Análise do classificador K-Nearest Neighbors

Gustavo Padilha Polleti*

Escola politécnica - Universidade de São Paulo
gustavo.polleti@usp.br

17 de março de 2019

Resumo

*Este relatório tem o objetivo de descrever o projeto de um classificador binário K-Nearest Neighbors baseado em dados tabulares, bem como estudar o modelo resultante. A análise de resultados apresentará a relação entre parâmetros de qualidade tradicionais (Root Mean Squared Error, accuracy e F1-score) ao variar o hyperparameter **K**.*

I. Introdução

O classificador *K-Nearest Neighbor* atribui a um ponto a mesma classe da maioria dos seus *K* vizinhos mais próximos, os quais advém da base de treino.

A proposta deste trabalho é aplicar o modelo KNN para construir um classificador binário capaz de predizer se a renda de um indivíduo é maior ou menor a 50 mil dólares, dado suas informações demográficas.

II. Resultados Experimentais

Para avaliar o modelo do classificador, construiu-se as curvas de *accuracy*, *root mean squared error* e *f1-score* ao variar o *hyperparameter* **K** de 1 a 100. Também foi feita a *confusion matrix*.

Pela figura 1, pode-se verifica uma expressiva queda do erro acompanhada do aumento da acurácia e do *f1-score* até o **K = 27**, com ***accuracy = 83,35%, rmse = 33,66% e f1-score = 62%***

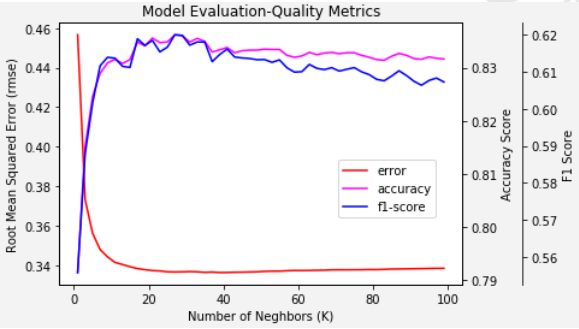

Figura 1: *Curva de métricas de qualidade*

A partir desse ponto, o erro cresce suavemente juntamento com a ligeira queda da acurácia, já o *f1-score* também apresenta queda, porém bem mais acentuada.

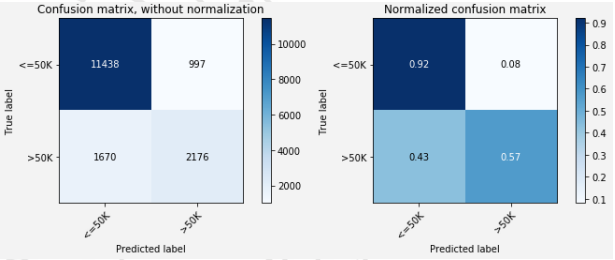

Figura 2: *Matriz de confusão para **K = 27***

Pela matriz de confusão calculada, para a *label* >**50k**, além das anteriores, pode-se calcular outras métricas: ***precision = 68,58%***, ***recall = 56,58%***, ***false positive rate = 8%*** e ***true negative rate = 92%***.

III. Conclusão

Pode-se observar uma significativa desproporcionalidade entre as *labels* (>50k: 23,93% , <=50k: 76,07%), o que provavelmente é a causa de um viés a favor da *label* <**=50k**. Embora a acurácia do classificador seja alta (**83,35%**), não é muito confiável. Por exemplo, a probabilidade de uma instância com *label* >**50k** ser classificada como tal é praticamente 50% ! (*recall = 56,58%*).

*NUSP: 9345193