Universidade Federal de Juiz de Fora
Instituto de Ciências Exatas
Bacharelado em Ciência da Computação

Analisando a frente de Pareto para encontrar uma solução para regressão Ridge

Vitor Monteiro Andrade Goulart

JUIZ DE FORA
JULHO, 2023

Analisando a frente de Pareto para encontrar uma solução para regressão Ridge

Vitor Monteiro Andrade Goulart

Universidade Federal de Juiz de Fora
Instituto de Ciências Exatas
Departamento de Ciência da Computação
Bacharelado em Ciência da Computação

Orientador: Saulo Moraes Villela
Coorientador: Wilhelm Passarella Freire

JUIZ DE FORA
JULHO, 2023

Analisando a frente de Pareto para encontrar uma solução para regressão Ridge

Vitor Monteiro Andrade Goulart

MONOGRAFIA SUBMETIDA AO CORPO DOCENTE DO INSTITUTO DE CIÊNCIAS EXATAS DA UNIVERSIDADE FEDERAL DE JUIZ DE FORA, COMO PARTE INTEGRANTE DOS REQUISITOS NECESSÁRIOS PARA A OBTENÇÃO DO GRAU DE BACHAREL EM CIÊNCIA DA COMPUTAÇÃO.

Aprovada por:

Saulo Moraes Villela
Doutor em Engenharia de Sistemas e Computação

Wilhelm Passarella Freire
Doutor em Engenharia de Sistemas e Computação

Heder Soares Bernardino
Doutor em Modelagem Computacional

Carlos Cristiano Hasenclever Borges
Doutor em Engenharia Civil

Sandro Rodrigues Mazorche
Doutor em Engenharia Mecânica

JUIZ DE FORA
11 DE JULHO, 2023

Resumo

As regressões Ridge e Lasso são casos especiais de regressão linear. Com base na teoria da otimização multiobjetivo, ambas podem ser vistas como problemas de otimização biobjetivo. As frentes de Pareto resultantes desse problema oferecem uma variedade de modelos de regressão dentre os quais uma solução ideal pode ser selecionada, de acordo com uma estratégia de escolha pré-determinada entre as soluções não dominadas. Neste trabalho, foi utilizado um algoritmo para gerar pontos na frente de Pareto da regressão Ridge, considerada como um problema biobjetivo, e propôs-se uma heurística para a seleção de um ponto ótimo na frente de Pareto. A heurística proposta consiste em verificar se o ponto mais próximo ao ponto ideal apresenta soluções satisfatórias em relação a métricas de erro como MSE e MAE. Os resultados apontam para uma direção favorável a essa hipótese, visto que, em experimentos realizados em quatro conjuntos de dados, a escolha dos pontos desta maneira proporciona soluções com erro reduzido.

Palavras-chave: Otimização multiobjetivo, Regressão Ridge, Regressão Lasso, Frente de Pareto, Fronteira de Pareto.

Abstract

Ridge and Lasso regressions are special cases of linear regression. Based on multiobjective optimization theory, both can be seen as bi-objective optimization problems. The resulting Pareto fronts from this problem offer a range of regression models from which an ideal solution can be selected, according to a predetermined choice strategy among non-dominated solutions. In this work, an algorithm was used to generate points on the Pareto front of the Ridge regression, considered as a bi-objective problem, and a heuristic for the selection of an optimal point on the Pareto front was proposed. The proposed heuristic consists of checking if the point closest to the ideal point presents satisfactory solutions in relation to error metrics such as MSE and MAE. The results point in a favorable direction to this hypothesis, since, in experiments carried out on four data sets, the choice of points in this way provides solutions with reduced error.

Keywords: Multiobjective Optimization, Ridge Regression. Lasso Regression. Pareto Front, Pareto Frontier.

Agradecimentos

A Deus em primeiro lugar, por ter me dado saúde e força para superar as dificuldades.

Aos meus pais, irmão e minha namorada pelo amor, incentivo e apoio incondicional durante o meu percurso acadêmico.

Aos meus familiares e amigos que me apoiaram durante esse processo.

Aos meus orientadores Saulo Moraes Villela e Wilhelm Passarella Freire, pelo seu suporte, pelas suas correções e incentivos na elaboração deste trabalho.

Agradeço também aos membros da banca de avaliação final por aceitarem debater minha pesquisa.

Conteúdo

Lista de Figuras

Lista de Tabelas

1 Introdução

Otimização multiobjetivo é uma área da pesquisa operacional que lida com a otimização de problemas envolvendo duas ou mais funções objetivo conflitantes (DEB; DEB, 2013) Esses problemas surgem em muitas aplicações práticas, onde diferentes aspectos de uma solução precisam ser considerados simultaneamente. Em termos intuitivos, funções objetivo conflitantes são aquelas que, dependendo das mesmas variáveis, apresentam *trade-offs*, ou seja, a melhoria em uma delas geralmente resulta na piora da(s) outra(s).

Os problemas de otimização multiobjetivo têm um papel importante em diversos campos, como engenharia, economia, biologia, logística, entre outros. Eles são caracterizados por apresentarem múltiplos objetivos, que são conflitantes, e um conjunto de soluções viáveis, que atendem a todas as restrições do problema. O conceito de dominância de Pareto é frequentemente usado para comparar soluções viáveis e identificar aquelas que são preferíveis. Neste contexto, o "ponto ideal" na frente de Pareto é conceituado como a solução teórica que otimiza todos os objetivos simultaneamente (MARLER; ARORA, 2004). No entanto, muitas vezes este ponto é inatingível devido à natureza conflitante dos objetivos.

Na engenharia, por exemplo, problemas de projeto podem envolver a minimização do custo de fabricação e a maximização do desempenho de um produto. Na economia, pode ser necessário equilibrar a inflação e o desemprego. Na biologia, pode ser importante encontrar um equilíbrio entre a preservação do habitat e o desenvolvimento de recursos. Na logística, pode ser necessário minimizar o tempo de entrega e maximizar a satisfação do cliente.

A importância da Otimização Multiobjetivo é evidente, uma vez que muitos problemas do mundo real são melhor modelados considerando várias funções objetivo em conflito (DEB; DEB, 2013). Um exemplo disso é a possibilidade de transformar problemas de regressão linear Ridge e Lasso em problemas de Otimização Multiobjetivo.

1.1 Justificativa

Dada a relevância do problema de otimização multiobjetivo em diversas áreas, a geração de soluções que formam a frente de Pareto e a identificação dos melhores pontos é uma tarefa essencial. Estimar quais pontos representam as melhores soluções pode diminuir o tempo e o processamento computacional necessários para a identificação das soluções mais adequadas. Além disso, proporcionar um conjunto de soluções preferíveis permite aos tomadores de decisão avaliar as várias opções e escolher aquela que melhor atenda às suas necessidades e preferências.

Nos últimos anos, muitos algoritmos heurísticos e metaheurísticos foram desenvolvidos para resolver problemas de otimização multiobjetivo, incluindo algoritmos genéticos multiobjetivo (COELLO; LAMONT; VELDHUIZEN, 2007), otimização por enxame de partículas (ZHOU et al., 2011), otimização por colônia de formigas (DORIGO; STÜTZLE, 2019), entre outros. Esses algoritmos têm demonstrado sucesso em encontrar soluções de alta qualidade para uma ampla gama de problemas de otimização multiobjetivo.

1.2 Hipóteses

Se as funções objetivo não fossem conflitantes, a solução ideal seria aquela que minimiza isoladamente cada uma das funções. No entanto, quando as funções são conflitantes, surge uma pergunta: existe alguma relação entre a proximidade do melhor ponto ou dos melhores pontos da frente de Pareto com o ponto ideal? O ponto ideal é formado pela combinação das coordenadas dos ótimos de cada uma das funções objetivo consideradas isoladamente. Se essa relação existir, seria possível identificar os melhores pontos da frente de Pareto com base na proximidade a esse ponto ideal.

A hipótese central deste trabalho é que o ponto da frente de Pareto que guarda a menor distância do ponto ideal pode ser uma boa indicação de um bom modelo para o problema. Se essa hipótese for confirmada, seria possível propor uma metodologia para identificar os melhores pontos da frente de Pareto com base em sua proximidade ao ponto ideal, o que poderia acelerar a escolha do melhor modelo para regressão Ridge.

1.3 Objetivos

O objetivo principal deste trabalho é analisar uma heurística para a identificação dos melhores pontos na frente de Pareto, o que não é uma tarefa trivial, baseando-se em quão próximos eles estão do ponto ideal. Este ponto ideal é determinado pela combinação das coordenadas ótimas de cada função objetivo quando avaliadas individualmente. Para realizar isso, utilizamos um algoritmo de otimização conhecido como NFDA, que produz a frente de Pareto para regressões Ridge. Esse algoritmo foi utilizado por ter uma estrutura linear e por ter apresentado bons resultados em problemas de otimização.

Os testes desse trabalho foram concentrados na regressão Ridge. Durante o estudo, conduzimos uma série de experimentos para verificar a hipótese proposta. Com as soluções obtidas da frente de Pareto, calculamos a proximidade dos pontos de solução ao ponto ideal. Também avaliamos a qualidade dessas soluções usando diferentes métricas de erro, permitindo uma análise completa e precisa do problema.

Os objetivos específicos incluem:

- Revisar a literatura sobre otimização multiobjetivo e problemas de regressão Ridge;
- Desenvolver uma metodologia para estimar a proximidade dos pontos na frente de Pareto ao ponto ideal;
- Implementar e testar a metodologia proposta usando diferentes métricas de distância;
- Analisar os resultados obtidos.

Espera-se que a investigação deste trabalho possa contribuir para o avanço do conhecimento na área de otimização multiobjetivo e fornecer uma ferramenta útil para a identificação dos melhores pontos na frente de Pareto em problemas práticos. Além disso, o trabalho pode servir como base para futuras pesquisas e desenvolvimento de algoritmos e metodologias aprimoradas para resolver problemas de otimização multiobjetivo.

1.4 Organização do Trabalho

Este trabalho está estruturado em 6 capítulos, sendo este o primeiro.

No Capítulo 2 estabelece-se a fundamentação teórica necessária para um entendimento completo do trabalho. Definições de regressão linear, Ridge e Lasso são apresentadas, juntamente com a explicação de como essas regressões podem ser resolvidas como um problema de otimização sem restrição. Para isso, o Método das Somas Ponderadas e o Método da ϵ-Restrição são detalhados, juntamente com suas equivalências.

O Capítulo 3 apresenta o algoritmo NFDA, incluindo detalhes de como sua direção de busca é determinada e como o algoritmo funciona, terminando com a apresentação de seu pseudocódigo.

O Capítulo 4 detalha a metodologia empregada para desenvolver o presente estudo.

No Capítulo 5 são apresentados os resultados dos experimentos conduzidos, com a aplicação da regressão Ridge em quatro conjuntos de dados, além de uma análise do desempenho dos pontos mais próximos ao ponto ideal.

Finalmente, o Capítulo 6 sintetiza as conclusões tiradas do estudo como um todo.

2 Fundamentação Teórica

A modelagem de dados é fundamental em várias disciplinas científicas e setores comerciais. Dentro deste domínio, a regressão linear tem sido uma ferramenta essencial por muitos anos, permitindo que os pesquisadores examinem as relações entre variáveis e façam previsões informadas (HASTIE; TIBSHIRANI; FRIEDMAN, 2009a). No entanto, a regressão linear, por mais útil que seja, tem suas limitações. Para abordar essas limitações, técnicas como a regressão Ridge e Lasso foram desenvolvidas (HOERL; KENNARD, 1970; TIBSHIRANI, 1996). Neste capítulo, discutiremos essas técnicas, sua importância, suas vantagens e suas limitações.

2.1 Regressão Linear

A regressão linear é uma técnica de modelagem estatística que permite estimar a relação entre uma variável dependente e uma ou mais variáveis independentes (HASTIE; TIBSHIRANI; FRIEDMAN, 2009a). Ela tem sido a base para muitas análises estatísticas e é conhecida por sua simplicidade e eficácia. A regressão linear, no entanto, tem suas falhas. Por exemplo, pode sofrer *overfitting*, especialmente quando temos um grande número de variáveis independentes. Além disso, se as variáveis independentes estão altamente correlacionadas, um problema conhecido como multicolinearidade, as estimativas dos coeficientes de regressão podem se tornar instáveis e difíceis de interpretar.

