APRENDIZADO DE MÁQUINA BASEADO NA TEORIA DA INFORMAÇÃO

Palavras-Chave: Aprendizado de Máquina, Teoria da Informação, Regressão Linear

Autores(as):
JOÃO VITOR TREVIZOLI, FEEC – UNICAMP
Prof. Dr. ROMIS RIBEIRO DE FAISSOL ATTUX, FEEC – UNICAMP

INTRODUÇÃO:

A área da inteligência artificial (IA) está em franca expansão nos últimos anos [1]. Em especial, houve uma enorme explosão de aplicações de IA, que vão desde tratamento e criação de imagens e textos até o tratamento de áudio e a detecção facial [2]. Nesse contexto, é presente o tópico de aprendizado de máquina (ML, do inglês Machine Learning), uma área de pesquisa que surgiu como uma subespecialidade da inteligência artificial.

Com o grande aumento da quantidade de dados disponíveis atualmente, o conceito de aprendizado de máquina tomou uma maior importância, visto que consiste intrinsecamente em possibilitar um aprendizado a partir de dados, ou seja, possibilitar que modelos computacionais aprimorem suas habilidades de realizar determinadas tarefas ao serem expostos a eles [3].

Nesse contexto, os problemas ditos supervisionados, ou seja, aqueles baseados em uma etapa de treinamento sobre dados com um certo rótulo, são frequentemente agrupados em dois tipos de problemas: classificação e regressão [3].

Problemas de regressão envolvem a construção de um mapeamento contínuo a partir de uma amostra de dados, buscando prever valores numéricos. A solução desses problemas requer a escolha de um modelo de regressão, um critério para ajuste dos parâmetros do modelo e um método de otimização adequado. Em contraste, problemas de classificação visam categorizar amostras em classes distintas, ou seja, o conjunto de possíveis respostas do modelo é discreto e finito.

Ambos os problemas podem ser abordados com relação à teoria da informação, de onde surge o termo *information theoretic learning* (ITL); dessa forma, faz-se possível a utilização de métricas alternativas àquelas baseadas em estatísticas de segunda ordem..

Neste trabalho, busca-se realizar um estudo sobre os conceitos de ITL, juntando conceitos de aprendizado de máquina e teoria da informação, e fazer uma comparação e análise do problema de regressão linear utilizando um método clássico (Erro quadrático médio - EQM) e um método baseado na entropia de Rényi do erro.

METODOLOGIA:

A construção do problema utilizando o critério da entropia de Rényi foi feito utilizando o amplamente conhecido método do gradiente descendente. A principal questão suscitada é a caracterização da própria função de minimização e a construção de seu gradiente. Já o problema utilizando o erro quadrático médio foi feito com a solução fechada conhecida para a regressão linear.

Para a definição da entropia quadrática de Renyi e de seu gradiente, primeiro foi necessário a definição do chamado *estimador de Parzen*, que se trata de uma técnica não-paramétrica utilizada para estimar a função densidade de probabilidade de uma variável aleatória, particularmente útil quando se deseja obter uma representação contínua da densidade de probabilidade a partir de um conjunto de dados amostrais, por meio da utilização de uma função kernel (gaussiana no nosso caso). O estimador de Parzen foi feito a partir de (1) [4], utilizando o kernel gaussiano descrito em (2):

$$p(x) = \frac{1}{N}\sum_{i=1}^{N} K_{\sigma}\left(x - x_i\right) \qquad (1)$$

$$K_{\sigma}(x - x_i) = \frac{1}{\sqrt{2\pi}\sigma} exp\left(-\frac{(x-x_i)^2}{2\sigma^2}\right) \qquad (2)$$

Na Figura 1, é mostrada a estimativa de Parzen, com diferentes valores de desvio padrão da função kernel gaussiana, para uma variável aleatória com distribuição normal.

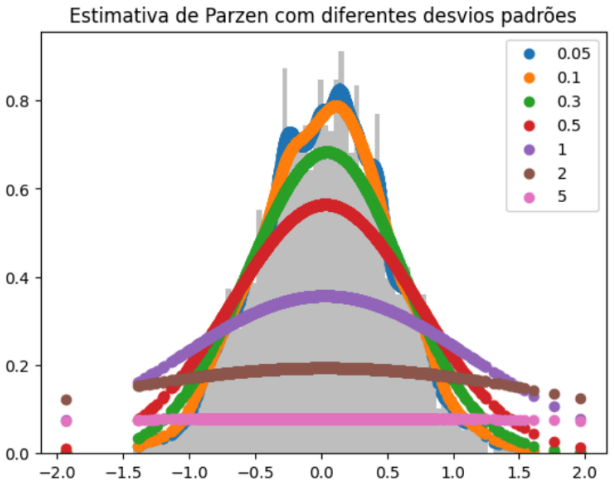

Figura 1: Estimador de Parzen

A definição da entropia de Rényi e seu gradiente, mostradas nas Equações (3) a (6) [4], foram feitas utilizando o estimador de Parzen. A estimativa de Parzen também ilustra o que podemos esperar de acordo com o valor do desvio padrão da função kernel.

