Proceedings of the V Brazilian Conference on Neural Networks - V Congresso Brasileiro de Redes Neurais
April 2–5, 2001 - Rio de Janeiro - RJ - Brazil

Comparação de Redes Neurais com Regressão Logística

Alcione Miranda dos Santos[1] , Basílio de Bragança Pereira[1,2], Kátia V. Bloch[2],
Carlos H. Klein[2], Nelson A. de S. e Silva[3], Armando da R. Nogueira[4], Lúcia H. S. Campos[4]
[1]COPPE-UFRJ
[2] Faculdade de Medicina - Núcleo de Estudos de Saúde Coletiva-UFRJ
[3] Faculdade de Medicina-UFRJ
[4] Hospital Universitário Clementino Fraga Filho-UFRJ
E-mails: amiranda@pep.ufrj.br,basilio@pep.ufrj.br

Abstract

In Medicine, there is a great interest in knowing what is the probability of an individual (patient) belonging to a specific diagnosis group. Several classic statistical techniques are used. Recently, neural artificial network has been used in the diagnosis and medical therapeutics.

This work aims at comparing the logisitic regression method and neural artificial network considering their prediction error. A data set, whose dependent variable assumes two categories, is used to compare the two methods.

1. Introdução

Uma aplicação comum de redes neurais está relacionado a problemas de classificação, onde o interesse está em decidir a qual grupo um determinado indivíduo pertence. Este tipo de problema pode ser resolvido através de técnicas estatísticas, entre elas, podemos citar *Regressão Logística*. Entretanto, redes neurais têm demonstrado, em certas situações, uma capacidade de classificação superior a estas técnicas.

Ultimamente, redes neurais artificiais vêm sendo utilizadas em um número crescente de tarefas de classificação de doenças, controle de equipamentos biomédicos inteligentes, processamentos de sinais e imagens, etc. Na área médica, onde se pretende utilizar as redes neurais como instrumento de apoio ao diagnóstico, esta metodologia é atraente e se tem mostrado eficiente.

O modelo de regressão logística é, em geral, utilizado para tratar problemas relacionados a dados dicotômicos em várias áreas de conhecimento. Em particular, regressão logística é bastante utilizada na área médica quando o interesse é saber qual a probabilidade de um indivíduo pertencer a um determinado grupo de diagnóstico.

Neste trabalho procuramos avaliar a capacidade de predição do modelo de regressão logística e o da rede neural artificial. O objetivo principal é identificar o modelo que possui menor erro de predição.

Para desenvolvermos nosso estudo, utilizamos um conjunto de dados referente a hipertensão arterial na Ilha do Governador (região administrativa do município do Rio de Janeiro). O conjunto de dados é composto de 1268 indivíduos adultos, de 20 anos e mais, residentes nos domicílios selecionados na Ilha do Governador. Características sócio-demográficas, medidas antropométricas, hábitos relacionados ao tabagismo, prática de exercícios físicos e pressão arterial foram observadas em tais indivíduos, entretanto, neste trabalho iremos apenas trabalhar com as seguintes variáveis: sexo, escolaridade , hábito relacionado ao tabagismo, prática de exercícios físicos, pressão arterial, índice de massa corporal, cor e a idade. Os indivíduos são classificados como hipertensos ou não hipertensos. O critério de hipertensão arterial utilizado foi o da Organização Mundial de Saúde(OMS), que determina valores da pressão arterial sistólica maiores ou iguais a 160mmHg ou da pressão arterial diastólica maiores ou iguais a 95mmHg como hipertensão arterial. A prevalência de hipertensos na amostra foi igual a 24,8%.

2. Materiais e Métodos

2.1. Análise via Rede Neural Artificial

Primeiramnte, a amostra em estudo foi dividida em dois subconjuntos. O primeiro, denominado *conjunto de treinamento*, composto de 1000 indivíduos selecionados aleatoriamente, é utilizado para a obtenção do modelo neural, e o *conjunto de teste* com 268 indivíduos, para a verificação do erro de predição.

Uma rede neural artificial feedforward([1][2])com sete neurônios na camada de entrada, cinco neurônios na camada escondida e um neurônio na camada de saída, visto que a variável resposta é binária, foi planejada. As variáveis de entrada foram reparametrizadas para receberem valores entre 0 e 1.

A rede foi treinada, utilizando o algoritmo backpropagation, por um número determinado de épocas, sendo que os pesos iniciais foram escolhidos aleatoriamente ente -0,1 e 0,1[3].

Para verificarmos o erro de predição da rede, utilizamos a amostra de teste, após o treinamento.

2.2. Análise via Regressão Logística

O modelo de regressão logística[4], também conhecido por modelo logístico, estabelece uma relação entre a probabilidade de ocorrência de cada um dos resultados de uma variável resposta de caráter binário, normalmente representados pelos termos *sucesso* e *fracasso*, e variáveis explicativas categóricas ou contínuas.

