FUNÇÃO DE CUSTO PARA APRENDIZADO ROBUSTO À INJUSTIÇA

Ygor de Mello Canalli

Exame de Qualificação de Doutorado apresentado ao Programa de Pós-graduação em Engenharia de Sistemas e Computação, COPPE, da Universidade Federal do Rio de Janeiro, como parte dos requisitos necessários à obtenção do título de Doutor em Engenharia de Sistemas e Computação.

Orientador: Geraldo Zimbrão da Silva

Rio de Janeiro
Outubro de 2020

FUNÇÃO DE CUSTO PARA APRENDIZADO ROBUSTO À INJUSTIÇA

Ygor de Mello Canalli

EXAME DE QUALIFICAÇÃO SUBMETIDO AO CORPO DOCENTE DO INSTITUTO ALBERTO LUIZ COIMBRA DE PÓS-GRADUAÇÃO E PESQUISA DE ENGENHARIA (COPPE) DA UNIVERSIDADE FEDERAL DO RIO DE JANEIRO COMO PARTE DOS REQUISITOS NECESSÁRIOS PARA A OBTENÇÃO DO GRAU DE DOUTOR EM CIÊNCIAS EM ENGENHARIA DE SISTEMAS E COMPUTAÇÃO.

Examinado por:

Prof. Geraldo Zimbrão da Silva, D.Sc.

Prof. Geraldo Bonorino Xexéo, D.Sc.

Prof. Leandro Guimaraes Marques Alvim, D.Sc.

Prof. Filipe Braida do Carmo, D.Sc.

RIO DE JANEIRO, RJ – BRASIL
OUTUBRO DE 2020

Àqueles cuja voz não pode ser ouvida.

Agradecimentos

Agradeço primeiramente a Deus, em quem reside minha esperança.

Agradeço à minha esposa Sâmara, por partilhar da vida ao longo desta extenuante caminhada, não me permitindo esmorecer e trazendo alento nos dias de desilusão. Também por sua sensibilidade, me ajudando a ouvir àqueles cuja voz se esvanece.

Agradeço a meu amigo Alexsander, pela amizade de valor inestimável, que extrapola os limites da universidade. Ao meu amigo Julio pelas ricas conversas entre experimentos e seminários.

Agradeço a meus pais, por desde cedo me ensinarem o gosto pela ciência e o valor de um trabalho feito com afinco. À minha irmã Yasmin, pelo afeto de sempre e a partilha da jornada acadêmica.

A meu orientador Geraldo Zimbrão, por sempre depositar confiança em meu trabalho, enriquecendo sempre as discussões com seus questionamento.

Resumo do Exame de Qualificação apresentado à COPPE/UFRJ como parte dos requisitos necessários para a obtenção do grau de Doutor em Ciências (D.Sc.)

FUNÇÃO DE CUSTO PARA APRENDIZADO ROBUSTO À INJUSTIÇA

Ygor de Mello Canalli

Outubro/2020

Orientador: Geraldo Zimbrão da Silva

Programa: Engenharia de Sistemas e Computação

O recente assunto de pesquisa de justiça em aprendizado de máquina tem apontado problemas e buscado soluções em cenários onde o uso de aprendizado de máquina na tomada de decisões com impacto social replica ou mesmo amplifica preconceitos e injustiças presentes na sociedade. Exemplos desta problemática podem ser vistos no acesso ao crédito e na justiça penal, onde certos grupos são sistematicamente desprivilegiados de acordo com gênero, raça, cor e etc. O presente trabalho propõe o uso de técnicas da literatura de classificação na presença de ruído de classe, especialmente de técnicas de correção de função de custo para mitigar o impacto negativo e viabilizar um aprendizado robusto à injustiça.

Abstract of Qualifying Exam presented to COPPE/UFRJ as a partial fulfillment of the requirements for the degree of Doctor of Science (D.Sc.)

FAIRNESS CORRECTION LOSS FUNCTION

Ygor de Mello Canalli

October/2020

Advisor: Geraldo Zimbrão da Silva

Department: Systems Engineering and Computer Science

The recent research subject of fairness in machine learning has appointed problems and searched for solutions where algorithmic decision-making reproduce and even amplifies prejudice and bias in scenarios with social impact. Examples are credit approval and criminal justice, where some groups are systematically underprivileged according to gender, race, color, etc. In this work, we use loss function correction techniques from classification in the presence of label noise literature to avoid negative impact and produce unfair robustness.

Sumário

Lista de Figuras

Lista de Tabelas

Capítulo 1

Introdução

O que fazer quando os dados utilizados no treinamento de estimadores e algoritmos de Aprendizado de Máquina carregam consigo viés, discrepâncias sociais e preconceito? Esta pergunta traz consigo um desafio para pesquisadores e profissionais, tanto de áreas técnicas e científicas, quanto ligados às humanidades e direito.

O uso industrial e comercial do Aprendizado de Máquina em aplicações que direta ou indiretamente influenciam em decisões com impacto social, como por exemplo acesso ao crédito, mercado de trabalho e educação, não apenas é uma realidade, como também gradativamente está se tornando o padrão. Embora o processo de automatização traga consigo grandes vantagens econômicas, existem também riscos que não podem ser ignorados. O uso de máquinas para tomada de decisões envolvendo atributos sensíveis carrega consigo o ideal de que algoritmos não são carregados de pré-concepções distorcidas e não julgam baseadas em preconceitos humanos, o que em parte é verdade. Embora algoritmos estatístico de fato não tenham, por si só, viés de injustiça social, frequentemente encontrada no julgamento humano, todo algoritmo de Aprendizado de Máquina faz uso de dados previamente adquiridos, a partir dos quais um estimador é treinado. O problema neste caso é mais sutil, porém não menos perigoso. Quaisquer discrepâncias sociais, injustiças e preconceitos presentes nos dados utilizados para treinamento são passíveis de serem assimilados e reproduzidos massivamente, sob o risco de não serem nem mesmo detectados.

Os assuntos de mineração de dados e classificação sem discriminação foram discutidos inicialmente por PEDRESCHI *et al.* (2008) e KAMIRAN e CALDERS (2009), respectivamente. O objetivo destes trabalhos é propor técnicas para evitar relações indesejadas envolvendo atributos ditos *sensíveis* (como por exemplo gênero, raça e religião) em decisões potencialmente danosas ou injustas do ponto de vista social. Os desdobramentos dos trabalhos acima citados conduziram a comunidade científica a um crescente número de publicação neste e em temas correlatos, em um ramo atualmente conhecido como *Fairness* (justiça), nomenclatura que utilizaremos no

decorrer de nosso trabalho.

Existem ainda outros assuntos, não menos relevantes, envolvendo Aprendizado de Máquina e a sociedade, dentre os quais destacamos *Accountability*, *Transparency* e *Data Privacy*. O ramo de pesquisa de *Transparency* (transparência) se debruça sobre a problemática de possibilitar que decisões tomadas por um estimador sejam passíveis de explicação, ao invés de funcionar como uma "caixa-preta". *Accountability*, por sua vez, é a capacidade que um modelo tem de ser auditado, de maneira que os entes envolvidos possam ser responsabilizados. Estes três assuntos frequentemente são unidos em eventos e publicações sob o acrônimo FAT (***F****airness-****A****ccountability-****T****ransparency*). Por último, destacamos também o ramo de pesquisa de *Data Privacy* (privacidade de dados), que se ocupa dos riscos à privacidade aos quais são expostos indivíduos cujos dados foram coletados, sobretudo para uso em aplicações de Aprendizado de Máquina. Muito embora os assuntos de *Data Privacy*, *Accountability* e *Transparency* muito se aproximem de *Fairness* em seu teor, o presente trabalho se limita a tratar apenes deste último.

O trabalho é dividido como segue. No Capítulo 2 realizamos uma breve revisão da literatura de Aprendizado de Máquina Justo, bem como de tópicos selecionados na área de Ruído em Aprendizado de Máquina necessários para nosso desenvolvimento. O Capítulo 3 apresenta e formaliza a proposta, que tem seus experimentos e resultados discutidos no Capítulo 4. Por fim, no Capítulo 5 apresentamos um resumo do conteúdo abordado e resultados obtidos, direcionamentos futuros e um cronograma de trabalho para desenvolvimento da Tese.

Capítulo 2

Revisão Bibliográfica

O Artigo 5ºda Constituição Brasileira garante a todos os residentes no país o direito à igualdade perante a lei. É assegurada liberdade de consciência e de crença, convicção filosófica ou política. Igualdade de gênero e raça também são garantidos pela Carta Magna, sendo inclusive o crime de racismo inafiançável e imprescritível. Da mesma forma o Artigo VII da Declaração Universal dos Direitos Humanos garante a todos igualdade diante da lei, e direito igual à proteção contra qualquer discriminação que viole outros direitos fundamentais.

Embora hajam dispositivos legais nacionais e internacionais que proíbam discriminação e que procurem garantir igualdade de acesso a oportunidades e qualidade de vida, é notório que existem discrepâncias sociais enormes para alguns grupos demográficos, seja por gênero, orientação sexual, etnia, cor, religião ou outras características. Com isso, todo processo de aquisição de dados envolvendo populações nas quais haja discriminação, discrepâncias sociais ou grupos desfavorecidos, está sujeito a viés de amostragem, seja este intencional ou não.

Assim, torna-se necessário o desenvolvimento de técnicas computacionais e estatísticas para garantir que em aplicações onde o uso de aprendizado de máquina interfere na tomada de decisões com impacto social, os direitos fundamentais de igualdade sejam respeitados.

2.1 Aprendizado de Máquina Justo

Antes da pesquisa em Aprendizado de Máquina abordar a questão da justiça, alguns tabalhos trouxeram à tona algumas discussões que colaboraram para o desenvolvimento da área. Uma desta é acerca do chamado impacto desigual injustificado (AYRES, 2002), mostrando que certos tomadores de decisão podem afetar de maneira desigual diferentes grupos sociais. A proposta é avaliar os resultados de tomadas de decição para verificar se grupos sociais desprivilegiados sofreram sistematicamente um impacto negativo mais acentuado que outros grupos. O autor então apresenta

alguns exemplos chave para comprensão do problema: concessão de crédito, progressão de pena, aceitação editorial e contratação.

Outro trabalho relevante conceitua o *redlining effect* (SQUIRES, 2003), examiando o papel do perfil racial no mercado de seguros imobiliários. O autor mostra como as práticas de mercado baseadas em estereótipos contribuíram para segregação racial desenvolvimento urbado desigual.

O primeiro trabalho acadêmico a tratar da questão de justiça em aprendizado de máquina foi PEDRESCHI *et al.* (2008). Neste trabalho o autor não apenas discorre sobre o assunto de maneira conceitual, mas propõe uma técnica computacional para contornar o problema no âmbito de mineração de regras de associação. Primeiramente, o autor argumenta que a solução simplista de não utilizar os atributos potencialmente discriminatórios não soluciona a questão. Ao eliminar os atributos sensíveis a tendência é que o estimador encontre preditores indiretos em outros atributos, ocorrendo o *redlining effect*.

Utilizando o exemplo de SQUIRES (2003), o local de moradia se tornaria um preditor indireto para a raça, caso esta fosse removida do treinamento. Neste caso o estimador estaria aprendendo indiretamente o atributo sensível através de correlação. Os riscos são grandes, com potencial de agravar a situação de desfavorecimento. Além do problema do *redline effect*, eliminar estes atributos pode comprometer significativamente o desempenho do modelo, o que seria prejudicial a todos impactados pelas previsões do sistema.

