Logistic regression and artificial neural network classification models: a methodology review

(Dreiseitl & Ohno-Machado, 2002)

Abstrato

A regressão logística e as redes neurais artificiais são os modelos de escolha em muitas tarefas de classificação de dados médicos. Neste de um ponto de vista técnico, resumimos as diferenças e semelhanças destes modelos, e comparamo-los com outros algoritmos de aprendizagem de máquinas. Fornecemos considerações úteis para avaliar criticamente a qualidade dos modelos e os resultados baseados sobre estes modelos. Finalmente, resumimos as nossas conclusões sobre como os critérios de qualidade para a regressão logística e rede neural artificial Os modelos são encontrados numa amostra de artigos da literatura médica.

1. Introdução

Os modelos preditivos são utilizados numa variedade de médicos domínios para tarefas de diagnóstico e prognóstico. Estes os modelos são construídos a partir da "experiência", que constitui dados adquiridos a partir de casos reais. Os dados podem ser pré-processados e expressos num conjunto de regras, como é frequentemente o caso em sistemas especializados baseados no conhecimento, ou servir como dados de formação para modelos estatísticos e de aprendizagem de máquinas. Entre as opções nesta última categoria, os modelos mais populares na medicina são a regressão logística (LR) e as redes neurais artificiais (ANN).

Estes modelos têm as suas origens em duas comunidades diferentes (estatística e informática), mas partilham muitas semelhanças.

Neste artigo, mostramos que a regressão logística e as redes neurais artificiais partilham raízes comuns no reconhecimento de padrões estatísticos, e como o último modelo pode ser visto como uma generalização do primeiro. Comparamos brevemente estes dois métodos com outros algoritmos de classificação populares do campo da aprendizagem de máquinas, tais como vizinhos k-nearest, árvores de decisão, e máquinas vectoriais de apoio (SVM).

2.1. Máquinas vectoriais de apoio

Estes modelos são implementações algorítmicas de ideias da teoria da aprendizagem estatística [2], que diz respeito com o problema da construção de estimadores consistentes a partir de dados: como pode o desempenho de um modelo sobre um conjunto de dados desconhecidos ser estimado, dadas apenas as características do modelo, e o desempenho num conjunto de formação? Algoritmicamente, as máquinas vectoriais de apoio constroem fronteiras de separação óptimas entre conjuntos de dados, resolvendo um problema de optimização quadrática limitada [3,4].

Ao utilizar diferentes funções do núcleo, vários graus de a não linearidade e a flexibilidade podem ser incluídas no modelo. Porque podem ser derivadas de estatísticas avançadas ideias, e os limites sobre o erro de generalização podem ser calculadas para eles, as máquinas vectoriais de apoio têm recebido um interesse de investigação considerável ao longo dos últimos anos. Desempenhos iguais ou superiores aos de outros algoritmos de aprendizagem de máquinas têm sido relatados no literatura médica.

A desvantagem das máquinas vectoriais de apoio é que o resultado da classificação é puramente dicotómico, e não é dada a probabilidade de pertencer a uma classe.

2.2. k-Nearest vizinhos

A classificação baseada no algoritmo k-nearest vizinho difere dos outros métodos aqui considerados, uma vez que este algoritmo utiliza os dados directamente para a classificação, sem construir primeiro um modelo [5,6]. Como tal, nenhum detalhe da construção do modelo precisa de ser considerado, e o único parâmetro ajustável no modelo é k, o número de vizinhos mais próximos a incluir na estimativa de membros da classe: o valor de P(x,y) é calculado simplesmente como a proporção de membros da classe y entre os k vizinhos mais próximos de x. Ao variar k, o modelo pode ser mais ou menos flexível (valores pequenos ou grandes de k, respectivamente).

A vantagem que os k-nearest vizinhos têm sobre outros algoritmos é o facto de os vizinhos poderem fornecer uma explicação para o resultado da classificação; isto Uma explicação baseada em casos pode proporcionar uma vantagem em áreas onde os modelos de caixa negra são inadequados. A maior desvantagem dos vizinhos mais próximos do k-nearest reside no cálculo da vizinhança do caso: para isso, é necessário para definir uma métrica que meça a distância entre itens de dados. Na maioria das áreas de aplicação, não é claro como para, a não ser por tentativa e erro, definir uma métrica em tal a forma como a importância relativa (mas desconhecida!) dos dados componentes é reflectida na métrica.