Considere uma situação onde observamos uma variável binária Y e um vetor $x = (x_1, x_2, ..., x_p)$ de variáveis explicativas para cada um dos n indivíduos da amostra. O modelo de regressão logística relaciona Y a x supondo que:

$$log\left(\frac{\pi_i(x)}{1-\pi_i(x)}\right) = \beta_0 x_{i1} + \beta_1 x_{i2} + \cdots + \beta_p x_{ip} \quad (1)$$

onde $\pi_i(x) = P\{Y = 1\}$ refere-se à probabilidade de sucesso para a variável resposta, x_{ij} é a j-ésima variável explicativa observada para o indivíduo i e $\beta_0, \beta_1, \beta_2, \cdots, \beta_p$ são parâmetros desconhecidos, devendo assim serem estimados.

Um particular interesse, é determinarmos a probabilidade de sucesso para a variável resposta. Assim, a probabilidade $\pi_i(x)$ de sucesso para a variável resposta é dada por:

$$\pi_i(x) = \frac{1}{1 + exp(-Z_i)} \quad (2)$$

onde $Z_i = \beta_0 x_{i1} + \beta_1 x_{i2} + \cdots + \beta_p x_{ip}$.

Alguns aspectos importantes a serem considerados consistem no ajuste do modelo, na seleção de variáveis explicativas e na verificação do ajuste. O processo de estimação dos parâmetros é feito através do método de máxima verossimilhança[4].

Primeiramente, ajustamos um modelo de regressão logística para amostra de treinamento. Tal modelo foi composto de seis variáveis explicativas de caráter binário (sexo, escolaridade , hábito relacionado ao tabagismo, prática de exercícios físicos, cor e índice de massa corporal) e de uma variável contínua (idade). A variável resposta, denotada por Y_i, do modelo logístico assume os seguintes valores:

$$Y_i = \begin{cases} 1, & \text{se o indivíduo é hipertenso} \\ 0, & \text{caso contrário} \end{cases}$$

O modelo logístico resultante é dado por:

$$log\left(\frac{\pi_i(x)}{1-\pi_i(x)}\right) = Z_i \quad (3)$$

onde $Z_i = -0,258x_{i1} + 0,056x_{i2} + 0,398x_{i3} + 0,220x_{i4} - 0,829x_{i5} - 0,108x_{i6} + 0,298x_{i7} - 3,486$ e x_{ij} é a j-ésima covariável observada para o indivíduo i.

Assim, a probabilidade do indivíduo i ser hipertenso, dadas as variáveis explicativas, é dada por:

$$\pi_i(x) = \frac{1}{1 + exp(-Z_i)} \quad (4)$$

Uma vez ajustado o modelo logístico para o conjunto de treinamento, podemos utilizá-lo para predizer o valor para a variável resposta de um novo indivíduo oriundo da mesma população que gerou a amostra de treinamento utilizada na estimação dos parâmetros do modelo logístico. Assim, dadas as variáveis explicativas $x = (x_1, x_2, ..., x_p)$ de um novo indivíduo, podemos predizer a probabilidade deste indivíduo ser hipertenso.

3 Resultados

Como dito anteriormente, o desempenho do modelo é avaliado em função do número de indivíduos classificados incorretamente na amostra de teste. Primeiramente, determinamos, para a amostra de treinamento, o número de indivíduos classificados incorretamente, em ambos os métodos. Temos que, o modelo logístico classifica apenas 29% dos indivíduos incorretamente, enquanto o modelo neural classifica 26% indevidamente.

Para medirmos o perfomance de predição do modelo logístico, aplicamos a eq.(3) na amostra de teste.

A Tabela 1 apresenta a proporção de indivíduos da amostra de teste que foram classificados incorretamente.

Modelo	**Proporção (%)**
RNA	32
Regressão Logística	36

Tabela 1: *Desempenho do Modelo.*

Observamos que a rede neural artificial mostrou-se superior a regressão logística, visto que o erro de predição para a RNA foi inferior ao erro da regressão logística. Sendo assim, podemos afirmar que a RNA possui um poder de generalização melhor do que a regressão logística.

4 Conclusões

Redes Neurais têm sido bastante utilizadas em problemas de predição e classificação, quando a ênfase é dada à precisão da predição ou da classificação. Entretanto,quando estamos interessados em identificarmos os fatores de risco que possuem associação com a variável resposta binária, regressão logística é bastante utilizada. A vantagem desta metodologia em relação a rede neural é o fato de haver uma teoria matemática por trás de sua fundamentação, o que nos permite obter estimadores para os coeficientes das covariáveis analisadas, e a partir deles, podemos obter intervalos de confiança, por exemplo

Com relação ao conjunto de dados em estudo, percebemos que a rede neural foi o melhor classificador, desde que ela superou a regressão logística em relação ao número de elementos classificados corretamente.

Na área médica, onde se pretende utilizar as redes neurais como instrumento de apoio ao diagnóstico, esta metodologia é atraente e se tem mostrado eficiente.

Referências

[1] S. Haykin. *Neural Networks: A comparative Foundation*. Prentice Hall, New Jersey, 1991.

[2] D. W. Hosmer and S. Lemeshow. *Applied Logistic Regression*. Wiley-Interscience, New York, 1998.

[3] B. B. Pereira and C. V. S. Rodrigues. *Redes Neurais em Estatística. XIII Simpósio Brasileiro de Probabilidade e Estatística*. Caxambú, 1998.

[4] O. M. da Saúde. Hipertensão arterial. In *Série de Informes Técnicos*, page 61, 1978.