Sua modelagem do problema é feita através das *regras potencialmente discriminatórias*, que são regras do tipo $\mathbf{A}, \mathbf{B} \rightarrow \mathbf{C}$ que contenha um conjunto de itens potencialmente discriminatórios $\mathbf{A}$ em suas premissas. A ideia é medir o ganho relativo de confiança devido à presença de itens potencialmente discriminatórios, utilizando um conceito denominado α*-protetor*, onde o parâmetro α determina o nível de proteção desejada.

Posteriormente, os mesmos autores apresentam um *framework* para medição de discriminação (PEDRESCHI *et al.*, 2009) através da análise de um registro histório de decisões de uma tarefa com impacto social (por exemplo aprovação de cédito). Os autores generalizam a apordagem de seu trabalho anterior (PEDRESCHI *et al.*, 2008) e mostram como um repertório de métricas discriminatórias compreensíveis podem ser definidas em termos da confiança das regras de associação que descrevem o fenômeno em análise.

Outro trabalho pioneiro no assunto é KAMIRAN e CALDERS (2009), onde os autores propõem uma técnica de classificação binária sem discriminação. Sua técnica consiste em um pré-processamento dos dados, chamado *data massaging*, onde instâncias potenciamente discriminatórias têm suas classes alteradas de forma a impactar o mínimo possível. Utilizando um exemplo de análise de crédito, as

instâncias são ordenadas de forma que os que tem maior probabilidade de serem *victims* ou *profiters* tem suas classes invertidas. Para manter as proporções das previsões, sempre que uma instância do grupo de *victms* é promovida outra dos *profiters* é rebaixada. Com esta técnica foi possível reduzir a discriminação com baixo impacto na acurácia.

Estre trabalho foi expandido em CALDERS *et al.* (2009), onde a técnica de *data massaging* foi extendida e uma nova técnica denominada *reweighing* é proposta. O *reweighing* consiste em atribuir pesos às instâncias da amostra, de acordo com critérios análogos aos utilizados para promoção e rebaixamento no *data massaging*. Com os pesos atribuídos às instâncias, uma amostragem com substituição é realizada respeitando os pesos atribuídos e as proporções de positivo e negativo. Esta é uma abordagem menos intrusiva que a anterior, mas que também é capaz de reduzir as correlações indesejadas minimizando a perda de acurácia. Neste trabalho também podemos ver uma avaliação experimental mais completa, comparando diferentes técnicas de modificação do *dataset* com diferentes estimadores utilizados para previsão.

Os trabalhos KAMIRAN e CALDERS (2009) e CALDERS *et al.* (2009) se propõem a realizar classificação sem discriminação com técnicas de alteração do *dataset*, em uma fase de pré-processamento. Em CALDERS e VERWER (2010) temos uma técnica que altera o estimador propriamente dito. São apresentadas três diferentes alternativas para modificar um classificador *Naive Bayes* de forma que este passe a considerar as correlações indevidas durante a fase de treinamento. A primeira alternativa consiste em modificar a probabilidade condicional. Ao invés de utilizar a probabilidade da classe dado o atributo sensível, passa-se a considerar a probabilidade do atributo sensível dado a classe. A segunda consiste em treinar um modelo para cada valor possível de atributo sensível e utilizá-los de acordo com o atributo sensível no momento da previsão. A terceira alternativa consiste em adicionar uma variável latente que representa a classe correta, isto é, a classe sem o viés social inedevido. Os parâmetros do modelo são otimizados utilizando *Expectattion Maximization*.

Em (KAMIRAN *et al.*, 2010) autores de CALDERS *et al.* (2009) apresentam outra técnica de classificação sem discriminação, utilizando Árvores de Decisão. A técnica atualiza o estado-da-arte de classificação sem discirminação com menor discriminação que os trabalhos anteriores para uma perda reduzida de acurácia. A proposta consiste em alterar os critérios de divisão da árvore de decisão, bem como as estratégias de poda. Para tanto, na divisão de cada nó avalia-se não apenas sua contribuição para acurácia, mas também para a discriminação, a partir de três diferentes critérios. Tendo a estrutura da árvore organizada, aplica-se um processo de remarcação, onde as novas instâncias a serem classificadas não recebem a classe

da maioria das instâncias da região onde foi localizada na árvore. Ao invés disso avalia-se a possibilidade de atribuir novas classes às folhas de maneira a diminuir a discirminação e melhorar a acurácia.

Outra abordagem possível é tratar o problema de classificação justa como um problema de otitimização. Esta abordagem é utilizada em DWORK *et al.* (2012), onde os autores propõem uma definição de justiça que possibilita a que o problam seja tratado como um problema de otimização é minimizado com uma função de custo arbitrária. Apresenta-se uma metrica que avalia em que nível os indivíduos são similares de acordo com uma certa tarefa de classificação, apresentado formalmente como uma condição *Lipschitz*. Também argumenta-se que paridade estatístiaca por si só não é suficiente para garantir justiça, porém investigam condições sob as quais a noção proposta de justiça implica também em paridade estatística. O conteito de paridade estatística consiste em garantir que a distribuição demográfica em cada atribuição de classifiação reflete a demografia de toda a população. Por fim, se discute a relação entre justiça e privacidade, verificando-se quando justiça implica em privacidade e como técnicas do contexto de privacidade diferencial podem ser aplicadas no problema de justiça.

Todas as propostas até então apresentadas pela literatura exigem ou modificação nos dados ou ajustes nos algoritmos de classificação. Visando propor uma alternativa, KAMIRAN *et al.* (2012) apresenta duas formas de contornar a injustiça sem alterar os dados ou os algoritmos, utilizando teorida da decisão. A primeira solução, chamada *Reject Option based Classification*, se utiliza das probabilidades a posteriori por um ou mais estimadores para identificar instâncias para classificação de uma maneira que neutralize os efeitos da discirminação. A segunda, chamada *Discrimination-Aware Ensemble*, realiza um *ensemble* de classificadores para encontrar instâncias discordantes, sendo estas consideradas candidatas a remarcação. Ambas apresentando excelente *tradeoff* entre acurácia e discriminação, atualizando o estado-da-arte de então.

Três causas de injustiça são apontadas em KAMISHIMA *et al.* (2012): preconceito, *underfitting* e legado negativo. O preconteito é descrito como sendo qualquer dependência indevida envolvendo um atributo sensível. O preconceito também é dividido em três categorias. Os autores consideram preconceito direto quanto o atributo sensível é diretamente utilizado no modelo para tomada de decisão. Considera-se preconceito indireto quando a classe alvo apresenta dependência estatística em relação ao atributo sensível. Já o preconceito latente se caracteriza quando há dependência estatística entre o atributo sensível e um outro atributo utilizado na classificação. Outra situação onde ocorre a injustiça é em casos que o modelo não convergiu completamente, gerando uma situação não intencional onde o *underfitting* produz injustiças. Por fim, o legado negativo se mostra quando as discrepâncias

sociais estão presentes nos dados por causa de uma amostragem injusta ou mesmo classificações indevidas. Para todos as causas de injustiça abordadas são definidas métricas correspondentes. Além de contribuir para a caracterização do fenômeno, os autores apresentam uma técnica de regularização que se propõe a reduzir os efeitos do preconceito indireto. A vantagem de utilizar um regularizador para reduzir o preconceito é que este pode ser aplicado a uma ampla variadade de estimadores. Por fim, uma avaliação experimental é conduzida comparando-se as tecnicas apresentadas em CALDERS e VERWER (2010) através das métricas propostas para cada tipo de injustiça.

Um breve e compreensível *survey* da literatura disponível até a presente época pode ser encontrado em SAINDANE e KOLHE (2014).

Após uma discussão rica sobre as complexidades do assunto de aprendizado de máquina e injustiça, com as complicações legais e técnicas envolvidas, BAROCAS e SELBST (2016) apresenta uma detalhadamente uma série de etapas ao se lidar com um problema de predição potencialmente discirminatória, abordando cada uma de maneira cuidadosa e ampla. Alguns dos assuntos tratados são: diferença entre variável alvo e a etiqueta da classe, aquisição de dados, seleção de características, atributos *proxy* e diferença entre disparidade de tratamento e disparidade de impacto. Dentre estes assuntos, destacamos a diferença entre disparidade de tratamento e disparidade de impacto, ambas previstas em lei. Disparidade de tratamento ocorre quando um atributo sensível é utilizado na previsão, uma concepção análoga à definida como preconceito direto em KAMISHIMA *et al.* (2012). A disparidade de impacto ocorre quando as saídas desproporcionalmente prejudicam (ou beneficiam) algum dos grupos, similar ao conceito de paridade estatística. A problemática reside, sobretudo, no fato de que eliminar a disparidade de tratamento, isto é, não disponibilizar o atributo sensível ao estimador, pode conduzir o estimador a produzir uma disparidade de tratamento, ocorrendo portanto preconceito indireto.

Como forma de quantificar e restringir o nível de desparidade de impacto, ZAFAR *et al.* (2017a) propõe o uso de p%-*rules* combinado a classificadores baseados em margem convexa, como por exemplo regressão logística e *Support Vector Machines*. A p%-*rule* estabelece um limite percentual de disparidade, medindo-se a proporção de previsões positivas atribuídas a membros do grupo protegido (ou preterido) do atributo sensível em relação ao percentual do grupo preferido. Assim é possível estabelecer uma restrição percentual (por exemplo 80%-*rule*) para um problema de otimização de acurácia, ou mesmo fixar um nível de acurácia e otimizar p.

Posteriormente, o conceito de disparidade de enganos é proposto em ZAFAR *et al.* (2017b) como forma de avaliar a problemática além da perspectiva de dipatidade de tratamento e impacto. A métrica proposta para medir a disparidade de enganos consiste em aferir a proporcionalidade com a qual o modelo erra nos dife-

rentes grupos do atributo sensível. Ou seja, deseja-se verificar se um estimador erra com mais frequência no grupo demográfico preterido. Os autores utilizam as taxas de falso positivo e falso negativo para criar restrições ao problema de minimização da acurácia em classificadores baseados em margem convexa, de modo similar ao desenvolvido em ZAFAR *et al.* (2017a).

O conceito de justiça é quantificado na maioria dos trabalhos como alguma medida de paridade estatística entre os grupos, como por exemplo acurácia, proporção de positovos, taxas de falsos positivos e falsos negativos, entre outros. O argumento de KEARNS *et al.* (2018a) é que restrições como esta estão suscetísveis a falha e fraude. Um classificador pode aparentar ser justo atendendo critérios de paridade entre diferentes grupos, enquanto subgrupos, com combinações de diferentes atributos protegidos, podem estar sendo tratados de maneira injusta. Os autores propõem que as métricas de paridade sejam verificadas em todas as combinações possíveis de atributos protegidos para que se possa garantir justiça. Naturalmente, aboradar o problema desta maneira traz de imadiato desafios do ponto de vista computacional, dada a natureza exponencial da definição de justiça. Prova-se que o problema é equivalente ao problema de Aprendizado Agnóstico Fraco KEARNS *et al.* (1994), que nos pior dos casos é computacionalmente difícil. Para viabilizar a solução do problema, é proposta uma heurística que computa uma solução aproximada para um problema do tipo *min-max* em um jogo de soma zero. No jogo há um jogador buscando minimizar (estimador) e outro maximizar (auditor).

Uma avaliação experimetal desta técnica é apresentada em KEARNS *et al.* (2018b). O algoritmo é avaliado em quatro diferentes *datasets* reais, onde geralmente converge rapidamente, com ganhos significativos de justiça e perdas moderadas em termos de acurácia. Também são apresentados resultados que mostram que apenas otimizar a acurácia sujeito apenas à critérios de justiça apenas de grupos isolados leva o classificador a tratar substancialmente subgrupos de maneira injusta.

