Equidade em Aprendizado de Máquina

Renan Del Buono Brotto
renanbrotto@gmail.com

I. Introdução

As técnicas de Aprendizado de Máquina estão em rápida ascensão em diversas áreas, como processamento de linguagem natural e reconhecimento de fala [1], processamento de imagem [2], problemas de classificação [3], reconhecimento de padrões [4] entre outras.

O objetivo central do Aprendizado de Máquina é projetar modelos computacionais capazes de aprender uma determinada tarefa a partir de dados, sem que exista uma programação específica para a tarefa em questão [5]. Para tanto, apresentamos um conjunto de dados disponíveis, denominado conjunto de treinamento, a fim de que o modelo capture as informações relevantes para o problema investigado [6].

Além de ter um bom desempenho, segundo uma métrica adequada, frente aos dados de treinamento, o modelo ajustado deve ser capaz de apresentar um bom desempenho frente a dados desconhecidos, pertencentes a um conjunto normalmente denominado conjunto de teste. Este segundo requisito configura a capacidade de generalização do modelo [4], [7].

Contudo, durante o processo de aprendizagem, o modelo pode capturar informações irrelevantes, ou mesmo indesejadas, para a tarefa considerada, configurando uma situação de sobreajuste [7]. Neste tipo de situação, temos um bom desempenho frente ao conjunto de treinamento, mas o modelo se comporta mal quando avaliado frente a novos dados [4]. Dentre as informações indesejadas aprendidas, podemos citar ruído presente sobre os dados [4] ou mesmo algum tipo de viés apresentado nos dados de treinamento.

O viés contido nos dados de treinamento desempenha um papel fundamental quando aplicamos os modelos de Aprendizado de Máquina em tarefas com impacto social direto, tais como concessão de crédito [8], aprovação em universidades [8] ou a concessão de liberdade condicional [9]. Neste tipo de problema, as técnicas clássicas de aprendizado incorporam as tendências observadas nos dados, o que pode dar origem a atributos discriminatórios, como por exemplo o gênero em um problema de concessão de crédito ou a etnia em um problema de concessão de liberdade condicional. Uma vez aprendida esta característica, o modelo ajustado passa a produzir resultados que podem contribuir ainda mais com a disparidade do problema [8].

II. Objetivo

Nosso objetivo neste trabalho é aplicar as técnicas de ICA (*Independent Component Analysis*) [10] sobre os atributos do nosso conjunto de dados visando a promover a equidade na classificação. Para tanto, vamos investigar a descorrelação linear, tendo como base o trabalho [11], e a descorrelação não-linear, que é uma condição mais próxima da independência.

Usaremos o Adult Data Set (Disponível em: http://archive.ics.uci.edu/ml/datasets/adult e de domínio público). Este conjunto de dados é formado por 14 atributos, tanto categóricos quanto numéricos, anonimizado. Dentre estes atributos temos idade, tipo de trabalho, gênero, etnia, estado civil, horas trabalhadas por semana entre outros. Para cada indivíduo do conjunto de dados, temos um rótulo, indicando uma renda anual superior ou a $50 k ou inferior a este limiar. Convertemos os atributos categóricos em atributos numéricos, a fim de uma representação adequada para os classificadores utilizados.

Para melhor entendermos a inequidade presente nos dados, vamos comparar, na sequência, o percentual de indivíduos com "Alta Renda" na população completa, bem como nas populações feminina e masculina:

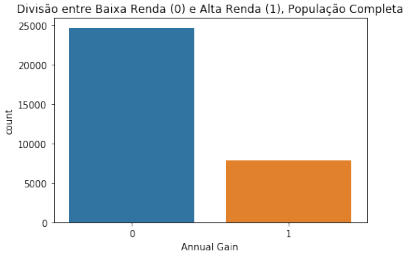

Figura 1: Divisão entre as classes na população completa.

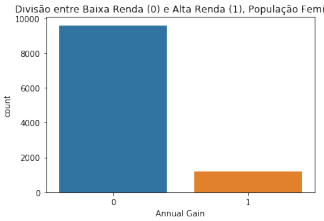

Figura 2: População Feminina.

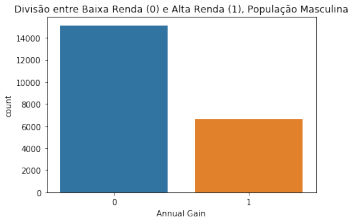

Figura 3: População Masculina.