No modelo de regressão linear, a variável dependente é expressa como uma combinação linear das variáveis independentes, junto com um termo de erro. Sua formulação pode ser expressa na forma:

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \varepsilon_i = x_i^T \beta + \varepsilon_i,\ i = 1, \ldots, n, \tag{2.1}$$

onde $x_i^T \beta$ denota o produto interno entre o vetor $x_i = (x_{i1}, \ldots, x_{ip})$ transposto e o vetor de parâmetros $\beta = (\beta_1, \ldots, \beta_p)$. Nesse contexto, x_i representa um ponto i do conjunto

de dados que se está modelando e no qual para cada ponto x_i, que contem as variáveis independentes, temos um ponto y_i correspondente com a variável dependente ou resposta esperada para o conjunto x_i. Outra forma de escrever esse problema para um conjunto de dados com n amostras e p variáveis é:

$$y = X\beta + \varepsilon, \tag{2.2}$$

onde

$$y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, X = \begin{bmatrix} x_1^T \\ x_2^T \\ \vdots \\ x_n^T \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & \dots & x_{1p} \\ 1 & x_{21} & \dots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \dots & x_{np} \end{bmatrix}, \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix}, \varepsilon = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}, \tag{2.3}$$

sendo y a variável resposta, variável dependente ou valores observados, X é a matriz com as amostras contendo as variáveis explicativas ou variáveis independentes. O valor 1 na primeira coluna da matriz X aparece para resultar no termo constante da equação chamado de intercepto ou coeficiente linear. β é o vetor dos parâmetros ou coeficientes da regressão que queremos encontrar. Por fim, ε é o termo de erro que representa o erro cometido pelo nosso modelo.

A solução do ajuste da equação da regressão consiste em, dado um conjunto de dados, estimar os coeficientes β de modo a minimizar o termo $\varepsilon = y - X\beta$. Por exemplo, pode-se utilizar a soma dos erros ao quadrado $\|\varepsilon\|_2^2$ como medida que se deseja minimizar.

2.2 Regressão Ridge e Lasso

A regressão Ridge, introduzida por Hoerl e Kennard (1970), é uma extensão da regressão linear que introduz um termo de regularização. Esse método adiciona uma penalidade ao quadrado dos coeficientes à função de custo. Esta penalidade tem como objetivo reduzir a magnitude dos coeficientes, levando a uma simplificação do modelo. Este procedimento torna a regressão Ridge especialmente útil em cenários de multicolinearidade, um fenômeno onde as variáveis independentes estão altamente correlacionadas entre si. Isso pode levar à instabilidade nas estimativas dos coeficientes na regressão linear simples,

e a regressão Ridge pode proporcionar uma solução mais estável e robusta. Contudo, a regressão Ridge não é uma solução universal, ou seja, não é a resposta para todos os problemas de modelagem e tem suas próprias limitações, como a escolha apropriada do hiperparâmetro de regularização. Este hiperparâmetro controla a severidade da penalidade aplicada aos coeficientes, determinando assim o grau de "encolhimento" dos coeficientes. Ou seja, a medida em que os coeficientes são diminuídos para evitar o overfitting e criar um modelo mais robusto e menos complexo.

A formulação matemática da regressão Ridge é dada por:

$$\hat{\beta}^{ridge} = \underset{\beta}{\operatorname{argmin}} \left\{ \sum_{i=1}^{n} (y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \right\}, \tag{2.4}$$

onde $\lambda \geq 0$ é o hiperparâmetro de *tuning* ou de regularização que controla a força ou severidade da penalidade.

A regressão Lasso (*Least Absolute Shrinkage and Selection Operator*), introduzida por Tibshirani (1996), é uma extensão da regressão linear que introduz um termo de penalidade na função de custo. Diferentemente da regressão Ridge, a penalidade na regressão Lasso é proporcional ao valor absoluto dos coeficientes. Esta particularidade pode levar alguns coeficientes a se tornarem zero, resultando na exclusão das correspondentes variáveis independentes do modelo. Essa capacidade de excluir variáveis torna a regressão Lasso uma ferramenta útil na redução da complexidade do modelo e na prevenção de *overfitting*, especialmente quando se trabalha com conjuntos de dados de alta dimensionalidade.

A formulação matemática da regressão Lasso é dada por:

$$\hat{\beta}^{lasso} = \underset{\beta}{\operatorname{argmin}} \left\{ \sum_{i=1}^{n} (y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right\}, \tag{2.5}$$

onde, assim como na regressão Ridge, $\lambda \geq 0$ é um hiperparâmetro de *tuning* que controla a força da penalidade.

Os problemas Ridge e Lasso podem ser expressos de uma outra forma como um

problema de otimização com restrição

$$\begin{cases} \text{minimizar} & \|Ax-b\|_2^2 \\ \text{sujeito a} & \|x\|_q^q \leq t \end{cases}, \; x \in \mathbb{R}^p \tag{2.6}$$

onde A é uma matriz $n \times p$, b é um vetor de dimensão n, $t > 0$ e temos um problema de regressão Ridge quando $q = 2$ e Lasso quando $q = 1$.

O problema descrito em (2.6) é um problema restrito e convexo. Além disso, no caso do Ridge é um problema diferenciável e no Lasso é não diferenciável. Se denotarmos ls como sendo a solução da regressão linear com os mínimos quadrados, para um $0 < t < \|ls\|_q^q$, a solução para esse problema se localiza na fronteira da esfera $\{x \in \mathbb{R}^n : ||x||_q^q = t\}$. Quando $t \geq \|ls\|_q^q$, a solução é exatamente ls. Por outro lado, se $t = 0$, a solução é o vetor nulo. O vetor nulo e ls geram os chamados pontos extremos da frente de Pareto (a definição será feita na seção 2.6). A medida que t varia entre 0 e $\|ls\|_q^q$, as soluções de (P2.1) evidenciam um *trade-off* entre o viés e a variância dos modelos (CHARKHGARD; ESHRAGH, 2019).

Associado ao problema da Equação (2.6) está a sua forma Lagrangiana semelhante a apresentada anteriormente

$$\text{minimizar} \quad \|Ax-b\|_2^2 + \lambda\|x\|_q^q \,, \; \lambda \geq 0. \tag{2.7}$$

A Dualidade Lagrangiana (OSBORNE; PRESNELL; TURLACH, 2000; HIRIART-URRUTY; LEMARECHAL, 1993) garante uma correspondência biunívoca entre o problema original (2.6) e sua forma Lagrangiana (2.7), isto é, para cada t existe um λ que leva à mesma solução e vice-versa. Muitas vezes é mais vantajoso resolver na forma Lagrangiana em vez de sua forma de otimização com restrição. No entanto, λ, assim como t, devem ser definidos antes de minimizar o Lagrangiano, o que não é uma tarefa trivial. Esse ajuste prévio pode ser contornado por meio da Otimização Multiobjetivo.

Em resumo, a regressão linear Ridge e Lasso são técnicas poderosas e versáteis para modelar a relação entre uma variável dependente e uma ou mais variáveis independentes. Cada um tem suas vantagens e desvantagens, e a escolha entre eles geralmente

depende das características específicas do problema em questão. Em particular, a regressão Ridge pode ser útil quando existe multicolinearidade, enquanto a regressão Lasso pode ser preferida quando há um grande número de variáveis independentes e acreditamos que apenas um subconjunto delas seja relevante.

2.3 Métricas de Erro

Várias métricas são utilizadas para avaliar o desempenho das soluções da frente de Pareto nos problemas de regressão Ridge que vamos resolver nesse trabalho. As quatro métricas utilizadas são detalhadas abaixo.

O coeficiente de determinação, também conhecido como R quadrado (*R Squared* – R^2), fornece uma medida de quão bem as previsões do modelo se ajustam aos verdadeiros valores. Ele é definido como:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}, \tag{2.8}$$

onde y_i é o valor verdadeiro, $\hat{y}_i$ é o valor previsto, e $\bar{y}$ é a média dos valores verdadeiros.

O erro absoluto médio (*mean absolute error* – MAE) é a média das diferenças absolutas entre os valores verdadeiros e previstos. Ele é definido como:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|. \tag{2.9}$$

O erro quadrático médio (*mean squared error* – MSE) é a média dos quadrados das diferenças entre os valores verdadeiros e previstos. Ele é definido como:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2. \tag{2.10}$$

O erro percentual absoluto médio (*mean absolute percentage error* – MAPE) é a média dos valores absolutos das diferenças percentuais entre os valores verdadeiros e

previstos. Ele é definido como:

$$MAPE = \frac{100\%}{n} \sum_{i=1}^{n} \left| \frac{y_i - \hat{y}_i}{y_i} \right|. \tag{2.11}$$

2.4 Métricas de Distância

A principal hipótese deste estudo é que o ponto mais próximo do ponto ideal é uma boa escolha entre os pontos da frente de Pareto para problemas de otimização biobjetivo em problemas de regressão.

Foram consideradas 6 diferentes medidas para o cálculo da distância dos pontos para o ponto ideal. Cada distância tem suas características próprias e pode ser mais adequada para diferentes situações. As distâncias ou normas utilizadas para o cálculo das distâncias foram:

- Norma 1 (Norma L_1 ou Manhattan): Soma dos valores absolutos das diferenças entre as coordenadas dos pontos. Fórmula geral:

$$||\mathbf{x}||_1 = \sum_{i=1}^{n} |x_i|. \tag{2.12}$$

- Norma 2 (Norma L_2 ou Euclidiana): Raiz quadrada da soma dos quadrados das diferenças entre as coordenadas dos pontos. Fórmula geral:

$$||\mathbf{x}||_2 = \sqrt{\sum_{i=1}^{n} x_i^2}. \tag{2.13}$$

- Norma infinito (L_∞): Valor máximo dos valores absolutos das coordenadas dos pontos. Fórmula geral:

$$||\mathbf{x}||_\infty = \max_{1 \leq i \leq n} |x_i|. \tag{2.14}$$

- Distância ou similaridade de Hassanat (HASSANAT, 2014) (HasD): Fórmula geral:

$$d_{\text{HasD}}(x, y) = \sum_{i=1}^{n} \begin{cases} 1 - \frac{1+\min(x_i,y_i)}{1+\max(x_i,y_i)}, & \text{se } \min(x_i, y_i) \geq 0, \\ 1 - \frac{1}{1+\max(x_i,y_i)+|\min(x_i,y_i)|}, & \text{se } \min(x_i, y_i) < 0, \end{cases} \tag{2.15}$$

onde x e y são os vetores de pontos.

- Distância Lorentzian (LD): Fórmula geral:

$$d_{\mathrm{LD}}(x, y) = \sum_{i=1}^{n} \ln(1 + |x_i - y_i|), \tag{2.16}$$

onde x e y são os vetores de pontos.

- Norma Nuclear (Nuc): Obtida através da soma dos valores singulares da matriz obtida a partir do vetor transformado em uma matriz diagonal. O cálculo dessa norma no trabalho foi realizado utilizando a biblioteca de algebra linear no *numpy* na linguagem *Python* (numpy.linalg).

2.5 Definições Básicas

Nesta seção, apresentaremos algumas definições essenciais para facilitar a compreensão do capítulo dedicado ao algoritmo NFDA.

Definição 2.5.1. O epígrafo de uma função $f : X \subset \mathbb{R}^n \to \mathbb{R}$ é o conjunto de pares ordenados (x, y) em $\mathbb{R}^n \times \mathbb{R}$, onde y é maior ou igual ao valor da função f em x. Pode ser denotado por $\mathrm{epi}(f)$ e definido por

$$\mathrm{epi}(f) = \{(x, y) \in X \times \mathbb{R} \,|\, y \geq f(x)\} \tag{2.17}$$

Definição 2.5.2. O Interior do Epígrafo de uma função $f : X \subset \mathbb{R}^n \to \mathbb{R}$ é dado por

$$(\mathrm{epi}(f))^0 = \{(x, y) \in X \times \mathbb{R} \,|\, y > f(x)\} \tag{2.18}$$

Definição 2.5.3. O Subdiferencial de uma função $f : D \subset \mathbb{R}^n \to \mathbb{R}$ no ponto $a \in D$, denotado por $\partial f(a)$, é o conjunto de vetores que fornece os hiperplanos de apoio ao gráfico de f no ponto a. Um vetor $v \in \partial f(a)$ é chamado de Subgradiente de f em a e, para todos os pontos x em D, temos:

$$f(x) \geq f(a) + \langle v, (x - a) \rangle \tag{2.19}$$

Na otimização matemática, trabalhamos com uma função objetivo e várias restrições que delimitam a região factível de um problema. Essas restrições são geralmente expressas como desigualdades ou igualdades envolvendo uma função g_i e uma variável de decisão x.

Definição 2.5.4. Uma restrição $g_i(x) \leq 0$ é dita ativa no ponto x se a restrição é satisfeita como uma igualdade, isto é, $g_i(x) = 0$.

Em outras palavras, uma restrição é ativa se o ponto x estiver exatamente na fronteira da região factível definida pela restrição.

