Uma disputa classificatória: KNN vs Regressão Logística

Lia Sucupira Furtado, Gabriel Alves das Neves, Gabriel da Rocha Silva e André Luis Dantas Gadelha

Universidade Federal do Ceará, Dept. de Teleinformática Campus do Pici, Av. Mister Hull, Brasil

Resumo. Os modelos de classificação preveem as categorias dos dados "etiquetando" as amostras para uma determinada classe. O modelo preditivo mapeia uma função que recebe os dados de entrada e fornece como saída valores qualitativos ou categóricos. Neste trabalho foram utilizados o modelo de classificação linear Regressão Logística e o modelo de classificação não-linear KNN. Foi utilizado um conjunto de dados com 8708 observações relacionadas a pedidos de subsídios entre os anos de 2005 a 2008. A performance dos modelos foram analisadas e comparadas por meio da matriz de confusão. Dos modelos selecionados, é possível observar tanto melhor precisão no modelo linear, quanto menor tempo de processamento, enquanto que o modelo não linear escolhido para comparação não apresentou o mesmo desempenho.

1 Introdução

A seleção de um modelo de classificação ideal para um conjunto de dados é de extrema importância para uma boa predição dos dados futuros e, por isso, torna-se um desafio devido à grande quantidade de modelos possíveis.

Dessa forma, escolhe-se analisar dois modelos diferentes, para concluir se a maior complexidade de um modelo não-linear irá prover uma precisão maior do que modelos mais simples, os lineares.

O modelo linear escolhido foi a Regressão Logística, que segundo M. Kuhn e K. Johnson [1] é um reconhecido modelo de classificação que, por predefinição, gera limites de classificação lineares.

Para o modelo não-linear foi utilizado o K-ésimo vizinho mais próximo, pois é um modelo de fácil compreensão por sua metodologia de classificação com base na distância geométrica das amostras escolhendo o número de vizinhos a serem analisados.

2 Metodologia

O conjunto de dados possui 1882 preditores e 8190 amostras. Destes, apenas 252 foram utilizados para o processamento, bem como as primeiras 6633 amostras são utilizadas como dados de treino, por serem dados de antes de 2008. As amostras restantes são dados de 2008 em diante.

Para o caso em que os dados do conjunto de observações (representado por X) estiverem entre duas categorias, a regressão logística prevê a probabilidade (P(X)) de uma variável qualitativa pertencer a uma categoria de dados. As probabilidades de pertinência são calculadas e é pré-determinado um limite "p"bem como as categorias. No caso de duas categorias, por exemplo, pode-se determinar que: se $P(X) > p$ implica na categoria 1; se $P(X) < p$ implica na categoria 2.

A função que descreve o modelo não deve ser linear tendo em vista que, em valores próximos ao zero teremos probabilidades negativas. A fim de evitar esse problema deve-se usar uma função que limite as saídas (Y) entre 0 e 1. Um exemplo de função com esse comportamento é a função logística:

$$P(X) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 * x1 + ...)}} \tag{1}$$

Onde x1, x2, etc, são os preditores do conjunto de dados de entrada para o modelo. Para ajustar o modelo usa-se o método da máxima verossimilhança. Manipulando a equação da função temos log(P(X)/[1 - P(X)]) = β0 + β1*X. Os coeficientes da parte linear da equação são estimados de modo que essas estimativas no

modelo para P(X), produzem um número próximo a 1 para todos as variáveis na categoria 1, e um número próximo de zero para todos as variáveis na categoria 2.

O método de classificação K-Vizinhos mais Próximos utiliza-se do conceito de aproximação geométrica, no qual o novo resultado a ser calculado depende dos seus vizinhos geográficos. Ou seja, KNN prevê a classificação para a nova amostra se utilizando das K-amostras mais próximas do conjunto de treino. A "proximidade"pode ser definida por uma distância métrica, como Euclidiana(2).[1]

$$(\sum_{j=1}^{P}(x_{aj} - x_{bj})^2)^{\frac{1}{2}} \tag{2}$$

Para que preditores com grande variação de valores não influenciem demais na capacidade de classificação do método KNN, todos os preditores devem ser centralizados e escalonados antes de executar o método.

