Trabalho de Redes Neurais Feedforward

Hugo Diniz Rebelo
COPPE - Universidade Federal do Rio de Janeiro

Trabalho desenvolvido para a matéria de Redes Neurais Feedforward (CPE 721).

1. INTRODUÇÃO

1.1 Objetivo

O presente trabalho visa uma implementação de um sistema que saiba se o paciente teve ou não doença cardiovascular, a partir de três atributos: colesterol total, idade e glicemia sérica em jejum.

1.2 Organização do Trabalho

Este trabalho foi organizado da seguinte forma:

—Capitulo 2: Será visto a metodologia e o Dataset.

—Capitulo 3: Será visto os experimentos e resultados com uma discussão dos mesmos.

—Capitulo 4: Uma Breve Conclusão do que foi feito.

2. O TRABALHO

Nesta Seção veremos como foi desenvolvido o sistema que tem por objetivo saber se o paciente teve ou não doença cardiovascular, vendo a metodologia para a definição de qual modelo utilizar e como foi utilizado o Dataset no processo.

2.1 Metodologia

Para definir qual modelo e parâmetros dos mesmos foi definido duas métricas de comparação que são a Especificidade e Sensibilidade. A sensibilidade reflete o quanto o modelo é eficiente em identificar corretamente aqueles que apresentam o resultado de interesse. Ela é representada por $\frac{VerdadeiroPositivo}{VerdadeiroPositivo+FalsoNegativo}$. Já especificidade é forma de identificar aqueles no qual não tenham o resultado de interesse, ela é representada por: $\frac{VerdadeiroNegativo}{VerdadeiroNegativo+FalsoPositivo}$.[Lalkhen and McCluskey 2008]

Nós procuraremos melhores resultados para as duas métricas dando um pouco mais de prioridade a Sensibilidade, pois ela é que define quanto de acerto o modelo tem para os caso da doença cardiovascular,pois ao deixar um paciente sem um devido suporte médico ao sofrer uma doença Cardiovascular pode leva-lo ao óbito, já dar suporte médico a uma pessoa que não terá uma doença Cardiovascular no momento é prejudicial ao sistema de saudê pois pode sobrecarrega-lo mas não estará levando um paciente um paciente ao óbito de forma direta.

2.1.1 *Ferramentas utilizadas.* Para esse trabalho foi escolhido utilizar a Linguagem Jalia na versão 0.5 com as Bibliotecas PyCall, IJulia, Plots para fazer o Estudo do Dataset. Para os experimentos com a Rede Neural foi utilizado Linguagem Python 3.5.2 com as bibliotecas e Frameworks IPython, numpy,pickle, SciKit-learn 0.18 para a regularização e normalização e a implementação da Rede Neural da biblioteca Theanets 0.73.

2.2 Dataset

O Dataset de que nós usamos possui 874 amostras com três features iniciais que são Colesterol Total, Idade e Glicemia. O Dataset foi dividido em treino, teste e validação, os dois primeiro foram utilizados para todos os experimentos e o segundo chamado de validação, só será utilizado no experimento chamado de validação para garantir que o modelo desenvolvido no trabalho não sofra com o problema do Data Snooping, que é quando o olharmos os dados, nós acabamos criando um sistema que ”sem querer”decore os dados e não tenha uma boa generalização.[Abu-Mostafa et al. 2012]

Primeiramente foi dividido em treino-teste e Validação, tendo o primeiro 90% e segundo 10% do dataset original. Após isso foi dividido na execução do K-Fold o dataset em treino e teste, tendo o primeiro 87,5% e segundo 12,5% do dataset treino-teste.

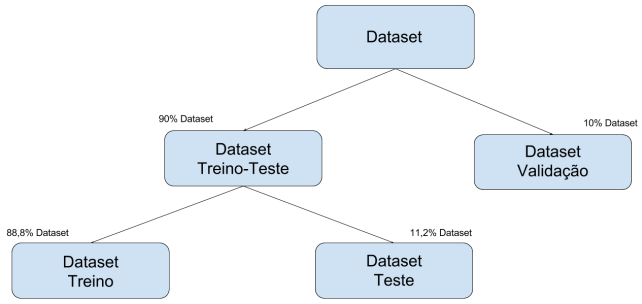

Fig. 1. Divisão do Dataset.

2.2.1 *Entendendo o dataset.* Foi utilizado o dataset de treino para uma avaliação de como estão distribuídos as amostras do dataset, ao plotar esses dados percebemos nitidamente a formação de 4 classes, e cada uma das classes tem uma probabilidade diferente de uma amostra ter a Doença Cardiovascular.