$$V_2(X) = \int p^2(x)dx = \frac{1}{N^2}\sum_{j=1}^{N}\sum_{i=1}^{N} K_{\sigma\sqrt{2}}\left(x_j - x_i\right) \qquad (3)$$

$$H_2(X) = -\log V_2(x) \qquad (4)$$

$$F_2\left(e_j|e_i\right) = \nabla K_{\sigma}\left(e_j - e_i\right) \qquad (5)$$

$$\frac{\partial H_2(X)}{\partial w} = \sum_{j=1}^{N}\sum_{i=1}^{N} \frac{F_2\left(e_j|e_i\right)}{-N^2 V_2(e)}\left(\frac{\partial e_j}{\partial w} - \frac{\partial e_i}{\partial w}\right) \qquad (6)$$

Feitas essas definições, resta a caracterização dos problemas para teste e análise. Para verificar o impacto do desvio padrão do kernel, que já é esperado devido à estimativa de Parzen, fez-se um primeiro problema com um conjunto de dados simples. Foram criados quatro conjuntos de dados com 100 elementos cada, sendo dois para treinamento e dois para teste, com valores variando entre 0 e 1 por uma distribuição uniforme. Os dados de entrada utilizados para treinamento e teste foram a concatenação dos respectivos dois conjuntos, formando elementos de entrada com dois parâmetros. Os rótulos de saída dos dados foram feitos pela soma dos dois conjuntos de dados

originais, o que estabelece um problema de regressão linear simples. A Figura 2 mostra um esquema da montagem do conjunto.

$$A = [a_1, a_2, \ldots\ldots a_{100}] \qquad a_i \in \{0, 1\} \qquad C = [c_1, c_2, \ldots\ldots c_{100}] \qquad c_i \in \{0, 1\}$$

$$B = [b_1, b_2, \ldots\ldots b_{100}] \qquad b_i \in \{0, 1\} \qquad D = [d_1, d_2, \ldots\ldots d_{100}] \qquad d_i \in \{0, 1\}$$

$$X_{Treino} = [(a_1, b_1), (a_2, b_2), \ldots\ldots (a_{100}, b_{100})] \qquad X_{Teste} = [(c_1, d_1), (c_2, d_2), \ldots\ldots (c_{100}, d_{100})]$$

$$Y_{Treino} = [(a_1 + b_1), (a_2 + b_2), \ldots\ldots (a_{100} + b_{100})] \qquad Y_{Teste} = [(c_1 + d_1), (c_2 + d_2), \ldots\ldots (c_{100} + d_{100})]$$

Figura 2: Formulação dos dados para análise do impacto do desvio padrão da função kernel

Para a comparação dos modelos baseados no erro quadrático médio e na entropia de Rényi do erro, utilizaram-se 4 casos. Fez-se novamente o processo de criação de dois conjuntos de dados para treinamento e mais dois para teste, mas agora com uma distribuição uniforme com valores entre -1 e 1, e contendo 500 elementos. O mesmo processo de concatenação mostrado anteriormente definiu os dados de entrada de treinamento e de teste. As mesmas distribuições uniformes e os mesmos dados de entrada foram utilizados em todos os 4 casos, sendo a diferenciação de cada um apenas no mapeamento das saídas. Os casos são detalhados na Figura 3.

A e B: *distribuições uniformes geradas para treinamento/teste*

Caso 1:

$$Y_1 = \alpha A + \beta B \qquad \alpha \text{ e } \beta\text{: coeficientes aleatórios} \in \{-1, 1\}$$

Caso 2:

$$Y_2 = \alpha A + \beta B + \eta \qquad \eta\text{: ruído}$$

Caso 3:

$$Y_3 = \alpha A^1 + \beta B^3$$

Caso 4:

$$Y_4 = tanh(A + B)$$

* *OBS: As operações se referem a cada elemento das distribuições A e B*

Figura 3: Mapeamentos para cada caso

RESULTADOS E DISCUSSÃO:

Para a análise do impacto do desvio padrão da função kernel, foram observados os rótulos verdadeiros dos dados de entrada de teste contra os rótulos previstos pelo modelo criado baseado na entropia do erro de Rényi, como mostrado na Figura 4.

