Maximização da metodologia de regressão linear através balanceamento e manipulação dos dados

Rodney Rick
RA 181.814
Disciplina MO444
Prof. Dr. Anderson Rocha

***Resumo*—A aplicação da técnica de regressão lineares (simples ou multivariadas) dentro de um conjunto de dados, normalmente é a forma mais simples de atender rapidamente o objetivo, mas nem sempre a mais eficaz, pelo fato de necessitar supervisão. Para esse trabalho abordaremos, a fim de atender esse tema, como selecionar dados e filtrá-los para posterior aplicação de metodologias estatísticas da área, usando técnicas de aproximação e exatatidão. E o entendimento dos resultados obtidos, nas etapas descritas.**

Palavras-chave: Seleção de dados, Regressão Linear, Gradiente Descendente, Equação Normal.

I. Introdução

A aplicação de métodos estatísticos é primordial para várias áreas de ciências. Justifica-se o uso através da abordagem dos seguintes itens:

(a) Previsão, por exemplo, tentar prever se amanhã irá chover baseado em leituras diárias de dados meteorológicos e comparar com dados passados;
(b) identificação de uma ressonância magnética se a pessoa está com câncer ou pode desenvolver no futuro, baseado em alguns fatores como alimentação.

Dado os dois exemplos acima, o próximo ponto consiste na mineração de dados (do inglês *data mining*). Este descreve o processo exploratório do grande conjunto de entrada de dados, o qual, normalmente, apresenta padrões na amostragem. Sendo assim, possível criar um subconjunto válido e conceber um modelo estatístico capaz de intepretar novas entradas e prever qual seria sua classificação. Por exemplo, na meteorologia, prever que amanhã existe uma chance 76% chuva. Mais detalhes exploratórios são descritos na seção II.

Após tratamento e seleção dos dados conforme primeira parte do livro de Carvalho, A. recomenda, "Preparação de Dados", este artigo trata, através da utilização da técnica de regressão linear, a classificação da música. Dado uma música e seus atributos como entrada e a aplicação do modelo estatístico gerado, descobrir qual o ano da música.

Elaborado o esperado do modelo regressivo, este artigo é dividido em:

1) manipulação e tratamento de dados;
2) explicação da técnica estatístico e como foi aplicada para interpretação dos dados;
3) abordagem da falhas e erros esperados nessa modelagem para este conjunto de dados;
4) conclusão.

II. Manipulação dos dados

Esta seção está dividida em dois passos. A primeira subseção é o entendimento do conceitos descritos nos primeiros capítulos do livro [1]. A segunda refere-se em como foi aplicado os conceitos na base de dados de música.

A. Aplicação de conceitos de preparação de dados

Abaixo seguem passos importantes para concepção de uma entrada, no mínim adequada, para garantir melhoria contínua da modelagem estatística.

Seleção dos dados
: Definir uma amostragem aleatória é importante pelo motivo ajudar a anular os efeitos de fatores não observados. Por exemplo, suponha que deseja-se calcular a altura média das pessoas em uma cidade e fazer a sua amostragem em um bairro qualquer, isto não obtém uma boa estimativa. Pois ocorre que as alturas estão condicionados a um determinado valor de fator "regional", sendo neste caso não adequado ao cálculo da altura das pessoas da cidade como um todo.

Redução de dimensionalidade
: Trabalhar com um escopo menor de dados, ou seja, um subconjunto torna-se mais viável ao tentar criar uma modelagem estatística ao invés de todo o conjunto de dados.

Balanceamento dos dados
: Caso a amostra de dados seja desigual, necessita-se uma adaptaçao no momento de seleção dos dados para que não seja favorecido muito somente um subconjunto de dados pequeno e prejudicial para a amostragem total. Para isso, é possível verificar através da aplicação de um gráfico de histograma. Outros estudos podem ser considerados como avaliação da dispersão do conjunto dos dados, aplicando desvio padrão para posterior seleção do melhor subconjunto de dados a ser elaborado.

Transformação dos dados
: Para dados numéricos, torna-se viável a normalização. Para dados descritivos, pode-se aplicar o uso de etiquetas numéricas para modelagem, mas não é regra.

Ainda não terminado, de acordo com Andrew Y. Ng em um de seus artigos [2] e o livro de Hastie et al. [3], torna-se necessário a confecção de 3 bases importantes perante seus dados originais de entrada:

Treino

A fase de treino confere a entrada de dados já selecionados e adaptados (por exemplo, normalizados), servir para treino do modelo e o arranje dos parâmetros e, por fim, adaptar o melhor emparelhamento de entrada com saída esperada.