Um trabalho notório na área de Aprendizado de Máquina Justo (CORBETT-DAVIES *et al.*, 2018) defende que as métricas e definições formais de justiça apresentadas em quase toda a literatura não só podem ser insuficientes para garantir equidade, como também são capazes de perniciosamente prejudicar os grupos aos quais se propuseram a defender. Estas definições formais baseiam-se majoritariamente em pelo menos um destes conceitos:

anti-classificação: atributos protegidos e seus preditores indiretos não são explicitamente utilizados na tomada de decisão

paridade de classificação: métricas convencionais de performance preditiva, como taxas de falso positivo, falso negativo, acurácia e taxa de são iguais para os grupos definidos pelos atributos protegidos

calibração: resyltados devem ser independentes dos atributos protegidos após um controle do risco estimado

Um exemplo de como a anti-classificação pode prejudicar um grupo é o caso de reincidência em crimes violentos, onde pode-se desejar que mulheres não sejam desfavorecidas em relação aos homens. Entretanto, a decisão de eliminar este atributo do modelo preditivo pode agravar a situação. Isso se dá pois tipicamente mulheres reincidem menos em crimes violentos. Sendo assim, um modelo que ignora o gênero provavelmente irá superestimar o risco de rés femininas e subestimar o de réus masculinos. Da mesma forma, quando os riscos diferem entre os diferentes grupos de um atributo protegido, espera-se que haja diferença entre métricas justamente pelo estimador capturar as peculiaridades do fenômeno por trás dos dados. Este fenômeno estatístico é conhecido como infra-marginalidade (AYRES, 2002, SIMOIU *et al.*, 2017).

A calibração, por sua vez, embora seja uma característica desejável para um sistema preditivo justo, por si só não é capaz de garantir justiça. Um exemplo de modelo considerado calibrado é o caso onde réus brancos e negros classificados em um mesmo nível de risco (por exemplo alto risco) de fato apresentem taxas de reincidência compatível. Entretanto, é possível que o limiar de risco para o qual cada um deles é considerado de alto risco seja diferente, fazendo com que o modelo seja injusto mesmo mantendo a calibração.

Como tratado anteriormente, um dos causadores de injustiça em aprendizado de máquina é o viés introduzido no conjunto de dados. Este problema é abordado em BUOLAMWINI e GEBRU (2018), mostrando como a falta de instâncias de um determinado grupo pode prejudicar os resultados de maneira significativa. Uma avaliação comparativa da *performance* de três sistemas comerciais de idenfiticação de gênero a partir de fotografias foi feito utilizando dois *datasets* de referência da literatura. Estes *datasets* são majoritariamente ($79,6\%$ e $86,2\%$) compostos por faces mais claras (conforme sistema de classificação de tipo de pele de Fitzpatrick). Nestas circunstâncias, os sistemas comerciais apresentaram uma taxa máxima de erro para homens de pele clara de $0,8\%$, enquanto para mulheres de pele escura a taxa de erro chega a $34,7\%$. Este é mais um exemplo de como problemas de injunstiça passíveis de solução podem ser negligenciados, criando circunstâncias desfavoráveis para um uso justo da tecnologia.

Uma avaliação experimental comparativa é apresentada em FRIEDLER *et al.* (2019), onde um benchmark é desenvolvido a partir de diversos *datasets* públicos de *fairness* e diversos algoritmos são avaliados de acordo com inúmeras métricas. O estudo mostra que embora hajam muitas diferenças em termos de formulação, os resultados são fortemente correlacionados uns com os outros. Outro ponto observado

é que grande parte das técnicas é sensível a flutuação nos dados, o que foi verificado através de validação cruzada, mostrando a fragilidade das técnicas.

Tomando um rumo metodológico diferente, PASSI e BAROCAS (2019) apresenta um estudo extremamente relevante sobre a formulação dos problemas de aprendizado de máquina como questão determinante para a justiça. Os autores realizam uma pesquisa etnográfica ao longo de seis meses com uma equipe corporativa de cientistas de dados. Utilizando-se de reflexões da sociologia, história da ciência e textos clássicos da área de descoberta de conhecimento em bancos de dados, descreve-se a complexidade das interações entre os atores envolvidos na formulação dos problemas de aprendizado de máquina. A pesquisa mostrar que a especificação e operacionalização do problema são sempre negociáveis, elásticas e raramente tratadas com considerações normativas em mente. O objetivo é mostrar que as decisões do dia a dia contam muito, e que ações normativas efetivas precisam atender à complexidade do processo de formulação do problema.

Uma perspectiva históricas das noções de justiça e como mensurá-las dos últimos 50 anos é apresentada em HUTCHINSON e MITCHELL (2019). São apresentadas noções e definições de justiça e injustiça de acordo com o contexto social e cultural, bem como as medidas utilizadas recentemente na literatura específica da área. Com um estudo comparativo, apresentam-se direções a serem seguidas na pesquisa e medição de justiça incorporando-se contribuições de pensamento do passado.

Um *survey* apresentando diferentes fontes de viés, categorização de injustiças, definições e métricas de justiça, etapas de solução e técnicas disponíveis na literatura pode ser encontrado em MEHRABI *et al.* (2019).

2.2 Aprendizado imperfeito e ruído

A atividade humana ao longo do tempo tem gerado uma quantidade cada vez maior de informação. Este imenso volume de informção é um dos principais fatores que contribui para a popularização do Aprendizado de Máquina e área correlatas. Entretanto, ao mesmo tempo que há uma quantidade enorme de informação disponível, quando se trata de Aprendizado Supervisionado, os respectivos rótulos de classe para o treinamento geralmente são oriundos de esforço humano direto. Assim, não é raro se deparar com problemas de Aprendizado de Máquina onde há sim abundância de dados, mas os respectivos rótulos de treinamento são escassos, pouco confiáveis ou ruidosos. Por isso o campo de estudo de Aprendizado Fracamente Supervisionado (*Weakly Supervised Lerning*), nome dado a problemas com estas características, tem atraído muitos pesquisadores, praticantes e interessados em Aprendizado de Máquina (PATRINI *et al.*, 2016).

Um dos elementos que pode levar um problema para a categoria de Fracamente

Supervisionado é o ruído, neste caso, o ruído de classe. Naturalmente pode haver atuação de fenômenos ruidosos nas características (*features*), entretanto o ruído de classe geralmente apresenta maiores impactos no processo de aprendizado. Intuitivamente, podemos pensar que existem diversas características para se aprender, muitos dos quais podem ter pouca correlação com a classe alvo. Por sua vez, o rótulo da classe é a única referência disponível para a saída do estimador.

Partindo do pressuposto de que há atuação de fenômenos ruidosos sobre o rótulo observado, é necessário distingui-lo da classe alvo, o valor de classe correto com o qual deseja-se aprender.

Em FRENAY e VERLEYSEN (2014) podemos encontrar um *survey* compreensível sobre a vasta literatura de Ruído de Classe. Primeiramente, são abordadas as definições e fontes de ruído de classe, e uma taxonomia é proposta. Também são abordadas possível consequências do ruído de classe. Posteriormente são abordados modelos robustos a ruído, limpeza de dados e algoritmos tolerantes a ruído. Adicionalmente, a metodologia experimental é discutida para algumas circunstâncias.

2.2.1 Taxonomia de ruído

No presente trabalho utilizamos a taxonomia de Ruído de Classe proposta por FRENAY e VERLEYSEN (2014). São abordados três tipos de ruído: Ruído Completamente Aleatório (NCAR - *Noisy Completely at Random*), Ruído Aleatório (NAR - *Noisy at Random*) e Ruído Não Aleatório (NNAR - *Noisy not at Random*). Para uma representação gráfica da taxonomia de Ruído em Classe veja a Figura 2.1. Considere:

X o vetor das características;

Y a classe verdadeira;

$\widetilde{Y}$ o rótulo observado;

E uma variável binária de erro ($Y \neq \widetilde{Y}$);

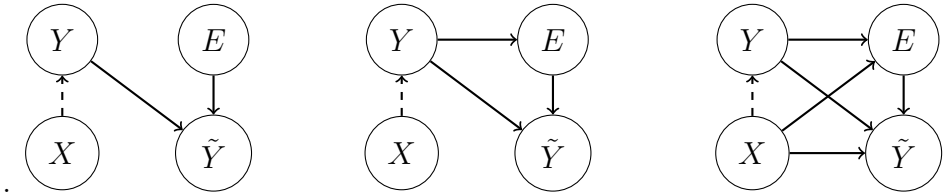
c o número de diferentes classes possíveis;

$\mathcal{Y} = \{y_1, y_2, \ldots, y_c\}$ o conjunto das classes possíveis.

No Ruído Completamente Aleatório (NCAR) (2.1a) o relacionamento entre Y e $\widetilde{Y}$ é considerado completamente aleatório pois a ocorrência de um erro E é independente de quaisquer outras variáves aleatórias. Neste caso, temos a probabilidade de erro

$$p_e = P(E = 1) = P(Y \neq \widetilde{Y}). \tag{2.1}$$

Figura 2.1: Taxonomia de Ruído de Classe segundo FRENAY e VERLEYSEN (2014). As setas correspondem às dependências estatísticas. A dependência entre X e Y foi representada como uma seta tracejada. (a) Ruído Completamente Aleatório (NCAR). (b) Ruído Aleatório (NAR). (c) Ruído Não Aleatório (NNAR).



(a) NCAR (b) NAR (c) NNAR

No caso do Ruído Aleatório (NAR) (2.1b) o rótulo observável $\widetilde{Y}$ depende da classe real Y. A ocorrência de erro E neste caso depende apenas de Y, caracterizando um ruído assimétrico. São casos onde a probabilidade de erro no rótulo varia de acordo com a classe verdadeira. Neste caso, a probabilidade de erro é dada por

$$p_e = P(E = 1) = \sum_{y \in \mathcal{Y}} P(Y = y) P(E = 1 | Y = y). \tag{2.2}$$

Por último, no caso do Ruído Não Aleatório (NNAR) (2.1c) a ocorrência do erro E depende não apenas da classe Y, mas também dos atributos X. Assim, a probabilidade $Y \neq \widetilde{Y}$ varia de acordo com os atributos em X, sendo mais ou menos provável em determinadas regiões de $\mathcal{X}$, o espaço vetorial de X. Este é o caso mais geral, o mais difícil de estimar e também o que mais nos interessa. Assim, em caso de X contínuo, temos a probabilidade de erro

$$p_e = P(E = 1) \tag{2.3}$$

$$= \sum_{y \in \mathcal{Y}} P(Y = y) \times \int_{x \in \mathcal{X}} P(X = x | Y = y)\, P(E = 1 | X = x, Y = y)\, dx. \tag{2.4}$$

Entretanto, como em casos práticos é comum que p_e seja próximo de zero em muitas regiões de $\mathcal{X}$, apresentando valores relevantes em outras regiões, podemos definir a probabilidade de erro como

$$p_e(x, y) = P(E = 1 | X = x, Y = y). \tag{2.5}$$

2.2.2 Correção de função de custo

Uma das formas de lidar com o problema de Aprendizado Fracamente Supervisionado é desenvolvendo modelos que incluam o ruído em sua modelagem de forma

explícita. Uma técnica que apresenta este tipo de abordagem pode ser visto em PATRINI *et al.* (2016). Os autores mostram que grande parte das funções de custo utilizadas podem ser decompostas atavés de uma agregação de termos lineares independentes do rótulo, de forma que possam ser expressas através de somas do mesma função de custo, conceito denominado Faforação de Função de Custo.