2.6 Otimização Multiobjetivo

A otimização multiobjetivo é um subcampo da otimização matemática que lida com problemas que envolvem mais de um objetivo a ser otimizado simultaneamente. Esses problemas são matematicamente representados da seguinte maneira:

$$\underset{x}{\text{minimizar}} \quad f(x) = (f_1(x), f_2(x), \ldots, f_k(x)) \quad x \in X, \tag{2.20}$$

onde $f : X \subset \mathbb{R}^n \to \mathbb{R}^k$ é o vetor objetivo, x é o vetor de decisão, e $X \subseteq \mathbb{R}^n$ é a região de viabilidade ou região factível. Em geral, $X = \{x \in \mathbb{R}^n | g_i(x) \leq 0, h_j(x) = 0, i = 1, \ldots, m, j = 1, \ldots, p\}$ com $g_i, h_j : \mathbb{R}^n \to \mathbb{R}$ contínuas. Quando o problema se limita a duas funções, ou seja $k = 2$, ele é chamado de problema de Otimização Biobjetivo. Neste estudo, o problema de otimização multiobjetivo analisado é o biobjetivo.

Definição 2.6.1. A Região Objetivo Viável é o conjunto de todos os vetores objetivo que podem ser alcançados ao variar x dentro do conjunto factível. Ou seja, é o conjunto $Z = f(X) = \{f(x) : x \in X\} \subset \mathbb{R}^k$. Os seus elementos $z = (z_1, z_2, \ldots, z_k) \in \mathbb{R}^k$ são chamados vetores objetivo e cada $z_i = f_i(x)$ é chamado valor objetivo.

Definição 2.6.2. Um vetor de decisão $x^* \in X$ é um Ótimo de Pareto se não existe outro vetor de decisão $x \in X$ tal que $f_i(x) \leq f_i(x^*), i = 1, 2, \ldots, k$ e $f_j(x) < f_j(x^*)$ para pelo menos um índice j. Em outras palavras, um Ótimo de Pareto é um ponto que não pode ser melhorado em nenhum dos objetivos sem piorar pelo menos um dos outros. Um vetor

objetivo $z^* \in Z$ é Ótimo de Pareto se seu vetor de decisão correspondente é um Ótimo de Pareto.

Definição 2.6.3. Um vetor de decisão $x^* \in X$ é um Ótimo Fraco de Pareto se não existe outro vetor de decisão $x \in X$ tal que $f_i(x) < f_i(x^*)$, $i = 1, 2, \ldots, k$. Em outras palavras, um Ótimo Fraco de Pareto é um ponto que não pode ser estritamente melhorado em nenhum dos objetivos. Um vetor objetivo $z^* \in Z$ é Ótimo Fraco de Pareto se seu vetor de decisão correspondente é um Ótimo Fraco de Pareto.

Note que a principal diferença entre o ótimo de Pareto e o Ótimo Fraco de Pareto está nas desigualdades. No Ótimo de Pareto temos um ponto que não pode ser melhorado (i.e., $f_i(x) \leq f_i(x^*)$) sem piorar pelo menos um dos outros objetivos, enquanto no Ótimo Fraco de Pareto, temos um ponto que não pode ser estritamente melhorado (i.e., $f_i(x) < f_i(x^*)$).

Um vetor é chamado de ponto de Pareto se é um ótimo de Pareto ou um Ótimo Fraco de Pareto.

Definição 2.6.4. A Frente de Pareto é o subconjunto da Região Objetivo Viável em que todos os pontos são Ótimos de Pareto ou Ótimos Fracos de Pareto.

Em resumo, o objetivo da otimização multiobjetivo é encontrar a Frente de Pareto. De uma forma mais específica, encontra-se uma aproximação da frente de Pareto mas por vezes nos referimos a essa aproximação da frente como a frente de Pareto em si. Isso nos fornece uma representação do *trade-off* entre os diferentes objetivos, e permite a tomada de decisão da melhor solução entre aquelas na frente com base nas preferências do decisor. Neste trabalho, o objetivo é construir essa frente de Pareto para a regressão Ridge. Cada ponto da Frente será uma solução do problema de otimização Multiobjetivo equivalente ao problema (2.6).

A técnica de escalarização (DUTTA; KAYA, 2011; PARDALOS et al., 2017; BURACHIK; KAYA; RIZVI, 2017) é frequentemente aplicada para resolver problemas de otimização multiobjetivo. Ela opera transformando a função vetorial associada ao problema multibojetivo (2.20) em uma função escalar para minimização. Iremos utilizar dois métodos de escalarização, sendo eles o Método das somas Ponderadas e o Método

da ϵ-Restrição. Será mostrada a relação entre esses métodos e a equivalência entre o problema (2.6) e sua forma Lagrangiana (2.7).

Considerando que a Regressão Ridge e Lasso podem ser vistas como problemas de otimização biobjetivo, fixamos, para simplificação, $k = 2$ nas próximas seções.

2.7 Método das Somas Ponderadas

O método das somas ponderadas é uma abordagem comum para a resolução de problemas de otimização multiobjetivo (MIETTINEN, 1998; BRANKE et al., 2008). Neste método, a função objetivo do problema multiobjetivo principal (2.20) é transformada em uma função escalar que consiste em uma soma ponderada das funções objetivo de (2.20), conforme expresso por:

$$\text{minimizar} \sum_{i=1}^{k} w_i f_i(x) \quad x \in S, \tag{2.21}$$

onde $w_i \geq 0$ para todo $i = 1, \ldots, k$, e tipicamente $\sum_{i=1}^{k} w_i = 1$.

De acordo com Miettinen (1998), as soluções do problema (2.21) são Ótimos Fraco de Pareto e, além disso, são ótimos de Pareto se $w_i > 0$ para todo $i = 1, \ldots, k$.

Apesar da simplicidade do método das somas ponderadas, ele apresenta algumas limitações. Em particular, qualquer solução ótima de Pareto pode ser encontrada ao alterar os pesos somente se o problema for convexo. Assim, pode ocorrer que algumas soluções ótimas de Pareto de problemas não convexos não possam ser encontradas, independentemente de como os pesos são selecionados.

Além disso, é importante normalizar os objetivos para que magnitudes diferentes não influenciem indevidamente o método. Um conjunto distribuído uniformemente de pesos não necessariamente produzirá uma representação distribuída uniformemente do conjunto ótimo de Pareto, mesmo se o problema for convexo.

Em suma, o método das somas ponderadas, que é conhecido e aplicado em muitos casos de otimização multiobjetivo (MARLER; ARORA, 2004), pode fornecer soluções úteis, mas é importante estar ciente de suas limitações e aplicá-lo com cautela.

Desta forma, o problema transformado na forma escalar fica:

$$\text{minimizar } w_1 f_1(x) + w_2 f_2(x), x \in X, w_1, w_2 > 0, w_1 + w_2 = 1. \tag{2.22}$$

A seguir é apresentado o teorema que garante que toda solução ótima de Pareto de um problema de otimização multiobjetivo convexo pode ser encontrado pelo método das somas ponderadas.

Teorema 2.7.1. *(Teorema 3.1.4 de Miettinen (1998))*
Suponha que o problema multiobjetivo seja convexo. Se $x^* \in X$ é ótimo de Pareto, então existem coeficientes positivos w_1, w_2 tais que x^* é uma solução do problema de ponderação (2.22).

2.8 Método da ϵ-Restrição

O Método da ϵ-Restrição é uma técnica utilizada para resolver problemas de otimização multiobjetivo (MIETTINEN, 1998). Na essência, este método transforma um problema de otimização multiobjetivo em um conjunto de problemas de otimização de objetivo único, onde todas as funções objetivo, exceto uma, são convertidas em restrições com um limite superior fixado em um valor ϵ. Ou seja, o problema (2.20) gera k problemas escalares da seguinte forma:

$$\begin{aligned} &\text{minimizar} && f_i(x) \\ &\text{sujeito a} && f_j(x) \leq \epsilon_j, \quad \forall j \neq i \\ & && x \in X. \end{aligned} \tag{2.23}$$

Considerando nosso problema com duas funções objetivo, $k = 2$, os problemas abordados pelo Método da ϵ-Restrição fica da forma:

$$\begin{aligned} &\text{minimizar} \quad f_1(x) \quad \text{sujeito a} \quad f_2(x) \leq \epsilon, x \in X \text{ e} \\ &\text{minimizar} \quad f_2(x) \quad \text{sujeito a} \quad f_1(x) \leq \epsilon, x \in X. \end{aligned} \tag{2.24}$$

De acordo com Miettinen (1998), uma solução desse problema é Ótimo Fraco de

Pareto (Teorema 3.2.1). Além disso, a vetor de decisão $x^* \in X$ é Ótimo de Pareto se e somente se for uma solução do problema acima para um valor específico de ϵ (Teorema 3.2.2). Isso é resumido pelo Teorema 2.8.1.

Teorema 2.8.1. *(Teoremas 3.2.1 e 3.2.2 de Miettinen (1998))*
(1) A solução do problema ϵ-restrição (2.24) *é Ótimo Fraco de Pareto.*
(2) Suponha que $\epsilon = f_2(x^*)$. Então o vetor de decisão $x^* \in X$ é ótimo de Pareto se e somente se for uma solução do problema (2.24).

A afirmação (2) do Teorema 2.8.1 possui um caráter mais teórico, uma vez que, na prática, não é viável conhecer o vetor x^* a priori para atribuir $\epsilon = f_2(x^*)$. No entanto, em um contexto mais específico envolvendo regressão Ridge e Lasso, observaremos que é possível fazer uma escolha adequada para ϵ de tal forma que $\epsilon = f_2(x^*)$.

2.9 Relação entre o Método das Somas Ponderadas e o Método da ϵ-Restrição

Nessa seção é explicitada a relação entre o Método das Somas Ponderadas e o Método da ϵ-Restrição.

Teorema 2.9.1. *(Teorema 3.2.5 de Miettinen (1998))*
Seja $x^* \in X$ uma solução de (2.22) *e $w_1, w_2 \geq 0$.*
(1) Se $w_1 > 0$, então x^* é uma solução de (2.24) *para f_1 como função objetivo e $\epsilon = f_2(x^*)$ ou*
(2) Se x^* é a única solução de (2.22)*, então x^* é uma solução de* (2.24) *com $\epsilon = f_2(x^*)$ e f_1 como função objetivo.*

Teorema 2.9.2. *(Teorema 3.2.6 de Miettinen (1998))*
Seja um problema de otimização multiobjetivo convexo. Se $x^* \in X$ é uma solução de (2.24) *para f_1 como função objetivo e $\epsilon = f_2(x^*)$, então existem $w_1, w_2 \geq 0$ com $w_1 + w_2 = 1$ de tal forma que x^* também seja uma solução de* (2.22).

Os Teoremas 2.9.1 e 2.9.2 apresentam a equivalência entre o Método das Somas Ponderadas e o Método da ϵ-Restrição para problemas convexos. Analisa-se os problemas

Ridge e *Lasso* (lembrando que neste trabalho será analisado apenas o problema Ridge) adotando uma abordagem diferente, observando-os como problemas de otimização multiobjetivo. Essa perspectiva nos permitirá entender melhor suas características e relações com os métodos acima. Se definirmos $f_1(x) = \|Ax - b\|_2^2$ e $f_2(x) = \|x\|_q^q$ temos o seguinte problema de otimização biobjetivo convexo

$$\text{minimizar} \quad \{\|Ax - b\|_2^2, \|x\|_q^q\}, x \in \mathbb{R}^n. \tag{2.25}$$

Aplicando o Método da ϵ-Restrição, o problema fica da seguinte forma

$$\text{minimizar} \quad \|Ax - b\|_2^2 \quad \text{sujeito a} \quad \|x\|_q^q < \epsilon \tag{2.26}$$

que é o problema original (2.6) se substituir t por ϵ.

Pelos Teoremas 2.9.1 e 2.9.2, para resolver (2.26) podemos utilizar o Método das Somas Ponderadas e resolver equivalentemente

$$\text{minimizar} \quad w_1\|Ax - b\|_2^2 + w_2\|x\|_q^q, x \in \mathbb{R}^n \tag{2.27}$$

onde w_1 e w_2 são fixados previamente.

Nesse trabalho, o problema Ridge resolvido é na forma da Equação (2.27). Esse problema é resolvido com o algoritmo NFDA, apresentado no próximo capítulo. Apesar do algoritmo NFDA não ser foco desse trabalho, desde o início pretendia-se utilizar tal algoritmo e por conta disso o problema Ridge resolvido é na forma das somas ponderadas (2.27).

3 Algoritmo NFDA

O Algoritmo de Direções Viáveis para Otimização Convexa não diferenciável (*Nonsmooth Feasible Directions Algorithm for Convex Optimization* – NFDA), desenvolvido originalmente por Freire (2005), e posteriormente estudado em detalhes por Herskovits et al. (2011), é uma ferramenta para resolver problemas de otimização não diferenciáveis. A principal característica deste algoritmo está na sua direção de busca, que é inspirada pelo Algoritmo de Pontos Interiores e Direções Viáveis (*Feasible Directions Interior Point Algorithm for Nonlinear Optimization* – FDIPA), proposto por Herskovits (1998), e destinado a problemas diferenciáveis.

