

k -Vizinhos mais Próximos

Advanced Institute for Artificial Intelligence – AI2

<https://advancedinstitute.ai>

Uma das técnicas mais tradicionais em aprendizado de máquina é conhecida por k -Vizinhos mais Próximos, do inglês *k-Nearest Neighbours* - k -NN. Esta técnica é uma generalização de outra mais antiga conhecida por Vizinhos mais Próximos, do inglês *Nearest Neighbours* - NN. Ambas são abordagens bastante simples, pois **não existe etapa de treinamento**, muito embora tenhamos ainda o conjunto de treinamento.

Definição do problema: dado um conjunto de treinamento rotulado com m amostras $\mathcal{X}^1 = \{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \dots, (\mathbf{x}_m, y_m)\}$, queremos classificar corretamente uma amostra $\mathbf{x} \in \mathcal{X}^2$.

Objetivo: dada uma amostra \mathbf{x} qualquer do conjunto de teste, o seu rótulo y será o mesmo da amostra mais próxima do conjunto de treinamento, ou seja, aquela que satisfaz a seguinte equação:

$$y = \arg \min_{y_i | \mathbf{x}_i \in \mathcal{X}^1} \|\mathbf{x} - \mathbf{x}_i\|_2. \quad (1)$$

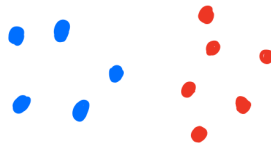
Veamos um exemplo do funcionamento da técnica NN.



Conjunto de treinamento

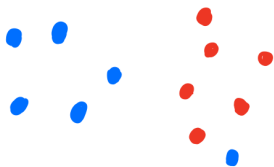


Amostra a ser classificada

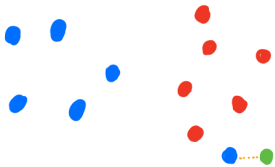


Amostra classificada

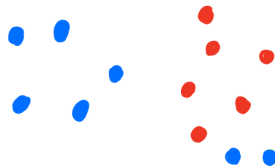
Quando o conjunto de dados está "bem comportado", NN é uma das melhores técnicas a serem utilizadas. No entanto, isso nem sempre acontece. Problema? "Ruídos" no conjunto de dados de treinamento.



Conjunto de treinamento

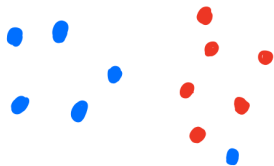


Amostra a ser classificada

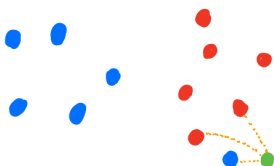


Amostra classificada

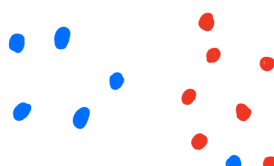
Uma generalização da técnica NN seria, então, conectar a amostra de teste aos seus k vizinhos mais próximos, dando origem ao classificador k -NN. Desta forma, considerando o exemplo anterior, a amostra seria corretamente classificada caso considerássemos $k = 3$, por exemplo (geralmente utilizamos valores ímpares para k para evitarmos desempates).



Conjunto de treinamento



Amostra a ser classificada

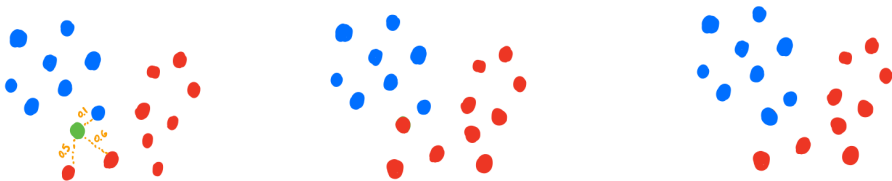


Amostra classificada

A técnica k -NN é interessante para problemas de **recomendação** e **recuperação**, dado que faz uso das amostras mais próximas para tomada de decisão. Esses dados podem ser, então, utilizados para fins de recomendação. Outro ponto importante diz respeito à **regressão** por k -NN, que também é bastante simples. Neste caso, ao conectar à amostra de teste aos seus k vizinhos mais próximos, basta utilizar, por exemplo, o valor médio de suas saídas como sendo o valor a ser estimado.

Uma variante conhecida da técnica k -NN é a sua versão **ponderada**, conhecida por *weighted k -NN*. A ideia consiste em associar pesos à cada um dos k vizinhos mais próximos, que podem ser, por exemplo, o **inverso de sua distância** para a amostra em questão. Esses pesos são normalizados e utilizados para ponderar a decisão. A ideia é que amostras mais longes tenham menos influência durante o processo de decisão.

Vejamos um exemplo do k -nn ponderado versus a sua versão tradicional.



Conjunto de treinamento Classificação por k -NN Classificação por k -NN ponderado

Seja $\mathbf{w} \in \mathbb{R}^3$ o vetor de pesos, tal que $w_1 = 1/0.1$ (classe azul), $w_2 = 1/0.5$ (classe vermelha) e $w_3 = 1/0.6$ (classe vermelha). Normalizando os mesmos, temos que $w_1 = 0.74$, $w_2 = 0.17$ e $w_3 = 0.09$. Muito embora a amostra verde esteja ligada às duas amostras da classe vermelha, o peso desta classe ($w_2 + w_3 = 0.26$) é menor do que aquele dado pela amostra da classe azul, ou seja, $w_1 = 0.74$.