Naive Bayes

April 29, 2025

0.1 Naive Bayes

Classificadores Naive Bayes são uma coleção de algoritmos de classificação baseados no Teorema de Bayes. Não é um único algoritmo, mas uma família de algoritmos que compartilham um princípio comum: cada par de características sendo classificadas é independente uma da outra.

Para começar, vamos considerar um conjunto de dados. Um dos algoritmos de classificação mais simples e eficazes, o classificador Naive Bayes, auxilia no rápido desenvolvimento de modelos de aprendizado de máquina com capacidades de previsão rápidas.

O algoritmo Naive Bayes é usado para problemas de classificação e é amplamente utilizado na classificação de textos. Em tarefas de classificação de textos, os dados possuem alta dimensão (já que cada palavra representa uma característica nos dados). Ele é usado em filtragem de spam, detecção de sentimentos, classificação de avaliações, etc.

A vantagem de usar o Naive Bayes é sua velocidade. Ele é rápido e facilita a previsão com alta dimensão de dados. Este modelo prevê a probabilidade de uma instância pertencer a uma classe com um conjunto de valores de características. É um classificador probabilístico, pois assume que uma característica no modelo é independente da existência de outra característica. Em outras palavras, cada característica contribui para as previsões sem relação entre si. No mundo real, essa condição raramente é satisfeita.

0.2 Por que é chamado de Naive Bayes?

A parte Naive do nome indica a suposição simplificadora feita pelo classificador Naive Bayes. O classificador assume que as características usadas para descrever uma observação são condicionalmente independentes, dado o rótulo da classe. A parte "Bayes" do nome refere-se ao Reverendo Thomas Bayes, um estatístico e teólogo do século XVIII que formulou o Teorema de Bayes.

Considere um conjunto de dados fictício que descreve as condições climáticas para jogar golfe. Dadas as condições climáticas, cada tupla classifica as condições como adequadas ("Sim") ou inadequadas ("Não") para jogar golfe.

Aqui está uma representação do nosso conjunto de dados:

#	Tempo	Temperatura	Umidade	Vento	Jogar golfe
0	Chuvoso	Quente	Alto	Falso	Não
1	Chuvoso	Quente	Alto	Verdadeiro	Não
2	Nublado	Quente	Alto	Falso	Sim
3	Ensolarado	Amena	Alto	Falso	Sim
4	Ensolarado	Frio	Normal	Falso	Sim

#	Tempo	Temperatura	Umidade	Vento	Jogar golfe
5	Ensolarado	Frio	Normal	Verdadeiro	Não
6	Nublado	Frio	Normal	Verdadeiro	Sim
7	Chuvoso	Amena	Alto	Falso	Não
8	Chuvoso	Frio	Normal	Falso	Sim
9	Ensolarado	Amena	Normal	Falso	Sim
10	Chuvoso	Amena	Normal	Verdadeiro	Sim
11	Nublado	Amena	Alto	Verdadeiro	Sim
12	Nublado	Quente	Normal	Falso	Sim
13	Ensolarado	Amena	Alto	Verdadeiro	Não

O conjunto de dados é dividido em duas partes: a matriz de características e o vetor de resposta.

- Matriz de características: Contém todos os vetores (linhas) do conjunto de dados, onde cada vetor consiste no valor das características dependentes. No conjunto de dados acima, as características são 'Outlook', 'Temperature', 'Humidity' e 'Windy'.
- Vetor de resposta: Contém o valor da variável de classe (previsão ou saída) para cada linha da matriz de características. No conjunto de dados acima, o nome da variável de classe é 'Play Golf'.

0.3 Suposições do modelo Naive Bayes

A suposição fundamental do Naive Bayes é que cada característica é:

- Independência das características: As características dos dados são condicionalmente independentes umas das outras, dado o rótulo da classe.
- Características contínuas são distribuídas normalmente: Se uma característica é contínua, então assume-se que ela é distribuída normalmente dentro de cada classe.
- Características discretas têm distribuições multinomiais: Se uma característica é discreta, então assume-se que ela tem uma distribuição multinomial dentro de cada classe.
- Características são igualmente importantes: Todas as características são assumidas como contribuindo igualmente para a previsão do rótulo da classe.
- Sem dados ausentes: Os dados não devem conter valores ausentes.

Em relação ao nosso conjunto de dados, este conceito pode ser entendido da seguinte forma:

- Assumimos que nenhum par de características é dependente. Por exemplo, a temperatura ser 'Hot' não tem nada a ver com a umidade, ou o outlook ser 'Rainy' não afeta os ventos. Portanto, as características são assumidas como independentes.
- Cada característica é dada o mesmo peso (ou importância). Por exemplo, saber apenas a temperatura e a umidade sozinhas não pode prever o resultado com precisão. Nenhum dos atributos é irrelevante e assume-se que todos contribuem igualmente para o resultado.