Posteriormente, PATRINI *et al.* (2017) apresenta duas técnicas de correção de função de custo para robustez à ruído de classe em redes neurais profundas. As técnicas se utilizam de uma matriz de transição T, composta pelas probabilidades de ocorrer corrupção de uma classe para outra, considerando-se ruído do tipo NAR. Formalmente, a matriz de transição $T \in [0, 1]c \times c$, tal que

$$T_{i,j} = P(\tilde{Y} = y_j | Y = y_i), \forall i, j. \tag{2.6}$$

O procedimento de correção *backward*, conforme descrito em PATRINI *et al.* (2017), pode ser descrito a partir de um função de cust ℓ qualquer e uma matriz de transição T da seguinte maneira.

$$\ell^{\leftarrow}(P(\tilde{Y}|X)) = T^{-1}\ell(P(\tilde{Y}|X)), \tag{2.7}$$

onde T^{-1} é a inversa de T.

Além deste, os autores também propõem o procedimento de correção *forward*, conforme descrito abaixo.

$$\ell^{\rightarrow}(P(\tilde{Y}|X)) = \ell(T^{\intercal}P(\tilde{Y}|X)), \tag{2.8}$$

onde $T^{\intercal}$ é a transposta de T.

BRAIDA DO CARMO (2017)

Capítulo 3

Proposta

O objetivo do presente trabalho é modelar o problema de *Fairness* como um problema de ruído natural de classe. Para tanto, consideramos que houve influência de preconceito e viés social durante a coleta de dados, alterando disproporcionalmente a classe de acordo com algum atributo sensível. Por exemplo, um potencial cliente pode ser registrado como mal pagador equivocadamente por estar presente em um grupo desprivilegiado. Embora não seja possível afirmar que se tratem de fenômenos totalmente compatíveis, neste caso o problema de *Fairness* se aproxima do problema de ruído natural na classe. Dentre as diferenças exitentes nos dois fenômenos, destacamos o fato de o problema de ruído natural pressupor que as alterações nas observações não são intencionais, o que não necessariamente ocorre em nosso problema alvo. Entretanto, para tirar proveito da literatura de ruído em problemas de *Fairness*, assumimos que se tratam de fenômenos análogos.

No contexto de ruído natural, a caracterização que mais se aproxima do problema de *Fairness* é a do tipo NNAR (*Noisy Not at Random*), onde a probabilidade de ocorrência de ruído na classe depende de X e Y. Assim, no caso de um problema de viés social, as probabilidades de ocorrência de falso positivo podem ser elevadas para o grupo com atributo sensível privilegiado. Simultaneamente, o grupo desprivilegiado pode possuir uma taxa de falsos negativas mais elevada que de seus pares. Tal combinação de fatores não meramente aleatórios, mas com influência de problemas sociais, levaria a uma situação onde o grupo privilegiado possui uma vantagem indevida, caracterizando de fato um problema de *Fairness*.

Nossa proposta consiste, objetivamente, de utilizar-se das técnicas de correção de função de custo propostas por PATRINI *et al.* (2017) para ruído do tipo NAR, conforme vimos no Capítulo 2, para construir um estimador robusto à preconceito e viés social. O termo robusto à preconceito, ou robusto à injustiça trata-se de uma extrapolação do conceito de robustez à ruído apresentado na seção 2.2. Baeando-se na suposição de que se tratam de fenômenos compatíveis, podemos utilizar duas matrizes de transição distintas para representar um ruído do tipo NNAR, uma para

cada grupo, de forma a contemplar a situação proposta acima. Considere T_+ a matriz de transição do grupo privilegiado, bem como T_- a matriz de transição do grupo desprivilegiado. Analogamente, temos fp_+, fp_-, fn_+ e fn_- as taxas de falso positivo do grupo privilegiado, falso positivo do grupo desprivilegiado, falso negativo do grupo privilegiado e falso negativo do grupo desprivilegiado, respectivamente. Assim, teríamos

$$T_+ = \begin{bmatrix} 1 - fp_+ & fp_+ \\ fn_+ & 1 - fn_+ \end{bmatrix}, \tag{3.1}$$

$$T_- = \begin{bmatrix} 1 - fp_- & fp_- \\ fn_- & 1 - fn_- \end{bmatrix} \tag{3.2}$$

.

Em posse dos valores de fp_+, fp_-, fn_+ e fn_-, podemos utilizar as matrizes T_+ e T_- para corrigir as discrepâncias durante a fase de treinamento, através das fatorações *forward* e *backward*. Na prática, a técnica funcionaria como um ponderador, dando mais credibilidade para a informação de certo grupo subestimado e menos para a de outro grupo superestimado.

Naturalmente, para que tal proposta seja efetiva em um problema prático, seria necessário saber quais as taxas de falso positivo e negativo para cada grupo, o que raramente é uma informação disponível. Assim, para viabilidade prática desta proposta, seria necessário uma maneira de estimar fp_+, fp_-, fn_+ e fn_-, de forma construir as matrizes de transição T_+ e T_-. Tal questão consiste no problema central proposta, o qual ainda não foi solucionado. Assim, o presente texto se limita a apresentar indícios de que, caso seja possível estimar T_+ e T_-, podemos utilizar as técnicas de fatoração de função de custo para construir estimadores robustos à injustiça.

Imediatamente, nos deparamos com o problema de lidar com duas matrizes de transição distintas, uma para cada grupo, o que não ocorre no trabalho desenvolvido em PATRINI *et al.* (2017). Para isso, utilizamos as matrizes de transição para produzir funções de custo específicas para cada grupo.

Formalmente, seja $\mathcal{D}$ um dataset com um atributo sensível s, temos $\mathcal{D}_+$ o subconjunto das instâncias cujo atributo s é do grupo privilegiado. Analogamente, tomamos $\mathcal{D}_-$, o subconjunto das instâncias cujo atributo s é do grupo desprivilegiado. Seja X o conjunto de todos os atributos em $\mathcal{D}$ e y suas classes. Por fim, seja X_+, y_+ os subconjuntos dos atributos e classes em $\mathcal{D}_+$ e X_-, y_- os subconjuntos dos atributos e classes em $\mathcal{D}_-$, respectivamente. Utilizamos a matriz de transição T_+ para construir uma função de custo ℓ_+, e a matriz T_- para construir uma função de ℓ_-. Com estas duas funções de custo, basta treinar as instâncias em $\mathcal{D}_+$ aplicando a função ℓ_+, bem como as instâncias em $\mathcal{D}_-$ aplicando ℓ_-.

A construção das funções ℓ_+ e ℓ_- pode ser feita através da fatoração *forward* da seguinte forma:

$$\ell_+(p(y_+|X_+)) = \ell(T_+^\intercal p(y_+|X_+)), \tag{3.3}$$
$$\ell_-(p(y_-|X_-)) = \ell(T_-^\intercal p(y_-|X_-)), \tag{3.4}$$

onde $p(y_+|X_+)$ são as probabilidades condicionais de y_+ dado X_+, $p(y_-|X_-)$ as probabilidades condicionais de y_- dado X_-, $T_+^\intercal$ é a matriz de transição T_+ transposta, $T_-^\intercal$ a transposta de T_- e ℓ um função de custo qualquer.

De forma análoga, podemos construir $\ell_+()$ e $\ell_-()$ através da fatoração *backward*:

$$\ell_+(p(y_+|X_+)) = T_+^{-1}\ell(p(y_+|X_+)),$$
$$\ell_-(p(y_-|X_-)) = T_-^{-1}\ell(p(y_-|X_-)),$$

onde, T_+^{-1} é a inversa da matriz de transição T_+ e T_-^{-1} a inversa de T_-.

Assim, estimando T_+ e T_- demos utilizar os procedimentos de correção *forward* e *backward* para correção de injustiças. Os detalhes metodológicos para implementação desta técnica são apresentados no Capítulo 4.

Capítulo 4

Avaliação experimental

Neste capítulo descreveremos os experimentos realizados, apresentando a metodologia, ambiente de execução, *datasets* e resultados obtidos. Nosso objetivo é demonstrar que a decomposição de erro na função de custo é capaz de contribuir para a robustez à injustiças. Como dissemos anteriormente, parte da problemática reside no fato de que não sabemos, a priori, como falsos positivos e falsos negativos se distribuem na maioria dos *datasets*. Assim, para viabilizar nosso estudo experimental preliminar, partiremos de um *dataset* original (supostamente justo) e então introduziremos artificialmente falsos positivos e falsos negativos selecionados aleatoriamente para inversão de classe de acordo com taxas estabelecidas previamente. Tais taxas serão utilizadas na função de custo para verificar a capacidade do modelo de ponderar incertezas nos dados distribuídas de forma injusta dado que se conhece a distribuição desta incerteza de acordo com o atributo sensível, seja por conhecimento a priori ou através de estimativa. Em outras palavras, escolhemos taxas de falso positivo e falso negativo, contaminamos o *dataset* com inversões de classe e então utilizamos estas mesmas taxas para construir a matriz de transição e aplicar a técnica de fatoração de erro para criar robustez á injustiça.

O ambiente de execução utilizado foi um Intel® Core™ i7-9700K com 3.6 GHz e 16 GB RAM. Os experimentos foram codificados utilizando-se do ambiente científico do Python 3, incluindo SciPy (VIRTANEN *et al.*, 2020), NumPy (VAN DER WALT *et al.*, 2011), Matplotlib (HUNTER, 2007), Pandas (MCKINNEY, 2010), Scikit-learn (PEDREGOSA *et al.*, 2011), TensorFlow (ABADI *et al.*, 2015) e Keras (CHOLLET *et al.*, 2015).

Em todos os nossos experimentos utilizamos como base a função de custo Entropia Cruzada (JAYNES, 1968). Assim, comparamos o desempenho da função original com sua versão modificada através da decomposição de erro de PATRINI *et al.* (2016). Tal estudo pode ser realizado tanto com a decomposição *forward* quanto a *backward*. Entretanto, em neste estudo nos limitamos a avaliar o desempenho da técnica apenas com a decomposição *forward*.

4.1 Experimentos no *Adult Income Dataset*

O *Adult Income Dataset* apresenta um problema de classificação binária, tendo por atributo alvo a renda do indivíduo, alta (> 50k) ou baixa (<= 50k). Possui um total de 48842 instâncias, divididas em teino e teste com as seguintes proporções:

Total: 48842

Treino: 32561 (66, 66%)

Teste: 16281 (33, 33%)

Positivo: 11687 (23.92%)

Treino: 7841 (16, 05%)

Teste: 3846 (7, 87%)

Negativo: 37155 (76, 07%)

Treino: 24720 (50, 61%)

Teste: 12435 (25, 45%)

Os atributos do *dataset* são descritos a seguir em sua lingua original.

age continuous;

workclass Private, Self-emp-not-inc, Self-emp-inc, Federal-gov, Local-gov, State-gov, Without-pay, Never-worked;

fnlwgt contínuo;

education Bachelors, Some-college, 11th, HS-grad, Prof-school, Assoc-acdm, Assoc-voc, 9th, 7th-8th, 12th, Masters, 1st-4th, 10th, Doctorate, 5th-6th, Preschool;

education-num contínuo;

marital-status Married-civ-spouse, Divorced, Never-married, Separated, Widowed, Married-spouse-absent, Married-AF-spouse;

occupation Tech-support, Craft-repair, Other-service, Sales, Exec-managerial, Prof-specialty, Handlers-cleaners, Machine-op-inspct, Adm-clerical, Farming-fishing, Transport-moving, Priv-house-serv, Protective-serv, Armed-Forces;

relationship Wife, Own-child, Husband, Not-in-family, Other-relative, Unmarried;

race White, Asian-Pac-Islander, Amer-Indian-Eskimo, Other, Black;

Figura 4.1: Arquitetura da Rede Neural utilizada nos experimentos no *Adult Income dataset*

108 neurônios (entrada) → 128 neurônios
→ Função de Ativaçao ReLU
→ 2 neurônios → Função softmax (saída)

sex Female, Male;

capital-gain contínuo;

capital-loss contínuo;

hours-per-week contínuo;

native-country United-States, Cambodia, England, Puerto-Rico, Canada, Germany, Outlying-US(Guam-USVI-etc), India, Japan, Greece, South, China, Cuba, Iran, Honduras, Philippines, Italy, Poland, Jamaica, Vietnam, Mexico, Portugal, Ireland, France, Dominican-Republic, Laos, Ecuador, Taiwan, Haiti, Columbia, Hungary, Guatemala, Nicaragua, Scotland, Thailand, Yugoslavia, El-Salvador, Trinadad&Tobago, Peru, Hong, Holand-Netherlands.

