USO DE TÉCNICAS DE REGRESSÃO APLICADOS NOS DATASETS REAL ESTATE E MNIST. PARA USO EM SITUAÇÕES PROBLEMAS

Carlos Alex Martins Oliveira [1]
Débora Maciel de Oliveira Veras [2]

RESUMO

Este trabalho expõe a resolução de duas situações problemas. Esses dois problemas abordam conteúdos de Regressão Linear e Redes Neurais. Nos problemas serão implementados modelos de regressão linear múltipla de mínimos quadrados, rede Adaline e rede MLP (1 e 2 camadas ocultas). Tais problemáticas são apresentadas em duas situações: o problema 1 terá uma abordagem classificatória, já no problema 2 a apresentação tratará da regressão, ambos sustentados pela utilização de redes neurais. Esses problemas são uma simulação das situações que podemos encontrar no nosso cotidiano. Além disso, contém também uma contextualização comentada dessa temática, a partir de bibliografias diversas, por meio de livros, artigos, monografias, dissertações e teses analisadas.

Palavras-chave: Redes Neurais, Regressão Linear, Inteligência Computacional Aplicada

INTRODUÇÃO

Diante de situações problemas que encontramos em eventos reais, precisamos de embasamentos para chegar a uma dada decisão. A partir disto, usamos modelos computacionais relacionados a estatística para alcançar uma conclusão destes problemas.

A Matemática precisa estabelecer conexões com as técnicas mais avançadas de estatística para melhorar a elaboração de habilidades que envolvem tanto o ensino como a aprendizagem em todos os ramos.

Segundo Primi (2012) os computadores ficaram bem mais comuns e as análises estatísticas estão com uma maior acessibilidade, logo se precisa ter uma maior formação nas pós-graduações.

O método utilizado para a resolução destes problemas é a Regressão Linear Múltipla, onde usamos simulações para obter a melhor acurácia e as Redes Neurais

[1] Doutorando do Curso de Engenharia de Teleinformática da Universidade Federal do Cerará - UFC, calexmo@hotmail.com;
[2] Mestranda do Programa de Pós Graduação em Engenahria e Ciência de Materiais da Universidade Federal do Cerará - UFC, deb.maciel@yahoo.com.br;

Artificiais, que nos ajuda por ter como vantagem a não obrigatoriedade da linearidade entre as variáveis.

Na primeira situação, temos um conjunto de várias imagens representadas em um formato 28 x 28 *pixels*, em tons de cinza. Este formato representa uma matriz, onde cada elemento analisado está variando de 0 a 255, sendo que esses valores são todos inteiros. Sendo o 0 correspondendo a cor preta absoluta e o 255 será o branco.

A ideia do problema é reconhecer, a partir de uma figura, o número que ela representa.

As figuras são separadas em dois conjuntos, sendo um para treinamento e o outro para teste na proporção de 75% e 25%, respectivamente. Na etapa de treinamento, as imagens do conjunto de referenciado são apresentadas à rede neural a fim de que os pesos de cada neurônio sejam ajustados de modo a reconhecer o dígito que aquela imagem representa.

Na etapa de testes, após a rede estar com todos os seus pesos definidos, são apresentadas imagens, até então inéditas ao modelo, a fim de avaliar a taxa de acerto na classificação das imagens.

A prática de segregação dos dados em conjunto de treinamento e teste tem como objetivos avaliar a capacidade de extrapolação do modelo para novos dados e evitar o problema de overfitting, ou seja, situação na qual o modelo apresenta bom desempenho nos dados usados para o ajuste de pesos e um desempenho ruim para novos dados. Foram usadas duas topologias, uma MLP *(MultiLayerPerceptron*) de uma e duas camadas.

Na segunda situação, temos uma avaliação imobiliária com seis características na venda de um imóvel e uma com preço de venda por unidade de área: Data da venda do imóvel, Idade do imóvel, Lojas de conveniência, distância para a estação de trem mais próxima, Localização de latitude, Localização de Longitude e o Valor do imóvel por unidade de área.