3.1 Formulação do Problema

Consideramos aqui um problema de minimização não restrito, convexo e não necessariamente diferenciável

$$\text{minimizar} \quad F(x)\,,\; x \in \mathbb{R}^n \tag{3.1}$$

onde $F : \mathbb{R}^n \longrightarrow \mathbb{R}$ é uma função convexa.

Nota-se que o problema (3.1) pode ser transformado em um problema restrito da forma

$$(P) \quad \begin{cases} \text{minimizar} & f(x,z) = z \\ \text{sujeito a} & F(x) \leq z \end{cases}, \; (x,z) \in \mathbb{R}^n \times \mathbb{R} \tag{3.2}$$

3.2 Metodologia

O NFDA começa com um ponto inicial (x^1, z^1) situado no interior do epígrafo de F ($(x^1, z^1) \in (\text{epi}(F))^0$). Na iteração k, o algoritmo constrói um hiperplano de suporte h_k ao o epigrafo de F no ponto $(x^k, F(x^k))$ dado pela equação $h_k(x) = F(x^k) + \langle s^k, (x - x^k)\rangle$ onde $s^k \in \partial F(x^k)$ é um subgradiente.

Utilizando os hiperplanos suporte calculados até então, define-se o seguinte pro-

blema auxiliar com restrições lineares

$$(P) \quad \begin{cases} \text{minimizar} & f(x,z) = z \\ \text{sujeito a} & g^k(x,z) \leq 0 \end{cases}, (x,z) \in \mathbb{R}^n \times \mathbb{R} \tag{3.3}$$

onde a função $g^k = (g_1, \ldots, g_k) : \mathbb{R}^{n+1} \longrightarrow \mathbb{R}^k$ é uma função vetorial com $g_i : \mathbb{R}^{n+1} \longrightarrow \mathbb{R}$ dado por

$$g_i(x,z) = h_i(x) - z. \tag{3.4}$$

Seja $p^* = (x^*, z^*)$ um ponto regular de (3.3). As condições de primeira ordem de Karush-Kuhn-Tucker (KKT) podem ser expressas da seguinte forma: se p^* representa um mínimo local em (3.3), então existe um $\lambda^* \in \mathbb{R}^k$ tal que

$$\begin{aligned} \nabla f(p^*) + \nabla g^k(p^*)\lambda^* &= 0 \\ G^k(p^*)\lambda^* &= 0 \\ \lambda^* &\geq 0 \\ g^k(p^*) &\leq 0 \end{aligned} \tag{3.5}$$

onde $G(p)$ denota a matriz diagonal tal que $G(p)_{ii} = g_i(p)$ para todo $i \in \{1, 2, \ldots, m\}$.

Consideremos o seguinte sistema linear extraído de (3.5):

$$\nabla f(p^*) + \nabla g(p^*)\lambda^* = 0 \tag{3.6}$$

$$G(p^*)\lambda^* = 0 \tag{3.7}$$

Definindo

$$y = \begin{pmatrix} p \\ \lambda \end{pmatrix} \quad \text{e} \quad \Phi(y) = \begin{pmatrix} \nabla f(p) + \nabla g(p)\lambda \\ G(p)\lambda \end{pmatrix} \tag{3.8}$$

o jacobiano de $\Phi(y)$ é dado por:

$$J\Phi(y) = \begin{pmatrix} H(p,\lambda) & \nabla g(p) \\ \Lambda \nabla g(p) & G(p) \end{pmatrix} \tag{3.9}$$

onde Λ é uma matriz diagonal de tal que $\Lambda_{ii} = \lambda_i$ para todo $i \in \{1, 2, \ldots, m\}$.

Considerando o ponto $y^k = (p^k, \lambda^k)$ na iteração k, encontra-se $y^{k+1} = (p^{k+1}, \lambda^{k+1})$ com uma iteração de Newton para resolver o sistema de equações lineares $\Phi(y) = 0$ definido pelas Equações (3.6) e (3.7) que pode ser escrito da seguinte forma:

$$J\Phi(y^k)(y^{k+1} - y^k) = -\Phi(y^k). \tag{3.10}$$

A Equação (3.10) pode ser reescrita, obtendo-se:

$$\begin{pmatrix} B^k & \nabla g^k(p^k) \\ \Lambda^k \nabla g^k(p^k)^t & G^k(p^k) \end{pmatrix} \begin{pmatrix} p - p^k \\ \lambda - \lambda^k \end{pmatrix} = - \begin{pmatrix} \nabla f(p^k) + \nabla g(p^k)\lambda^k \\ G^k(p^k)\lambda^k \end{pmatrix} \tag{3.11}$$

onde $B^k \equiv \nabla^2 f(p^k) + \sum_{i=1}^{k} \lambda_i^k \nabla^2 g_i(p^k)$ é a hessiana da função Lagrangiana $L(p, \lambda) = f(p) + \lambda^T g(p)$ ou uma aproximação quasi-Newton, que precisa ser simétrica e definida positiva para garantir a convergência (HERSKOVITS, 1998). Por vezes, nos referimos a $y^{k+1} = (p^{k+1}, \lambda^{k+1})$ apenas como $y = (p, \lambda)$.

Pondo $d = p - p^k$, podemos reescrever o sistema (3.11) da seguinte forma:

$$B^k d + \nabla g^k(p^k)\lambda = -\nabla f(x^k) \tag{3.12}$$

$$\Lambda^k \nabla g^k(p^k)^T d + G^k(p^k)\lambda = 0. \tag{3.13}$$

A solução do sistema definido pelas Equações (3.12) e (3.13) nos fornece uma tupla (d_1^k, λ_1^k) onde d_1^k é uma direção e λ_1^k uma estimativa para λ. Em (HERSKOVITS, 1998) foi provado que $(d_1^k)^t \nabla f(p^k) < 0$, isto é, d_1^k é uma direção de descida para a função f. Entretanto, não há nenhuma garantia, até o momento, que a direção d_1^k seja viável.

De fato, caso o ponto p^k esteja "próximo" da curva $g_i(p^k) = 0$ para algum i, a direção d_1^k pode deixar de ser viável pois nesse caso, d_1^k tende a uma direção tangente ao conjunto viável. Com efeito, se reescrevermos a Equação (3.13) temos que :

$$\lambda_i \nabla g_i(p^k) d_1^k + g_i(p^k)\lambda_i = 0, \qquad i = 1, 2, ..., m \tag{3.14}$$

e isso implica que $\nabla g_i(p^k) d_1^k = 0$ para todo i tal que $g_i(p^k) = 0$, ou seja, d_1^k é tangente a

curva $g_i(p) = 0$ e portanto, é uma direção que sai da região viável.

Uma solução para esse problema, é realizar o cálculo de uma nova direção de restauração para impedir que a direção original seja inviável. Para isso, é definido um novo sistema linear em d e $\overline{\lambda}$ a partir do sistema das Equações (3.12) e (3.13) adicionando uma matriz $-\rho_k\Lambda^k$, com $\rho_k > 0$, no lado direito da Equação. O sistema fica

$$B^k d + \nabla g(p^k)^t \overline{\lambda} = -\nabla f(p^k) \tag{3.15}$$

$$\Lambda^k \nabla g^k(p^k) d + \quad G^k(p^k)\overline{\lambda} = -\rho_k \Lambda^k. \tag{3.16}$$

Agora, a Equação (3.16) é equivalente a

$$\lambda_i^k \nabla g_i(p^k)^t d^k + g_i(p^k)\overline{\lambda}_i = -\rho_k \lambda_i^k, \quad i = 1, 2, \ldots, m \tag{3.17}$$

e com isso, $\nabla g_i^t(x^k)d = -\rho_k \omega_i^k$ para as restrições ativas em x^k. Sendo $\rho_k > 0$ para todo i, temos que $-\rho_k\omega_i^k < 0$ e, portanto, a direção d calculada pelo novo sistema é viável uma vez que $\nabla g_i^t(x^k)d < 0$. Esse novo sistema, definido pelas Equações (3.15) e (3.16) produz uma direção d que é viável mas que pode não ser de descida para a função f. Para resolver esse problema, o sistema perturbado é desacoplado, dando origem aos seguintes sistemas:

$$B^k d_1^k + \nabla g^k(p^k)^t \lambda_1^k = -\nabla f(p^k) \tag{3.18}$$

$$\Lambda^k \nabla g^k(p^k) d_1^k + G^k(p^k)\lambda_1^k = 0 \tag{3.19}$$

cujas soluções são d_1^k e λ_1^k e

$$B^k d_2^k + \nabla g^k(p^k)^t \lambda_2^k = 0 \tag{3.20}$$

$$\Lambda^k \nabla g^k(p^k) d_2^k + G^k(p^k)\lambda_2^k = -\Lambda^k \tag{3.21}$$

que tem como solução d_2^k e λ_2^k. Com isso, definimos a direção d e o multiplicador $\overline{\lambda}$ na iteração k como $d^k = d_1^k + \rho_k d_2^k$ e $\overline{\lambda}^k = \lambda_1^k + \rho_k \lambda_2^k$.

Apresentaremos agora a condição para que d^k seja também de descida. Já sabe-

mos que $(d_1^k)^t \nabla f(p^k) < 0$. Da definição de d^k obtém-se:

$$(d^k)^t \nabla f(p^k) = (d_1^k)^t \nabla f(p^k) + \rho_k (d_2^k)^t \nabla f(p^k). \tag{3.22}$$

Se tivermos $(d_2^k)^t \nabla f(p^k) \leq 0$ então teremos que $(d^k)^t \nabla f(p^k) < 0$, $\forall \rho_k \in \mathbb{R}_+^*$. Portanto, devemos escolher ρ_k adequadamente no caso em que $(d_2^k)^t \nabla f(p^k) > 0$. Neste caso, impondo-se a desigualdade:

$$(d^k)^t \nabla f(p^k) \leq \xi (d_1^k)^t \nabla f(p^k) < 0 \text{ com } \xi \in (0,1) \tag{3.23}$$

obtém-se:

$$(d_1^k)^t \nabla f(p^k) + \rho_k (d_2^k)^t \nabla f(p^k) \leq \xi (d_1^k)^t \nabla f(p^k) < 0 \tag{3.24}$$

e, isolando ρ_k, vem:

$$\rho_k \leq (\xi - 1) \frac{(d_1^k)^t \nabla f(p^k)}{(d_2^k)^t \nabla f(p^k)}. \tag{3.25}$$

Escolhendo ρ_k de modo a respeitar a desigualdade em (3.25), para qualquer $\xi \in (0,1)$, conclui-se que d^k será uma direção viável e de descida para a função f.

Um tamanho de passo t_k é estabelecido como

$$t_k = \min\{t_{\max}, T\} \tag{3.26}$$

onde

$$t_{\max} = \max\{t : g_i^k(x^k, z^k) + t d^k \leq 0\} \tag{3.27}$$

e $T > 0$ é um parâmetro previamente definido.

Calcula-se um ponto auxiliar

$$(y^k, \omega^k) = (x^k, z^k) + \mu t_k d^k \tag{3.28}$$

onde $\mu \in (0,1)$.

Se $F(y^k) < \omega^k$ ou, equivalentemente, se $(y^k, \omega^k) \in (\text{epi}(F))^0$ então $(x^{k+1}, z^{k+1}) =$

(y^k, ω^k), o hiperplano

$$h_{k+1} = F(x^{k+1}) + \langle s^{k+1}, x - x^{k+1} \rangle \quad , \quad s^{k+1} \in \partial F(x^{k+1}), \tag{3.29}$$

é calculado e a restrição

$$g_{k+1}(x, z) = h_{k+1}(x) - z \leq 0 \tag{3.30}$$

é adicionada ao (P4.3). Este processo é nomeado de passo sério.

Se $F(y^k) \geq \omega^k$, então $(x^{k+1}, z^{k+1}) = (x^k, z^k)$. Este é conhecido como passo nulo. Nesse caso, o hiperplano de suporte h_{k+1} é definido por

$$h_{k+1}(x) = F(y^k) + \langle s, x - y^k \rangle \ \text{ com } \ s \in \partial F(y^k) \tag{3.31}$$

e, como antes,

$$g_{k+1}(x, z) = h_{k+1} - z \leq 0 \tag{3.32}$$

é adicionado para atualizar o problema auxiliar (P) (3.3).

Foi provado por Freire (2005) e Herskovits et al. (2011) que a direção d^k converge para zero à medida que k aumenta. Este comportamento oferece um critério para interromper a execução do algoritmo. Além disso, os pontos de acumulação da sequência limitada $(x^k, z^k) \in (\text{epi}(F))^0$ produzida pelo NFDA são soluções de (3.1), como também é demonstrado nas referências (FREIRE, 2005; HERSKOVITS et al., 2011).