O pré-processamento foi feito utilizando *one-hot-encoding* para atributos categóricos e normalização no intervalo $[0, 1]$ para os contínuos, totalizando 108 atributos após as transformações.

O modelo utilizado foi uma rede neural do tipo *Multi Layer Perceptron*, treinado durante 6 épocas, *batchs* de tamanho 32 e otimizador ADAM (KINGMA e BA, 2014), utilizando a arquitetura apresentada na Figura 4.1. A Tabela 4.1 apresenta o desempenho de referência do modelo no *dataset* sem qualquer injeção de ruído ou alteração na função de custo.

4.1.1 NAR

Como experimento introdutório, realizamos uma avaliação apenas com introdução de ruído do tipo NAR (*Noisy at Random*), onde a probabilidade de ocorrência de uma inversão de classe não depende de X, mas apenas da classe y, portando não dependento também do atributo sensível. Assim, temos apenas uma matriz de transição, com taxas de falso positivo e falso negativo, análoga à definida na Equação (3.1).

Para medir o desempenho da correção *forward* no problema, inserimos falsos positivos e falsos negativos no *dataset* original com taxas de $0,0$ a $0,5$ cada, com

Tabela 4.1: Matriz de confusão para o *Adult Income Dataset*

		Previsto		
		$<= 50\mathbf{k}$	$> 50\mathbf{k}$	**total**
Efetivo	$<= 50\mathbf{k}$	69.53% Verdadeiro negativo	6.84% Falso positivo	76.37%
	$> 50\mathbf{k}$	8.27% Falso negativo	15.34% Verdadeiro positivo	23.62%
	total	77.80%	22.19%	

incremento de $0,1$, produzindo um total de 36 combinações. A Figura 4.2 apresenta as taxas de erro de classificação verificadas nas previsões feitas pelo modelo sem a correção, sendo portanto nosso *baseline*. Tons vermelhos mais intensos representam taxas de erro mais acentuadas. Note que o modelo apresenta maior sensibilidade à introdução de falsos positivos, o que pode ser explicado pelo fato de que grande parte das classes originais são negativas (veja Tabela 4.1).

Em posse das taxas de falso positivo e falso negativo, repetimos o experimento utilizando-se da fatoração *forward* de função de custo para corrigir a Entropia Cruzada a partir da matriz de transição, conforme descrito anteriormente. A Figura 4.3 apresenta os resultados obtidos. Podemos ver uma consistente redução da taxa de erro em relação ao *baseline*. A Figura 4.4 apresenta a melhora relativa da correção em relação ao baseline. Tons azuis mais intensos representam melhoras mais acentuadas, assim como tons avermelhados piora. Nota-se a capacidade de mitigar os efeitos das inversões introduzidas, sobretudo a taxas mais elevadas. O único caso onde houve piora relevante foi o caso onde as taxas de falso positivo e falso negativo são $0,5$, configurando portanto um rótulo completamente aleatório. Assim, nos demais experimentos abandonamos a avaliação do nível de corrupção de classes em $0,5$.

4.1.2 NNAR

Embora o experimento anterior demonstre, em certo aspecto, os efeitos da decomposição *forward*, este não é capaz de simular uma situação onde há viés social ou injustiça, pois nenhum atributo sensível foi levado em conta nas taxas de introdução de falso positivo e falso negativo. Assim, neste experimento introduzimos um ruído

Figura 4.2: Taxa de erro de classificação do modelo sem correção (NAR).

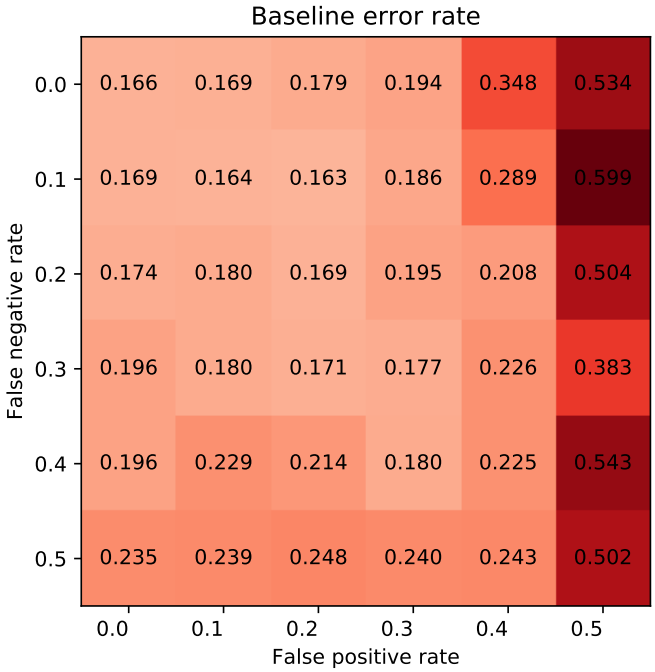


Figura 4.3: Taxa de erro de classificação do modelo com correção (NAR).

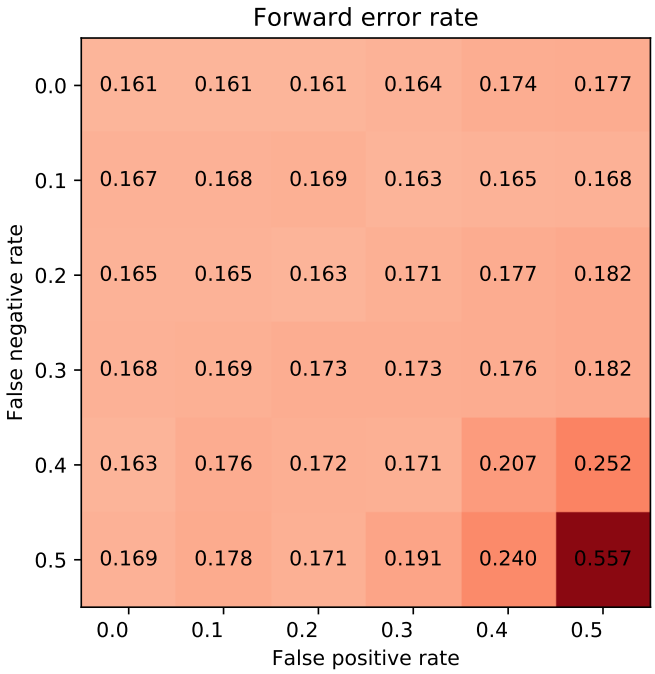
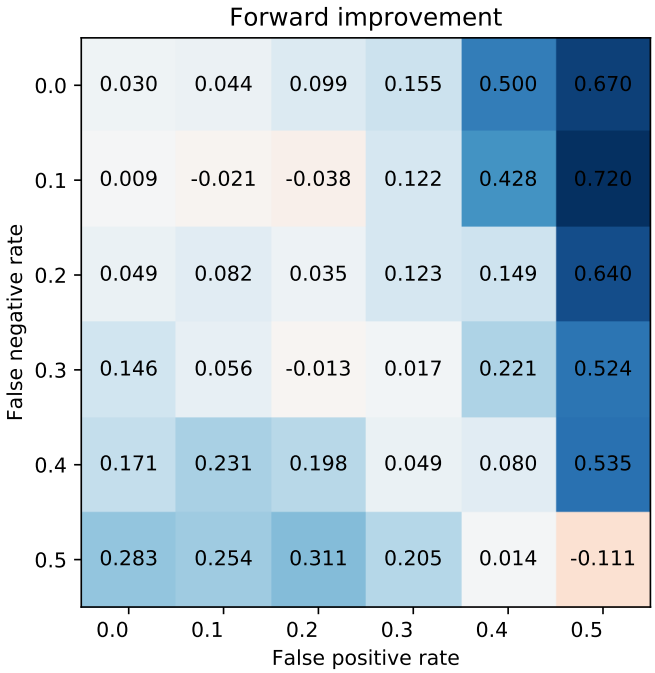


Figura 4.4: Melhora do modelo usando correção em relação ao *baseline* (NAR)



do tipo NNAR, onde não apenas a classe y influencia na probabilidade de inversão, mas também os atributos em X. Especialmente, utilizamos o atributo sensível de gênero (masculino/feminino), onde poderíamos considerar grupo masculino como privilegiado e o feminino como desprivilegiado.

Assim, temos não apenas uma matriz de transição com falsos positivos e falsos negativos, mas uma matriz para cada grupo estudado: masculino e feminino, conforme apresentado nas Equações (3.1) e 3.2. Para utilizarmos estas duas matrizes de transição, precisamos de taxas de falso positivo masculino (fp_m), falso negativo masculino (fn_m), falso positivo feminino (fp_f) e falso negativo feminino (fn_f), combinadas conforme vemos nas equações (4.1) e (4.2).

$$T_m = \begin{bmatrix} 1 - fp_m & fp_m \\ fn_m & 1 - fn_m \end{bmatrix}, \tag{4.1}$$

$$T_f = \begin{bmatrix} 1 - fp_f & fp_f \\ fn_f & 1 - fn_f \end{bmatrix}. \tag{4.2}$$

A utilização de duas matrizies de transição distintas traz consigo um problema do ponto de vista de implementação. Cada uma das matrizes de transição produziria uma nova função de custo, em nosso caso, duas variações da Entropia Cruzada. Então, precisamos fazer com que cada matriz de transição seja utilizada de acordo com sua instância correspondente, isto é, T_m para instâncias de gênero masculino e T_f para instâncias femininas. Para contornar este problema utilizamos a matriz

de transição em um formato tensorial, produzindo um volume de dimensões $b \times t \times t$, onde b é a quantidade de instâncias no *batch* e t é a dimensão da matrix de transição quadrada $t \times t$. Assim, ao aplicar o produto interno, temos cada instâncias correspondendo à uma das matrizes de transição de tamanho $t \times t$, de acordo com seu valor do atributo sensível, ao longo da dimensão b.