Antes de começar a analisar os dados nos dois modelos, realizou-se o pré-processamento dos dados, normalizando-os de modo a ficarem na faixa de 0 a 1, conforme ilustrado na Equação 01, em sequência.

$$\frac{[x_1 \ x_2 \ x_3 \ x_4 \ x_5 \ \dots \ x_n] - x_{mínimo}}{x_{máximo} - x_{mínimo}} \qquad (01)$$

A normalização é usada para situações do cotidiano como os dos problemas, onde os valores estão apresentados em intervalos muito distantes. Os resultados, normalmente, são mais eficientes quando é usado esse processo (PAIVA,2020).

Na regressão Linear Múltipla usamos o erro médio quadrático, onde encontramos um valor para o erro na análise dos dados.

Foram avaliados também modelos de Redes Neurais Artificiais (RNA's) e os seus resultados foram comparados ao modelo de regressão.

O objetivo geral deste trabalho, portanto, é construir um modelo para cada técnica, tendo vários disponíveis, e comparar os resultados obtidos em cada um deles. Sendo o problema 1, uma classificação e o problema 2, uma regressão. Deste modo, foram estruturados os seguintes objetivos específicos, visando alcançar o que foi disposto anteriormente.

- Comparar as técnicas de Redes Neurais Artificiais e Regressão Linear Múltipla para resolver os dois problemas propostos no artigo;

- Verificar a efciência dos métodos na resolução dos problemas.

- Ultilizar métodos estatísticos para resolução de problemas no dia a dia para tomada de decisões.

METODOLOGIA

A princípio, para este trabalho foi realizada uma pesquisa do tipo bibliográfica, de modo a embasar teoricamente o assunto aqui abordado, conforme explorada na seção seguinte a esta.

Em seguida, foi realizado uma comparação com métodos estatísticos para verificar o uso dos mesmos, pois é impressindivel o uso de métodos estatístivos em todos os tipos de pesquisa.

REFERENCIAL TEÓRICO

Nesta seção encontram-se dispostos o embasamento teórico deste trabalho.

3.1 Regressão Linear Múltipla

Segundo (BAPTISTELLA *et al.*, 2005), usamos Regressão Linear Múltipla para verificar a relação entre as variáveis independentes e a variável dependente.

O intuito da regressão múltipla é de melhorar o modelo que está sendo trabalhado, para explicar como está sendo o comportamento das variáveis que estão sendo utilizadas (SASSI,2020).

Foi utilizado, desse modo, o erro médio quadrático (RMSE) para se encontrar o menor erro possível. Esse método foi utilizado, pois já consideramos os valores absolutos, porque trabalhamos com o quadrado dos resíduos. Logo, chegaremos melhor no resultado que procuramos, com a seguinte Equação 02:

$$RMSE_{\Delta x} = \sqrt{\sum_{i}^{N} \Delta X_i^{\,2} / N} \qquad (02)$$

Onde RMSE é a Representação Matemática do erro médio quadrático.

Sabendo que ΔX é o resíduo correspondente e o N representa o número de amostras.

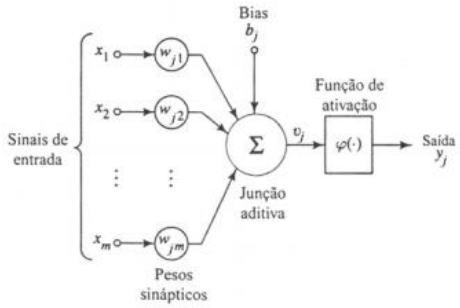
Exceção feita ao caso de Regressão Linear Múltipla, utilizaremos as Redes Neurais, com isso, falaremos de alguns modelos que serão usados para resolver os problemas.

Assim, as Redes Neurais Artificiais são estruturadas em camadas, podendo ser constituído por uma ou mais camadas.

3.2 Adaline

Nesse modelo, denominado ADALINE (*ADAptiveLInearNEuron*) o neurônio é uma entidade que possui n entradas ($x_1, x_2, x_3, ..., x_n$), cada uma delas associadas a pesos ($w_1, w_2, w_3, ..., w_n$) chamados sinapses. Conforme mostrado na figura abaixo, o neurônio realiza o somatório dos produtos $x_i w_i$, chegando a um valor j. Esse valor é aplicado na função de ativação e finalmente tem-se o valor de saída y . Esse modelo foi desenvolvido por Windrow e Holf em 1960. Essa saída é baseada no método do gradiente para minimização do erro (GUARNIERI, 2006).

