****

**Comparando os Resultados dos Algoritmos de Regressão Logística e K-Means.**

Tiago Merli Carile

Universidade Federal do ABC, Centro de Matemática, Computação e Cognição (E-mail: [tiago.merli@ufabc.edu.br](mailto:tiago.merli@ufabc.edu.br)).  
Rua Catequese, 242 – Santo André - SP, Brazil. CEP: 09090-400

****

**Resumo – O artigo compara a acurácia na utilização dos algoritmos de Regressão Logística e K-means, explorando o seu resultado.**

Palavras-chave: Regressão Logística, K-Means.

1. Regressão Logística

A regressão logística é uma técnica estatística que tem como objetivo produzir, a partir de um conjunto de observações, um modelo que permita a predição de valores tomados por uma variável categórica, frequentemente binária, a partir de uma série de variáveis explicativas contínuas e/ou binárias. [1] [2]

A regressão logística é amplamente usada em ciências médicas e sociais, e tem outras denominações, como modelo logístico, modelo logit, e classificador de máxima entropia. A regressão logística é utilizada em áreas como as seguintes:

* Em medicina, permite por exemplo determinar os factores que caracterizam um grupo de indivíduos doentes em relação a indivíduos sãos.
* No domínio dos seguros, permite encontrar fracções da clientela que sejam sensíveis a determinada política securitária em relação a um dado risco particular.
* Em instituições financeiras, pode detectar os grupos de risco para a subscrição de um crédito.
* Em econometria, permite explicar uma variável discreta, como por exemplo as intenções de voto em actos eleitorais.

O êxito da regressão logística assenta sobretudo nas numerosas ferramentas que permitem interpretar de modo aprofundado os resultados obtidos.

Em comparação com as técnicas conhecidas em regressão, em especial a regressão linear, a regressão logística distingue-se essencialmente pelo facto de a variável resposta ser categórica.

Enquanto método de predição para variáveis categóricas, a regressão logística é comparável às técnicas supervisionadas propostas em aprendizagem automática (árvores de decisão, redes neuronais, etc.), ou ainda a análise discriminante preditiva em estatística exploratória. É possível de as colocar em concorrência para escolha do modelo mais adaptado para um certo problema preditivo a resolver.

Trata-se de um modelo de regressão para variáveis dependentes ou de resposta binomialmente distribuídas. É útil para modelar a probabilidade de um evento ocorrer como função de outros fatores. É um modelo linear generalizado que usa como função de ligação a função logit.

A regressão logística analisa dados distribuídos binomialmente da forma

  
onde os números de ensaios de Bernoulli ni são conhecidos e as probabilidades de êxito pi são desconhecidas. Um exemplo desta distribuição é a percentagem de sementes (pi) que germinam depois de ni serem plantadas.

O modelo é então obtido na base de que cada ensaio (valor de i) e o conjunto de variáveis explicativas/independentes possa informar acerca da probabilidade final. Estas variáveis explicativas podem-se ver como um vector Xi k-dimensional e o modelo toma então a forma:



1. Algoritmo K-means

Em mineração de dados, agrupamento k-means é um método de Clustering que objetiva particionar n observações dentre k grupos onde cada observação pertence ao grupo mais próximo da média. Isso resulta em uma divisão do espaço de dados em um Diagrama de Voronoi.

O termo K-means foi empregado primeiramente por James MacQueen em 1967[3], embora a ideia remonta a Hugo Steinhaus em 1957.[4] O Standard algorithm foi proposto primeiramente por Stuart Lloyd em 1957 como uma técnica para modulação por código de pulso, embora não tenha sido publicada fora dos laboratórios Bell até 1982.[5] Em 1965, E.W.Forgy publicou essencialmente o mesmo método, é por isso que é por vezes referido também como Lloyd-Forgy.[6] Uma versão mais eficiente foi proposta e publicada em Fortran por Hartigan e Wong, no período entre 1975 e 1979.[7]

. O algoritmo inicia com a escolha dos k elementos que formaram as sementes iniciais. Esta escolha pode ser feita de muitas formas, entre elas:

* Selecionando as k primeiras observações;
* Selecionando k observações aleatoriamente; e
* Escolhendo k observações de modo que seus valores sejam bastante diferentes. Por exemplo, ao se agrupar uma população em três grupos de acordo com a altura dos indivíduos, poderia se escolher um indivíduo de baixa estatura, um de estatura mediana e um alto.