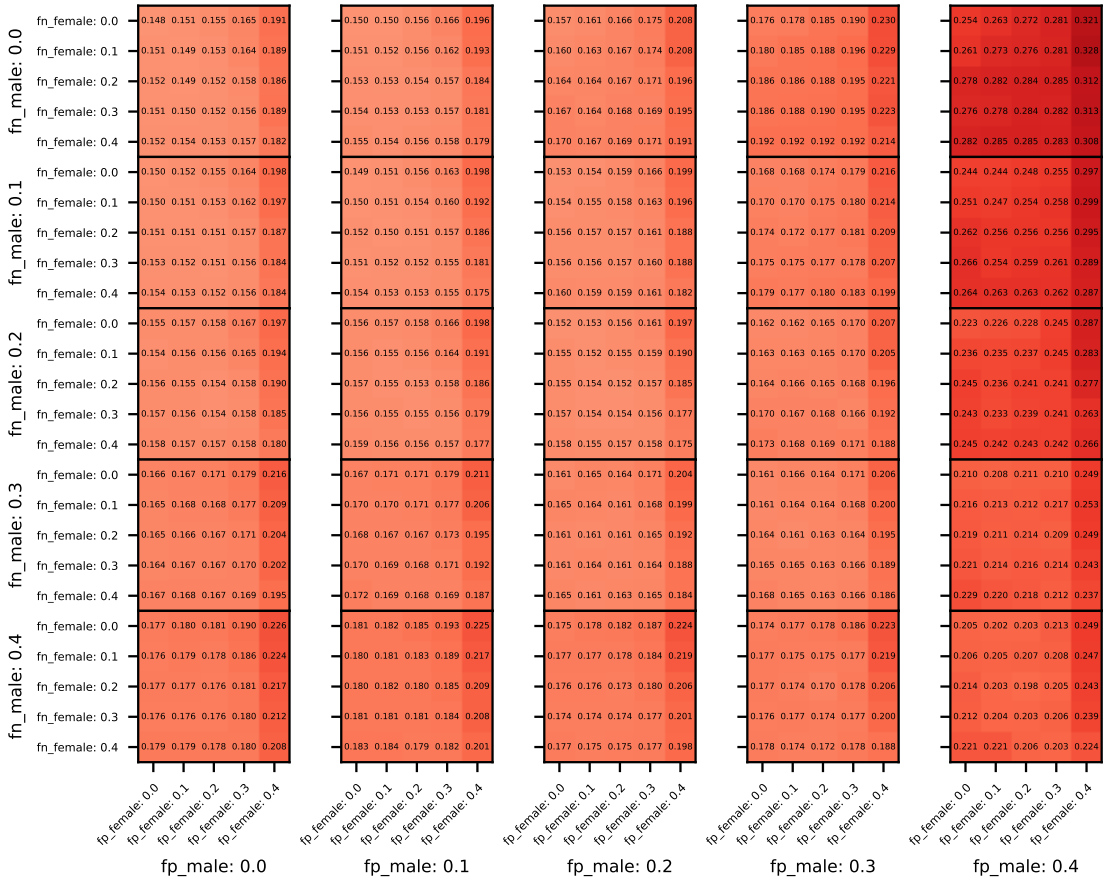
Assim, conforme fizemos no experimento anterior, inserimos inversões de classe no *dataset* com diferentes combinações, de $0,0$ a $0,4$ para cada taxa (fp_m, fn_m, fp_f, fn_f), com incremento de $0,1$, produzindo um total de 625 combinações. Como temos 4 diferentes taxas de inversão sendo combinadas, não é possível uma representação bidimensional, conforme apresentado no experimento NAR (veja Figura 4.5). Para contornarmos este problema, criamos uma combinação de dois eixos bidimensionais, onde temos um eixo externo com as taxas masculinas (fp_m, fn_m), e um interno com as taxas femininas (fp_f, fn_f). Desta maneira, para cada combinação de taxas de falso positivo e falso negativo masculino, temos um *grid* de tamanho 5×5 análogo ao do experimento NAR, porém com as diferentes taxas femininas. Em ambos casos, externo e interno, temos no eixo das abscissas taxas de falso positivo e no eixo das ordenadas as taxas de falso negativo. O resultado final é uma espécie de *grid* de *grids*.

A Figura 4.5 apresenta as taxas de erro obtidas no experimento sem correção, para os diferentes níveis de inversão de classe, servindo de *baseline* para os demais resultados. Assim como no experimento NNAR, prevalece a piora dos resultados com o aumento das taxas de falso positivo, tanto para o gênero masculino, quanto feminino. Além disso, a ocorrência de muitos falsos negativos simultâneamente com falsos positivos atenua o efeito de piora, se comparado aos casos em que os falsos negativos são menos frequentes. Podemos compreender este efeito como uma compensação para inversões que aumentariam o número de casos positivos, piorando o desempenho do modelo. Este efeito fica mais claro observando as taxas de falso positivo e falso negativo masculino, isto é, o eixo externo. Ademais, notamos efeitos similares na piora do desempenho com a introdução de ruído nos grupos de gênero masculino e feminino. Com o objetivo de viabilizar a comparação dos resultados, as figuras 4.5 e 4.6 se mantém na mesma escala de cores.

Passamos agora a avaliar os resultados obtidos utilizando a correção *forward* adaptada para ruído do tipo NNAR simulando situações de injustiça. Para tanto, temos na Figura 4.6 as taxas de erro de classificação do modelo. Comparando-se os resultados obtidos neste experimentos com o *baseline* (ver Figura 4.5) fica clara a atenuação da piora causada pela introdução de diferentes níveis de inversão de classe. Além disso, o efeito de atenuação se mantém em todos os diferentes níveis de corrupção avaliados.

Como recurso visual comparativo, apresentamos na Figura 4.7 a melhora relativa

Figura 4.5: Taxa de erro de classificação do modelo sem correção (NNAR).

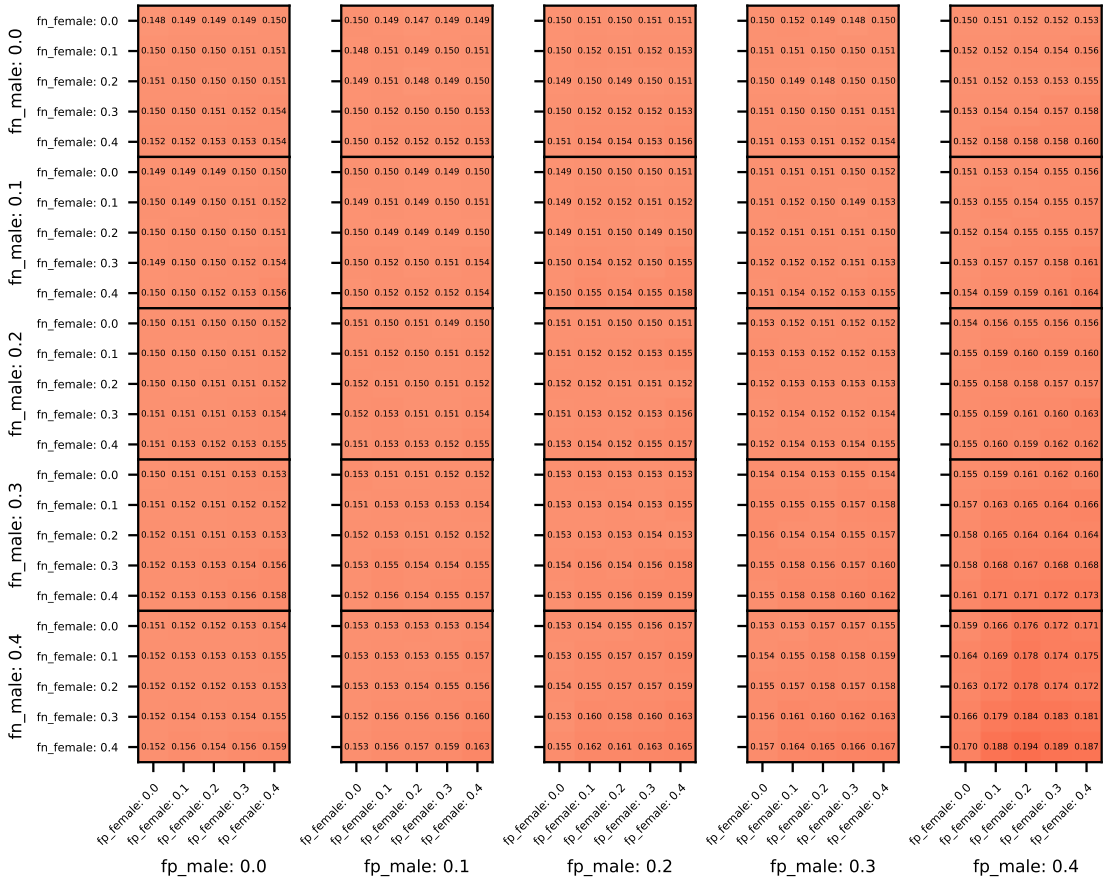


do presente experimento em relação ao *baseline*. Tons azulados apresentam melhora e avermelhados piora, a intensidade dos tons é proporcional à melhora/piora. Podemos perceber que houve melhora todos os casos, sobretudo nas situações onde o baseline apresentou maior queda de desempenho. Há também uma quantidade relevante de casos em que a melhora foi tênue ou mesmo nula, sobretudo ao longo da diagonal principal, situações onde a proporção de falsos positivos e falsos negativos se equipara.

Os resultados de melhora apresentados foram possíveis somente porque utilizamos na correção as mesmas taxas de falso positivo e falso negativo utilizadas para corremper os dados. Ter este conhecimento *apriori* é algo muito raro em casos reais, restando a alternativa da estimativa. Como abordado anteriormente, estamos avaliando o desempenho da reconstrução caso seja possível estimar com qualidade tais taxas. Supondo que esta estimativa seja viável, naturalmente haverá um erro em relação às taxas verdadeiras. Este erro de estimativa da matriz de transição poderia ser um fator complicador na qualidade da reconstrução, ou até mesmo torná-la inviabilizá-la.

Assim, com o objetivo de avaliar a sensibilidade da correção a erros de estimativa

Figura 4.6: Taxa de erro de classificação do modelo com correção (NNAR).