Com o pré-processamento realizado, cada nova amostra da predição terá sua classe definida pela quantidade de vizinhos do conjunto de treino que foram identificados de cada classe. A predição da classe da amostra será a classe com maior probabilidade, ou seja, a que tiver a maioria dos vizinhos sendo da mesma. Se duas ou mais classes tiverem a maior estimativa, um novo vizinho é procurado, e será utilizado para definir a classe da amostra.

Apesar do conceito do método KNN ser de fácil compreensão, deve-se ter em mente que tanto preditores não relevantes ou ruidosos reduzem o desempenho do método, como o caso da estrutura local dos preditores não for relevante para a modelar o método. Ou seja, se a organização dos preditores não influenciar no desempenho da predição. Portanto o pré-processamento é essencial para que o modelo tenha desempenho adequado.

Para medir a performance dos modelos foram usados duas métricas, a Matriz de Confusão e a Curva ROC.

A matriz de confusão, é uma tabulação cruzada simples das classes observadas e previstas para os dados. Células diagonais denotam casos em que as classes são corretamente preditas enquanto as anti-diagonais ilustram o número de erros para cada caso possível.

A curva ROC é criada avaliando as probabilidades de classe para o modelo em um continum de limites. Para cada limiar candidato, a taxa verdadeira-positiva resultante (ou seja, a sensibilidade) e a taxa de falso-positivo (uma menos a especificidade) são traçadas entre si.Um modelo perfeito que separa completamente as duas classes teria 100% de sensibilidade e especificidade.Um modelo completamente ineficaz resultaria em uma curva ROC que segue de perto a linha diagonal de 45 e teria uma área sob a curva ROC de aproximadamente 0,50.Dessa forma, o modelo com a maior área sob a curva ROC seria o mais efetivo.

3 Resultados

Na análise feita usando os modelos classificatórios no conjunto de dados explicitado, os seguintes resultados foram destacados.

3.1 Regressão Logística

Foi ajustado o modelo de Regressão Logística nos dados do conjunto de treino. Pelo método da máxima verossimilhança, os valores dos parâmetros da função linear, que representa o limite da classificação, foram calculados. Em seguida foi prevista a saída para o conjunto de dados de teste. Duas classes foram definidas, a de sucesso e fracasso. Os dados que tiveram probabilidade acima de 0.5 foram categorizadas como sucesso, e os outros como fracasso. A matriz de confusão está representada na Tabela 1 onde a diagonal principal indica um alto número de observações, então o modelo tem uma boa performance.

Calculando uma porcentagem de predições corretas temos uma taxa de aproximadamente 84%.

	Referência	
Predição	Sucesso	Fracasso
Sucesso	148	43
Fracasso	41	286

Tabela 1: Matriz de confusão

Além disso,plotamos a curva ROC ter essa representação gráfica que ilustra o desempenho do modelo.No gráfico da Figura 1, temos que a área abaixo da curva é 0.91, o que é outro indicativo que o modelo é preciso, diante do fato de que para modelos ótimos de predição, quando as predições são 100% corretas a área abaixo da curva do ROC é 1.0.