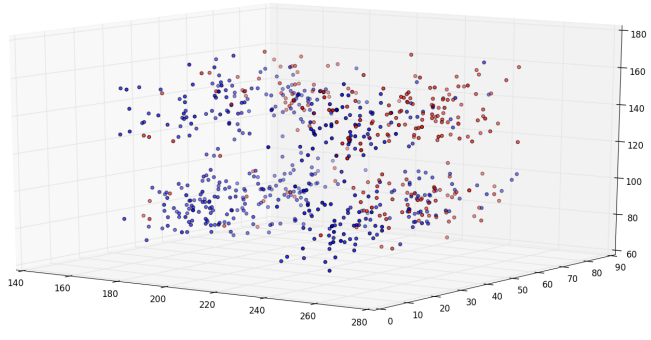

Fig. 2. Divisão do Dataset.

O percentual de pessoas com a doença cardiovascular e quantidade de pessoas sem a doença cardiovascular pode ser vista na tabela abaixo.

	Numero de Pessoas	Porcentagem
Desfecho Positivo	281	30,78%
Desfecho Negativo	593	69,22%
Total	874	100,00%

Analisando a tabela podemos ver uma diferença muito grande nos dois diferentes desfechos, apesar dessa diferença pela quantidade de amostras que tem no dataset, poderemos dividir os dados para o experimento de forma aleatória precisar ter a principio a preocupação com quantidade dos diferentes desfechos nas amostras escolhidas para treino, teste e validação.

Sabendo a distribuição dos dados do dataset, foi feita a matriz de correlação dos dados afim de saber como as features se relacionam entre si e sua relação com a saída

	Colesterol	Idade	Glicemia	Saída
Colesterol	1	0.0315	0.0083	0.3270
Idade	0.0315	1	0.069	0.3198
Glicemia	0.0083	0.069	1	0.233
Saída	0.3270	0.3198	0.233	1

Ao analisar a tabela de correlação, foi visto que as features Colesterol, Idade e Glicemia, possuem uma baixa correlação entre si e também em relação a saída.

3. EXPERIMENTOS E RESULTADOS

Nesta seção explicitaremos quais modelos utilizados, como foi os resultados dos mesmos em relação Especificidade e Sensibilidade fazendo uma analisa para saber a sua capacidade de aprendizado a partir dos diferentes modelos, qual peso dar para o erro, já que para o problema o falso positivo tem peso distinto do falso negativo e o efeito da retirado de uma feature para o problema.

3.1 Modelos

Foram definidas 7 modelos de Rede Neurais para avaliar sua capacidade de generalização e diversos aspectos, a diferenciação desses modelos se feitas pela a camada escondida, no qual são:

—Sem camada Escondida
—1 Camada de 1 Neurônios
—1 Camada de 2 Neurônios
—1 Camada de 3 Neurônios
—1 Camada de 4 Neurônios
—1 Camada de 5 Neurônios
—1 Camada de 7 Neurônios

Cada neurônio da camada intermediaria utilizou como função de ativação Tangente Hiperbólica, na camada de saída foram 2 neurônios e na camada de entrada 3 neurônios quando foi utilizado 3 features e 2 neurônios quando foi utilizado 2 features,A taxa de aprendizagem foi de 0.01, a camada de entrada e saída utilizam função de ativação Linear.

As três features tem escalas distintas, por isso foi utilizada uma normalização afim de deixar todas as 3 features na mesma escala de 0 a 1, onde 0 é valor minimo da escala por feature e 1 valor máximo por escala da feature.

Para o treinamento foi utilizado o algorítimo Gradiente descendente estocástico, com a Entropia Cruzada como função de erro(Função de perda).

3.2 Experimento utilizando as 3 features

Este experimento foi feito para saber a capacidade generalização dos nossos modelos escolhidos, para este experimento foi utilizado o método de validação cruzada K-Fold com 8 folds, utilizando somente os modelos: Sem camada Escondida, 1 Camada de 1 Neurônios, 1 Camada de 2 Neurônios, 1 Camada de 3 Neurônios, 1 Camada de 4 Neurônios, 1 Camada de 5 Neurônios e 1 Camada de 7 Neurônios.

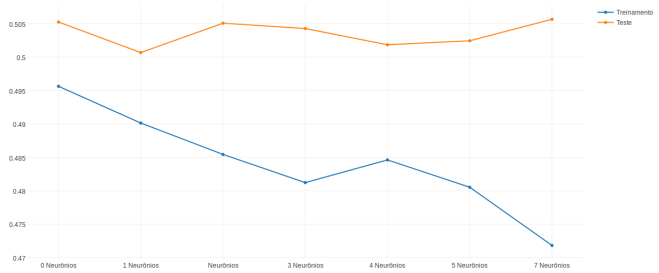

Fig. 3. Relação de erros ao aumento de neurônios na camada escondida no treino e na validação.