Figura 01 – Modelo de um neurônio de uma rede MLP



Fonte: Guarnieri, 2006.

Além disso, podemos ter Redes neurais com multicamadas, as quais serão apresentadas posteriormente.

3.3. Madaline

Na rede Adaline, o reconhecimento é somente de um padrão simples, a partir daí, surgiu a ideia de criar um elemento composto, o Madaline. Esse modelo apenas agrupa a resposta de dois ou mais Adalines (OSORIO, 1991).

3.4 MLP

Conhecida pela sigla MLP, com nome MultiLayerPerceptron, tem maior popularidade, geralmente com o algoritmo de retro propagação do erro (backpropagation), baseado na regra delta proposta por Widrow e Hoff, e suas variações.

Quando trabalhamos na rede MLP, tem-se uma vantagem, pois se de trabalhar de forma não linear.

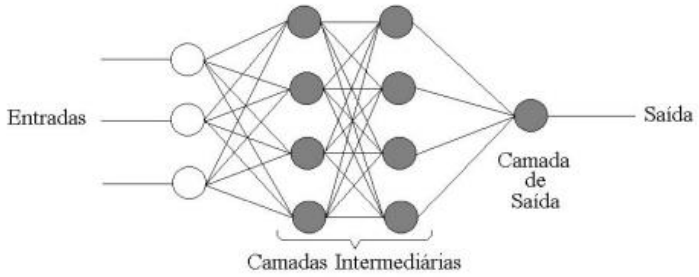
As redes multicamadas são constituídas de:

- Camada de entrada: Interface de entrada para recepção das informações do problema em análise, podendo ser constituída de um ou mais neurônios.

- Camadas Ocultas: uma rede MLP pode apresentar uma ou mais camadas intermediárias, com cada camada constituída por um ou mais neurônios. Nas camadas ocultas são realizados os processamentos que permitem a extração de padrões a fim de determinar a saída adequada, com base no processo de treinamento.

- Camada de saída: interface onde é apresentada a saída do modelo. Assim como as demais camadas, a quantidade de neurônio pode ser arbitrada.

Figura 02 – Exemplo de uma RNA-MLP com duas camadas intermediárias



Fonte: Guarnieri, 2006.

Conforme figura acima, temos uma rede MLP com duas camadas ocultas com quatro neurônios em cada camada e uma camada de saída com um neurônio.

Segundo Haykin (1994), o modelo MLP de treinamento supervisionado, é um dos modelos mais utilizados.

Em função da flexibilidade para escolha da quantidade de camadas, neurônios e funções de ativação, existe um elevado número de configurações possíveis. Sendo necessário a definição de critérios para escolha da configuração que melhor atende aos objetivos do problema avaliado.

3.4.1. Backpropagation

Também chamado de Retropropagação, esse algoritmo de treinamento supervisionado para redes MLP. Esse algoritmo é dividido em duas fases, sendo a primeira a propagação do sinal *feedforward*, onde nela, partindo das entradas, se propaga, mantendo fixos os pesos, até a geração de saída. Os pesos são ajustados com o intuito de se minimizar o erro, sendo os pesos ajustados pelo método de todas as camadas e não apenas na camada de saída e tendo como base a regra delta. O ajuste dos pesos acontece com a retropropagação do erro, ou seja, da saída para a entrada.

Para avaliação do erro é utilizado a função custo que é definida em *t* treinamentos, para cada neurônio *j* de saída, como o somatório de *J* sinais de erro:

$$E(t) = \frac{1}{2}\sum_{j=1}^{J} e_j^2(t) \qquad (03)$$

Onde RMSE é a Representação Matemática do erro médio quadrático.

É utilizado o método do gradiente para minimização do erro, conforme Equação 04.

$$\Delta w_{ji}(t) = \eta \delta_j(t) x_i(t) \quad (04)$$

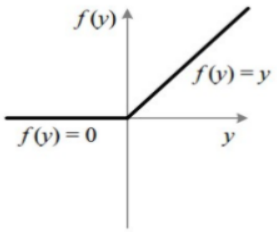
RESULTADOS E DISCUSSÃO

Após a etapa de treinamento da rede neural foram apresentados os dados do conjunto de testes, normalizamos o conjunto deixando em um intervalo de 0 a 1. Foi avaliado o grau de acerto, ou seja, a correta classificação da imagem, obtendo-se um percentual de 97,89%, onde foi feito com duas camadas, sendo uma com 128 neurônios e a segunda camada com 10 neurônios. Foi utilizado o otimizador Adam, que é uma otimização do algoritmo Backpropagation que o torna mais efetivo, sendo baseado no gradiente descendente estocástico.