É crucial enfatizar que o problema (P) (3.3) não é resolvido. Ele pode nem mesmo possuir uma solução. O NFDA aproveita a sua estrutura linear e, portanto, diferenciável, para obter uma direção de descida viável.

O NFDA basicamente necessita apenas da solução de dois sistemas lineares com a mesma matriz. Além disso, é robusto, pois não exige o ajuste de parâmetros e tem demonstrado bom desempenho em diversas aplicações, especialmente no contexto atual. Destacamos também que, na descrição do NFDA apresentada nesta seção, todos os hiperplanos de suporte são armazenados. Existem versões do NFDA nas quais apenas parte deles é mantida. Para discussões mais detalhadas sobre o NFDA, suas premissas, regras de atualização, convergência e outras características, o leitor pode consultar as referências

(FREIRE, 2005; HERSKOVITS et al., 2011; HERSKOVITS, 1998).

Por fim, para gerar a frente de Pareto da regressão Ridge e Lasso por meio do NFDA, deve-se definir a função F em (3.3) como sendo $w_1\|Ax-b\|_2^2 + w_2\|x\|_q^q$ e variar w_1 e w_2 para gerar a quantidade de pontos da frente de Pareto que se deseja.

O Algoritmo 1 apresenta o pseudocódigo do NFDA, onde B é uma matriz simétrica positiva definida.

Algoritmo 1: NFDA

Entrada: $\xi; \mu \in (0,1)$; $\varphi > 0$; $T > 0$; $\gamma > 0$;
Dados: $(x^1, z^1) \in (\text{epi}(F))^0$; $\lambda^1 > 0$; $B^1 \in \mathbb{R}^{n+1} \times \mathbb{R}^{n+1}$;

início
  enquanto $\|d^k\| \leq \gamma$ **faça**
    Calcule $p^k = (x^k, z^k) \in (\text{epi}(F))^0$, h_k, $\lambda^k > 0$;
    Compute $g_k(x,z) = h_k(x) - z$, $g^k = [g_1, g_2, \ldots, g_k]$ e $\nabla g^k = [\nabla g_1, \nabla g_2, \ldots, \nabla g_k]$;
    Encontre $d_1^k, d_2^k, \lambda_1^k$ e λ_2^k resolvendo os sistemas:

$$\begin{cases} B^k d_1 + \nabla g^k(p^k)\lambda_1 = -\nabla f(p^k) \\ \Lambda^k \nabla g^k(p^k)^T d_1 + G^k(p^k)\lambda_1 = 0, \end{cases}$$

$$\begin{cases} B^k d_2 + \nabla g^k(p^k)\lambda_2 = 0 \\ \Lambda^k \nabla g^k(p^k)^T d_2 + G^k(p^k)\lambda_2 = -\Lambda^k, \end{cases}$$

    onde $G^k(p^k)$ e Λ^k são matrizes diagonais com $G^k_{ii}(p^k) = g_i^k(p^k)$ e $\Lambda^k_{ii} = \max\{\lambda_i^{k-1}, \varphi\|d_1^k\|^k\}$;
    se $d_2^k \nabla f(p^k) > 0$ **então**
      Defina $\rho^k = \varphi\|d_1^k\|^2$;
    senão
      Defina $\rho^k = \min\left\{\varphi||d_1^k||^2, (1-\xi)\dfrac{(d_1^k)^T\nabla f(p^k)}{(d_2^k)^T\nabla f(p^k)}\right\}$;
    fim se
    Atualize $d^k = d_1^k + \rho^k d_2^k$ e $\lambda^k = \lambda_1^k + \rho^k\lambda_2^k$;
    Calcule $t_{\max} = \min\{t | g_i^k(x,x) + td^k \leq 0\}$ e o passo $t^k = \max\{t_{\max}, T\}$;
    Calcule o ponto auxiliar $(y^k, w^k) = (x^k, z^k) + \mu t^k d^k$;
    se $F(y^k) < w^k$ **então**
      $(x^{k+1}, z^{k+1}) = (y^k, w^k)$;
      Calcule $s^{k+1} \in \partial F(x^{k+1})$ e $h_{k+1} = F(x^{k+1}) + \langle s^{k+1}, x - x^{k+1}\rangle$;
    senão
      $(x^{k+1}, z^{k+1}) = (x^k, z^k)$;
      Calcule $s \in \partial F(y^k)$ e $h_{k+1} = F(y^k) + \langle s, x - y^k\rangle$;
    fim se
  fim enqto
fim

4 Metodologia Proposta

A principal hipótese deste estudo é que o ponto mais próximo do ponto ideal é uma boa escolha entre os pontos da frente de Pareto para problemas de otimização biobjetivo em problemas de regressão. Para testá-la, foi desenvolvida uma metodologia específica que envolveu a aplicação do algoritmo NFDA, utilizando a estratégia k-*fold* com $k = 10$. Essa etapa foi realizada considerando N pontos de Pareto em cada execução.

Em métodos de aprendizado de máquina, a técnica de k-*fold* é fundamental para a validação do desempenho dos modelos. Esta técnica divide o conjunto de dados em k subconjuntos de tamanhos iguais. O modelo é treinado e testado k vezes. Em cada iteração, o modelo é treinado em k-1 subconjuntos e testado no subconjunto restante. Dessa maneira, cada ponto de dados é empregado tanto no treinamento quanto no teste, proporcionando uma avaliação robusta do desempenho do modelo (HASTIE; TIBSHIRANI; FRIEDMAN, 2009b).

Para a implementação da metodologia proposta, escolheu-se a configuração de 10 *folds*, o que significa que para cada rodada de treinamento do algoritmo NFDA, utilizou-se 90% dos dados para treinamento e 10% para teste.

Para cada ponto encontrado na frente de Pareto, foi calculado o erro cometido nos conjuntos de treinamento e teste de cada *fold*. Para avaliar o desempenho de cada ponto da frente de Pareto, foram utilizadas quatro métricas diferentes: R^2 (coeficiente de determinação), MSE (erro quadrático médio), MAE (erro absoluto médio) e MAPE (erro percentual absoluto médio). O erro foi calculado considerando o erro cometido na regressão, ou seja, o quanto a previsão $\hat{y}$ estava próximo da resposta esperada y_{true}. Além do erro, para cada ponto foi calculado a distância para o ponto ideal considerando as métricas de distância e norma 1, 2, infinito, HasD, LD e Nuc.

Com esses resultados obtidos, foram selecionados os pontos (entre todos os pontos que fossem pontos de Pareto e que não fossem pontos extremos pois são os casos triviais otimizando individualmente cada função) com o menor erro em cada uma das métricas de erro. Também foram selecionados os pontos que tivessem a menor distância para o ponto

ideal em cada uma das métricas de distância. Para cada um desses pontos mais próximos ao ponto ideal foi realizado um cálculo para obter o erro relativo ao ponto com menor erro em cada uma das métricas de erro. Para facilitar, seja x_n o ponto da aproximação da frente de Pareto mais próximo ao ponto ideal com a norma 1. Seja $\text{erro}_n^{\text{MSE}}$ o erro na métrica MSE cometido por esse ponto x_n. Seja $\text{erro}_{\text{min}}^{\text{MSE}}$ o erro cometido pelo ponto x_k que possui o menor erro na métrica MSE. O erro percentual para a métrica MSE do ponto x_n é dado por:

$$\frac{\text{erro}_n^{\text{MSE}} - \text{min}^{\text{MSE}}}{\text{min}^{\text{MSE}}} \tag{4.1}$$

Em geral, o erro percentual pra métrica m para o ponto x_n que possui erro erro_n^m para a métrica m é dado por:

$$\text{erro-perc}_n^m = \frac{\text{erro}_n^m - \text{erro}_{\text{min}}^m}{\text{erro}_{\text{min}}^m} \tag{4.2}$$

onde $\text{erro}_{\text{min}}^m$ é o menor erro cometido entre os 200 pontos na métrica de erro m.

Com esse cálculo realizado para cada um dos pontos mais próximos ao ponto ideal em relação a cada métrica, pode-se obter o resultado final tirando a média entre os 10 *folds*. Com isso, para cada *dataset* foram selecionados 10 pontos (o mais próximo ao ponto ideal em cada *fold*) considerando cada uma das métricas de distância. Com esses 10 pontos para cada métrica, foi realizado um cálculo da média das métricas de erro percentual em cada uma das métricas de erro e com isso, obtem-se a informação de qual é o erro percentual relativo médio do ponto mais próximo ao ponto ideal em cada uma das métricas consideradas para o cálculo da distância. Um fluxograma simplificado com os passos da metodologia aplicado pode ser observado na Figura 4.1.

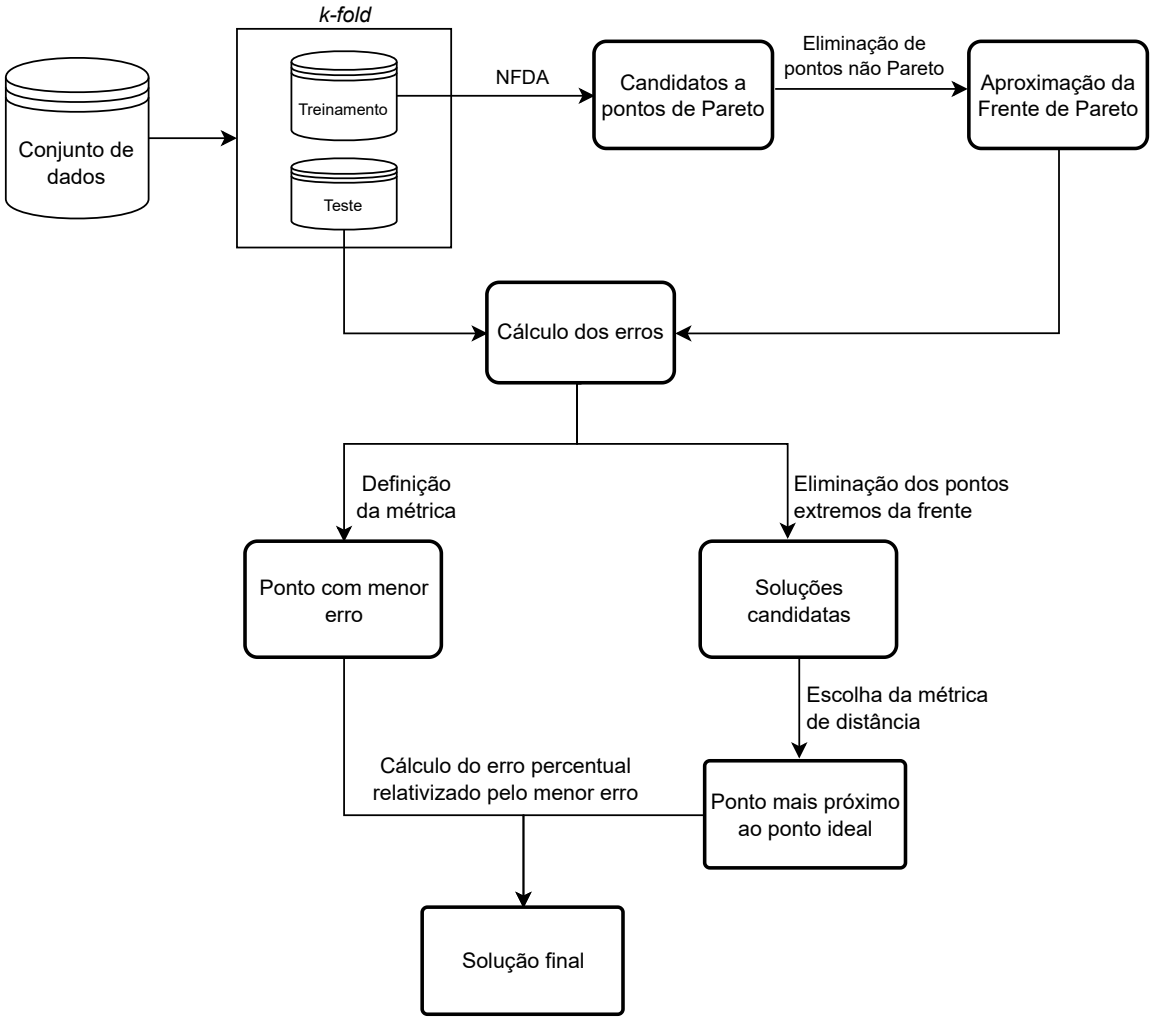


Figura 4.1: Fluxograma da metodologia realizada no trabalho.

5 Experimentos e Resultados

Neste capítulo são apresentados os resultados obtidos ao aplicar a metodologia descrita no capítulo anterior. Os resultados apresentados mostram as médias dos erros normalizados para os pontos mais próximos do ponto ideal, considerando cada uma das normas e métricas de erro analisadas. Nos experimentos realizados utilizamos 200 pontos de pareto para gerar a aproximação da frente de Pareto. Esse valor foi escolhido de modo que não fosse baixo o suficiente para termos uma baixa densidade da frente de Pareto e nem alto de modo que aumentaria o tempo para a realização dos testes.