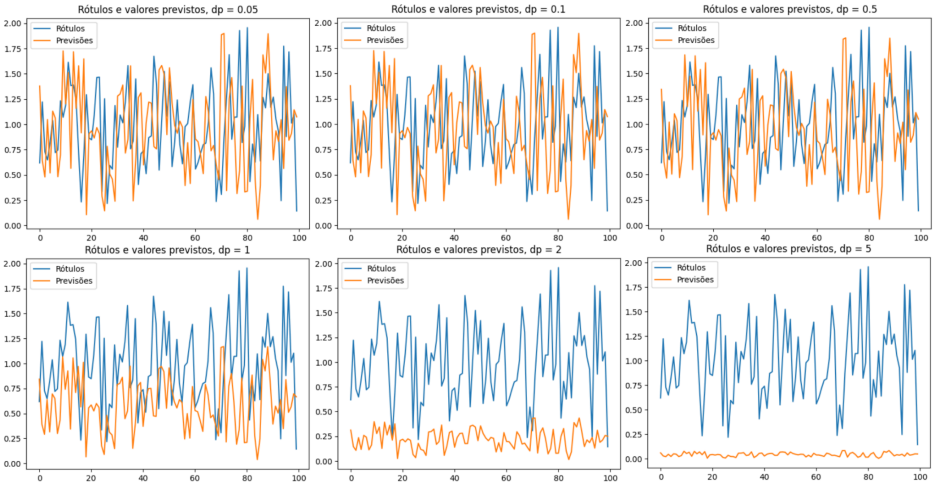

Figura 4: Rótulos verdadeiros x valores previstos para cada valor de desvio padrão

Como esperado, percebe-se a piora da previsibilidade do modelo quanto maior o valor de desvio padrão, tal fato também pode ser visto pela Figura 5, que relaciona a raiz do erro quadrático médio (*root mean squared error - RMSE*) entre os valores verdadeiros e previstos pelo modelo, com o valor do desvio padrão utilizado.

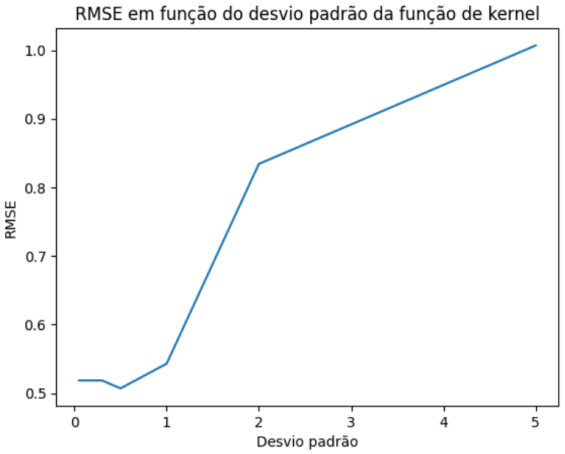

Figura 5: RMSE x Desvio padrão

Comparação dos RMSE		
	RMSE - Renyi	RMSE - Clássico
1.º Coef Aleatórios	3.64e-17	2.53e-16
2.º Ruído	0.0983	0.0985
3.º Polinomial	0.0637	0.0635
4.º Tanh	0.1265	0.1009

Tabela 1 - Comparação dos RMSE de cada caso

Na comparação entre o modelo clássico e o da entropia de erro de Rényi, foram medidos as raízes dos erros quadráticos médios para cada caso de problemas, que estão apresentados na Tabela 1.

Vemos que ambos os modelos têm desempenho melhor para o primeiro caso, já que se trata de um mapeamento linear sem ruído, o que melhor se encaixa com a técnica de regressão linear utilizada. O modelo baseado em Rényi teve um desempenho consideravelmente melhor nesse caso, mas no restante não houve grande diferença, indicando que a utilização de Rényi ainda é semelhante à abordagem clássica em problemas não ideais para a utilização de regressão linear. Entretanto, ao observar os histogramas de erro, na Figura 6, vemos que, com exceção do segundo caso (com ruído), ao utilizar o critério da entropia de Renyi do erro, foi possível obter mais amostras em torno do zero, ou seja, um histograma de erro menos espalhado. Isso possivelmente não ocorreu no segundo caso pois nenhum dos métodos tem um meio específico para se lidar com ruído.

