Algoritmo Classificação Bayesiana

O algoritmo de Classificação Bayesiana recebe este nome por ser baseado no teorema de probabilidade de Bayes. Também é conhecido por algoritmo de Bayes.

O algoritmo tem como objetivo calcular a probabilidade de uma amostra desconhecida pertencer a cada uma das classes possíveis, ou seja, predizer a classe mais provável. Este tipo de predição é chamada de classificação estatística, pois é baseada em probabilidades.

Este algoritmo requer um conjunto de dados prévio que já esteja classificado, ou seja, um conjunto onde as linhas estejam separadas em classes (ou clusters). Baseado neste conjunto de dados prévio o algoritmo recebe como entrada uma nova amostra desconhecida, que não possui classificação, e retorna como saída a classe mais provável para esta amostra de acordo com cálculos probabilísticos. Diferente do algoritmo K-Means, a Classificação Bayesiana não necessita de uma métrica para comparar a ‘distância’ entre as instâncias e nem classifica a amostra desconhecida automaticamente, pois é necessário um conjunto de dados já classificados. Devido a esta necessidade considera-se o algoritmo de Classificação Bayesiana como um algoritmo de mineração de dados supervisionado.

Passo a Passo

Passo	Descrição
Passo 01	Cálculos das probabilidades das classes. Cada classe do conjunto de treinamento possui sua probabilidade calculada. Na maioria das vezes, trabalhamos com apenas duas classes. Por exemplo: uma classe que indica se um determinado consumidor compra ou não um produto baseado em características demográficas. O cálculo é feito dividindo-se o número de instâncias da classe pelo número total de instâncias do conjunto de treinamento.
Passo 02	Cálculo das probabilidades da amostra desconhecida. Cada valor de cada atributo da amostra desconhecida possui sua probabilidade calculada para cada possível classe. Este passo é onde o processamento mais ‘pesado’ do algoritmo ocorre, pois, dependendo do número de atributos, classes e instâncias do conjunto de treinamento, é possível que muitos cálculos sejam necessários para se obter as probabilidades.
Passo 03	Calcular a probabilidades da amostra desconhecida. Neste passo, as probabilidades calculadas para os valores da amostra desconhecida de uma mesma classe são multiplicadas. Em seguida, o valor obtido é multiplicado pela probabilidade da classe calculada no Passo 01. Com as probabilidades de cada classe calculadas, verifica-se qual é a classe que possui maior probabilidade para a amostra desconhecida. Com isso, o algoritmo termina retornando a classe que possui maior probabilidade de conter a amostra desconhecida.

Exemplo

Imagine uma financeira que deseja prever se um cliente será inadimplente ou não. Para isso, a financeira deve levar em consideração a sua base histórica de clientes e alguns atributos. Para facilitar o entendimento do cenário e do modelo de dados vamos utilizar um conjunto de treinamento com quinze linhas e com quatro atributos. A imagem abaixo mostra o conjunto de treinamento que será utilizado neste exemplo, armazenado na tabela TB_FINANCEIRA.

CODIGO_CLIENTE	SEXO	ESTADOCIVIL	ESCOLARIDADE	RENDIMENTOS	INADIMPLENTE
1	FEMININO	SOLTEIRO	ENSINO MÉDIO INCOMPLETO	ACIMA DE TRÊS SALARIOS MÍNIMOS	NAO
2	FEMININO	SOLTEIRO	ENSINO MÉDIO INCOMPLETO	UM SALARIO MÍNIMO	NAO
3	MASCULINO	SOLTEIRO	ENSINO MÉDIO COMPLETO	ACIMA DE TRÊS SALARIOS MÍNIMOS	SIM
4	FEMININO	SOLTEIRO	ENSINO MÉDIO COMPLETO	UM SALARIO MÍNIMO	NAO
5	FEMININO	SOLTEIRO	SUPERIOR INCOMPLETO	DOIS SALARIOS MÍNIMOS	NAO
6	MASCULINO	CASADO	ENSINO MÉDIO COMPLETO	UM SALARIO MÍNIMO	NAO
7	FEMININO	CASADO	ENSINO MÉDIO COMPLETO	ACIMA DE TRÊS SALARIOS MÍNIMOS	NAO
8	MASCULINO	CASADO	ENSINO MÉDIO INCOMPLETO	ACIMA DE TRÊS SALARIOS MÍNIMOS	NAO
9	MASCULINO	CASADO	ENSINO MÉDIO INCOMPLETO	UM SALARIO MÍNIMO	NAO
10	MASCULINO	CASADO	SUPERIOR COMPLETO	DOIS SALARIOS MÍNIMOS	NAO
11	FEMININO	SOLTEIRO	ENSINO MÉDIO COMPLETO	ACIMA DE TRÊS SALARIOS MÍNIMOS	NAO
12	FEMININO	CASADO	ENSINO MÉDIO INCOMPLETO	DOIS SALARIOS MÍNIMOS	SIM
13	FEMININO	CASADO	SUPERIOR INCOMPLETO	UM SALARIO MÍNIMO	SIM
14	MASCULINO	CASADO	SUPERIOR INCOMPLETO	DOIS SALARIOS MÍNIMOS	SIM
15	FEMININO	SOLTEIRO	SUPERIOR COMPLETO	DOIS SALARIOS MÍNIMOS	NAO

Segue a descrição de cada coluna:

CODIGO_CLIENTE: possui um identificador inteiro sequencial.