Escolhidas as sementes iniciais, é calculada a distância de cada elemento em relação às sementes, agrupando o elemento ao grupo que possuir a menor distância (mais similar) e recalculando o centróide do mesmo. O processo é repetido até que todos os elementos façam parte de um dos clusters.

Após agrupar todos os elementos, procura-se encontrar uma partição melhor do que a gerada arbitrariamente. Para isto, calcula-se o grau de homogeneidade interna dos grupos através da Soma de Quadrados Residual (SQRes), que é a medida usada para avaliar o quão boa é uma partição.

Após o cálculo, move-se o primeiro objeto para os demais grupos e verifica-se se existe ganho na Soma de Quadrados Residual, ou seja, se ocorre uma diminuição no valor da SQRes. Existindo, o objeto é movido para o grupo que produzir o maior ganho, a SQRes dos grupos é recalculada e passa-se ao objeto seguinte. Depois de um certo número de iterações ou não havendo mais mudanças, o processo é interrompido.

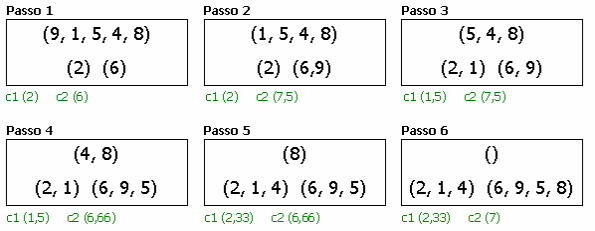


Figura 1: Algoritmo de Análise de Agrupamentos K-means

A figura 1 mostra a execução do algoritmo das k-médias para formar dois agrupamentos a partir da população composta pelos elementos {2, 6, 9, 1, 5, 4, 8}. Pode-se observar que como sementes foram escolhidos os dois primeiros e como critério para definir o valor do centróide após a união foi usada a média. Na figura, c1 e c2 apresentam os valores dos centróides de cada um dos agrupamentos após a adição de um novo elemento.

O algoritmo K-means é bastante escalar e confiável, porém apresenta alguns problemas. Os dois principais problemas são:

* Exige que as variáveis sejam numéricas ou binárias e frequentemente aplicações envolvem dados categorizados, neste caso, uma alternativa é converter os dados categorizados em valores numéricos ou utilizar uma das muitas variações do método; e
* É sensível a valores outliers, um único objeto com valor muito extremo pode modificar, substancialmente, a distribuição dos dados.

1. Preparação dos dados

Para a comparação, os dados foram organizados em quinze atributos de um arquivo texto: age, workclass, fnlwgt, education, educationNum, maritalStatus, occupation, relationship, race, sex, capitalGain, capitalLoss, hoursPerWeek, nativeCountry e target. Após a leitura, foi verificada se não haviam dados nulos e as respectivas distribuições dos dados.

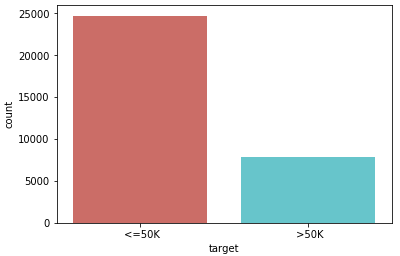


Figura 2: Distribuição entre a quantidade de ocorrências e o target.

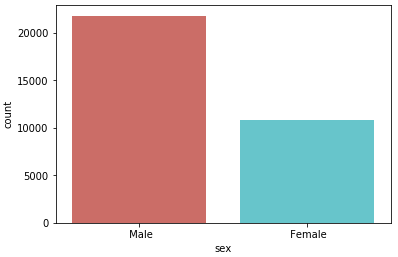


Figura 3: Distribuição entre a quantidade de ocorrências e o sexo.

Além disso, é essencial verificar se existem atributos não-numéricos e transformá-los em valores numéricos.

Após esta conversão é possível verificar a relação entre os atributos.

