(\kappa|u_i)$ é um subgrupo com os κ indivíduos mais semelhantes a u_i e y_i é a classificação para o indivíduo u_i.

$$C = 1 - \frac{1}{\kappa \times N} \sum_{i=1}^{N} \sum_{y_j \in D(\kappa|u_i)} |y_i - y_j| \tag{4.23}$$

Em relação à avaliação de resultados, valores próximos de 0 indicam uma forte evidência de discriminação direta, e resultados próximos de 1 podem indicar uma baixa existência ou inexistência de discriminação.

Vantagens e desvantagens

Vantagens:

- Boa métrica para se ter uma noção inicial do quão discriminatório é um determinado classificador;
- Cumpre bem os objetivos propostos pelas medidas situacionais.

Desvantagens:

- A interpretação dos resultados pode não ser trivial, pois na prática dificilmente todos os indivíduos e seus vizinhos mais próximos receberão as mesmas classificações, logo é necessário saber o quão próximo de 1 um valor deve ser para se considerar que não foi detectado discriminação;

- É necessário definir alguma métrica para quantificar a semelhança entre indivíduos;
- É necessário estimar um valor para κ, o que pode não ser trivial;
- Essa métrica apresenta problemas em situações em que existe uma correlação muito alta entre uma variável protegida e uma variável preditora. (ŽLIOBAITė, 2017)

5 Conclusão

Esse documento apresentou uma análise detalhada de 17 métricas usadas para a detecção de discriminação em classificadores e tomadores de decisão orientado a dados. Como é possível perceber pelo trabalho, o uso dessas técnicas não é algo trivial, pois muitas delas exploram alguns aspectos nos dados e deixam outros de lado, logo é necessário um estudo caso a caso para decidir quais métricas usar. Um outro aspecto importante é que o uso das técnicas pode ser aprimorado usando em conjunto duas ou mais medidas. Para contornar o problema de qual técnica usar, esse documento aponta as principais vantagens e desvantagens do uso de cada métrica. Assim, o usuário final é capaz de elaborar a sua própria metodologia para detecção de discriminação algorítmica.

Referências

BONCHI, F. et al. Exposing the probabilistic causal structure of discrimination. *International Journal of Data Science and Analytics*, v. 3, n. 1, p. 1–21, Feb 2017. ISSN 2364-4168. Disponível em: <https://doi.org/10.1007/s41060-016-0040-z>. Citado na página 8.

CALDERS, T. et al. Controlling attribute effect in linear regression. In: *2013 IEEE 13th International Conference on Data Mining*. [S.l.: s.n.], 2013. p. 71–80. ISSN 1550-4786. Citado na página 20.

CALDERS, T.; VERWER, S. Three naive bayes approaches for discrimination-free classification. *Data Mining and Knowledge Discovery*, v. 21, n. 2, p. 277–292, Sep 2010. ISSN 1573-756X. Disponível em: <https://doi.org/10.1007/s10618-010-0190-x>. Citado na página 18.

COHEN, J. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, v. 20, n. 1, p. 37–46, 1960. Disponível em: <https://doi.org/10.1177/001316446002000104>. Citado na página 19.

FUKUCHI, K.; SAKUMA, J.; KAMISHIMA, T. Prediction with model-based neutrality. In: BLOCKEEL, H. et al. (Ed.). *Machine Learning and Knowledge Discovery in Databases*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2013. p. 499–514. ISBN 978-3-642-40991-2. Citado na página 21.

KAMIRAN, F.; ŽLIOBAITĖ, I.; CALDERS, T. Quantifying explainable discrimination and removing illegal discrimination in automated decision making. *Knowledge and Information Systems*, v. 35, n. 3, p. 613–644, Jun 2013. ISSN 0219-3116. Disponível em: <https://doi.org/10.1007/s10115-012-0584-8>. Citado na página 8.

MANCUHAN, K.; CLIFTON, C. Combating discrimination using bayesian networks. v. 22, 06 2014. Citado na página 25.

MANN, H. B.; WHITNEY, D. R. On a test of whether one of two random variables is stochastically larger than the other. *The Annals of Mathematical Statistics*, Institute of Mathematical Statistics, v. 18, n. 1, p. 50–60, 1947. ISSN 00034851. Disponível em: <http://www.jstor.org/stable/2236101>. Citado na página 16.

ROTH, A. *Between "statistical"and "individual"notions of fairness in Machine Learning*. 2017. Disponível em: <https://aaronsadventures.blogspot.com.br/2017/11/between-statistical-and-individual.html>. Citado na página 25.

SILVA, T. Notes on the fairness in machine learning problem. 2017. Citado 7 vezes nas páginas 18, 20, 21, 22, 23, 24 e 25.

SPBHWEB. *Mann Whitney U Test*. Disponível em: <http://sphweb.bumc.bu.edu/>. Citado 2 vezes nas páginas 16 e 17.

STATTREK. *Dictionary*. Disponível em: <http://stattrek.com/statistics/dictionary.aspx>. Citado na página 7.

STATTREK. *Hypothesis Test: Difference Between Means*. Disponível em: <http://stattrek.com/hypothesis-test/difference-in-means.aspx?Tutorial=AP>. Citado na página 11.

STATTREK. *Hypothesis Test: Difference Between Proportions*. Disponível em: <http://stattrek.com/hypothesis-test/difference-in-proportions.aspx>. Citado na página 13.

STATTREK. *Hypothesis Test for Regression Slope*. Disponível em: <http://stattrek.com/regression/slope-test.aspx?tutorial=ap>. Citado 2 vezes nas páginas 9 e 10.

STATTREK. *Statistics and Probability*. Disponível em: <http://stattrek.com>. Citado na página 8.

TECHTARGET. *Data Science*. 2017. Disponível em: <http://searchcio.techtarget.com/definition/data-science>. Citado na página 5.

TRIPATHI, K. Analysis of direct and indirect discrimination discovery and prevetion algorithms in data mining. 2014. Citado na página 6.

WAMBA, S. et al. How 'big data' can make big impact: Findings from a systematic review and a longitudinal case study. *International Journal of Production Economics*, v. 165, p. 234–246, 2015. Citado na página 5.

ZLIOBAITE, I. On the relation between accuracy and fairness in binary classification. *CoRR*, abs/1505.05723, 2015. Disponível em: <http://arxiv.org/abs/1505.05723>. Citado na página 19.

ŽLIOBAITė, I. Measuring discrimination in algorithmic decision making. *Data Mining and Knowledge Discovery*, v. 31, p. 1060–1089, 2017. Citado 20 vezes nas páginas 2, 5, 6, 8, 9, 10, 11, 13, 15, 16, 17, 18, 19, 20, 21, 23, 24, 25, 26 e 28.