Deste modo, temos os seguintes percentuais de indivíduos de "Alta Renda" na população completa e nas populações masculina e feminina:

	Geral	Masculino	Feminino
Alta Renda [%]	24.08	30.57	10.95

Tabela I: Percentual de indivíduos com "Alta Renda"nas populações completa, masculina e feminina

Percebemos do resultado acima que o número de indivíduos na população feminina na classe "Alta Renda" é inferior ao percentual observado na população masculina e na população como um todo. Por outro lado, o percentual de indivíduos da população masculina pertencentes à categoria "Alta Renda" é superior ao observado na população geral.

O que notamos do resultado acima é uma inequidade entre indivíduos rotulados como "Alta Renda" nas populações masculina e feminina. Treinar um classificador da maneira tradicional, sem nenhum tratamento sobre os dados, fará com que o modelo de Aprendizado de Máquina incorpore as tendências observadas nos dados, *i.e.*, gênero masculino implica em alta renda, o que pode contribuir com a acentuação da inequidade observada.

Na sequência do texto, apresentamos a nossa proposta para promover a equidade na classificação.

III. Metodologia

Para a tarefa de classificação, empregaremos classificadores baseados em regressão logística e árvores de decisão. Para os nossos propósitos aqui, adotaremos o atributo "Gênero (*Sex*)" como atributo discriminatório. Ao todo, investigaremos 6 cenários de classificação:

1) Classificador 1: aplicaremos um classificador com regressão logística sobre os dados de treinamento originais.
2) Árvore de Decisão 1: esta será a estrutura de comparação com o Classificador 1. Também aplicaremos esta árvore sobre os dados originais.
3) Classificador 2 : neste caso, buscaremos descorrelacionar linearmente o atributo discriminatório das classes do problema (como em *"Fairness Constraints: Mechanisms for Fair Classification"*, Zafar, *et al.*, 2017). Em seguida, aplicaremos uma regressão logística sobre os dados.
4) Árvore de Decisão 2: usaremos esta segunda árvore para comparação com o Classificador 2, ou seja, quando promovemos a descorrelação linear entre os dados.
5) Classificador 3: novamente adotaremos uma estrutura baseada em regressão logística, mas agora promovendo a descorrelação não-linear entre o atributo discriminatório e as classes do problema.
6) Árvore de Decisão 3: esta útima estrutura será usada para a comparação com o Classificador 3. Neste caso, esta árvore de decisão também será aplicada após a promoção da descorrelação não-linear sobre os dados de treinamento.

Sintetizamos nossa metodologia de trabalho na figura abaixo:

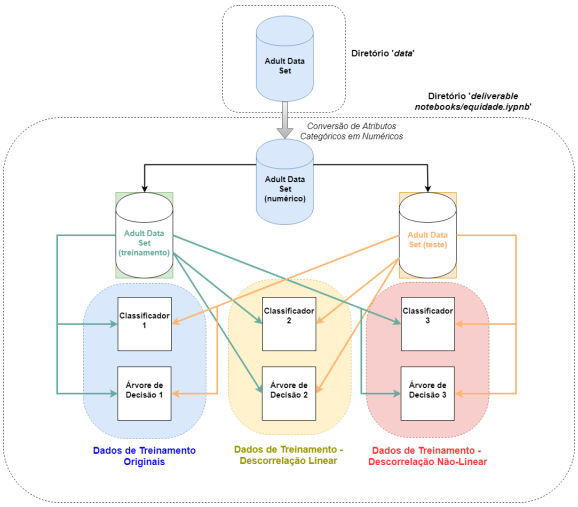

Figura 4: Workflow de trabalho.

Todo o código desenvolvido para este trabalho, juntamente com as instruções para execução do mesmo, pode ser encontrado no seguinte repositório: https://github.com/renanbrotto/ia369-reprodutibilidade-cientifica

IV. Resultados

Nesta seção, apresentamos nossos resultados na tarefa de classificação em termos de acurácia e equidade entre os indivíduos classificados como "Alta Renda" nas populações masculina e feminina. Abaixo, descrevemos os parâmetros adotados em cada um dos classificadores.

Ajustamos todos os classificadores baseados em regressão logística durante 5000 épocas e, para fins de comparação, fixamos o parâmetro *random-state* em 1 para as seis estruturas de classificação.