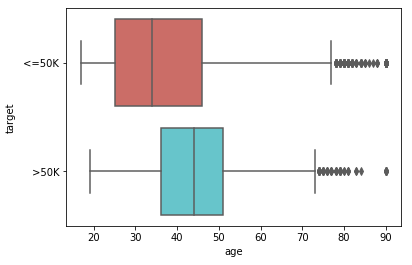


Figura 4: Relação entre idade e a classe.

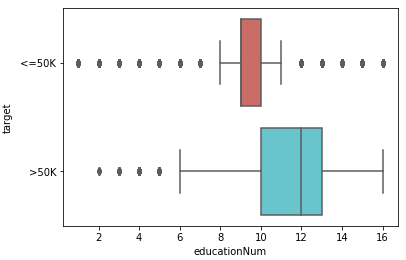


Figura 5: Relação entre educationNum e a classe.

Para evitar o enviesamento do algoritmo, como a raça inferir de forma determinante na renda e implicando diretamente em um resultado racista, alguns atributos foram desconsiderados: workclass, fnlwgt, education, maritalStatus, occupation, relationship, race, sex, nativeCountry e target.

1. Treinamento

Os dados foram separados da seguinte maneira:

* 70% dos dados para o treinamento
* 30% dos dados para os testes

1. CONCLUSÃO

Utilizando Regressão Logística, o resultado ficou tendencioso para a classe 0, aparentemente devido ao desbalanceamento do conjunto de treino, conforme demonstrado na Figura 6.

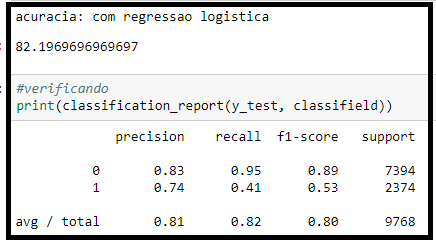


Figura 6: Resultado do Algoritmo de Regressão Logística

Quando comparamos a acurácia, ambos os modelos tiveram desempenho semelhante, porém o algoritmo K-means foi 0,2% melhor como mostra a figura 7.

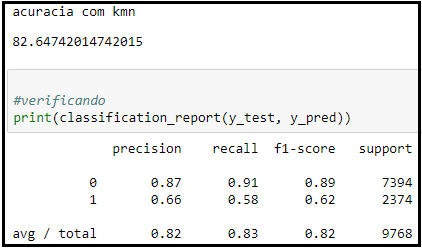


Figura 7: Resultado do Algoritmo K-means.

A média de precisão foi de 81% para regressão logística e 83% para K-means.

Considerando o recall, podemos dizer que o modelo com k-means foi melhor pois tem uma média de 83% enquanto na regressão logística temos uma média de 82%. Isso porque o K-means é menos sensível ao desbalanceamento do conjunto de treinamento.

Comparando o f1-score temos que na média o K-means foi 1% melhor.

Portanto, com os atributos utilizados e conjunto de treinamento para este problema é melhor utilizar o K-means.

REFERÊNCIAS

1. Hosmer, David W.; Stanley Lemeshow (2000). Applied Logistic Regression, 2nd ed. [S.l.]: New York; Chichester, Wiley. ISBN 0-471-35632-8
2. Logistic Regression. Disponível em: <http://userwww.sfsu.edu/efc/classes/biol710/logistic/logisticreg.htm>, 2018.
3. MacQueen, J. B. (1967). Some Methods for classification and Analysis of Multivariate Observations. Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability. 1. University of California Press. pp. 281–297. MR 0214227. Zbl 0214.46201. Consultado em 14 de novembro de 2012
4. Steinhaus, H. (1957). «Sur la division des corps matériels en parties». Bull. Acad. Polon. Sci. (em francês). 4 (12): 801–804. MR 0090073. Zbl 0079.16403
5. Lloyd, S. P. (1957). «Least square quantization in PCM». Bell Telephone Laboratories Paper Published in journal much later: Lloyd., S. P. (1982). «Least squares quantization in PCM» (PDF). IEEE Transactions on Information Theory. 28 (2): 129–137. doi:10.1109/TIT.1982.1056489. Consultado em 15 de abril de 2009
6. E.W. Forgy (1965). «Cluster analysis of multivariate data: efficiency versus interpretability of classifications». Biometrics. 21: 768-769
7. J.A. Hartigan (1975). Clustering algorithms. [S.l.]: John Wiley & Sons, Inc.

****