Trabalho apresentado no XLI CNMAC, Unicamp - Campinas - SP, 2022.

Proceeding Series of the Brazilian Society of Computational and Applied Mathematics
Preprint

Uma aplicação do Perceptron para previsão de mortalidade da COVID-19

João P. M. Terra, Fabiano S. Oliveira, Luerbio Faria[1]
UERJ, Rio de Janeiro, RJ

COVID-19 é uma doença causada pelo vírus coronavirus SARS-CoV-2, o qual é altamente contagioso. O primeiro caso oficial de COVID-19 surgiu em dezembro de 2019 em Wuhan, China, tendo assolado o mundo com mais de 6.130.000 de fatalidades até o momento.

O aprendizado de máquina tem sido usado para previsão de mortalidade do COVID-19 em diferentes abordagens. Nesse artigo, empregamos o aprendizado de máquina para realizar uma previsão do nível de gravidade de pacientes de COVID-19 aplicada à realidade brasileira. Nosso sistema surge em paralelo a várias outras publicações muito recentes que propõe ferramentas com objetivos semelhantes. Moulaei et al. [4] usaram aprendizado de máquina para a previsão de mortalidade de COVID-19, no qual o método com melhor resultado foi o *random forest* (95% de acurácia). Esse método desempenhou um pouco melhor que o perceptron multicamada, porém analisou apenas 1500 pacientes provindos da província do Khuzestan, no Irã. Borghi [1] et al. usaram o Perceptron com uma e várias camadas para prever número de mortes e infectados ao longo de seis dias. Os dados usados pelos autores foram de 185 países diferentes em um intervalo de 20 dias. No Brasil, temos um banco de dados nacional, o DATASUS, o qual centraliza as informações da evolução do COVID-19. Um dos dados fornecidos de cada paciente é a condição de remido/óbito. Elaboramos um sistema que, dado um formulário preenchido do DATASUS, estabelece com alta acurácia se o paciente é ou não de risco, baseando-se nos dados históricos das classes Óbito/Remissão, para servir como uma ferramenta auxiliar para o médico.

O algoritmo base escolhido foi o Perceptron [2], que é um algoritmo que recebe como entrada com conjunto $\mathbf{X} = \{X_1, X_2, \ldots, X_n\}$ com n pontos de $\mathbb{R}^d$, $X_i = (x_{i,1}, \ldots, x_{i,d})$ (sintomas do paciente X_i) e uma função $y : \mathbf{X} \to \{0, 1\}$ e produz uma função $f : \mathbb{R}^d \to \{0, 1\}$ para a classificação binária de pontos arbitrários. O Perceptron determina um vetor de pesos $W = (w_1, \ldots, w_d) \in \mathbb{R}^d$ e um valor de limiar B tal que $f(X_i) = y(X_i)$ para todo $i \in \{1, \ldots, n\}$, onde $f(X) = \text{sinal}(\sum_{j=1}^{d} x_j \cdot w_j + B)$ e $\text{sinal}(x) = 1$ se $x > 0$ ou $\text{sinal}(x) = 0$, caso contrário. A previsão para um outro ponto $X \in \mathbb{R}^d$ se dá usando a função f aplicada a um novo paciente X, isto é, $f(X) = 1$ tem remido como previsão, enquanto $f(X) = 0$ de óbito.

dataevolucao	dataentubado	surtosg	nosocomial	avesuino	febre	tosse	garganta	dispnea	desc_resp	hospital	uti	suport	raiox	amostra
diarreia	vomito	outrosin	puerpera	fatorrisc	saturacao	cardiopati	hematologi	sinddown	hepatica	amostra	pospcrflu	tpflupcr	histovgm	ps
asma	diabetes	neurologi	pneumopati	imunodepre	renal	obesidade	vacina	co_uni_not	sexo	dorabd	fadiga	perdolft	perdpala	tomo
idade	gestante	raça	escolaridade	uf	co_mun_res	cs_zona	antiviral	tpantivir	criterio	tptesan	res_an	tp_sor	igg	resposta

Na Tabela acima representamos os dados utilizados na fase de treinamento e classificação do Perceptron, sendo que a feature "Resposta" representa a evolução do paciente (remido ou óbito). Cada dado foi codificado em binário para o Perceptron. Um critério geral para dado faltante pode ser visto por exemplo, para febre codifica-se com (1,1), ausência (1,0) e não preenchido (0,0).

