01

Naive Bayes

Transcrição

Agora veremos outro algoritmo, o algoritmo de **Naive Bayes**.

Ele tem esse nome porque "*naive*" significa "inocente", e ele é baseado numa premissa inocente de que cada atributo da nossa base de dados tem o mesmo peso na hora da classificação, o que não é verdade em casos reais. É baseado também no Teorema de Bayes, inventado por Thomas Bayes, que tem relação com probabilidade. Vamos justamente trabalhar com probabilidade, pensando nas possibilidades de sim ou não para o depósito.

Teremos nesse momento uma base de dados enxuta, com apenas 14 clientes ou instâncias, 4 atributos (Estado civil, Educação, Habitação e Empréstimo) e uma classe. Podemos pensar que o número de vezes que "Sim" aparece para as 14 instâncias é de 9 vezes. Sendo assim, a probabilidade dele aparecer é de 9/14.

P(Sim) = 9/14, sendo P(C) ou P(Classe)

Para o "Não", teremos 5 respostas, ou seja, 5/14.

P(Não) = 5/14

Poderíamos avaliar também essas probabilidades para cada uma das respostas dos atributos. Então, para o caso do Estado Civil, a palavra "Solteiro" aparece 2 vezes no caso do "Sim" e 3 vezes no caso do depósito não ter sido feito.

P(Estado civil = Solteiro | Depósito = Sim) = 2/9

(Estado civil = Solteiro | Depósito = Não) = 3/5

Se considerarmos um cliente com as características de ser Solteiro, com ensino Primário, que tem uma Habitação e respondeu "Sim" para um Empréstimo, teríamos que fazer a avaliação para cada uma das características dele, no caso de ter feito e de não ter feito um depósito.

O caso de ser Solteiro e ter respondido "Sim" para o depósito ocorrerá por 2 vezes. Educação Primária e "Sim" para o depósito, 3 vezes. Ter Habitação e ter feito o depósito também acontecerá 3 vezes. Ter feito um Empréstimo e o Depósito, 3 vezes também.

P(Estado civil = Solteiro | Depósito = Sim) = 2/9

P(Educação = Primário | Depósito = Sim) = 3/9

P(Habitação = Sim | Depósito = Sim) = 3/9

P(Empréstimo = Sim | Empréstimo = Sim) = 3/9

P(Depósito = Sim) = 9/14

Para a situação de não ter feito um depósito, também seria necessário fazer a análise para cada uma das características do cliente. 3 vezes no caso do "Não" para ser Solteiro, 1 vez para a educação Primária, 4 vezes para ter Habitação e 3 vezes para ter feito um Empréstimo. Também devemos ter anotado que o "Não" aparece 5 vezes em 14 instâncias.

P(Estado civil = Solteiro | Depósito = Não) = 3/5

P(Educação = Primário | Depósito = Não) = 1/5

P(Habitação = Sim | Depósito = Não) = 4/5

P(Empréstimo = Sim | Empréstimo = Não) = 3/5

P(Depósito = Não) = 5/14

Teremos por fim essas duas situações. Tratando-se do "Sim", teremos ainda que multiplicar cada uma das probabilidades obtidas, de forma que obteremos um valor final de 0,0053.

P(x | Depósito = Sim) P(Depósito = Sim) = (2/9) * (3/9) * (3/9) * (3/9) * (9/14) = 0,0053

Para o "Não" também multiplicaremos todos os valores obtidos, e teremos como produto o valor 0,0206.

P(x | Depósito = Não) P(Depósito = Não) = (3/5) * (1/5) * (4/5) * (3/5) * (5/14) = 0,0206

Posteriormente, faremos uma normalização, pegando o número de vezes total que cada um desses atributos apareceu na base de dados. O Empréstimo apareceu um determinado número de vezes, 7, para 14 instâncias, então usaremos esses valores. Depois, multiplicaremos todos esses valores para obter o da normalização, de 0,02186.

Dividiremos o valor obtido para o "Sim" pelo 0,02186 e obteremos um novo valor, e faremos o mesmo para o "Não".

P(x) = (5/14) * (4/14) * (7/14) * (6/14) = 0,02186

P(Depósito = Sim | x) = 0,0053/0,02186 = 0,2424

P(Depósito = Não | x) = 0,0206/0,02186 = 0.9421

Assim, estaremos avaliando esse cliente que possivelmente não teríamos visto durante o treinamento da classificação e nessa ocasião, a probabilidade é maior para o "Não". Então, o classificaríamos como um cliente que não faria o depósito.

Retornando ao Weka, vamos clicar em "*Choose > bayes > NaiveBayes*" e após selecionar o algoritmo, clicaremos em "*Start*" novamente. Veremos que ele teve uma taxa de acerto mais baixa, 77%. Porém, ele se baseou justamente nas duas situações, a do "*Yes*" e a do "*No*" para o depósito.