Validação

Durante o procedimento de validação: um conjunto de exemplos usados para ajustar os parâmetros do modelo treinado a fim de estimar o quão bom foi elaborado o mesmo. Trabalhar com informações de erro médio determinando assim, caso necessário, um ponto de parada readaptação do modelo. Como demonstra a figura 1

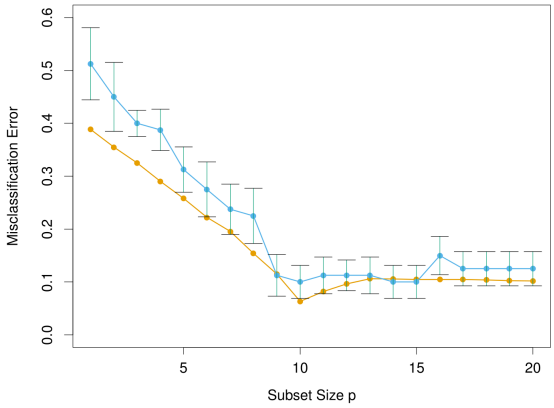

Figura 1: Validação do conjunto de treino usando a conjunto de validação, conforme medição de erros (figura retirada do livro [3], pág. 244).

Teste

Nesta etapa, o modelo final já está concluído e deve ser testado, realizando mesmos testes conforme aplicados na etapa anterior. E uma nova verificação, se houve pouco predição causando uma modelagem chamada de *underfitting* ou uma grande aproximação, porém se aplicado ao modelo, novos dados de testes, não confere nenhuma adaptação. Neste caso o modelo recebe o nome de *overfitting*. A figura 2 demonstra tal medição.

B. O repertório de dados

Atendido a essas premissas da sub-seção anterior, torna-se possível preparar uma base de entrada para criação do modelo estatístico com uma qualidade, afinal não está sendo beneficiado nenhum grupo majoritário.

Primeiro passo que deve-se seguir é compreender como é a distribuição dos dados. A figura 3 demonstra o histograma da base de entrada.

Após esse estudo quantitativo e entendimento, verifica-se que os dados em um primeiro momento necessitam em passar em uma triagem para balanceamento dos dados. O segundo passo, trabalhar comesse conceito. Para isso, dado que a base

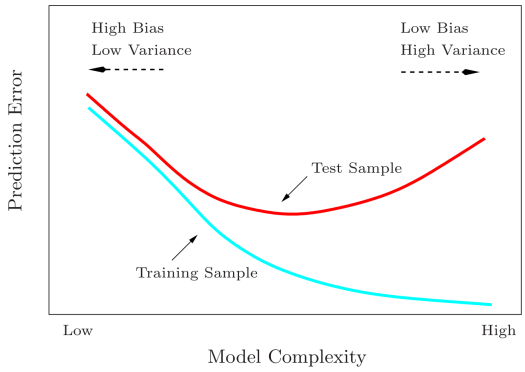

Figura 2: Acompanhamento da modelagem conforme medição de erros (figura retirada do livro [3], pág. 44).

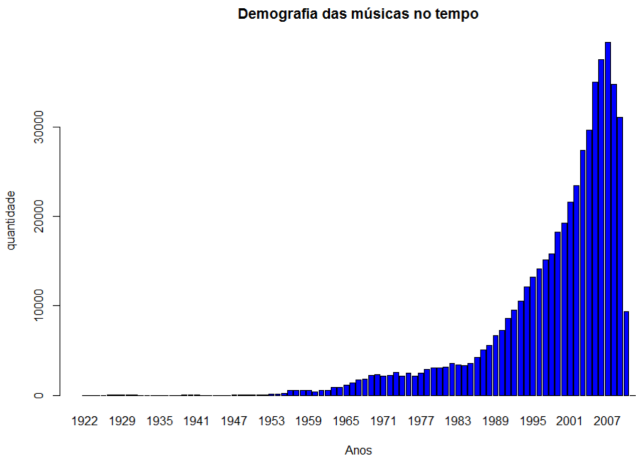

Figura 3: Histograma da distribuição de músicas.

de entrada é extremamente desigual, ou seja, existe anos os quais há uma imensa quantidade de músicas e deseja-se não favorecer na modelagem certos anos e depreciar o resultados de outros anos. Na tabela II, descreve uma adaptação na base de entrada.