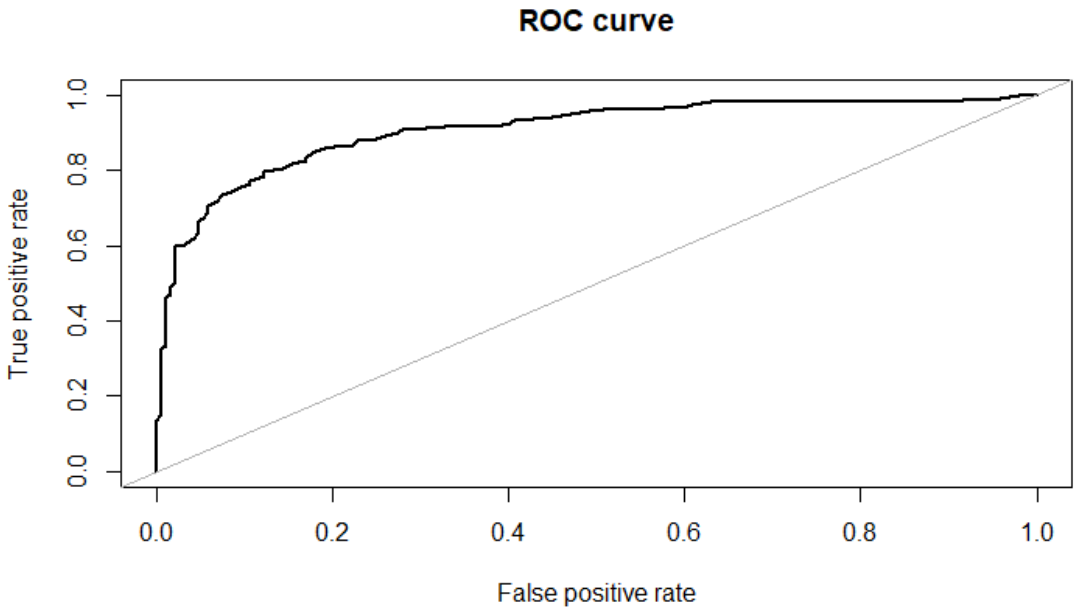

Figura 1: Curva ROC do Modelo de Regressão Logística

3.2 K Vizinhos mais Próximos - KNN

A partir do pré-processamento e modelagem do método com as amostras do conjunto de treino, os resultados indicam que um número k = 451 vizinhos entregam o melhor desempenho de predição, como pode ser observado na figura 2. A partir de 100 vizinhos, o ganho de desempenho não é tão expressivo, mas ainda assim válido. Acima do limite de k = 451 vizinhos, o método começa a perder desempenho.

A matriz de confusão teve uma taxa de 69,8% de predições corretas.

Já a área abaixo da curva ROC foi de 0.804, como observado na Figura 3. A matriz de confusão do método KNN[2] possui os seguintes resultados:

