Reproduzindo o Experimento de um Artigo Científico

Apresentação

O presente projeto foi originado no contexto das atividades da disciplina de pós-graduação, Ciência e Visualização de Dados em Saúde, oferecida no primeiro semestre de 2022, na Unicamp, e foi desenvolvido por Mariângela Lima Rodrigues, RA 183863, aluna de mestrado em Estatística.

Referência do Artigo

SILVA, Robson Mariano; LEAL, M. R. R.; LIMA, F. M. Predico do Câncer de Mama com Aplicação de Modelos de Inteligência Computacional. TEMA (São Carlos), v. 20, p. 229-240, 2019. Disponível em: https://www.scielo.br/j/tema/a/yJSLsVLQmfYph8SThYCj8Xh/?lang=pt.

Contextualização da Proposta

O objetivo deste estudo foi reproduzir o experimento desenvolvido no artigo científico, Predição do Câncer de Mama com Aplicação de Modelos de Inteligência Computacional de Silva et.al (2019).

Ferramentas

- Software R Studio versão 4.1.2

Resumo

O artigo científico no qual este projeto foi baseado trata sobre a aplicação de técnicas de inteligência computacional para a predição do câncer de mama.

O câncer é uma doença que vêm se tornando uma das principais causas de morte no mundo, e isso se deve, em partes, ao fato de que a forma como a sociedade tem escolhido viver não inclui um conjunto de hábitos saudáveis - que são fatores fundamentais na prevenção ao câncer, conforme aponta Paulinelli et.al (2003).

Essa doença tão temida pela população pode ser desenvolvida nas mais diferentes áreas do organismo humano. No cenário geral, o que tem sido observado é que 30% dos casos de tumores, entre as mulheres, são malignos e para além disso, a doença tem sido uma das maiores causas de morte nesse grupo, com isso questões estão surgindo com respeito ao que pode ser feito para evitar que a taxa de mortes devido ao câncer de mama continue crescendo.

Existe um conjunto de fatores que dão indícios sobre o desenvolvimento do câncer de mama e a partir da identificação destes fatores inicia-se um protocolo de investigação para a detecção da doença; neste protocolo, o principal exame para o diagnóstico é a mamografia. De acordo com o Instituto Nacional de Câncer, o exame de mamografia é o único que tem se mostrado eficaz na detecção do câncer de mama e, com isso, permite que sejam aplicados tratamentos de modo a reduzir as chances de mortalidade da doença. Nesse sentido, torna-se imprescindível a realização deste exame por mulheres que apresentam alguns dos sintomas, ou por mulheres de idade mais avançada, uma vez que estudos indicam que a mamografia é mais assertiva em mulheres cuja idade esteja entre 50 e 65 anos, e o diagnóstico precoce reduz a mortalidade em 20%

A mamografia é uma importante ferramenta para realizar o diagnóstico, neste exame é feito um raio-X da mama e a partir deste, busca-se por sinais da presença do câncer. No Brasil, o Sistema Único de Saúde (SUS) garante a acessibilidade ao exame para todas as mulheres desde que solicitado por um médico, o que é refletido em uma identificação prévia e redução na taxa de mortalidade. Entretanto ressalta-se que a acurácia garantida pelo exame não é de 100%, podendo gerar tanto resultados falsos-positivos quanto falsos-negativos. Além disso, esse exame requer uma habilidade extrema do profissional que o analisa, e falhar no diagnóstico é extremamente preocupante, dado que o câncer de mama é uma doença muito grave. Nesse sentido, é importante e necessário buscar meios que auxiliem o processo de diagnóstico e o tornem mais eficiente, diminuindo os erros de leitura de exames e diagnósticos incorretos.

Nesse contexto, a tecnologia tem se mostrado cada vez mais importante, possibilitando o diagnóstico e tratamento precoce do paciente. Sendo assim, a utilização de algoritmos de aprendizado de máquina podem potencializar novos desenvolvimentos e contribuir fortemente como aliados da medicina.

Método e Resultados

Dada a importância da tecnologia no processo de diagnóstico de doenças, o objetivo do presente trabalho é a reprodução da aplicação do algoritmo de Máquinas de Vetores Suporte (Support Vector Machines, SVM) para classificação, e para além disso a aplicação do SVM com o intuito de prever o diagnóstico de câncer de mama a partir do conjunto de dados Breast Cancer Wisconsin (Diagnostic) Data Set [1], disponível em domínio público, experimento realizado por Silva et.al (2019). No artigo de referência os autores compararam o SVM e a técnica de redes neurais, neste projeto foi estudado apenas o cenário do SVM.