5.1 Conjuntos de Dados

Os experimentos foram realizados em 4 conjuntos de dados: *Housing*, Servo, *HouseElectric* e *Kin40k*. Esses conjuntos de dados foram selecionados ao acaso entre os conjuntos de dados disponíveis no repositório em que foram obtidos. Os números de amostras e de atributos para cada conjunto de dados estão detalhados na Tabela 5.1. Os conjuntos de dados são problemas de regressão do repositório de aprendizado de máquina UCI preparados para estudos de *benchmarking* com as divisões de treino e teste em 10 *folds* já realizadas e padronizadas. Os *datasets* foram obtidos a partir de um repositório no GitHub[1].

Tabela 5.1: Descrição das bases de dados.

Base	Amostras	Atributos
Housing	506	13
Servo	167	4
HouseElectric	2.049.280	11
Kin40k	40.000	8

O conjunto de dados *Housing* contém informações coletadas pelo Serviço de Censo dos EUA sobre habitação na área de Boston. Compreende 506 amostras com 13 variáveis. Essas variáveis representam diversas características da habitação e do ambiente ao redor,

[1] ⟨https://github.com/treforevans/uci_datasets⟩

incluindo taxas de criminalidade per capita, proporções de zonas residenciais e não comerciais, número médio de quartos por habitação, idade das casas, distâncias a centros de emprego, acessibilidade a rodovias, taxas de imposto sobre a propriedade e a proporção de alunos por professor. Além dessas características, o conjunto de dados também inclui a variável resposta que representa o valor médio das casas.

O conjunto de dados *Servo* contem informações de simulações de um sistema servo, que é um tipo de sistema de controle de *feedback* com o objetivo de mover um motor em resposta a um sinal de comando. O conjunto de dados é composto por 167 amostras, cada uma com 4 variáveis. Esses variáveis representam características do sistema servo, incluindo o nome do motor, o tipo de parafuso, o ângulo de inclinação da imagem e o torque do motor. Cada uma dessas características fornece informações valiosas para entender a operação e o desempenho do sistema servo. As variáveis categóricas já são fornecidas transformadas em variáveis numéricas. Esse preprocessamento realizados nas variáveis categóricas podem não ter sido realizados da melhor maneira possível e isso pode ser melhorado em testes futuros de outros trabalhos.

O conjunto de dados *HouseElectric* consiste em medidas de consumo de energia elétrica de uma casa, coletadas ao longo de um período de quatro anos, com uma amostra sendo coletada a cada minuto. Ele contém um total de 2.049.280 amostras, e cada amostra é caracterizada por 11 variáveis. Essas variáveis representam várias medidas e características associadas ao consumo de energia, como a demanda global de energia, a intensidade da corrente e a voltagem. A análise desses dados permite entender o padrão de consumo de energia na residência, o que é fundamental para estudos de eficiência energética, previsão da demanda de energia e sistemas inteligentes de gestão de energia residencial. A variável resposta neste conjunto de dados é o consumo de energia que se deseja prever ou analisar com base nas 11 variáveis.

O conjunto de dados *Kin40k* é uma coleção de pontos de dados para um sistema de cinemática de um braço robótico, com o objetivo de modelar e analisar o movimento do braço. O mesmo é composto por 40.000 registros, cada uma com 8 atributos. Essas variáveis incluem três ângulos de junta do braço robótico e a posição do braço em um sistema de coordenadas tridimensional. A variável resposta deste conjunto de dados é a

distância do efetuador final (a parte do robô que interage com o ambiente) até um alvo.

5.2 Resultados

Os resultados obtidos são apresentados em percentuais, que representam o quão distante o resultado encontrado está do valor ótimo em cada uma das métricas de erro. Por exemplo, ao observar a Tabela 5.2, o primeiro valor na coluna R^2 para o conjunto de treinamento, sob a distância de Manhattan, é de $0,27$. Isso indica que o ponto encontrado usando a distância de Manhattan está $0,27\%$ pior em relação ao ponto que representa o melhor resultado para a métrica R^2 no conjunto de treinamento.

A Tabela 5.2 apresenta os resultados para o conjunto de dados *Housing*.

Tabela 5.2: Resultados para o *dataset* Housing.

Distância	Conjunto de treinamento				Conjunto de teste			
	R^2	MSE	MAE	MAPE	R^2	MSE	MAE	MAPE
Manhattan	**0,27**	**0,79**	**0,11**	0,44	**2,46**	**4,55**	**0,85**	8,06
Euclideana	0,33	0,96	0,18	0,35	2,54	4,73	0,92	7,89
Chebyshev	0,37	1,07	0,22	**0,31**	2,59	4,86	0,96	**7,80**
Hassanat	4,73	13,66	7,06	5,77	4,59	15,38	6,09	10,48
Lorentzian	4,73	13,66	7,06	5,77	4,59	15,38	6,09	10,48
Nuclear	0,35	1,00	0,19	0,33	2,56	4,79	0,94	7,86

Na Tabela 5.2 a Norma 1 foi a que obteve os melhores resultados na maioria das métricas. No conjunto de teste, o ponto mais próximo ao ponto ideal considerando a Norma 1 ficou $2,46\%$ pior que a melhor solução na métrica R^2, $4,55\%$ pior que a melhor solução na métrica MSE e $0,85\%$ pior que a melhor solução na métrica MAE. Apesar do resultado não ter sido o melhor também na métrica MAPE ainda assim o resultado foi bom ficando com um MAPE apenas $8,06\%$ maior que o da melhor solução. Nas demais normas os resultados para esse conjunto de dados também foram ótimos ficando abaixo de 10% pior que o melhor ponto nas métricas na maior parte dos resultados. Um destaque também pode ser observado nos resultados da Norma Inf que teve resultados muito bons, tão bons quanto os da Norma 1 porém teve o melhor resultado na métrica MAPE ficando apenas $7,80\%$ pior que a melhor solução em tal métrica.

Na Tabela 5.3 são apresentados os resultados para o conjunto de dados *Servo*.

Tabela 5.3: Resultados para o *dataset* Servo.

Distância	Conjunto de treinamento				Conjunto de teste			
	R^2	MSE	MAE	MAPE	R^2	MSE	MAE	MAPE
Manhattan	**0,04**	**0,07**	**0,04**	94,58	6,37	12,71	3,90	189,13
Euclideana	0,41	0,65	0,17	86,34	5,69	11,70	3,65	170,90
Chebyshev	1,20	1,91	0,50	77,84	5,61	11,66	**3,64**	152,76
Hassanat	44,44	70,34	20,37	**0,00**	46,07	79,23	23,08	**37,54**
Lorentzian	44,44	70,34	20,37	**0,00**	46,07	79,23	23,08	**37,54**
Nuclear	0,64	1,01	0,27	83,43	**5,60**	**11,62**	**3,64**	164,36

Na Tabela 5.3, pode-se observar que a Norma 1 atingiu os melhores resultados nas métricas de R^2, MSE e MAE para o conjunto de treino. Porém, no conjunto de teste, a Norma Inf teve a melhor performance nessas mesmas métricas, estando apenas $5,61\%$ e $3,64\%$ pior que a melhor solução nas métricas R^2 e MAE, respectivamente. As distâncias HASD e LD, por sua vez, foram destaque na métrica MAPE, obtendo o melhor resultado tanto no treino quanto no teste, evidenciando uma performance nula e $37,54\%$ pior que a melhor solução, respectivamente. As demais Normas também apresentaram resultados relevantes. A Norma 2, por exemplo, apresentou um desempenho sólido, com percentuais de erro inferiores a 12% em quase todas as métricas para o conjunto de teste. A Norma Nuc, apresentou o melhor desempenho nas métricas R^2, MSE e MAE no conjunto de teste obtendo erros percentuais abaixo de 12%. No geral, as normas 1, 2, Inf e Nuc foram boas nos conjuntos de treino e teste para as métricas R^2, MSE e MAE porém não se saíram tão bem na métrica MAPE obtendo valores acima de 100% maiores que a melhor solução. O contrário ocorreu para as distâncias HASD e LD pois tiveram os melhores resultados na métrica MAPE porém os piores nas demais métricas obtendo valores acima de 20%.

A Tabela 5.4 apresenta os resultados para o conjunto de dados *HouseElectric*.

Tabela 5.4: Resultados para o *dataset* HouseElectric.

Distância	Conjunto de treinamento				Conjunto de teste			
	R^2	MSE	MAE	MAPE	R^2	MSE	MAE	MAPE
Manhattan	**1,7e-08**	**1,2e-07**	3,4e-03	**1,8e-03**	4,6e-05	3,2e-04	3,5e-03	**1,9e-03**
Euclideana	6,0e-07	4,2e-06	3,8e-03	3,3e-03	**4,2e-05**	**2,9e-04**	3,8e-03	3,5e-03
Chebyshev	9,2e-06	6,3e-05	**3,1e-03**	1,1e-02	4,5e-05	3,1e-04	**3,1e-03**	1,1e-02
Hassanat	3,4e-05	2,4e-04	6,6e-03	2,0e-02	6,5e-05	4,4e-04	6,7e-03	2,0e-02
Lorentzian	3,4e-05	2,4e-04	6,6e-03	2,0e-02	6,5e-05	4,4e-04	6,7e-03	2,0e-02
Nuclear	1,9e-06	1,3e-05	3,3e-03	5,2e-03	4,7e-05	3,2e-04	3,4e-03	5,2e-03

Analisando a Tabela 5.4, que apresenta os resultados para o conjunto de dados HouseElectric, a Norma 1 destaca-se por obter os melhores resultados para as métricas R^2, MSE e MAPE no conjunto de treino, estando somente $0,000046\%$, $0,00032\%$ e $0,0019\%$ pior que a melhor solução para o conjunto de teste nessas métricas, respectivamente. Por outro lado, a Norma Inf apresentou a melhor performance na métrica MAE para o conjunto de treino, ficando apenas $0,31\%$ pior que a melhor solução tanto no conjunto de treino quanto no teste. Adicionalmente, para a métrica R^2 no conjunto de teste, a Norma 2 alcançou a melhor performance, ficando apenas $0,000042\%$ pior que a melhor solução. As demais normas, embora não tenham atingido o melhor desempenho em nenhuma métrica específica, apresentaram resultados significativos. Por exemplo, a Norma Nuc conseguiu manter a consistência nos resultados, estando menos de $0,032\%$ pior que a melhor solução nas métricas R^2, MSE e MAE no conjunto de teste. A distância HasD e a distância LD, por sua vez, apresentaram resultados piores que as demais porém ainda são resultados muito bons tendo seus erros percentuais próximos a 0 em todas as métricas de erro. Portanto, esses resultados não devem ser desconsiderados, já que diferentes normas podem funcionar melhor para diferentes tipos de dados ou cenários.

Finalmente, na Tabela 5.5 são apresentados os resultados para o conjunto de dados *Kin40k*.

Tabela 5.5: Resultados para o *dataset* Kin40k.

Distância	Conjunto de treinamento				Conjunto de teste			
	R^2	MSE	MAE	MAPE	R^2	MSE	MAE	MAPE
Manhattan	**2,3e-3**	**8,3e-7**	6,5e-4	0,11	12,08	1,2e-3	**8,6e-4**	0,16
Euclideana	**2,3e-3**	**8,3e-7**	6,5e-4	0,11	12,08	1,2e-3	**8,6e-4**	0,16
Chebyshev	3,6e-3	1,4e-6	**6,4e-4**	0,11	17,78	1,2e-3	8,9e-4	0,14
Hassanat	6,3e-2	2,5e-5	1,2e-3	**6,0e-4**	**5,97**	**3,0e-4**	1,1e-3	**0,07**
Lorentzian	6,3e-2	2,5e-5	1,2e-3	**6,0e-4**	**5,97**	**3,0e-4**	1,1e-3	**0,07**
Nuclear	**2,3e-3**	**8,3e-7**	6,5e-4	0,11	12,08	1,2e-3	**8,6e-4**	0,16

Ao analisar a Tabela 5.5, que apresenta os resultados para o conjunto de dados Kin40k, percebe-se que as normas 1, 2 e NUC se destacam ao obterem os melhores resultados para as métricas R^2, MSE e MAE no conjunto de treinamento, e também para a métrica MAE no conjunto de teste. As métricas R^2 e MSE para estas normas no conjunto de teste ficaram $12,08\%$ e $0,0012\%$ piores, respectivamente, do que a melhor solução, en-

quanto para a métrica MAE, a diferença para a melhor solução foi de apenas $0,086\%$. Por outro lado, a distância HASD e a distância LD destacam-se para as métricas R^2 e MSE no conjunto de teste, onde obtiveram os melhores resultados, ficando apenas $5,97\%$ e $0,0003\%$ piores, respectivamente, que a melhor solução. Além disso, essas duas normas também se destacaram na métrica MAPE tanto no conjunto de treinamento quanto no de teste, ficando apenas $0,0006\%$ e $0,07\%$ piores, respectivamente, do que a melhor solução. A Norma Inf, embora não tenha obtido o melhor desempenho em nenhuma métrica específica, apresentou o menor erro MAE no conjunto de treinamento, com uma diferença de apenas $0,00064\%$ para a melhor solução.