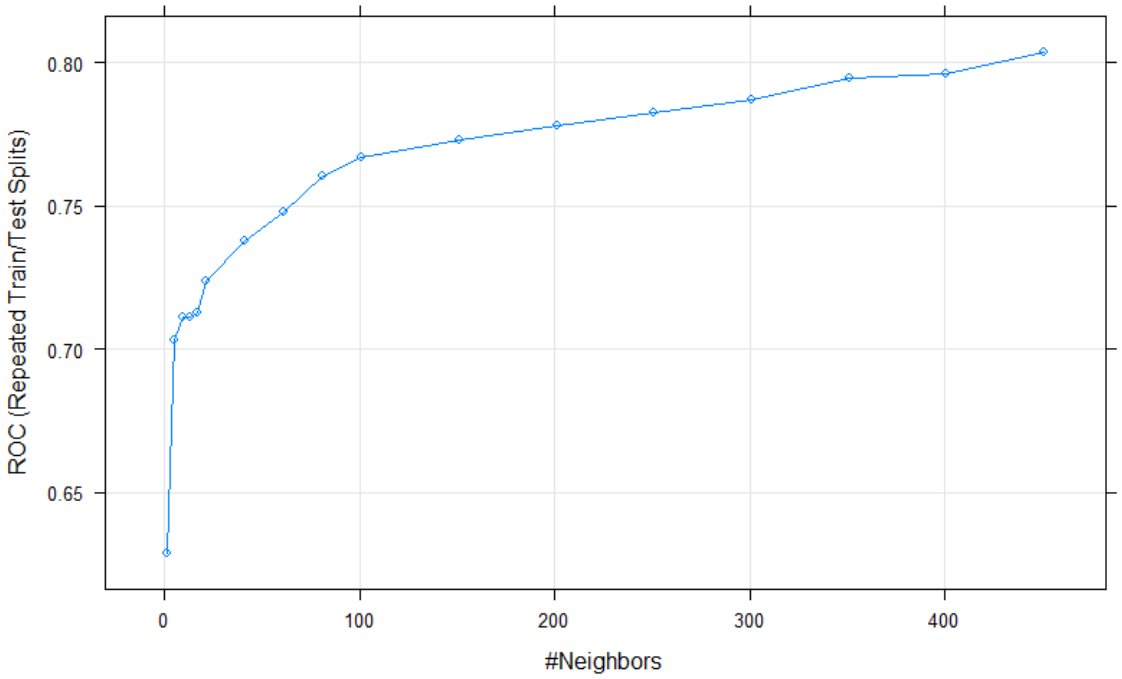

Figura 2: Variação de desempenho do modelo KNN a partir do número de vizinhos

	Referência	
Predição	Sucesso	Fracasso
Sucesso	58	26
Fracasso	131	303

Tabela 2: Matriz de Confusão - KNN

4 Conclusão

A despeito dos dois modelos utilizados para fazer a comparação observamos uma diferença de performance notável entre a regressão logística e o modelo KNN. O modelo não linear apresenta performance inferior à do linear, e isso pode ser devido a várias questões, como por exemplo podem haver muitos vizinhos que estão distantes da instância terem um peso muito grande e atrapalhar a classificação da instância, que poderia ser sanado utilizando o inverso da distancia Euclidiana para se processar os valores dos pesos dos vetores. No caso da performance do método linear ter sido superior, podemos constatar que a resposta segue um padrão também linear, evidenciado tanto pela tabela de confusão, quanto pela área ROC do gráfico. Fica claro que ambos os modelos são de fácil implementação, porém o modelo KNN apresentou degradação pela presença de ruído e linearidade. Outro fator a ser levado em consideração é o alto gasto computacional que o KNN apresenta para ser utilizado. No conjunto de dados utilizado, o KNN levou uma média de cinco minutos para processar a classificação das amostras. Portanto, este método pode ser considerado inviável para conjuntos de dados muito extensos, com enorme quantidade de preditores.

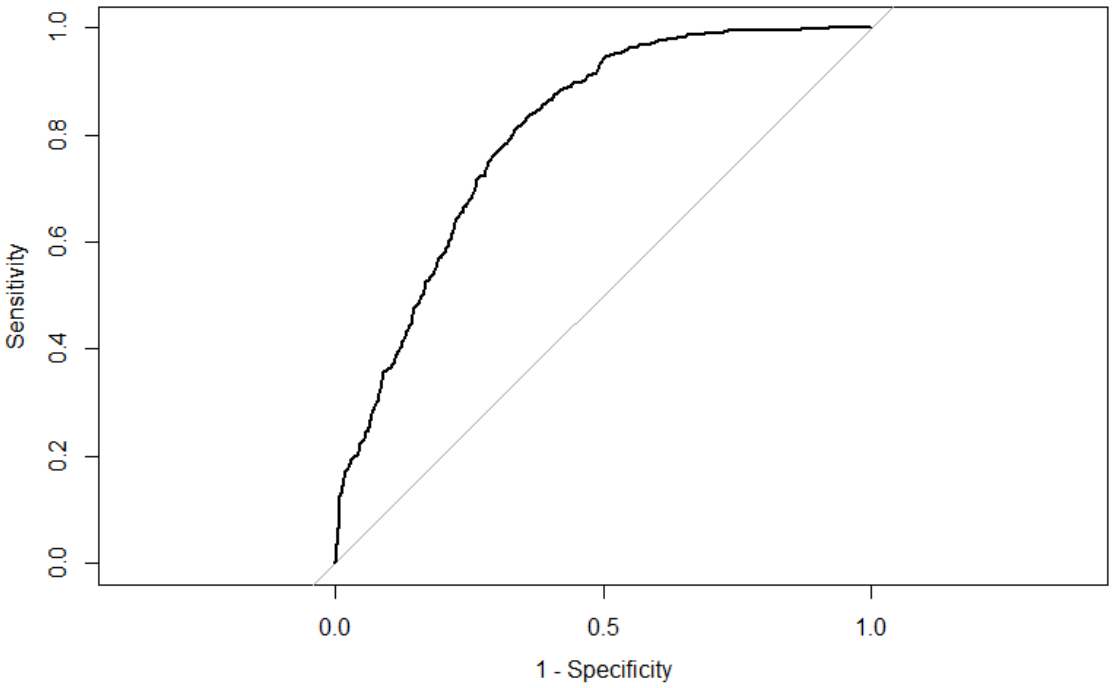

Figura 3: Curva ROC do Modelo de K-Vizinhos mais Próximos

Referências

[1] M. Kuhn and K. Johnson. *Applied Predictive Modeling* , 2014.

[2] G. James, H. Trevor, W. Daniela, T. Robert. *An Introduction to Statistical Learning with Applications in R*, 2013