As suposições feitas pelo Naive Bayes geralmente não estão corretas em situações do mundo real. Na verdade, a suposição de independência nunca está correta, mas muitas vezes funciona bem na prática.

Antes de passar para a fórmula do Naive Bayes, é importante conhecer o Teorema de Bayes.

0.4 Teorema de Bayes

O Teorema de Bayes encontra a probabilidade de um evento ocorrer, dada a probabilidade de outro evento que já ocorreu. O teorema de Bayes é expresso matematicamente pela seguinte equação:

$$P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)}$$

onde A e B são eventos e $P(B) \neq 0$.

Basicamente, estamos tentando encontrar a probabilidade do evento A, dado que o evento B é verdadeiro. O evento B também é chamado de evidência.

- $P(A)$ é a priori de A (a probabilidade anterior, ou seja, a probabilidade do evento antes de ver a evidência).
- $P(B)$ é a probabilidade marginal: probabilidade da evidência.
- $P(A|B)$ é a probabilidade a posteriori de B, ou seja, a probabilidade do evento após ver a evidência.
- $P(B|A)$ é a probabilidade de ver a evidência, dado que a hipótese é verdadeira.

Agora, em relação ao nosso conjunto de dados, podemos aplicar o teorema de Bayes da seguinte maneira:

$$P(y|X) = \frac{P(X|y) \cdot P(y)}{P(X)}$$

onde y é a variável de classe e X é um vetor de características dependentes (de tamanho n), onde:

$$X = (x_1, x_2, x_3, ... x_n)$$

Para esclarecer, um exemplo de um vetor de características e a variável de classe correspondente pode ser: (referindo-se à primeira linha do conjunto de dados)

- $$X = (\text{Chuvoso, Quente, Alta, Falso})$$
- y= Não

aqui significa a probabilidade de “Não jogar golfe” dado que as condições climáticas são “Chuva”, “Temperatura quente”, “Alta umidade” e “Sem vento”.

0.4.1 Suposições do Naive Bayes

- Independência das características: Assumimos que nenhum par de características é dependente. Por exemplo, a temperatura ser ‘Quente’ não tem nada a ver com a umidade, ou o tempo ser ‘Chuvoso’ não afeta os ventos. Portanto, as características são assumidas como independentes.

- Peso igual para cada característica: Cada característica é dada o mesmo peso (ou importância). Por exemplo, saber apenas a temperatura e a umidade sozinhas não pode prever o resultado com precisão. Nenhum dos atributos é irrelevante e assume-se que todos contribuem igualmente para o resultado.

As suposições feitas pelo Naive Bayes geralmente não estão corretas em situações do mundo real. Na verdade, a suposição de independência nunca está correta, mas muitas vezes funciona bem na prática.

Agora, é hora de aplicar uma suposição ingênua ao teorema de Bayes, que é a independência entre as características. Então, dividimos a evidência em partes independentes.

Se dois eventos A e B são independentes, então:

$$P(A|B) = P(A) \cdot P(B)$$

Assim, chegamos ao resultado:

$$P(y|x_1, \ldots, x_n) = \frac{P(x_1|y) \cdot P(x_2|y) \ldots P(x_n|y) \cdot P(y)}{P(x_1) \cdot P(x_2) \ldots P(x_n)}$$

Como o denominador permanece constante para uma entrada dada, podemos remover esse termo:

$$P(y|x_1, \ldots, x_n) \propto P(y) \cdot \prod_{i=1}^{n} P(x_i|y)$$

Agora, precisamos criar um modelo de classificador. Para isso, encontramos a probabilidade do conjunto de entradas dado para todos os valores possíveis da variável de classe y e escolhemos a saída com a maior probabilidade. Isso pode ser expresso matematicamente como:

$$y = argmax P(y) \cdot \prod_{i=1}^{n} P(x_i|y)$$

Finalmente, resta-nos a tarefa de calcular $P(y)$ e $P(x_i|y)$. Note que $P(y)$ também é chamada de probabilidade da classe e $P(x_i|y)$ é chamada de probabilidade condicional.

Os diferentes classificadores Naive Bayes diferem principalmente pelas suposições que fazem em relação à distribuição de $P(x_i|y)$.

Vamos tentar aplicar a fórmula acima manualmente em nosso conjunto de dados climáticos. Para isso, precisamos fazer alguns pré-cálculos em nosso conjunto de dados.

Precisamos encontrar $P(x_i|y_j)$ para cada x_i em X e y_j em y. Todos esses cálculos foram demonstrados nas tabelas abaixo.