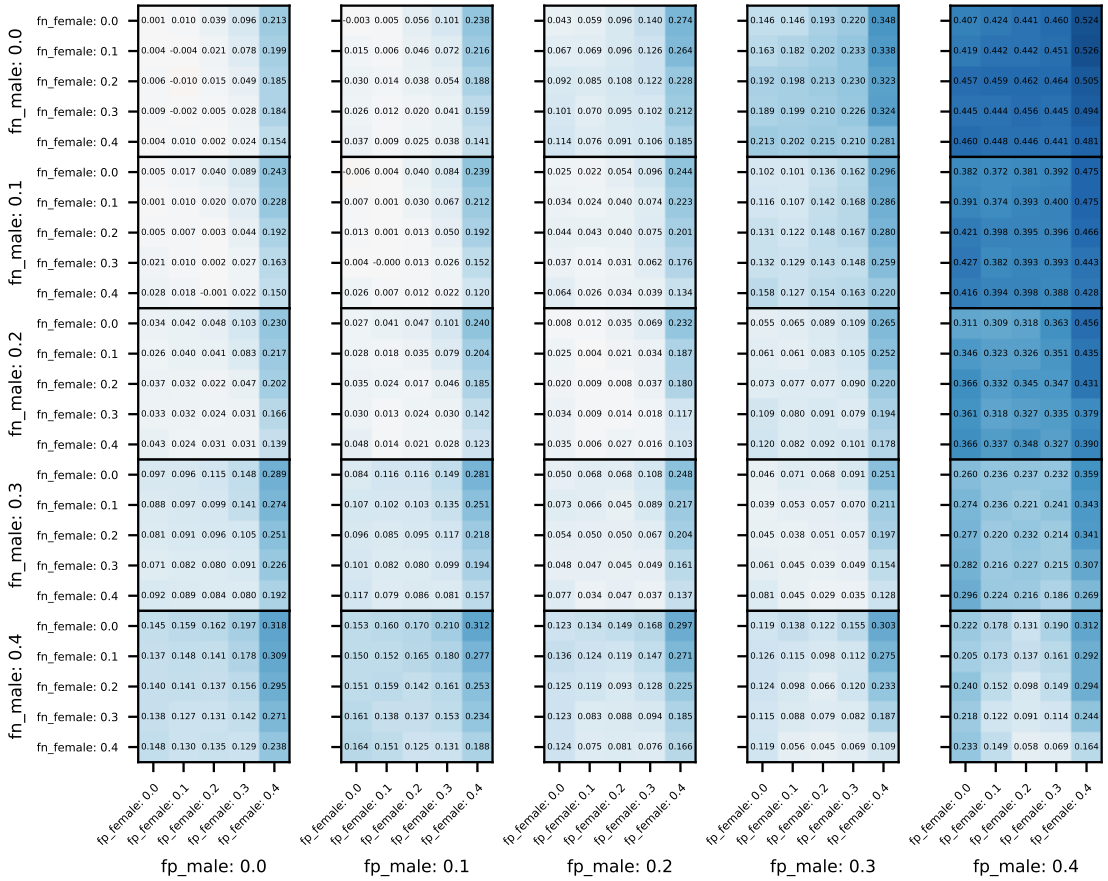


na matriz de transição, fixamos as taxas de falsos positivos e negativos utilizadas para corromper os dados e variamos a matriz de transição da correção com combinações de valores próximos. As taxas utilizadas para corromper os dados foram fixadas em $fp_m = 0.2, fn_m = 0.1, fp_f = 0.1, fn_f = 0.3$, um situação com claro desfavorecimento da classe feminina. A taxa de erro de classificação do modelo sem correção com estes níveis de contaminação dos rótulos é de 0.156 (veja Figura 4.5), sendo este nosso *baseline* para a avaliação de sensibilidade à desvios na matriz de transição utilizada na correção. Avaliamos matrizes de transição com combinações de desvios de $\pm 0,01,\ 0,05$ e $0,1$ em relação às taxas corretas, totalizando 2401 combinações.

A Figura 4.8 apresenta os resultados desta análise de sensibilidade na mesma escala de cor que o *baseline*. Nelas podemos notar que a técnica apresentou baixa perda da capacidade de correção para estes níveis erro de estimativa das taxas de falso positivo e falso negativo corretas (comparar com Figura 4.3). Neste sentido, a técnica apresenta boa tolerância a erros de estimativa.

Com o objetivo de identificar as circunstâncias nas quais houve menor tolerância a desvios nas taxas de corrupção corretas, apresentamos na Figura 4.9 os mesmos

Figura 4.7: Melhora relativa da correção em relação ao baseline (NNAR).



resultados comparados com o *baseline*. Tons azulados apresentam melhora e avermelhados piora, a intensidade dos tons é proporcional à melhora/piora. Podemos notar que em grande parte dos casos a melhora se mantém, com pequenas variações. Por outro lado, há uma piora mais intensa nos casos onde os falsos negativos são mais elevados e os falsos positivos são menores do que a taxa correta, corespondendo ao canto inferior esquerdo das figuras. Uma explicação para este fenômeno é que com falsos negativos superestimados mais instâncias passam a ser errôneamente consideradas positivas. Simultâneamente temos outra fonte de um aumento nos casos classificados como positivos, pois os falsos positivos subestimados. Esta combinação produz um aumento no volume total de previsões positivas, efeito discutido anteriormente como fator de piora no desempenho do modelo.

Figura 4.8: Taxa de erro de classificação do modelo utilizando correção com diferentes matrizes de transição para taxas de corrupção fixadas.

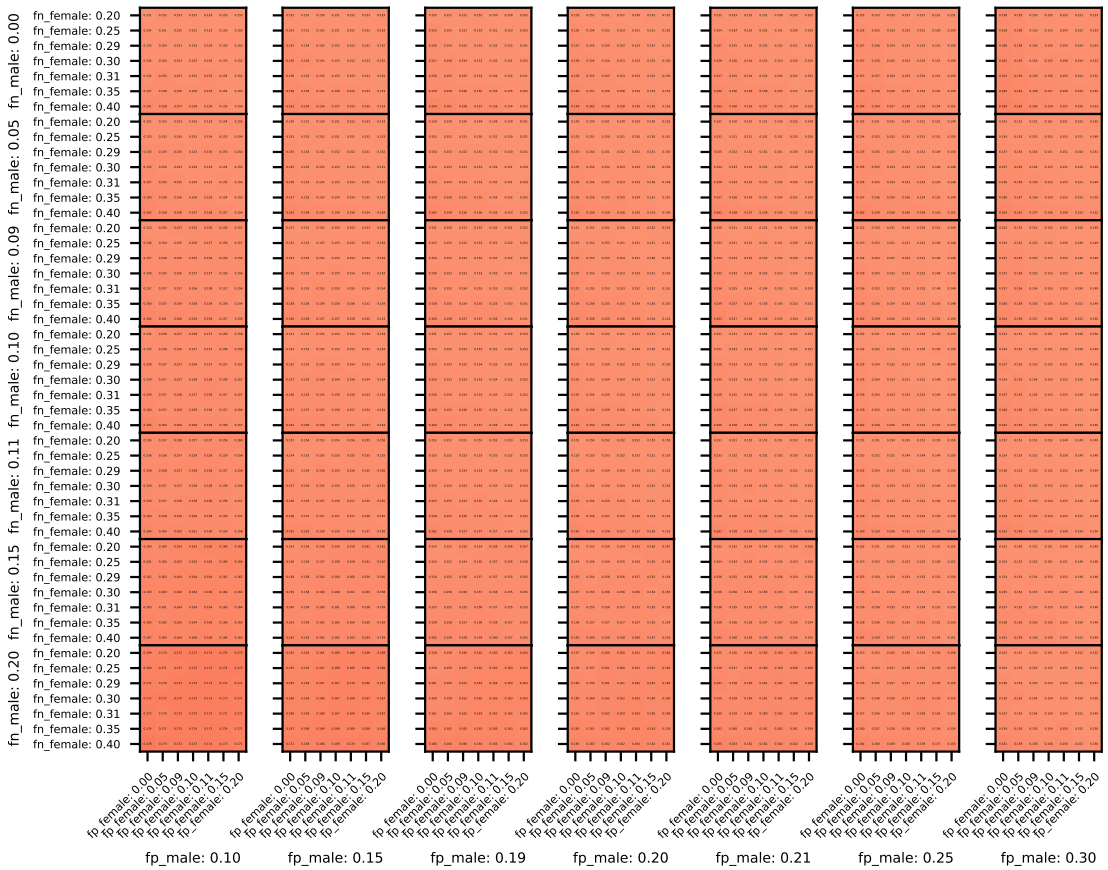
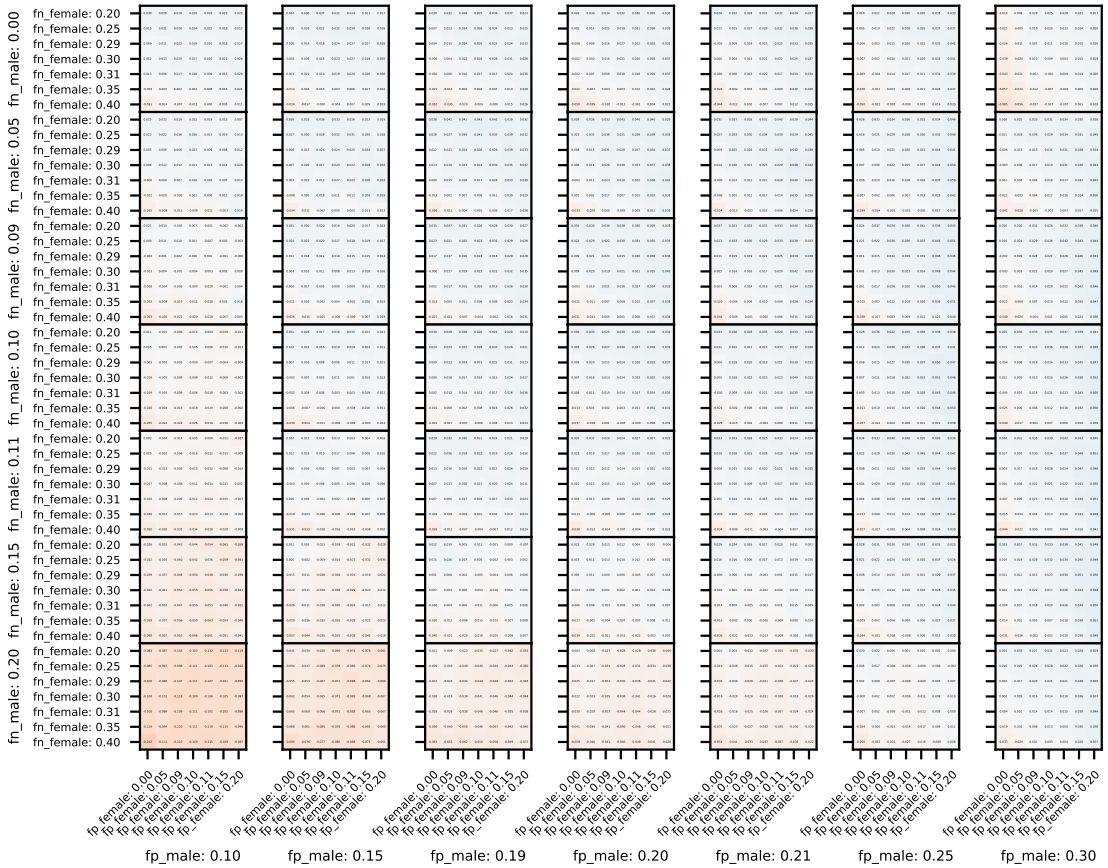


Figura 4.9: Melhora relativa utilizando correção com diferentes matrizes de transição para taxas de corrupção fixadas.



Capítulo 5

Conclusões

O presente trabalho mostra que a abordagem de tratar o problema injustiça em aprendizado de máquina como um problema de ruído natural em classe é promissora. Experimentos realizados no *dataset Adult Income* mostram que distorções introduzidas artificialmente puderam ser revertidas de maneira satisfatória através de técnicas de correção de função de custo. Entretanto, a abordagem somente foi possível pois pelo fato das distorções terem sido introduzidas artificialmente, suas distribuições de probabilidade são conhecidas, isto é, sabemos com que frequência as distorções ocorrem de acordo com a classe verdadeira e o atributo sensível.

Restam ainda avanços importantes a serem alcançados para validação da proposta. Primeiramente, o *dataset* utilizado para introdução de viés social já continha naturalmente seu próprio viés, sendo necessária a avaliação dos resultados encontrados neste estudo em circunstâncias onde o conjunto de dados original é justo, ou ao menos, próximo de ser justo. Para tanto, duas abordagens são possíveis: repetir o experimento em um *dataset* naturalmente justo e/ou gerar artificialmente um *dataset* que seja ao mesmo tempo justo e alinhado com a realidade.

Outro ponto fundamental para uso da técnica em problemas reais é a necessidade de se conhecer *a priori* a distribuição de probabilidade com a qual as injustiças ocorrem, informação que raramente está disponível. Assim, para o sucesso desta abordagem torna-se indispensável que esta distribuição de probabilidade seja estimada, e desta forma a correção da função de custo seja feita corretamente. Este problema mostra-se o maior desafio para a tese até o presente momento.

Tendo-se a capacidade de estimar as distribuições de probabilidade da injustiça em dados reais, a técnica podeira ser utilizada para corrigir o viés negativo existente. Neste ponto, métricas de injustiça deverão ser aferidas para comparação do desempenho com as técnicas disponível na literatura de aprendizado de máquina justo.