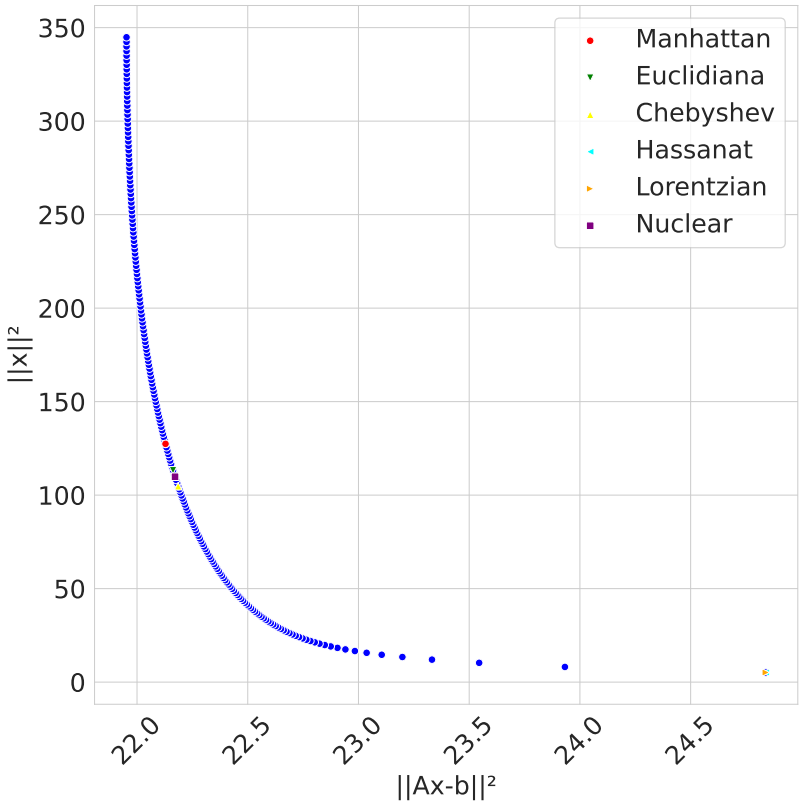
Os resultados destacam a influência significativa que a escolha da norma pode exercer sobre os resultados, reforçando a necessidade de considerar diferentes normas para diferentes tipos de dados e cenários de modelagem. Observa-se também uma heterogeneidade na performance das normas, quando avaliadas através de diferentes métricas e conjuntos de dados, sendo que todas as normas demonstraram um desempenho sólido no conjunto de dados *HouseElectric*. Apesar da variedade de desempenho, todas se mostraram valiosas em determinadas métricas, o que realça a importância de experimentar e explorar diversas abordagens de distâncias na análise dos dados. Essa prática não apenas otimiza os resultados, mas também oferece *insights* relevantes para a compreensão dos dados.

A Figura 5.1 apresenta a Frente de Pareto para o conjunto de dados *Housing*. Na Figura 5.1(a) é possível visualizar a frente de Pareto sem os pontos extremos e sem o ponto ideal destacado. Na Figura 5.1(b) é apresentada a Frente de Pareto com o ponto ideal destacado com sua projeção mostrando como é obtida. Em todas as figuras, além dos pontos da frente também são destacados os pontos que possuem a menor distância em cada uma das métricas de distância.

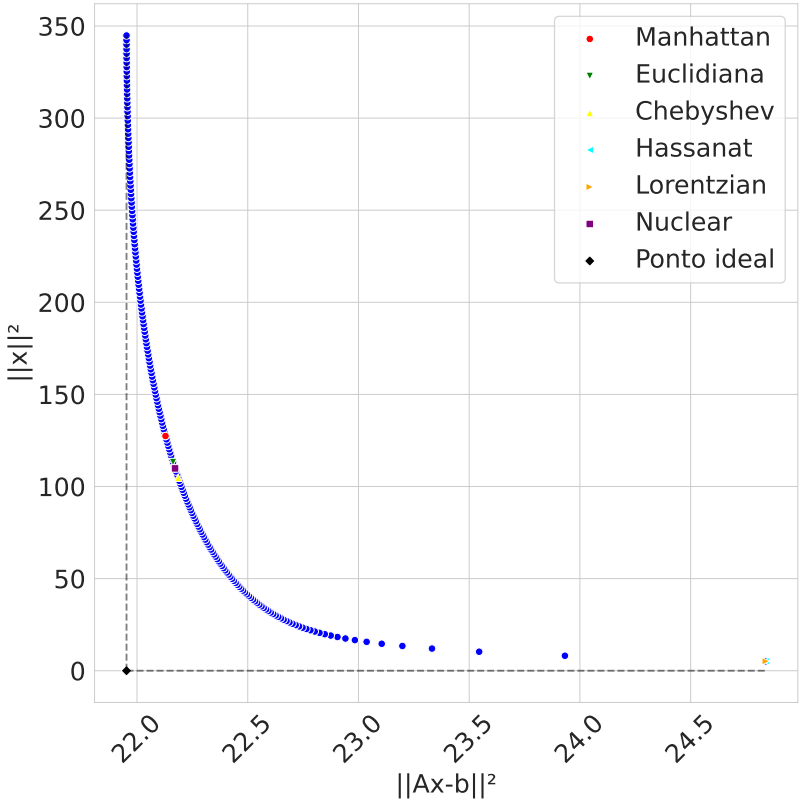
De forma análoga, na Figura 5.2 tem-se a Frente de Pareto para o conjunto de dados *Servo* com a Figura 5.2(a) apresentando apenas a Frente de Pareto e a Figura 5.2(b) destacando o ponto ideal.

As Figuras 5.3 e 5.4 apresentam as visualizações da Frente para os conjuntos *Houseelectric* e *Kin40k* respectivamente. Esses conjuntos tiveram uma menor quantidade

de pontos quando o processo de remoção de pontos não Pareto foi realizada e por conta disso a visualização não ficou muito suave. De forma semelhante ao que foi feito para os outros 2 conjuntos de dados as Figuras 5.3(a) e 5.4(a) apresentam apenas a frente de Pareto e as Figuras 5.3(b) e 5.4(b) destacam o ponto ideal dos conjuntos de dados *Houseelectric* e *Kin40k* respectivamente.

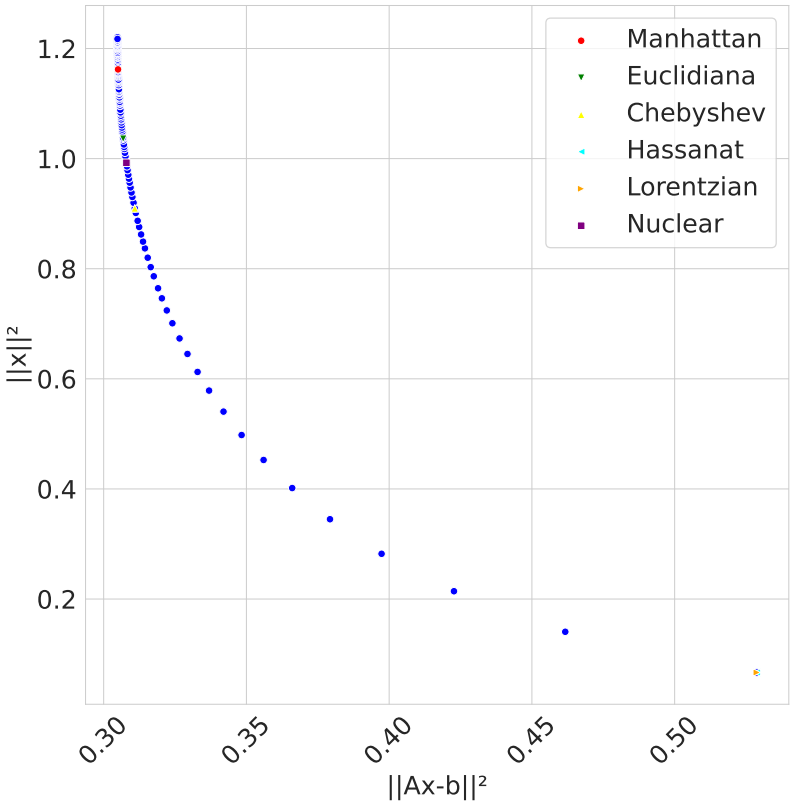


(a) Frente de Pareto.

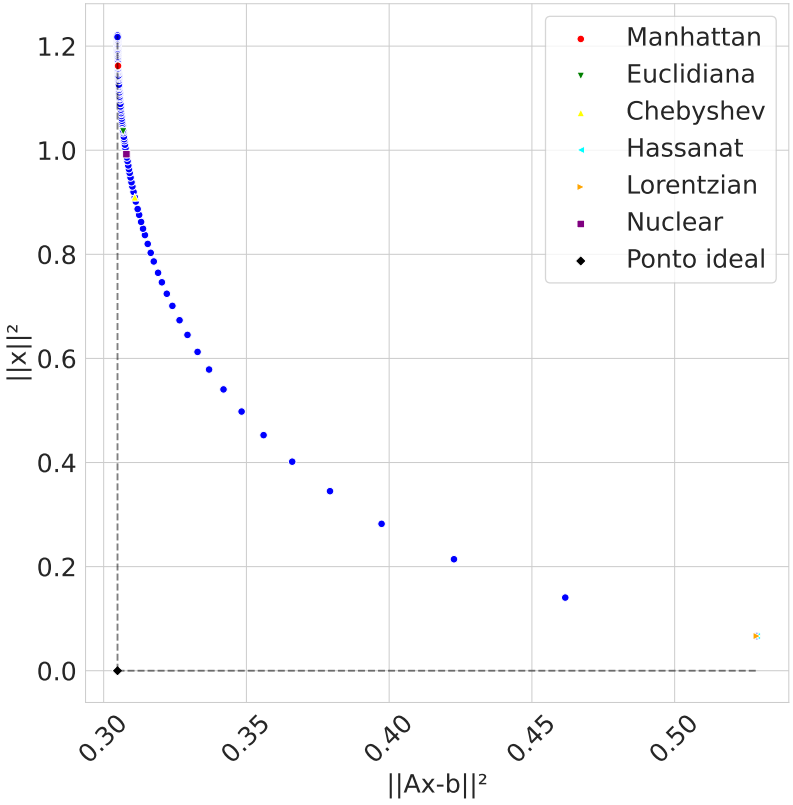


(b) Frente de Pareto com ponto ideal.

Figura 5.1: Frente de Pareto *Housing*.

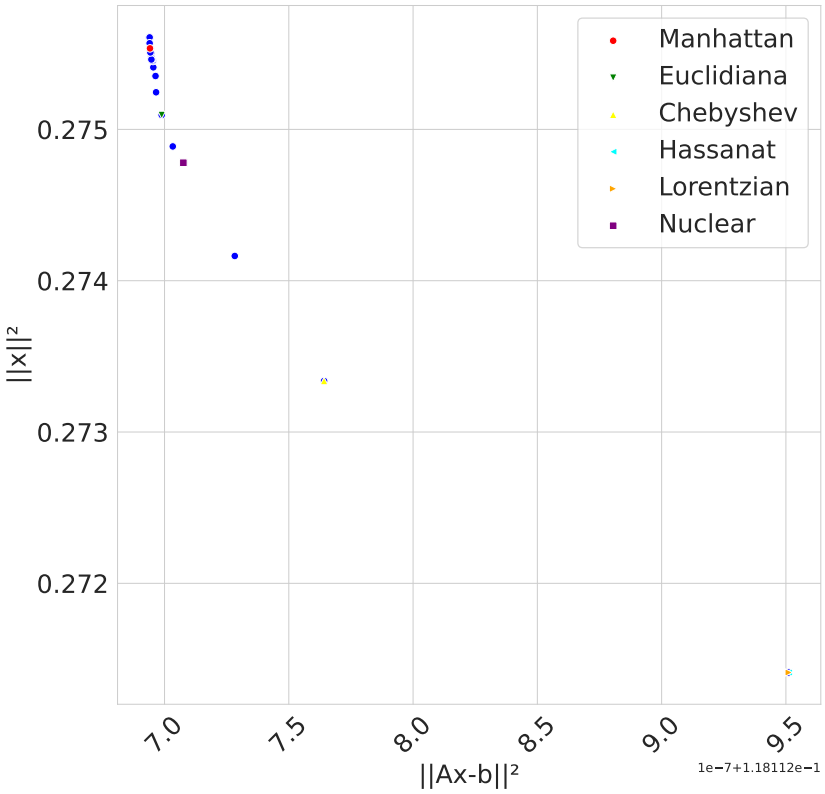


(a) Frente de Pareto.