O Perceptron é um separador linear, isto é, ele produz um hiperplano de dimensão $d - 1$, cujos pontos na região acima separada pelo hiperplano são classificadas como 1 enquanto abaixo como 0. O procedimento é iterativo: a partir de um conjunto de pesos e um limiar arbitrário, se o

[1]joaopedromterra@gmail.com, fabiano.oliveira@ime.uerj.br, luerbio@ime.uerj.br

hiperplano obtido não separa corretamente a entrada, então o hiperplano é modificado, com algum peso w_i ajustado, que passa a definir o hiperplano corrente e o processo se repete até a obtenção do hiperplano separador correto. Determinamos experimentalmente que o banco de dados de pacientes não é linearmente separável, ou seja, não passível de ser resolvido pela técnica padrão do Perceptron. Ao invés de modificar e descartar um hiperplano que não separou a entrada, armazena-se o hiperplano que melhor separou os dados até o momento corrente, semelhante ao procedimento do "pocket algorithm" em [3]. O hiperplano é devolvido, após um certo número de iterações.

Os resultados apresentados nesse artigo usam o Perceptron implementado na linguagem de programação Javascript. Utilizamos o banco de dados disponibilizado pelo DATASUS de 2021, com 1.185.228 pacientes. Cada registro do paciente é descrito por 80 campos. O banco de dados bruto foi tratado excluindo os pacientes com menos de 25 campos preenchidos e com resultados de PCR negativos. O banco tratado ficou com 440.915 pacientes: 271.685 remidos e 169.230 óbitos.

Os parâmetros finais do algoritmo foram ajustados experimentalmente, até a obtenção de resultados satisfatórios. Para utilizar o algoritmo gerador e classificador, particionamos o banco tratado em treinamento e validação. A parte de treinamento é formada por 20.000 pacientes aleatórios, dividida em 10.000 remidos e 10.000 óbitos. A parte para validação consiste dos 420.915 pacientes restantes. Na etapa de treinamento, faz-se a cada turno um total de 10.000 iterações do Perceptron e a cada iteração $i = 1, \ldots, 10000$ do algoritmo, calcula-se a porcentagem de acerto do hiperplano corrente para os 20.000 pacientes. No final das iterações, o algoritmo retorna o hiperplano com maior porcentagem de acerto total em relação aos 20.000 pacientes para a validação. Na etapa de validação, os demais 420.915 pacientes são usados de entrada para o hiperplano gerado e é verificado a porcentagem de acerto tanto dos remidos quanto dos óbitos. Os resultados de nosso experimento com o banco de dados gerado é estável, com média de acerto de 67% dos remidos e nos óbitos. Para melhorar a acurácia total do sistema, armazenamos um conjunto com os melhores 103 hiperplanos gerados. O resultado da validação foi redefinido para um processo de votação com vitória por maioria, no qual cada hiperplano "vota" com sua previsão particular. O conjunto de hiperplanos aumentou a acurácia para 77,5%, sendo 80,4% de acerto nos remidos e 74,6% de acerto nos óbitos. Todos os programas, dados e resultados estão disponíveis *online* `https://drive.google.com/drive/folders/12yiOaph3OgceJAmq4j3gTLzS8WlIsKxf?usp=sharing` Pretendemos modificar o sistema para previsão de internação e entubação, além de aumentar o número de camadas de neurônios aplicando outras técnicas de aprendizado para comparar os resultados.

Agradecimentos

Trabalho realizado com apoio da CAPES - Código de Financiamento 001, CNPq e FAPERJ.

Referências

[1] Pedro Henrique Borghi, Oleksandr Zakordonets e João Paulo Teixeira. "A COVID-19 time series forecasting model based on MLP ANN". Em: **Procedia Computer Science** 181 (2021), pp. 940–947.

[2] Luerbio Faria et al. **Ciência de dados: algoritmos e aplicações**. pt. 1ª ed. Vol. 5. 33º Colóquio Brasileiro de Matemática. Course of the 33º CBM, Aug 02–06, 2021. Rio de Janeiro: Editora do IMPA, 2021, p. 276.

[3] Stephen I. Gallant. "Perceptron Learning and the Pocket Algorithm". Em: **Neural Network Learning and Expert Systems**. 1993, pp. 63–94.

[4] Khadijeh Moulaei et al. "Comparing machine learning algorithms for predicting COVID-19 mortality". Em: **BMC medical informatics and decision making** 22.1 (2022), pp. 1–12.