Abaixo, sintetizamos a acurácia obtida, frente ao conjunto de teste, para cada um dos seis classificadores:

	Acurácia [%]
Cl 1	74.42
AD 1	82.10
Cl 2	75.88
AD 2	81.94
Cl 3	75.88
AD 3	81.94

Tabela II: Acurácia dos classificadores estudados.

Em termos de acurácia, notamos que as árvores de decisão superaram em todos os casos os classificadores baseados em regressão logistica de comparação.

Para melhor compreensão destas estruturas, apresentamos as árvores obtidas para um conjunto reduzido de dados (considerando 100 indivíduos, apenas para visualização):

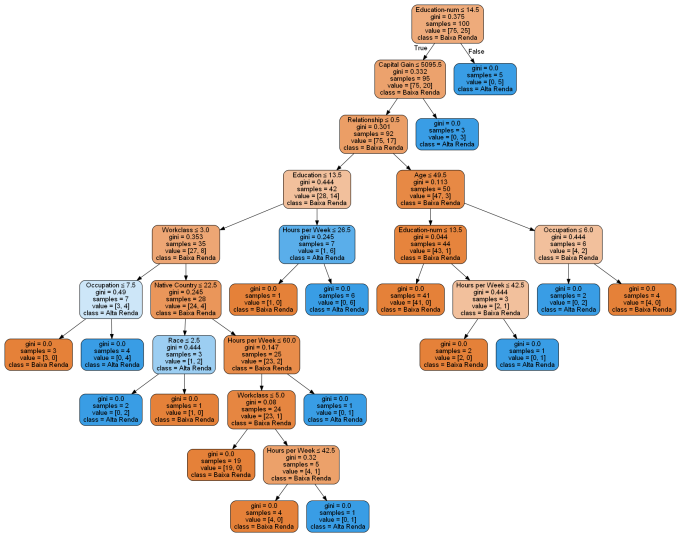

Figura 5: Árvore de Decisão 1 (dados originais).

Figura 6: Árvore de Decisão 2 (descorrelação linear).

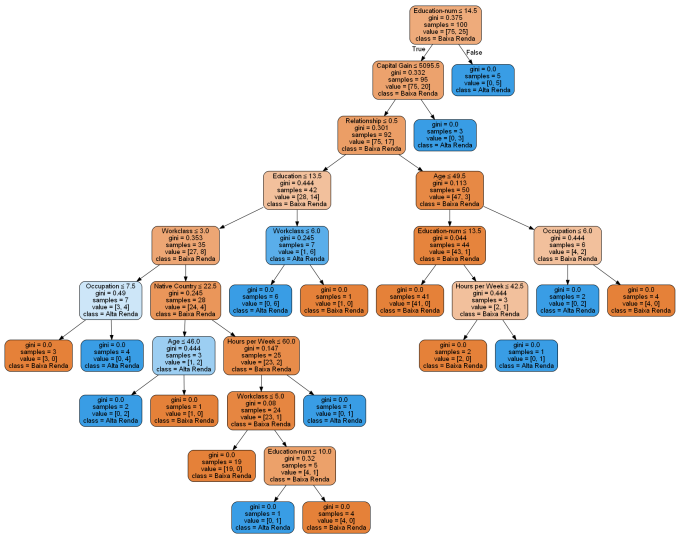

Figura 7: Árvore de Decisão 3 (Descorrelação não-linear).

Notamos nas três árvores acima a ausência do atributo discriminatório "Gênero" na formação direta da estrutura. Contudo, o mesmo atributo pode ter efeitos indiretos sobre os demais, como *Workclass*, por exemplo.

Ao aplicarmos as técnicas de ICA, podemos constatar que os classificadores 2 e 3 apresentaram uma acurácia superior ao classificador 1. Já para o caso das árvores de decisão, notamos o comportamento inverso: comparada com as demais, a árvore 1 apresenta a melhor acurácia. Ainda com relação à ICA, notamos que tanto a descorrelação linear quanto a não-linear tiveram efeitos muito parecidos na classificação, uma vez que a acurárica do classificador 2 é igual à do classificador 3 e o mesmo pode ser dito em relação às árvores 2 e 3.