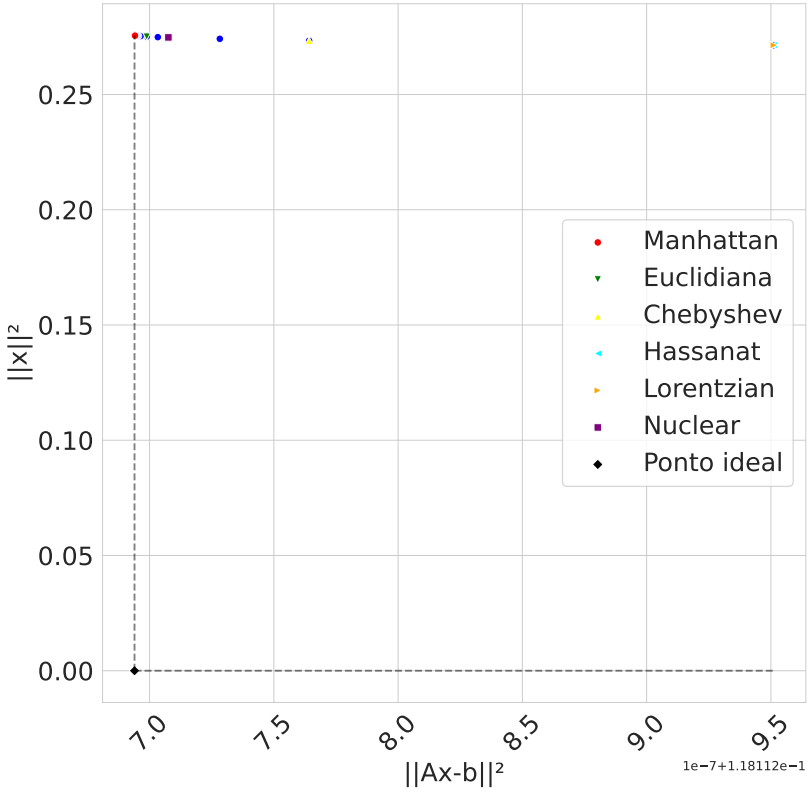


(b) Frente de Pareto com ponto ideal.

Figura 5.2: Frente de Pareto *Servo*.

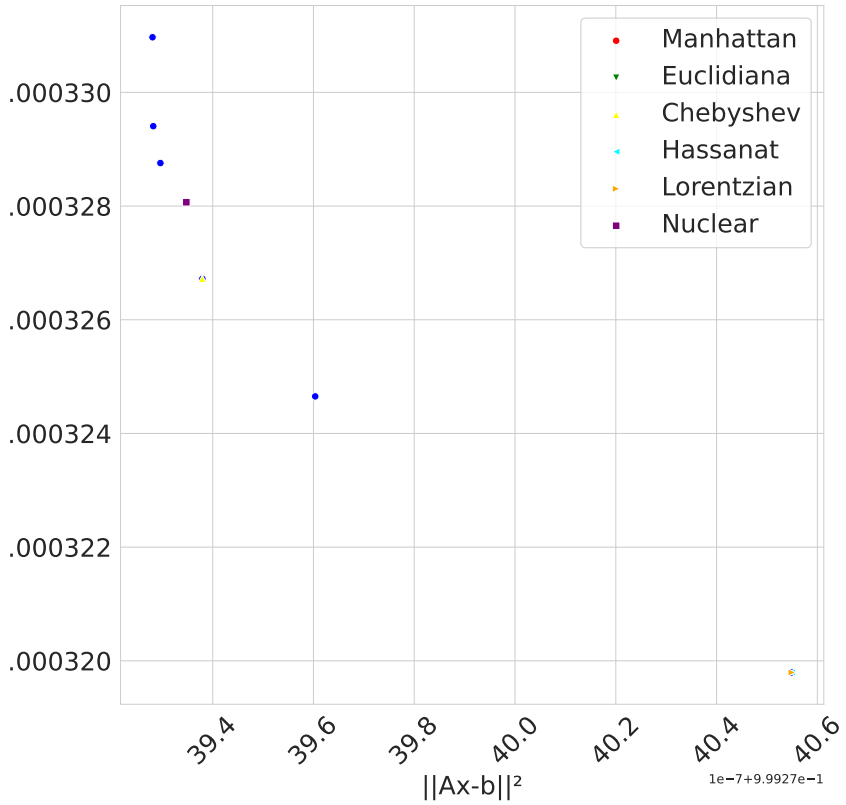


(a) Frente de Pareto.

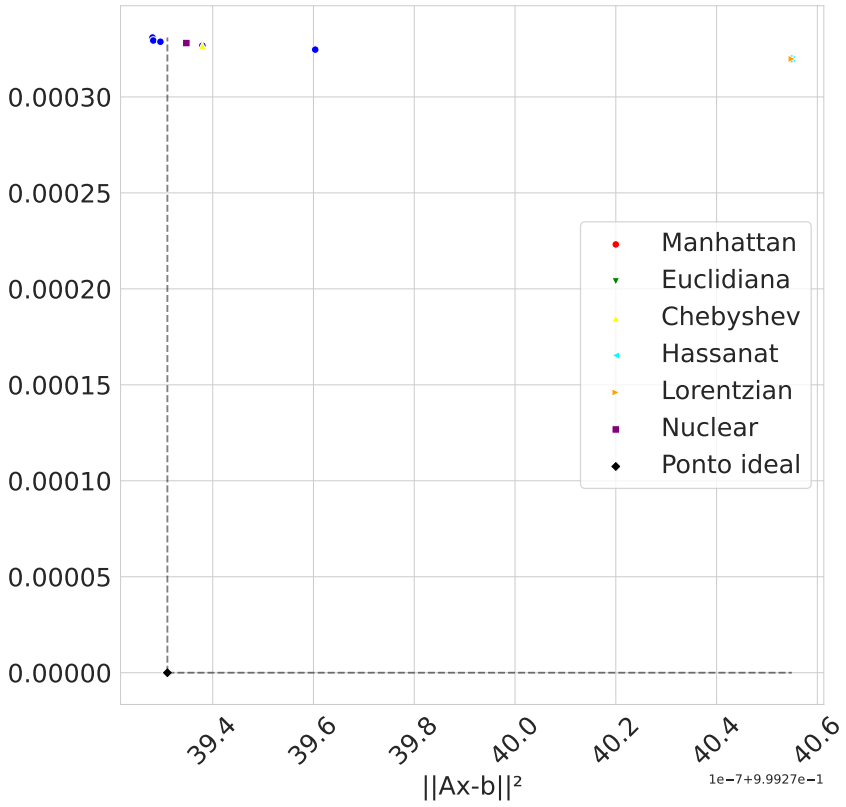


(b) Frente de Pareto com ponto ideal.

Figura 5.3: Frente de Pareto *Houseelectric*.



(a) Frente de Pareto.



(b) Frente de Pareto com ponto ideal.

Figura 5.4: Frente de Pareto *Kin40k*.

6 Conclusão e Trabalhos Futuros

Neste capítulo final, volta-se a atenção para os resultados alcançados no decorrer deste estudo e como eles ajudam a verificar a hipótese desse trabalho. Durante o processo de análise, a hipótese foi testada pela seleção de pontos que estavam mais próximos ao ideal, de acordo com as normas escolhidas. Com os resultados obtidos, verificou-se que esses pontos apresentam boas soluções nas métricas de erro se comparadas as melhores soluções dos problemas.

Esse padrão reforça uma compreensão da relação entre a proximidade ao ponto ideal e a diminuição do erro, fornecendo uma verificação inicial da hipótese principal desse trabalho. Os resultados encontrados apresentam um panorama de relevância não apenas para este trabalho, mas também para trabalhos futuros estendendo essa verificação com outras metodologias e outros problemas de regressão mais interessantes como a regressão Lasso.

Em resumo, pode-se adotar a escolha do ponto mais próximo ao ponto ideal considerando a norma 1 ou norma 2, por exemplo, entre os pontos de uma frente de Pareto para se obter uma boa solução. Com isso, não é necessário uma validação custosa com a geração de diversas frentes de Pareto e com a verificação da qualidade de cada um dos pontos da frente. Mais estudos com diferentes problemas de regressão e com conjuntos de dados maiores e problemas mais complexos são necessários para reforçar ainda mais a hipótese principal desse trabalho.

Portanto, este trabalho verificou que pontos da frente de Pareto mais próximos ao ponto ideal são boas escolhas de solução quando se deseja obter um ponto com baixo erro. Contudo, é importante ressaltar que, embora os resultados sejam interessantes, outros estudos devem ser realizados para aprofundar ainda mais o conhecimento sobre este assunto, com objetivo de confirmar ou melhorar tais conclusões.

A próxima etapa será realizar o mesmo estudo considerando a regressão Lasso. O problema Lasso é não diferenciável e portanto um problema mais complexo de se resolver. Será utilizado o mesmo algoritmo NFDA para encontrar a frente de Pareto e a hipótese

será verificada com a mesma metodologia utilizada neste trabalho. Espera-se com esse estudo conseguir uma nova validação de que é possível obter uma boa solução escolhendo o ponto mais próximo ao ponto ideal na frente de Pareto em problemas de regressão.

Bibliografia

BRANKE, J. et al. *Multiobjective optimization: Interactive and evolutionary approaches*. [S.l.]: Springer Science & Business Media, 2008. v. 5252.

BURACHIK, R. S.; KAYA, C. Y.; RIZVI, M. A new scalarization technique and new algorithms to generate pareto fronts. *SIAM Journal on Optimization*, SIAM, v. 27, n. 2, p. 1010–1034, 2017.

CHARKHGARD, H.; ESHRAGH, A. A new approach to select the best subset of predictors in linear regression modelling: bi-objective mixed integer linear programming. *The ANZIAM Journal*, Cambridge University Press, v. 61, n. 1, p. 64–75, 2019.

COELLO, C. A. C.; LAMONT, G. B.; VELDHUIZEN, D. A. V. *Evolutionary algorithms for solving multi-objective problems*. [S.l.]: Springer, 2007. v. 5.

DEB, K.; DEB, K. Multi-objective optimization. In: *Search methodologies: Introductory tutorials in optimization and decision support techniques*. [S.l.]: Springer, 2013. p. 403–449.

DORIGO, M.; STÜTZLE, T. *Ant colony optimization: overview and recent advances*. [S.l.]: Springer, 2019.

DUTTA, J.; KAYA, C. Y. A new scalarization and numerical method for constructing the weak pareto front of multi-objective optimization problems. *Optimization*, Taylor & Francis, v. 60, n. 8-9, p. 1091–1104, 2011.

FREIRE, W. *A Feasible Directions Algorithm for Convex Nondifferentiable Optimization*. Tese (Doutorado) — Federal University of Rio de Janeiro, 2005. Disponível em: ⟨http://www.optimize.ufrj.br/files/WilhelmPassarellaFreire.pdf⟩.

HASSANAT, A. B. *Dimensionality Invariant Similarity Measure*. 2014.

HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. *The elements of statistical learning: data mining, inference, and prediction*. [S.l.]: Springer Science & Business Media, 2009.

HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. [S.l.]: Springer, 2009.

HERSKOVITS, J. Feasible directions interior point technique for nonlinear optimization. *Journal of Optimization Theory and Applications*, v. 99, n. 1, p. 121–146, 1998.

HERSKOVITS, J. et al. A feasible directions method for nonsmooth convex optimization. *Structural and Multidisciplinary Optimization*, v. 44, n. 3, p. 363–377, 2011.

HIRIART-URRUTY, J.; LEMARECHAL, C. *Convex Analysis and Minimization Algorithms I, II*. [S.l.: s.n.], 1993.

HOERL, A. E.; KENNARD, R. W. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, v. 12, n. 1, p. 55–67, 1970.

MARLER, R. T.; ARORA, J. S. Survey of multi-objective optimization methods for engineering. *Structural and multidisciplinary optimization*, Springer, v. 26, p. 369–395, 2004.

MIETTINEN, K. *Nonlinear multiobjective optimization*. [S.l.]: Springer Science & Business Media, 1998. v. 12.

OSBORNE, M. R.; PRESNELL, B.; TURLACH, B. A. On the lasso and its dual. *Journal of Computational and Graphical statistics*, Taylor & Francis, v. 9, n. 2, p. 319–337, 2000.

PARDALOS, P. M. et al. *Non-convex multi-objective optimization*. [S.l.]: Springer, 2017.

TIBSHIRANI, R. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, v. 58, n. 1, p. 267–288, 1996.

ZHOU, A. et al. Multiobjective evolutionary algorithms: A survey of the state of the art. *Swarm and Evolutionary Computation*, Elsevier, v. 1, n. 1, p. 32–49, 2011.