Tabela I: Tabela de adaptação dos dados.

Quantidade	**Adaptação**
Valores abaixo de 1k	Manteve-se seu valor
Entre 1k e 10k	1500
Entre 10k e 20k	2000
Entre 20k e 30k	2500
Acima de 30k	3000

Dado agora que a distribuição de músicas ao longo do anos torna-se mais justa, a figura 4 descreve a nova distribuição dos dados musicais, conforme sua frequência de exemplo.

Na literatura, existe uma divergência, na quantidade de dados usados para a elaboração das bases de treino, validação e teste, mas não é demasiado discordante. Normalmente é aplicado para treino, validação e teste a proporção, respec-

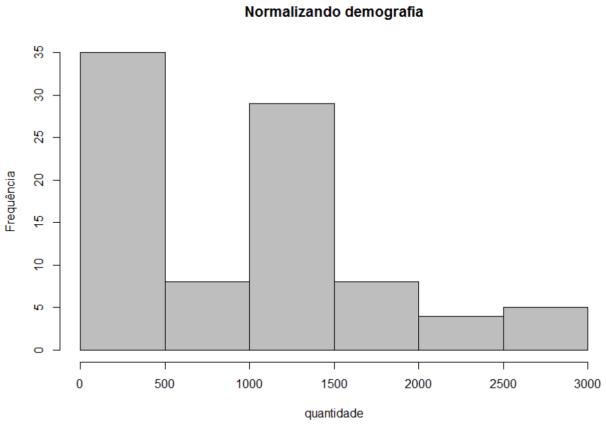

Figura 4: Frequência da distribuição dos anos.

tivamente, 50%, 25% e 25%. Para esta modelagem de dados musicais, dados as premissas atendidas em cada subconjunto de dados, foi utilizado um conjunto de 40 mil dados conforme para treino, 20 mil para validação e 20 mil para teste.

III. Regressão Linear

A. A metodologia

A aplicação da técnica de regressão linear é simples e muito usada na área estatística, onde o objetivo é avaliar a relação de uma variável de interesse (ou objetivo) Y,também chamada variável dependente, em relação a k variáveis X_j (variável independente ou covariável), $j = [1, k]$.

De acordo com Bishop, em seu livro [4], transcreve que esse processo concebe-se através do envolvimento das variáveis X_j, ou seja, através de uma combinação linear entre as variáveis. Portanto, um possível modelo para avaliar essa relação pode ser dado por:

$$y_i = \beta_0 + \sum \beta_i x_{i,k} + \varepsilon \tag{1}$$

Sendo $i = [1, \ldots, n]$ e ε o erro acumulado durante o processo de modelagem.

A fim de minimizar ε, necessita-se de um método de minimização de o valor obtido e valor alvo, nomeada função de custo, para isso a seguir demonstra-se a fórmula da técnica *Gradient Descent*:

$$J(\theta_0, \theta_1) = \frac{1}{2m} * \sum_{i=1}^{m} [h_\theta(x^i) - y^i]^2 \tag{2}$$

A função de custo, obtida pós-processamento da 1, cálcula o erro médio para a aplicação da equação. Pode ser adaptada e ser calculada 2. A função de custo ajuda a elaboração da fórmula incial desta seção, combinando, linearmente, da melhor maneira as variáveis independentes e diminuindo o erro médio obtido.

É possível aplicar uma metodologia para cálculo exato, mas isso tem seu custo computacional.

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

Na tabela II, apresenta uma comparação entre as técnicas:

Tabela II: Tabela comparativa entre *Gradient Decent* e Equação Normal

Gradient Descent	***Normal Equation***
Precisa escolher a taxa de aprendizagem	Não precisa escolher a taxa de aprendizagem
Precisa muitas iterações	Não precisa de iterações
Funciona bem para n grande ($n >= 10^6$)	Necessário calcular $(X^T X)^{-1}$
-	Lento se n grande

Obs: n é o número de atributos de um registro

B. Aplicação aos dados

Com o conhecimento das técnicas na sub-seção anterior, o objetivo é aplicar na base de repertório musical tal metodologia. Alcançar a melhor combinação linear das variáveis independentes. E, com o o auxílio do *Gradient Descent*, em cada iteração, encontrar o melhor arranjo de combinações possíveis a ponto de minimizar o valor esperado para o objetivo, ou seja, minimizar a função de custo e obter o melhor resultado para a variável indepedente.