Foi visto que ao adicionar a camada escondida e adicionar neurônios nela gera um melhor resultado no treinamento mas isso não se reflete na validação, nos levando a crê que o aumento de neurônios está fazendo o modelo generalizar melhor e pode até está começando a decorar os dados, sintoma do Overffiting.

Como foi visto um certa estabilidade do modelo que não contem camada escondida e até modelo com 5 neurônios com uma pequeno aumento no erro no caso do modelo com 7 neurônios na camada escondida, foi escolhido o modelo sem camada intermediaria como o modelo a ser utilizado no problema. A seguir será visto somente o modelos sem camada utilizando as métricas dos próximos experimentos.

Modelo	Sensibilidade	Especificidade
Sem camada Escondida	5.62e-01	8.74e-01

3.3 Experimento utilizando as 2 features

Este experimento foi feito para saber a capacidade generalização dos nossos modelo escolhido, no caso de uma perda de uma feature, analisando o impacto que cada feature tem em relação a aprendizagem de cada modelo. Para este experimento foi utilizado o método de validação cruzada K-Fold com 8 folds e sem peso.

Features	Erro	Sensibilidade	Especificidade
Colesterol e Idade	5.33e-01	4.78e-01	8.71e-01
Colesterol e Glicemia	5.60e-01	4.31e-01	8.58e-01
Glicemia e Idade	5.65e-01	4.28e-01	8.52e-01

O erro do quando só se utiliza as features Colesterol e Idade é 0,533, maior 0,03 do que o erro visto no Experimento da utilizando as 3 Features, Já nos experimentos utilizando Colesterol e Glicemia assim como Glicemia e Idade nós temos um aumento de 0,06 para o experimento com as 2 features, sendo que podemos considerar algo pequeno pois essa variação acontece já na segunda casa decimal, mas o modelo começa a ter uma sensibilidade muito baixa, ou seja ele está tendendo a dizer que ninguém vai ter a doença cardiovascular algo que pode acontecer já que no nosso dataset temos uma maior quantidade de resultados de pessoas que tiveram a doença é relativamente menor do que as que não tiveram.

3.4 Alterando o peso dos erros

Este experimento foi feito para saber qual peso dar para cada erro, pois ele afetara diretamente os resultados, caso seja muito discrepante, a rede irá tender a dar todos casos com positivo, aumentando a Sensibilidade em detrimento do Especificidade. Para este experimento foi utilizado o método de validação cruzada K-Fold com 8 folds.

Peso	Sensibilidade	Especificidade
1	5.64e-01	8.73e-01
2	7.68e-01	7.04e-01
3	8.60e-01	6.17e-01
4	9.00e-01	5.22e-01
8	9.40e-01	2.80e-01

O resultado demonstra que ao aumentar o peso 1 até o 2 nós vemos uma queda na Especificidade para uma aumento considerável na Sensibilidade chegando peso 2 já com números próximos, já quando se olha para peso 3 em diante vemos uma queda considerável no da Especificidade e um pequeno aumento na Sensibilidade, levando a a conclusão que para uma escolha que visa o maior acerto nos casos positivos da doença o peso 3 tende ser a melhor escolha.

3.5 Validação

Após a execução do K-Fold com o intuito de escolher o modelo e suas variações "ideal", utilizaremos ele no Dataset de Validação, um dataset onde até o exato momento não tínhamos visto seus dados, para podermos confirmar a partir de uma simulação que irá avaliar como modelo reage a novos dados até então desconhecidos.

Modelo	Sensibilidade	Especificidade
Sem camada Escondida	9,09e-01	6.06e-01

Apesar de uma redução na Especificidade em relação ao resultado na escolha de pesos, o resultado na saiu com um ótimo resultado sendo sua variação natural já que não havíamos visto estas amostras até então.

4. CONCLUSÃO

O presente trabalho visou a implementação de um sistema que saiba se o paciente teve ou não doença cardiovascular. Como podemos perceber através da implementação e dos experimentos, o algoritmo se comportou da maneira muito boa, tendo uma Sensibilidade e Especificidade, sem demonstrar uma grande variação na Validação.

Apêndice

Referências

Yaser S Abu-Mostafa, Malik Magdon-Ismail, and Hsuan-Tien Lin. 2012. *Learning from data*. Vol. 4. AMLBook Singapore.

Abdul Ghaaliq Lalkhen and Anthony McCluskey. 2008. Clinical tests: sensitivity and specificity. *Continuing Education in Anaesthesia, Critical Care & Pain* 8, 6 (2008), 221–223.