Contudo, nosso interesse neste trabalho não é apenas com relação à acurácia (ou outra métrica de desempenho análoga) dos classificadores. Desejamos também avaliar a capacidade das estruturas testadas em prover a equidade. Para tanto, na sequência analisamos o percentual de indivíduos classificados como "Alta Renda" nas populações masculina e feminina com base na classificação realizada pelas 6 estruturas:

	Masc. [%]	Fem. [%]	Dif [%]
Cl 1	35.37	6.49	28.88
ÁD 1	21.05	3.83	17.22
Cl 2	30.33	7.26	23.08
ÁD 2	20.89	4.29	16.61
Cl 3	30.33	7.26	23.08
ÁD 3	20.89	4.29	16.61

Tabela III: Percentual de "Alta Renda" para cada uma das subpopulações obtido por cada um dos classificadores.

Como primeiro resultado da tabela acima, notamos que a descorrelação linear levou aos mesmos resultados que a descorrelação não-linear, como podemos constatar pela igualdade dos percentuais de cada uma das subpopulações obtidos pelos classificadores 2 e 3, assim como observamos a igualdade entre os percentuais das árvores 2 e 3.

Comparando as 3 árvores de decisão, observamos que a aplicação das técnicas de ICA levaram a um pequeno incremento na equidade, uma vez que observamos um decréscimo de indivíduos classificados como "Alta Renda" na população masculina e um acréscimo na mesma categoria na população feminima. A mesma promoção de equidade notamos nos classificadores baseados em regressão logística.

Analisando apenas a população feminina, notamos que a classificação baseada em regressão logística levou aos maiores percentuais de "Alta Renda". Contudo, quando analisamos as distâncias entre os pecentuais de cada subpopulação, notamos que a menor distância ocorre para as árvores de decisão 2 e 3.

Para o caso particular deste trabalho e da base de dados estudada, as árvores de decisão apresentaram um melhor compromisso entre acurácia e equidade. Também notamos que para o nosso problema e com a função não-liner adotada em particular, tanto a decorrelação linear quanto a não-linear levaram a resultados idênticos.

V. Conclusões

Neste trabalho avaliamos como algumas das técnicas de ICA, *i.e*, descorrelação linear e descorrelação não-linear, podem ser usadas para prover a equidade de gênero na tarefa de classificação de um conjunto de dados específico. Aplicamos estas técnicas tanto em classificadores baseados em regressão logística quanto em árvores de decisão. Para a base de dados que investigamos, as árvores de decisão se mostraram mais adequadas, mantendo um bom compromisso entre acurácia e equidade.

No caso que estudamos, as árvores de decisão não mostraram uma dependência explícita do atributo discriminatório. Contudo, o atributo "Gênero (*Sex*)" pode estar correlacionado com outros atributos, como "Categoria de Trabalho (*Workclass*), o que também contribui para a inequidade na classificação. Como perspectiva de trabalhos futuros, podemos investigar como a independência entre atributos críticos e discriminatórios pode influenciar na equidade de classificação.

Uma outra vertente bastante interessante é estudar como outras métricas de independência, como a Informação Mútua, por exemplo, podem colaborar na redução do viés contido nos dados.

Referências

[1] U. Kamath, J. Liu, and J. Whitaker, *Deep Learning for NLP and Speech Recognition*. Springer, 1ª ed., 2019.

[2] R. Cipolla, S. Battiato, and G. M. Farinella, *Machine Learning for Computer Vision*. Springer, 1ª ed., 2013.

[3] R. O. Duda, D. G. Stork, and P. E. Hart, *Pattern Classification*. Wiley-Interscience, 2ª ed., 2000.

[4] C. M. Bishop, *Pattern Recognition and Machine Learning*. Springer, 1ª ed., 2006.

[5] A. L. Samuel, "Some studies in machine learning using the game of checkers," *IBM Journal of Research and Development (Volume: 3 , Issue: 3 , July 1959)*, pp. 210–229, 1959. DOI:https://doi.org/10.1147/rd.33.0210.

[6] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*. Springer, 2ª ed., 2009.

[7] S. Haykin, *Neural Networks and Learning Machines*. Pearson, 2008.

[8] C. O'Neil, *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. Penguin, 1ª ed., 2016.

[9] A. Chouldechova, "Fair prediction with disparate impact: A study of bias in recidivism prediction instruments," *Big Data*, vol. 5, 6 2017.

[10] A. Hyvarinen, J. Karhunen, and E. Oja, *Independent component analysis*. John Wiley & Sons, 2001.

[11] M. B. Zafar, I. Valera, M. G. Rodriguez, and K. P. Gummadi, "Fairness constraints: Mechanisms for fair classification," *Proceedings of the 20th Artificial Intelligence and Statistics (20-22 April 2017, Fort Lauderdale, FL, USA)*, 2017.