5.1 Cronograma e Metas

Abaixo, apresentamos uma lista resumida de atividades previstas até o termino doutorado. Embora a presente pesquisa seja desenvolvida na modalidade tempo parcial, e portanto sem bolsa, nos baseamos na decisão da CAPES de prorrogar as bolsas em até 6 meses por conta de pandemia para estimar um tempo total de duração de 4 anos e meio, 6 meses a mais da previsão inicial. Assim, estimamos a conclusão da pesquisa até setembro de 2021.

- A1: Aferição de métricas de *Fairness* nos experimentos com correção.
- A2: Busca e/ou construção de *datasets* justos.
- A3: Realização dos experimentos propostos no presente trabalho em conjuntos de dados justos.
- A4: Elaboração de um artigo para congresso (preferencialmente online) com os resultados obtidos.
- A5: Avaliação experimental da correção *backward*.
- A6: Revisão Bibliográfica: atualizar revisão sistemática da literatura de aprendizado justo.
- A7: Revisão Bibliográfica Específica: encontrar outras propostas de correção de função de custo para aprendizado justo.
- A8: Revisão Bibliográfica Específica: pesquisar técnicas de correção de função de custo e estimadores de matriz de transição.
- A9: Desenvolver e avaliar técnicas para estimar a matriz de transição.
- A10: Formular nova proposta de aprendizado justo utilizando o estimador de matriz de transição e correção de função de custo.
- A11: Avaliação experimental da proposta comparando com técnicas competitivas da literatura de aprendizado justo.
- A12: Elaboração de um artigo com os resultados obtidos.
- A13: Escrita da tese e preparação para defesa.

Tabela 5.1: Cronograma das tarefas que serão realizadas.

	A1	A2	A3	A4	A5	A6	A7	A8	A9	A10	A11	A12	A13
10/2020	•	•	•	•									
11/2020		•	•	•									
12/2020				•	•	•							
01/2021						•	•	•	•				
02/2021							•	•	•	•			
03/2021									•	•	•		
04/2021										•	•	•	
05/2021											•	•	
06/2021						•	•	•			•		•
07/2021						•					•		•
08/2021											•		•
09/2021													•

Referências Bibliográficas

FRENAY, B., VERLEYSEN, M. "Classification in the presence of label noise: A survey", *IEEE Transactions on Neural Networks and Learning Systems*, v. 25, n. 5, pp. 845–869, 2014. ISSN: 21622388. doi: 10.1109/TNNLS.2013.2292894.

PEDRESCHI, D., RUGGIERI, S., TURINI, F. "Discrimination-aware data mining", *Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD 08*, p. 560, 2008. ISSN: 0309-0167 (Print). doi: 10.1145/1401890.1401959. Disponível em: <`http://dl.acm.org/citation.cfm?doid=1401890.1401959`>.

KAMIRAN, F., CALDERS, T. "Classifying without discriminating", *2009 2nd International Conference on Computer, Control and Communication, IC4 2009*, 2009. doi: 10.1109/IC4.2009.4909197.

AYRES, I. "Outcome Tests of Racial Disparities in Police Practices", 2002. Disponível em: <`https://journals.sagepub.com/doi/pdf/10.3818/JRP.4.1.2002.131`>.

SQUIRES, G. D. "Racial Profiling, Insurance Style: Insurance Redlining and the Uneven Development of Metropolitan Areas", *Journal of Urban Affairs*, v. 25, n. 4, pp. 391–410, 2003.

PEDRESCHI, D., RUGGIERI, S., TURINI, F. "Measuring Discrimination in Socially-Sensitive Decision Records", *Proceedings of the 2009 SIAM International Conference on Data Mining*, pp. 581–592, 2009. doi: 10.1137/1.9781611972795.50. Disponível em: <`http://epubs.siam.org/doi/abs/10.1137/1.9781611972795.50`>.

CALDERS, T., KAMIRAN, F., PECHENIZKIY, M. "Building classifiers with independency constraints", *ICDM Workshops 2009 - IEEE International Conference on Data Mining*, pp. 13–18, 2009. doi: 10.1109/ICDMW.2009.83.

CALDERS, T., VERWER, S. "Three naive Bayes approaches for discrimination-free classification", *Data Mining and Knowledge Discovery*, v. 21, n. 2, pp. 277–292, 2010. ISSN: 13845810. doi: 10.1007/s10618-010-0190-x.

KAMIRAN, F., CALDERS, T., PECHENIZKIY, M. "Discrimination aware decision tree learning", *Proceedings - IEEE International Conference on Data Mining, ICDM*, pp. 869–874, 2010. ISSN: 15504786. doi: 10.1109/ICDM.2010.50.

DWORK, C., HARDT, M., PITASSI, T., etãl. *Fairness Through Awareness*. Relatório técnico, 2012. Disponível em: <https://arxiv.org/pdf/1104.3913.pdf>.

KAMIRAN, F., KARIM, A., ZHANG, X. "Decision theory for discrimination-aware classification", *Proceedings - IEEE International Conference on Data Mining, ICDM*, pp. 924–929, 2012. ISSN: 15504786. doi: 10.1109/ICDM.2012.45.

KAMISHIMA, T., AKAHO, S., ASOH, H., etãl. "Fairness-aware classifier with prejudice remover regularizer", *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, v. 7524 LNAI, n. PART 2, pp. 35–50, 2012. ISSN: 03029743. doi: 10.1007/978-3-642-33486-3_3.

SAINDANE, H. M., KOLHE, V. L. "Survey of Approaches for Discrimination Prevention in Data Mining", v. 5, n. 6, pp. 8114–8118, 2014.

BAROCAS, S., SELBST, A. D. "Big Data's Disparate Impact", *Ssrn*, v. 671, pp. 671–732, 2016. ISSN: 9780262327343. doi: 10.2139/ssrn.2477899.

ZAFAR, M. B., VALERA, I., RODRIGUEZ, M. G., etãl. "Fairness Constraints: Mechanisms for Fair Classification", v. 54, 2017a. ISSN: 15523098. doi: 10.1109/TRO.2009.2019886. Disponível em: <http://arxiv.org/abs/1507.05259>.

ZAFAR, M. B., VALERA, I., RODRIGUEZ, M. G., etãl. "Fairness Beyond Disparate Treatment & Disparate Impact: Learning Classification without Disparate Mistreatment", 2017b. doi: 10.1145/3038912.3052660. Disponível em: <http://arxiv.org/abs/1610.08452http://dx.doi.org/10.1145/3038912.3052660>.

KEARNS, M., NEEL, S., ROTH, A., etãl. "Preventing Fairness Gerrymandering: Auditing and Learning for Subgroup Fairness", nov 2018a. Disponível em: <http://arxiv.org/abs/1711.05144>.

KEARNS, M., SCHAPIRE, R. E., SELLIE, L. "Toward efficient agnostic learning", *Machine Learning*, v. 17, pp. 115–141, 1994.

KEARNS, M., NEEL, S., ROTH, A., etãl. *An Empirical Study of Rich Subgroup Fairness for Machine Learning*. Relatório técnico, 2018b. Disponível em: <https://arxiv.org/pdf/1808.08166.pdf>.

CORBETT-DAVIES, S., GOEL, S., CHOHLAS-WOOD, A., etãl. *The Measure and Mismeasure of Fairness: A Critical Review of Fair Machine Learning **. Relatório técnico, 2018. Disponível em: <https://arxiv.org/pdf/1808.00023.pdf>.

SIMOIU, C., CORBETT-DAVIES, S., GOEL, S. "The problem of infra-marginality in outcome tests for discrimination", *Annals of Applied Statistics*, v. 11, n. 3, pp. 1193–1216, 2017. ISSN: 19417330. doi: 10.1214/17-AOAS1058.

BUOLAMWINI, J., GEBRU, T. "Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification", *Conference on Fairness, Accountability, and Transparency*, 2018. ISSN: 11786930. doi: 10.2147/OTT.S126905.

FRIEDLER, S. A., SCHEIDEGGER, C., VENKATASUBRAMANIAN, S., etãl. "A comparative study of fairness-enhancing interventions in machine learning * ACM Reference Format", 2019. doi: 10.1145/3287560.3287589. Disponível em: <https://doi.org/10.1145/3287560.3287589>.

PASSI, S., BAROCAS, S. "Problem Formulation and Fairness", jan 2019. doi: 10.1145/3287560.3287567. Disponível em: <http://arxiv.org/abs/1901.02547http://dx.doi.org/10.1145/3287560.3287567>.

HUTCHINSON, B., MITCHELL, M. "50 Years of Test (Un)fairness: Lessons for Machine Learning", 2019. doi: 10.1145/3287560.3287600. Disponível em: <https://doi.org/10.1145/3287560.3287600>.

MEHRABI, N., MORSTATTER, F., SAXENA, N., etãl. *A Survey on Bias and Fairness in Machine Learning*. Relatório técnico, 2019. Disponível em: <https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing>.

PATRINI, G., NIELSEN, F., RICHARD NOCK, L., etãl. *Loss Factorization, Weakly Supervised Learning and Label Noise Robustness*. Relatório técnico, 2016. Disponível em: <http://proceedings.mlr.press/v48/patrini16.pdf>.

PATRINI, G., ROZZA, A., MENON, A. K., etãl. "Making deep neural networks robust to label noise: A loss correction approach", *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, v. 2017-Janua, pp. 2233–2241, 2017. doi: 10.1109/CVPR.2017.240.

BRAIDA DO CARMO, F. *Considerando o ruído no aprendizado de modelos preditivos robustos para a filtragem colaborativa.* Tese de Doutorado, Universidade Federal do Rio de Janeiro, 2017. Disponível em: <`https://www.cos.ufrj.br/uploadfile/publicacao/2871.pdf`>.

VIRTANEN, P., GOMMERS, R., OLIPHANT, T. E., etãl. "SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python", *Nature Methods*, 2020. doi: https://doi.org/10.1038/s41592-019-0686-2.

VAN DER WALT, S., COLBERT, S. C., VAROQUAUX, G. "The NumPy Array: A Structure for Efficient Numerical Computation", *Computing in Science Engineering*, v. 13, n. 2, pp. 22–30, March 2011. ISSN: 1558-366X. doi: 10.1109/MCSE.2011.37.

HUNTER, J. D. "Matplotlib: A 2D Graphics Environment", *Computing in Science Engineering*, v. 9, n. 3, pp. 90–95, May 2007. ISSN: 1558-366X. doi: 10.1109/MCSE.2007.55.

MCKINNEY, W. "Data Structures for Statistical Computing in Python". In: van der Walt, S., Millman, J. (Eds.), *Proceedings of the 9th Python in Science Conference*, pp. 51 – 56, 2010.

PEDREGOSA, F., VAROQUAUX, G., GRAMFORT, A., etãl. "Scikit-learn: Machine Learning in Python ", *Journal of Machine Learning Research*, v. 12, pp. 2825–2830, 2011.

ABADI, M., AGARWAL, A., BARHAM, P., etãl. "TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems". 2015. Disponível em: <`http://tensorflow.org/`>. Software available from tensorflow.org.

CHOLLET, F., OTHERS. "Keras". `https://github.com/fchollet/keras`, 2015.

JAYNES, E. T. "Prior Probabilities", *IEEE Transactions on Systems Science and Cybernetics*, v. 4, n. 3, pp. 227–241, Sep. 1968. ISSN: 2168-2887. doi: 10.1109/TSSC.1968.300117.

KINGMA, D., BA, J. "Adam: A method for stochastic optimization", *arXiv:1412.6980 [cs.LG]*, pp. 1–15, 2014. Disponível em: <`http://arxiv.org/abs/1412.6980`>.