TEMPO	Sim	Não	P(sim)	P(não)	Temperatura	Sim	Não	P(sim)	P(não)
Ensolarado	3	2	3/9	2/5	Quente	2	2	2/9	2/5
Nublado	4	0	4/9	0/5	Amena	4	2	4/9	2/5
Chuvoso	2	3	2/9	3/5	Fria	3	1	3/9	1/5

TEMPO	Sim	Não	P(sim)	P(não)	Temperatura	Sim	Não	P(sim)	P(não)
Total	9	5	100%	100%	Total	9	5	100%	100%

Humidade	Sim	Não	P(sim)	P(não)	Vento	Sim	Não	P(sim)	P(não)
Alto	3	4	3/9	4/5	Falso	6	2	6/9	2/5
Normal	6	1	6/9	1/5	Verdadeiro	3	3	3/9	3/5
Total	9	5	100%	100%	Total	9	5	100%	100%

Jogar	Sim	P(sim)/P(não)
Sim	9	9/14
Não	5	5/14
Total	9	100%

[3]:
```
from IPython.display import Image
```

0.5 Classificadores Naive Bayes

Na figura acima, calculamos $P(x_i|y_j)$ para cada x_i em X e y_j em y manualmente nas tabelas acima. Por exemplo, a probabilidade de jogar golfe dado que a temperatura está fria, ou seja,

$$P(temp. = frio|JogarGolfe = Sim) = \frac{3}{9}$$

Também precisamos encontrar as probabilidades das classes $P(y)$, que foram calculadas na tabela 5. Por exemplo, $P(JogarGolfe = Sim) = \frac{9}{14}$.

Agora, terminamos nossos pré-cálculos e o classificador está pronto!

Vamos testá-lo em um novo conjunto de características (vamos chamá-lo de hoje):

- hoje = (Ensolarado, Quente, Normal, Sem vento)

A probabilidade de jogar golfe é dada por:

$$P(Sim|hoje) = \frac{P(Ensolarado|Sim) \cdot P(Quente|Sim) \cdot P(Normal|Sim) \cdot P(SemVento|Sim) \cdot P(Sim)}{P(hoje)}$$

E a probabilidade de não jogar golfe é dada por:

$$P(No|hoje) = \frac{P(Ensolarado|No) \cdot P(Quente|No) \cdot P(Normal|No) \cdot P(SemVento|No) \cdot P(No)}{P(hoje)}$$

Como $P(hoje)$ é comum em ambas as probabilidades, podemos ignorar $P(hoje)$ e encontrar as probabilidades proporcionais como:

$$P(Sim|hoje) \propto \frac{3}{9} \cdot \frac{2}{9} \cdot \frac{6}{9} \cdot \frac{6}{9} \cdot \frac{9}{14} \approx 0,02116$$

e

$$P(No|hoje) \propto \frac{2}{5} \cdot \frac{2}{5} \cdot \frac{1}{5} \cdot \frac{2}{5} \cdot \frac{5}{14} \approx 0,0046$$

Agora, como $P(Sim|hoje) + P(No|hoje) = 1$, esses números podem ser convertidos em uma probabilidade normalizando a soma para 1:

$$P(Sim|hoje) = \frac{0,02116}{0,02116 + 0,0046} \approx 0,821$$

e

$$P(No|hoje) = \frac{0,0046}{0,02116 + 0,0046} \approx 0,179$$

Como $P(Sim|hoje) > P(No|hoje)$, a previsão é que o golfe será jogado, ou seja, 'Sim'.

O método que discutimos acima é aplicável para dados discretos. No caso de dados contínuos, precisamos fazer algumas suposições sobre a distribuição dos valores de cada característica. Os diferentes classificadores Naive Bayes diferem principalmente pelas suposições que fazem em relação à distribuição de $P(x_i|y)$.