Na primeira camada, foi utilizada a função de ativação ReLU (Função de Ativação Linear Retificada), foi usada porque, por muitos, ela se tornou uma função de ativação padrão para muitos tipos de redes neurais, pois ela geralmente tem melhores resultados, além de ser mais fácil o seu treino.

Além de tudo, a função não terá problema de desempenho como a Sigmoid e Tanh, visto que ela considera o primeiro quadrante apenas, os demais são retirados.

Figura 03 – Função ReLU



Fonte: Vaz, 2018.

Já na segunda camada, utilizamos a função softmax, como visto, uma vez que o problema 1 é de classificação, e essa é uma excelente função quando se está trabalhando com problemas de classificação de várias classes.

$$\frac{e^{z_i}}{\sum_{j=1}^{k} e^{z_j}} \quad (05)$$

No caso do problema 1, a métrica utilizada para a avaliação do modelo foi a acurácia, definida como a correta classificação da imagem entre os dez dígitos possíveis.

Foi utilizada a rede MLP com 1 camada oculto com 128 neurônios. A função de ativação foi a *Softmax* e a Acurácia foi de 0.9778(97,78%).

Quando foi modelado com a rede MLP com 2 camadas, se chegou a um valor bem próximo de Acurácia igual a 0.9783 (97,83%)

Figura 04 – Classificação MLP 1 camada

Fonte: Autor, 2021.

Quando se faz uma análise do problema 2, é feita uma regressão dos dados do problema a fim de realizar previsões do preço do imóvel a partir de suas características já citadas.

As técnicas utilizadas foram, a Regressão Linear Múltipla, a Gradiente Descendente Estocástico, a rede ADALINE, a MLP 1 e 2 camadas.

O primeiro modelo foi o de Regressão Linear Múltipla. Se chegou na seguinte função de regressão:

$$y = -14305{,}4 + 5{,}072097x_1 - 0{,}2675x_2 - 0{,}00447x_3 + 1{,}150848x_4 + 226{,}7277x_5 - 12{,}5379x_6$$

Onde se chegou aos seguintes resultados, conforme Tabela 01.

Tabela 01 – Estatística de Regressão

Estatística de regressão	
R múltiplo	0,763891
R-Quadrado	0,583529
R-quadrado ajustado	0,577375
Erro padrão	8,856257
Observações	413

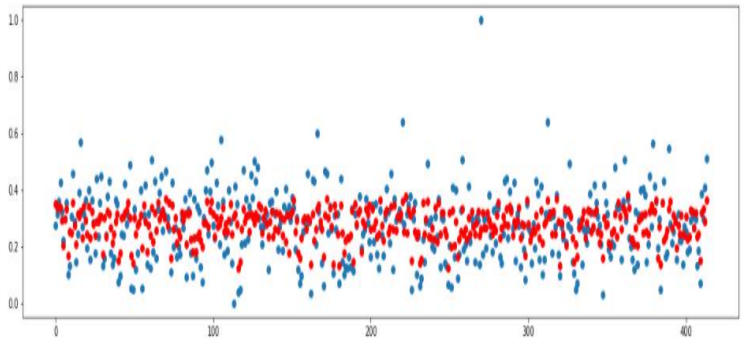
Fonte: Autor, 2021.

O R-Quadrado tem a capacidade de explicação do resultado. Assim como as seis entradas tem a capacidade de explicar 60% do preço.

Para se fazer um comparativo com os outros modelos, foi colocado os valores no erro médio quadrático, já falado inicialmente, no qual se chegou ao resultado de 0,01076.

O segundo modelo foi o Gradiente Descendente Estocástico, em que se chegou ao seguinte resultado: 0.00886465094637318, deixado no mesmo parâmetro.