2.3. Árvores de decisão

Este algoritmo divide repetidamente o conjunto de dados de acordo com um critério que maximiza a separação dos dados, resultando numa estrutura em forma de árvore [7,8]. O critério mais comum utilizado é o ganho de informação; isto significa que em cada divisão, a diminuição da entropia devido a esta divisão é maximizada. A estimativa de P(x,y) é a proporção de elementos de classe y sobre todos os elementos do nó da folha que contém o item de dados x.

Uma grande desvantagem das árvores de decisão é dada pelo processo de construção ganancioso: em cada passo, é selecionada a combinação da melhor variável única e ponto de divisão óptimo; no entanto, um lookahead de múltiplos passos que considera combinações de variáveis pode obter resultados diferentes (e melhores). Uma outra desvantagem reside no facto de as variáveis contínuas serem implicitamente discretizadas pelo processo de divisão, perdendo informação ao longo do caminho. Em comparação com os outros métodos de aprendizagem de máquinas aqui mencionados, as árvores de decisão têm a vantagem de não serem modelos de caixa negra, mas podem facilmente ser expressas como regras. Em muitos domínios de aplicação, esta vantagem pesa mais do que os inconvenientes, de modo que estes modelos são amplamente utilizados na medicina.

3. Regressão logística vs. modelos artificiais de redes neurais

Para os seguintes, deixar que todos os vectores de dados xi contenham um componente adicional 1. Isto facilitará a anotação em permitindo-nos escrever um simples produto de ponto a x para um combinação linear de componentes vectoriais em vez da mais incómodo a x þ a0. Geralmente, um modelo de regressão logística calcula a probabilidade de adesão à classe para uma das duas categorias do conjunto de dados:

Pð1jx; aÞ ¼ 1

1 þ eðaxÞ ;

e Pð0jx; aÞ ¼ 1 Pð1jx; aÞ. Aqui, escrevemos Pð1jx; aÞ para fazer a dependência da distribuição posterior da parâmetros a explícitos. Pode ser demonstrado que este modelo é corrigir quando ambas as densidades condicionais de classe pðxj1Þ e pðxj0Þ são multinormais com matrizes de covariância iguais [6].

O hiperplano de todos os pontos x satisfaz a equação um x ¼ 0 forma a fronteira de decisão entre os dois classes; estes são os pontos para os quais Pð1jx; aÞ ¼ Pð0jx; aÞ ¼ 0:5. Um modelo de regressão logística que inclui apenas os covariáveis originais são chamados modelos de efeitos principais; incluindo termos de interacção, tais como produtos, faz com que o modelo não-linear nas covariáveis, e portanto mais flexível. Embora uma maior flexibilidade possa ser desejável em geral, traz consigo um risco mais elevado de sobreposição de modelos (''memorizar os casos de formação''), o que pode potencialmente reduzir a precisão de um modelo em modelos anteriores casos invisíveis. Na modelação preditiva, adequando a formação casos é apenas parte da tarefa: classificar correctamente os novos casos é o objectivo mais importante.

Estimativa da probabilidade máxima dos valores óptimos dos parâmetros

Q a exige a maximização de n

i¼1 Pðyijxi; aÞ. Embora as formas funcionais para regressão logística e modelos artificiais de redes neuronais sejam bastante diferente, uma rede sem uma camada oculta é na realidade idêntica a um modelo de regressão logística se for utilizada a função de activação logística (sigmoidal) [9,10].

As redes neurais artificiais são agregações de perceptrons. Para redes de alimentação multi-camadas, as a saída é

oN ¼ 1

1 þ eðboHþb0Þ ;

e esta saída é novamente tomada como Pð1jx; b; b0; aÞ. Aqui, oH é um vector de saídas de perceptron, cada um com a sua parâmetros; estes perceptrons são normalmente chamados de ocultos neurónios. Devido à não linearidade destes neurónios ocultos, a saída oN de uma rede neural artificial é uma função não-linear dos inputs. Numa classificação isto significa que a fronteira da decisão pode ser não lineares também, tornando o modelo mais flexível em comparação com a regressão logística. Na Secção 4, resumimos uma amostra de publicações da revista biomédica campo para avaliar se este maior grau de flexibilidade resulta numa melhor precisão de classificação no mundo real conjuntos de dados.