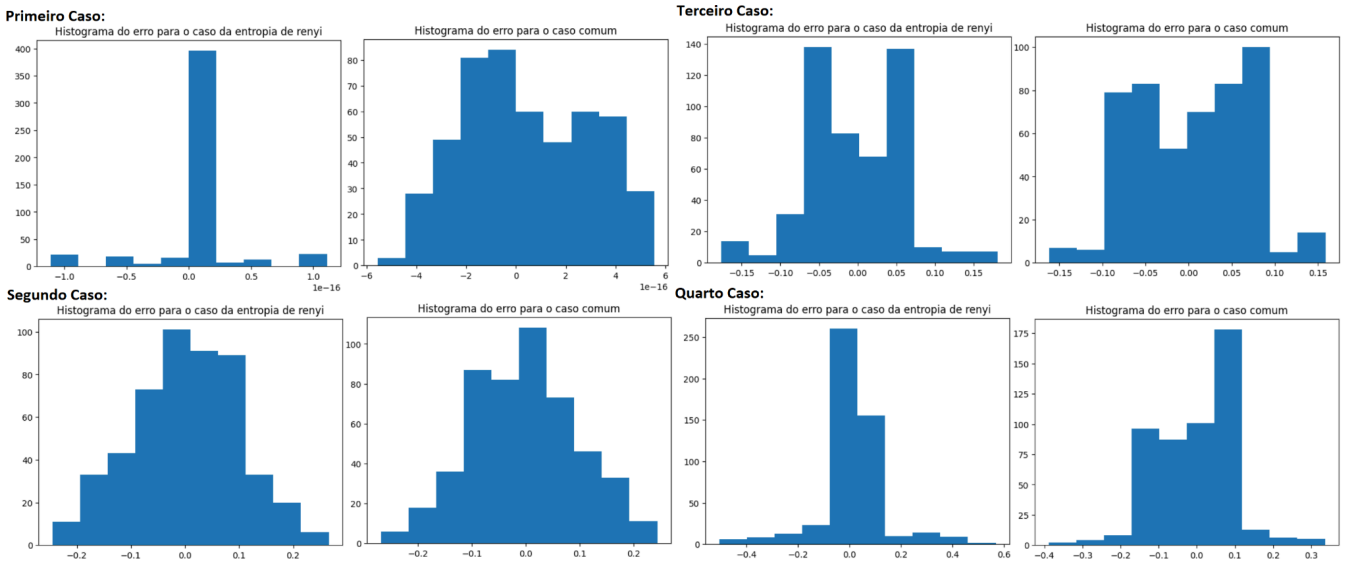

Figura 6: Histogramas do erro

CONCLUSÕES:

Utilizando uma técnica de ITL com o critério de minimização sendo a entropia de Rényi do erro, foi possível analisar a eficiência de um modelo de aprendizado de máquina feito a base do método do gradiente descendente. A validação do conteúdo teórico necessário para a construção do modelo passou pela definição das equações e métricas como o desvio padrão da função kernel, que mostrou impactar diretamente o processo de previsão.

Além disso, a comparação entre a técnica de ITL e a técnica clássica, feita com a solução fechada baseada na matriz pseudo-inversa, utilizando como critério de minimização o erro quadrático médio, possibilitou o vislumbre de impactos positivos na sua utilização.

A competitividade do modelo de ITL constatada pelo critério de avaliação RMSE, critério este que é o próprio alvo de minimização do método clássico, além da menor dispersão dos histogramas do erro, com amostras mais centradas em zero, mostram a possibilidade de aplicação desse método com alguma vantagem. Com isso, fica evidenciado a possibilidade de utilização de métodos de ITL em problemas de regressão linear, e a importância do estudo desse tema para avançar a utilização prática em diversas áreas de aprendizado de máquina e inteligência artificial.

BIBLIOGRAFIA

[1] I. Goodfellow, Y. Bengio, A. Courville. **Deep Learning**. MIT Press, 2016.

[2] A. Géron. **Hands-On Machine Learning with Scikit-Learn, Keras & Tensorflow**. O'Reilly, 2023.

[3] L. Boccato e R. Attux. **Notas de Aula do Curso IA048 – Aprendizado de Máquina**. FEEC/UNICAMP, 2020.

[4] D. Erdogmus e J. Principe. **From linear adaptive filtering to nonlinear information processing - The design and analysis of information processing systems**. IEEE Signal Process, 2006.

[5] A. Barua, M. U. Ahmed, S. Begum. **A Systematic Literature Review on Multimodal Machine Learning: Applications, Challenges, Gaps and Future Directions**. IEEE Access, Vol. 11, pp. 14804 – 14831, 2023.

[6] C. M. Bishop. **Pattern Recognition and Machine Learning**. Springer, 2006.

[7] D. J. C. MacKay. I**nformation Theory, Inference and Learning Algorithms**. Cambridge University Press, 2005.

[8] J. C. Principe. I**nformation Theoretic Learning: Rényi's Entropy and Kernel Perspectives**. Springer, 2010.