Figura 05 – Gráfico Gradiente Descendente Estocástico



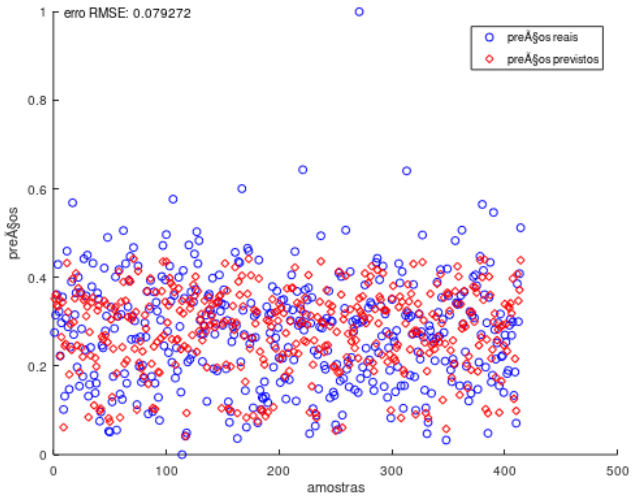
Fonte: Autor, 2021.

Representação visual dos dados, com azul para os preços reais e vermelho para os preços previstos.

Na rede ADALINE, foram feitos vários testes para se chegar a um erro mais próximo dos demais modelos, demandando muito tempo e uma quantidade elevada de testes.

Foi colocado uma taxa de aprendizagem de 0,1 e com o número de 5000 iterações, chegando-se em um erro de 0,079272.

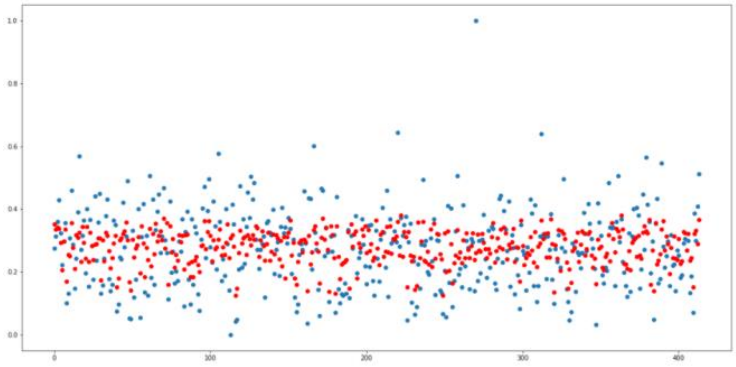
Figura 06 – Gráfico ADALINE



Fonte: Autor, 2021.

Na camada MLP com 1 camada oculta, chegamos ao erro de 0,006474283385152889.

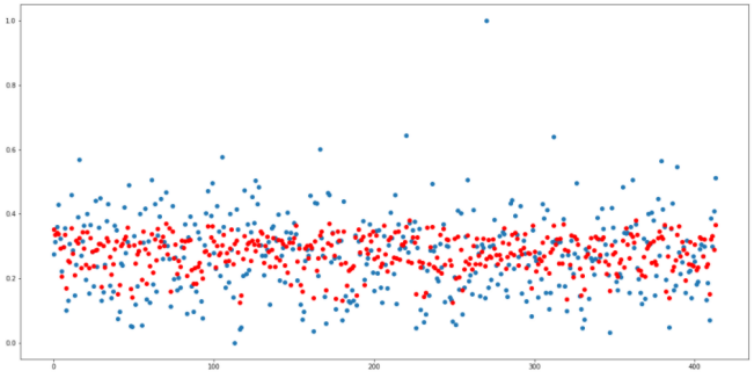
Figura 07 – Gráfico MLP com 1 camada



Fonte: Autor, 2021.

No que se refere à camada MLP com 2 camadas ocultas, chegamos ao erro de 0,0060110004802975585

Figura 08 – Classificação MLP 2 camadas



Fonte: Autor, 2021.

Colocando os resultados em uma tabela para uma melhor visualização, tem-se:

Tabela 01 –Resultados

Resultados	
Problema 1	
Rede Neural	**Acurácia**
MLP 1 camada	97,78%
MLP 2 camada	97,83%
Problema 2	
Regressão / Rede Neural	**RMSE**
Regressão	0,01076
Gradiente Desc. Esto.	0,00886
ADALINE	0,07927
MLP 1 camada	0,00647
MLP 2 camada	0,00601

Fonte: Autor, 2021.

