03

Vizinhos mais próximos

Transcrição

Veremos o algoritmo dos "*Nearest neighbours*" ou **Vizinhos mais próximos**.

Vamos pensar que temos uma série de clientes que já classificamos entre os que fizeram ou não um depósito. Teremos um gráfico com os clientes que não fizeram um depósito representados pelas bolinhas vermelhas e as bolinhas azuis serão os que depositaram. Esse será um gráfico 2d, apesar do ideal ser um gráfico multidimensional, pois ele teria todos os atributos de clientes.

Teremos no Eixo x do gráfico a Idade e no y, o Emprego. Haverá agora um novo cliente a ser classificado. Para fazer isso com esse algoritmo, vamos nos basear na distância dele com relação a seu vizinho mais próximo. Vamos classificá-lo da mesma forma que o cliente que estiver mais próximo dele.

Poderemos medir a distância de forma Euclidiana.

> Uma medida com base na equação Euclidiana nesse exemplo seria dada pela diferença entre os empregos do clientes 1 e 2 elevada ao quadrado, somada à diferença entre as idades dos clientes 1 e 2 elevada ao quadrado. O resultado será a raiz de tudo isso.

Com a distância, sabemos que se nosso cliente estiver mais perto de um cliente representado em vermelho no gráfico, ele também se tornará vermelho.

O ideal é pegarmos mais um caso de vizinho mais próximo. Podemos considerar que nossos dados são ruidosos e a base de dados dos clientes está desatualizada, por exemplo, com relação à idade. Para diminuir esses erros, pegaremos mais clientes, como 2 vizinhos mais próximos em vez de apenas 1. K representará o número de vizinhos.

Pegando dois vizinhos, poderíamos dizer que K = 2 no nosso caso em que temos duas classes, a dos clientes que fizeram e a dos que não fizeram o depósito? Não seria uma ideia muito inteligente, pois poderíamos ter vizinhos mais próximos com a mesma distância, tão próximos de "Não depositou" quanto de "Depositou". Então, teríamos um impasse. Precisaríamos de pelo menos um K = 3, então. Num caso mais genérico, não devemos ter um K múltiplo do número de classes que temos.

Retornando ao Weka, veremos um algoritmo de classificação pelos K vizinhos mais próximos. Eles ficam em "*Choose* > *"lazy*". Encontraremos o algoritmo como ***Ibk***, que significa "*Instance-based K nearest neighbours*" ou seja, os vizinhos mais próximos com base nas instâncias, sendo cada cliente uma instância de classificação.

Faremos uma classificação com esse algoritmo e veremos que os resultados no caso dele foram similares aos do Naive Bayes, em torno dos 71%. Poderemos modificar as opções dele também, e veremos que a primeira serão os "*KNN*", vizinhos mais próximos.

A opção padrão está como 1, utilizaremos 3 para ver se os resultados se modificarão. Clicaremos em "Ok" na janela das opções e em "*Start*" no classificador mais uma vez. Veremos que as *folds* serão percorridas e a taxa de acerto aumentará um pouco, para 73% no novo resultado calculado utilizando mais Ks. Conseguimos essa melhoria por tentativa e erro.