Relatório de Desafio de Dados

Autor:
Vitor Martins Barbosa

12 de dezembro de 2021

Conteúdo

1 Introdução

Uma das grandes problemáticas nos hospitais é a detecção tardia da deterioração clínica. O agravamento do quadro clínico de pacientes pode ter várias consequências; como a sepse, por exemplo, que soma mais de 6 milhões de mortes por ano mundialmente. Muitas dessas mortes poderiam ser evitadas com o reconhecimento antecipado do risco, isto é, da deterioração clínica. Os hospitais usam protocolos com base, principalmente, nos sinais vitais para atribuir um escore de risco aos pacientes. No entanto, esses protocolos, como o NEWS (National Early Warning Score) e MEWS (Modified Early Warning Score), assim como outros protocolos, têm muitos falsos positivos e baixa precisão para identificar os pacientes em risco.

Por outro lado, a predição da deterioração clínica dos pacientes também pode ser feita através de algoritmos de Aprendizado de Máquina; com o potencial de serem muito mais precisos em detectar os pacientes em risco do que os protocolos convencionais

2 Objetivo

O objetivo deste algoritmos é realizar a construção de algoritmos de preditividade para óbito, baseado nos dados passados.

3 Metodologia

3.1 Separação de Base de Dados

A base de dados para o sistema desenvolvido considera duas classes "Melhora"e "Óbito". Entretanto, afim de realizar a validação dos dados, consideramos uma base equilibrada com 766 dados (sendo 50% para cada classe).

Para realizar o treinamento e validação, a base de dados foi exposta a aleatoridade e o mesmo foi treinada com aproximadamente 70% do seu total e validada com 30%.

3.2 Classificação

Classificadores podem ser entendidos como uma técnica de *machine learning*, reconhecendo ou identificando classes de padrões existentes. O classificador consiste em três estágios: treinamento, validação e operação. Na etapa de treinamento se emprega uma base de dados rotulada para "ensinar"o sistema, ou seja, para encontrar os hiperplanos ou curvas que separam ou delimitam cada uma das classes. Num segundo momento, uma outra partição da base de dados, também rotulada, é utilizada para validar se o sistema de discriminação gerado é válido e estimar o desempenho na discriminação de dados desconhecidos.

Na validação pode-se projetar a taxa de acerto que o classificador deve apresentar durante a sua operação, na qual a base de dados de entrada não é rotulada, e portanto não é simples verificar se houve acerto ou erro do sistema. Neste trabalho, foi implementado o classificador linear (método dos quadrados mínimos) e não linear (Rede Neural Artificial).

3.2.1 Classificador Linear

No classificador linear se utiliza uma função discriminante, dada por:

$$y(X) = X^T w \tag{1}$$

Onde X é a matriz de características e w é o vetor de pesos, que pode ser estimado utilizando o método dos mínimos quadrados (MMQ).

O MMQ consiste em minimizar a seguinte função de custo:

$$J(w) = \sum_{i=1}^{N} (y_i - X_i^T w)^2 \tag{2}$$

cuja solução resulta em:

$$w = (X^T X)^{-1} X^T y \tag{3}$$

Após o treinamento do sistema com a determinação de w, a validação é realizada pelo cálculo de y. O resultado final do sistema de classificação é dado pelo maior valor obitdo pela expressão 1, sendo que:

- $y > 0$ indica que o dado pertence a classe
- $y < 0$ indica que ele não pertence a classe.

3.2.2 Classificador Não Linear

A RNA é um dos métodos matemáticos mais eficientes utilizados para classificar um conjunto de dados, tendo uma boa acurácia, porém alto custo computacional. A demanda do cálculo é pelo fato do sistema implementado realizar o seu treinamento com base em dados passados, num esquema de realimentação.

A RNA é divida em diversas unidades de processamento (neurônios) que, em geral, são conectadas através de pesos. Uma rede pode ter diversas entradas e diversas camadas de neurônios. Para realizar a classificação de um sistema, são realizados os seguintes procedimentos:

- Os sinais que serão classificados são inseridos na entrada da RNA (X_n);
- A entrada aplicada recebe um valor do peso;
- É realizada uma soma ponderada no sistema que resultará na resposta da atividade;
- A resposta da atividade é comparada com a função de ativação, gerando uma resposta para entrada.

4 Resultados

Para análise dos resultados de ambos os classificadores os mesmos foram submetidos a 20 iterações, considerando assim a taxa de acerto média e o desvio padrão do mesmo.

Os resultados para o classificador linear se mostraram promissores, considerando todos os dados relevantes a sintomas para o tratamento de dados, a taxa de acerto foi de aproximadamente 70% e desvio padrão de 9.5%.

Considerando a rede neural, a taxa de acerto foi de 68.5% com um desvio padrão de 24.5% o que torna um algoritmo não muito seguro de ser utilizado, mas ainda deve ser trabalhado para realizar melhorias.

5 Conclusão

Fato a ser considerado, a base de dados se mostra com um grande potencial para ser tratada e salvar muitas vidas através de classificação.

A resposta para o classificador linear mostra que o mesmo é possível de ser tratado linearmente, assim, uma abordagem que pode ser considerada seria abordar outros classificadores lineares. Além disso, considerando a rede neural artificial, a mesma pode ser melhor trabalhada, afim de realizar as devidas variações de parâmetro, como função de ativação e número de neurônios.

O código pode ser clonado através do link do github ou acessado através do drive:

- <https://github.com/vitor-martinsb/classifier_disease.git>;
- <https://drive.google.com/drive/folders/1sD5id5aCTgnL5fXyhptnQmgVTzrzANad?usp=sharing>.