CONSIDERAÇÕES FINAIS

Este trabalho teve como objetivo comparar as técnicas de Redes Neurais Artificiais e Regressão Linear Múltipla para resolver dois problemas. O primeiro teve como objetivo uma melhor classificação em reconhecimento de um número, onde quando se foi testado a MLP de 1 camada, já se chegou em um valor bem satisfatório de resultado, porém, ainda assim, foi feito o teste com uma MLP de duas camadas, chegando a conclusão inicial, que não é necessário ser feito, pois a acurácia teve um aumento insignificante, quando se é analisado um montante grande de exemplos.

No problema 2, os resultados ficaram bem próximos uns dos outros na comparação dos modelos, porém alguns demandaram mais trabalho e exigiram mais tempo para se chegar a um valor tão aproximado. Dessa forma, podemos perceber que existem outliers, em que muitos valores reais se dispersam dos valores previstos. Nesse problema, a Regressão foi o que ficou com maior erro, enquanto a rede MLP com duas camadas ocultas teve o menor erro.

AGRADECIMENTOS

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Código de Financiamento 001.

REFERÊNCIAS

BAPTISTELLA, Marisa; STEINER, Maria Teresinha Arns; NETO, Anselmo Chaves. *O uso de redes neurais e regressão linear múltipla na engenharia de avaliações: Determinação dos valores venais de imóveis urbanos.* **Diss., Universidade Federal do Paraná**, 2005.

BROWNLEE, Jason. *A gentle introduction to the rectified linear unit (relu)*. **Machine Learning Mastery. https://machinelearningmastery. com/rectified-linear-activation-function-fordeep-learning-neural-networks**, 2019.

GUARNIERI, Ricardo André. *Emprego de redes neurais artificiais e regressão linear múltipla no refinamento das previsões de radiação solar do modelo* **Eta. Insituto Nacional de Pesquisas Espaciais**, 171pp, 2006.

HAYKIN, S. Neural networks: *a comprehensivefoundation*. **New York: MacmillanCollegePublishingCompany,** 1994.

HAYKIN, S. *Redes neurais: princípios e prática, 2. ed.* **Porto Alegre: Bookman Companhia Editora,** 2001. 900p.

KINGMA, Diederik P.; BA, Jimmy. *Adam: A method for stochastic optimization*. **arXiv preprint arXiv:1412.6980,** 2014.

KOVÁCS, Zsolt László. *Redes neurais artificiais.* **Editora Livraria da Fisica,** 2002.

OSORIO, Fernando Santos. *Um estudo sobre reconhecimento visual de caracteres através de Redes Neurais***.** 1991.

PAIVA, Cloves. *Porquê e quando é necessário normalizar os dados. Medium.* 2019. **Disponível em: <https://medium.com/tentando-ser-um-unic%C3%B3rnio/porqu%C3%AA-e-quando-%C3%A9-necess%C3%A1rio-normalizar-os-dados**>. Acesso em: 28 Set. 2020.

PRIMI, Ricardo. *Psicometria: fundamentos matemáticos da Teoria Clássica dos Testes*. **Avaliação Psicológica, v. 11, n. 2, p. 297-307**, 2012.

SASSI, Cecília P. et al. *Modelos de regressão linear múltipla utilizando os softwares R e STATISTICA: uma aplicação a dados de conservação de frutas*. 2012.

VIEIRA, Helder Barros Gama; GENRO, Rafael Santos. *Estimativa da acurácia posicional de documentos cartográficos na Petrobras a partir do erro máximo provável inferido do erro médio quadrático e da respectiva variância propagada*. **XVI Simpósio Brasileiro de Sensoriamento Remoto-SBSR, Foz do Iguaçu, PR,** 2013.

VAZ, Arthur Lamblet. **Introdução teórica a Neural Network – Deep Learning.** 2018. Disponível em: <https://medium.com/data-hackers/neural-network-deep-learning-parte-1-introdu%C3%A7%C3%A3o-te%C3%B3rica-5c6dcd2e5a79>. Acesso em: 01 Out. 2020.