Assim, torna-se possível, através de uma entrada, da modelagem, com a formulação do melhor alinhamento entre as variáveis independentes, garantir o menor erro. No caso do exemplo musical, dado a entrada de parâmetros que caracterizam a música, o algoritmo deverá encontrar o melhor resultado e classificar o ano correto da música.

A figura 5 apresenta o uso de regressão linear. O eixo X, está associado com o valor do erro médio e o eixo Y, com as iterações. Para esse gráfico, foi utilizado para a saída de cada mil iterações um valor de saída. Logo, para a execução de 60 mil iterações, houve 60 saídas contendo o valor do erro médio. O último valor é representado por 8.2 anos de erro, pelo modelo e a partir da base de treino. Ou seja, dada uma música, existe um erro ao classificá-la no ano correto, divergindo em até 8.2 anos. Porém, o modelo apresenta um erro maior que esse valor quando aplicado para a base de teste, conforme a figura.

No entanto, mesmo com esse modelo, que apresenta um erro médio pouco menor que uma década, esta formulação erra ao predizer para uma entrada musical do começo da década 20 até 60. Isso acontece pois, a volumetria de músicas para esse intervalo de tempo é muito pequena (conforme figura 3), e mesmo com o balanceamento dos dados, não torna-se suficiente para o modelo estimar um valor adequado.

IV. Conclusão

A partir deste trabalho, entende-se que a aplicação de técnicas para manipulação dos dados e de regressão linear, esta última para previsão de dados. Métodos importantes e também auxiliares para determinação dos melhores recursos, respectivamente para melhores exemplos de registros para construção da base de treino, validação e teste, mas também os atributos mais adequados (para efetuar as combinações das variáveis independentes) para a elaboração do modelo em si.

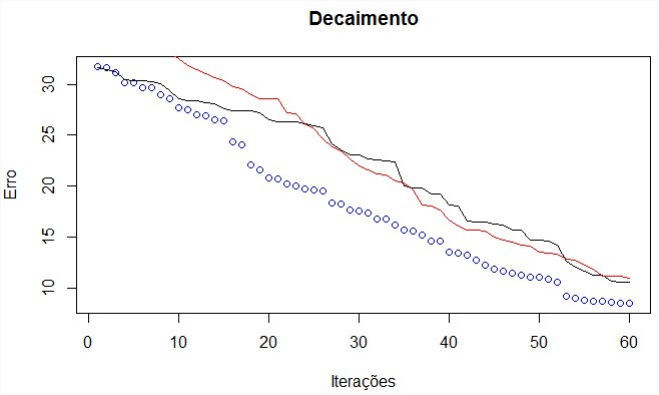

Figura 5: A ilustrção representa entrada de treino, base de validação e base de teste, respectivamente, o círculo azul, linha preta e linha vermelha.

A aplicação da regressão linear, com auxílio de *Gradient Descent* ou Equação Normal (usa-se e ao invés de ou quando existe a necessidade de validação entre as técnicas), é de importante uso quando existe o desejo de predizer. Isto é, dado vários parâmetros de entrada, como explicar a saída esperada para validação de forma supervisionada posteriormente, pois necessita-se da validação dos resultados.

Ainda, deve-se ressaltar, que aplicar uma regressão linear não é conclusivo, mas sim, auxiliar. Logo, para melhores resultados, necessita-se acompanhar os resultados e efetuar a supervisão da técnica, ou ainda a aplicação outras técnicas, para que, em conjunto, representem um resultado melhor ou mais adequado conforme uma saída esperada.

REFERÊNCIAS

[1] Carvalho, A., *Inteligência Artificial: Uma Abordagem de Aprendizado de Máquina*. LTC. [Online]. Available: https://books.google.com.br/books?id=4DwelAEACAAJ

[2] Andrew Y. Ng, “Preventing "overfitting"of cross-validation data,” in *In Proceedings of the Fourteenth International Conference on Machine Learning*. Morgan Kaufmann, 1997, pp. 245–253.

[3] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition*, ser. Springer Series in Statistics. Springer, 2009. [Online]. Available: https://books.google.com.br/books?id=tVIjmNS3Ob8C

[4] C. M. Bishop, *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Secaucus, NJ, USA: Springer-Verlag New York, Inc., 2006.