O algortimo Support Vector Machines (SVM), em um caso binário, parametriza as variáveis respostas categóricas como 1 e -1, e a partir disto enxerga cada vetor de variável explicativa como um ponto em um espaço multidimensional de tamanho $(n-1)$, onde n representa o número de variáveis disponíveis. Na sequência, é traçado um hiperplano que separa as respostas marcadas como 1 daquelas marcadas como -1 (fronteira de decisão).

A equação do hiperplano é dada por dois parâmetros: um vetor de pesos w de mesma dimensionalidade do vetor de variáveis explicativas e um número real b tal que:

$$wx - b = 0,$$

onde $wx = w(1)x(1) + ... + w(m)x(m)$, sendo m a dimensão do vetor x.

Assim, a predição para qualquer vetor de variáveis explicativas x é dada por:

$$y = sign(wx - b),$$

onde $sign$ é um operador matemático que retorna 1 se a observação pertence a classe dessa referência e -1 caso contrário.

O SVM busca potencializar o conjunto de dados e encontrar os melhores valores w^* e b^* para os parâmetros w e b, respectivamente. Um vez encontrados, o modelo $f(x)$ passa a ser definido por:

$$f(x) = sign(w^*x - b^*).$$

Algumas restrições devem ser levadas em conta ao otimizar w e b. Uma vez que cada observação do conjunto de dados é dada como um par (x_i, y_i), onde x_i é o vetor de variáveis explicativas da observação i e y_i é a resposta dada pelos valores -1 ou 1, a primeira restrição a ser levada em conta é definida por:

$$wx_i - b \geq 1, se \;\; y_i = 1$$

[1] https://www.kaggle.com/uciml/breast-cancer-wisconsin-data

$$wx_i - b \leq -1, se \;\; y_i = -1$$

Outra restrição é encontrar um hiperplano que separe as observações com a máxima margem, sendo que a margem é a distância entre a observação mais próxima de cada uma das duas categorias. A fronteira de decisão será equidistante às margens. Uma margem maior contribui para uma melhor generalização, que é o quão bem o modelo classifica novas observações.

Portanto encontrar w^* e b^* consiste em minimizar a norma(w) de tal forma que $y_i(wx_i - b) \geq 1$ para $i = 1, ..., m$ (forma compacta de representar as duas restrições apresentadas anteriormente.

Na Figura (1) é possível visualizar o funcionamento do SVM quando há duas variáveis explicativas. Dispondo das categorias azul (1) e laranja (-1), são traçadas as respectivas margens e a fronteira de decisão é a linha vermelha. Observações acima desta linha são classificadas como azuis e abaixo, como laranjas.

Note também, que as margens definem hiperplanos paralelos e a distância entre elas é dada por $\frac{2}{norma(w)}$. Assim, quanto menor for a $norma(w)$, maior será a distância entre as margens e tenho uma maior separabilidade dos dados.

Esta versão do algoritmos é chamada de modelo linear, pois a fronteira de decisão consiste em uma linha reta (plano ou hiperplano).

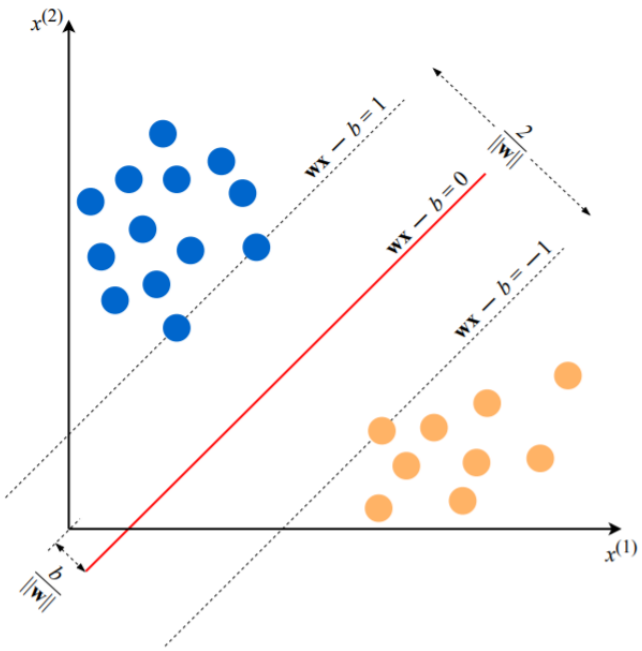

Figura 1: Exemplo de um modelo SVM para vetores binários de variáveis explicativas. Fonte: BURKOV, Andriy (2019).