SEXO: identifica o sexo do cliente. Pode assumir apenas os valores MASCULINO e FEMININO.

ESTADOCIVIL: informações sobre o estado civil do cliente. Pode assumir apenas os valores CASADO e SOLTEIRO.

ESCOLARIDADE: informações sobre a escolaridade do cliente. Pode assumir apenas quatro valores diferentes: ENSINO MÉDIO INCOMPLETO, ENSINO MÉDIO COMPLETO, SUPERIOR INCOMPLETO e SUPERIOR COMPLETO.

RENDIMENTOS: informações sobre a faixa salarial do cliente. Pode assumir apenas os valores UM SALÁRIO MÍNIMO, DOIS SALÁRIOS MÍNIMOS e ACIMA DE TRÊS SALÁRIOS MÍNIMOS.

INADIMPLENTE: é a coluna que apresenta a classificação das amostras. Neste exemplo a classificação indica se o cliente é devedor, isto é INADIMPLENTE=SIM, ou se o cliente não é devedor, isto é INADIMPLENTE=NAO. Para facilitar a visualização, os clientes do conjunto de treinamento que são inadimplentes foram marcados em vermelho e os clientes que não são inadimplentes foram marcados em azul.

Passo a Passo do Exemplo

Vamos executar a Classificação Bayesiana para a amostra desconhecia a seguir:

SEXO	ESTADOCIVIL	ESCOLARIDADE	RENDIMENTOS	INADIMPLENTE
MASCULINO	SOLTEIRO	ENSINO MÉDIO INCOMPLETO	UM SALARIO MÍNIMO	?

Passo 01: Cálculos das probabilidades das classes.

Existem apenas duas classes, uma que indica que o cliente é devedor (INADIMPLENTE=SIM) e outra que indica que o cliente não é devedor (INADIMPLENTE=NÃO). Calculando as probabilidades das classes tempos:

Probabilidade INADIMPLENTE=SIM: 4/15= 0,2667
Probabilidade INADIMPLENTE=NÃO: 11/15= 0,7334

Passo 02: Cálculo das probabilidades da amostra desconhecida.

Para o primeiro atributo da amostra desconhecida SEXO=MASCULINO, vamos calcular a probabilidade de INADIMPLENTE=SIM:

Probabilidade de SEXO=MASCULINO e INADIMPLENTE=SIM: 2/4 = 0,5

E para o caso onde o cliente é masculino e não é devedor temos:

Probabilidade de SEXO=MASCULINO e INADIMPLENTE=NAO: 4/11 = 0,3636

Para os demais valores dos atributos da amostra desconhecida, temos:

Probabilidade de ESTADOCIVIL=SOLTEIRO e INADIMPLENTE=SIM: 1/4 = 0,25
Probabilidade de ESTADOCIVIL=SOLTEIRO e INADIMPLENTE=NÃO: 6/11 = 0,5455

Probabilidade de ESCOLARIDADE= ENSINO MÉDIO INCOMPLETO e INADIMPLENTE=SIM: 1/4 = 0,25
Probabilidade de ESCOLARIDADE = ENSINO MÉDIO INCOMPLETO e INADIMPLENTE=NÃO: 4/11 = 0,3636

Probabilidade de RENDIMENTOS= UM SALÁRIO MÍNIMO e INADIMPLENTE=SIM: 1/4 = 0,25
Probabilidade de RENDIMENTOS = UM SALÁRIO MÍNIMO e INADIMPLENTE=NÃO: 4/11 = 0,3636

Passo 03: Calcular a probabilidades da amostra desconhecida.

Multiplicando as probabilidades da amostra desconhecida para o caso de INADIMPLENTE=SIM pela probabilidade de inadimplência calculada no PASSO 1 temos:

0,5 x 0,25 x 0,25 x 0,25 x 0,2667 = 0,0021

Multiplicando as probabilidades da amostra desconhecida para o caso de INADIMPLENTE=NÃO pela probabilidade de não inadimplência calculada no Passo 01, temos:

0,3636 x 0,5455 x 0,3636 x 0,3636 x 0,7334 = 0,0192

Como 0,0192 > 0,0021, o algoritmo classifica a amostra desconhecida como INADIMPLENTE=NÃO, ou seja, este novo cliente tem uma probabilidade maior de não se tornar devedor do que de se tornar inadimplente, de acordo com os dados históricos e a Classificação Bayesiana.

Vejamos na prática!