De forma análoga ao caso linear, o SVM não-linear constrói fronteiras de decisão que não são definidas por retas, planos ou hiperplanos, mas sim superfícies capazes de distinguir um grupo de outro com base nas varíaveis descritivas fornecidas ao modelo.

É importante salientar que este algoritmo também é conhecido como Support-Vector Networks, uma vez que permite entender as relações estabelecidas em rede entre as covariáveis do modelo de modo a identificar um comportamento padrão que pode ser utilizado, como neste caso, para classificação da presença de câncer de mama ou não.

Resultados

Para reproduzir o experimento, foi utilizado o diagnóstico de câncer de mama para 569 indivíduos diferentes, sendo esse diagnóstico benigno (B), para o qual foi observado um total de 357 resultados, e maligno (M), com um total de 212 observações. O conjunto de dados utilizado, foi o mesmo usado no artigo de Silva et.al, e estes trazem informação com respeito a variáveis observadas a partir da leitura de uma imagem digitalizada do aspirado por agulha fina (FNA) da massa mamária de cada indivíduo, e essas variáveis são:

- Identificador do indivíduo;
- Diagnóstico;
- Raio (média das distâncias do centro aos pontos no perímetro);
- Textura (desvio padrão dos valores da escala de cinza);
- Perímetro;
- Área;
- Suavidade (variação local nos comprimentos dos raios);
- Compacidade;
- Concavidade (severidade das porções côncavas do contorno);
- Pontos côncavos (número de porções côncavas do contorno);
- Simetria;
- Dimensão fractal.

O processo de reprodução do experimento se deu através da aplicação da função `svm()` do pacote e1071 disponível no Software R Studio. Para esta reprodução foi considerada uma aplicação como o treinamento e teste de um modelo SVM, a fim de observar a performance do resultado obtido e comparar estes resultados com o experimento original. Conforme a seção de código abaixo, o conjunto de dados foi separado em um banco de treinamento (75%) e outro de teste (25%), e então aplicou-se a técnica SVM não linear.

```
set.seed(003) #semente para reprodução nas mesmas configurações

treino_id <- sample(569,569*0.75)
treino <- Cancer[treino_id,] #Cancer é o nome a base de dados utilizada

teste <- Cancer[-treino_id,]

modelsvm <- svm(Diagnóstico ~.  ,
                treino %>% dplyr::select(-id),
                kernel = "radial")
```

A partir do resultado obtido, foi construída a matriz de confusão, apresentada na Tabela 1, usando os dados de teste do modelo.

Tabela 1: Matriz de Confusão

		Valor Predito	
		B	M
Valor Real	B	92	1
	M	1	49

A partir da Tabela 1 é possível extrair métricas do modelo ajustado e observa-se que a acurácia obtida foi de 98.6%, a sensibilidade foi de 98% e a especificidade foi igual a 98.9%, resultados estes que refletem o obtido pelo artigo de Silva et.al a partir do uso de SVM como forma de predição de câncer de mama.

Com isto, concluí-se a reprodução do experimento do artigo científico Predição do Câncer de Mama com Aplicação de Modelos de Inteligência Computacional.

Referências Bibliográficas

[1] Livro de ML: BURKOV, Andriy. The hundred-page machine learning book. Canada: Andriy Burkov, 2019.

[2] PAULINELLI, Régis Resende et al. A situação do câncer de mama em Goiás, no Brasil e no mundo: tendências atuais para a incidência e a mortalidade. Revista Brasileira de Saúde Materno Infantil, v. 3, n. 1, p. 17-24, 2003.

[3] SILVA, Robson Mariano; LEAL, M. R. R.; LIMA, F. M. Predico do Câncer de Mama com Aplicação de Modelos de Inteligência Computacional. TEMA (São Carlos), v. 20, p. 229-240, 2019. Disponível em: https://www.scielo.br/j/tema/a/yJSLsVLQmfYph8SThYCj8Xh/?lang=pt.