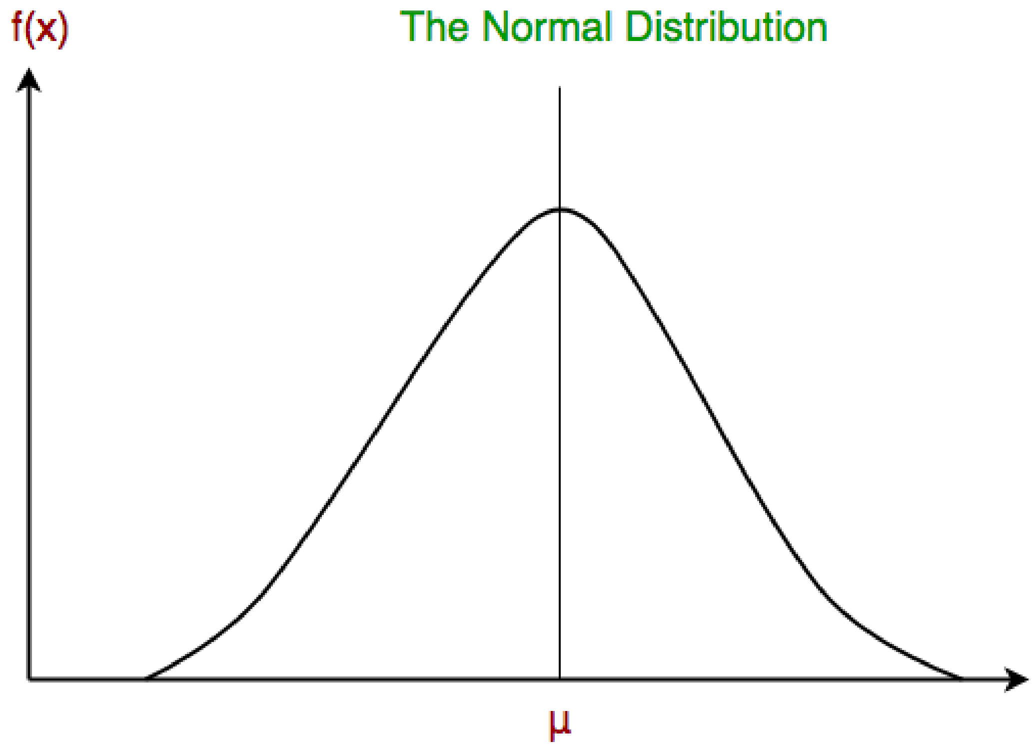
0.6 Tipos de Modelos Naive Bayes

Existem três tipos de modelos Naive Bayes:

- Classificador Naive Bayes Gaussiano: No Naive Bayes Gaussiano, os valores contínuos associados a cada característica são assumidos como distribuídos de acordo com uma distribuição Gaussiana. Uma distribuição Gaussiana também é chamada de distribuição Normal. Quando plotada, ela dá uma curva em forma de sino que é simétrica em relação à média dos valores das características.

[4]:
```
Image(filename='naive-bayes-classification-1.png')
```

[4]:



A probabilidade das características é assumida como Gaussiana, portanto, a probabilidade condicional é dada por:

$$P(x_i|y) = \frac{1}{\sqrt{2\pi\sigma_y^2}} \exp\left(-\frac{(x_i-\mu_y)^2}{2\sigma_y^2}\right)$$

	Sim	Não	P(Sim)	P(Não)
Ensolarado	3	2	3/9	2/5
Chuvoso	4	0	4/9	0/5
Nublado	2	3	2/9	3/5
Total	9	5	100%	100%

[ ]:
```
Agora, vamos ver uma implementação do classificador Naive Bayes Gaussiano␣
↪usando scikit-learn.
```

[2]:
```
# load the iris dataset
from sklearn.datasets import load_iris
iris = load_iris()
```

```
# store the feature matrix (X) and response vector (y)
X = iris.data
y = iris.target

# splitting X and y into training and testing sets
from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.4,␣
↪random_state=1)

# training the model on training set
from sklearn.naive_bayes import GaussianNB
gnb = GaussianNB()
gnb.fit(X_train, y_train)

# making predictions on the testing set
y_pred = gnb.predict(X_test)

# comparing actual response values (y_test) with predicted response values␣
↪(y_pred)
from sklearn import metrics
print("Gaussian Naive Bayes model accuracy(in %):", metrics.
↪accuracy_score(y_test, y_pred)*100)
print(metrics.classification_report(y_test, y_pred))
print(metrics.confusion_matrix(y_test, y_pred))
```

```
Gaussian Naive Bayes model accuracy(in %): 95.0
              precision    recall  f1-score   support

           0       1.00      1.00      1.00        19
           1       0.95      0.90      0.93        21
           2       0.90      0.95      0.93        20

    accuracy                           0.95        60
   macro avg       0.95      0.95      0.95        60
weighted avg       0.95      0.95      0.95        60

[[19  0  0]
 [ 0 19  2]
 [ 0  1 19]]
```

0.6.1 Naive Bayes Multinomial

Os vetores de características representam as frequências com que certos eventos foram gerados por uma distribuição multinomial. Este é o modelo de evento tipicamente usado para classificação de documentos.

0.6.2 Naive Bayes Bernoulli

No modelo de evento Bernoulli multivariado, as características são booleanas independentes (variáveis binárias) que descrevem as entradas. Como o modelo multinomial, este modelo é popular para tarefas de classificação de documentos, onde a ocorrência binária de termos (ou seja, uma palavra ocorre em um documento ou não) é usada em vez das frequências dos termos (ou seja, a frequência de uma palavra no documento).

0.6.3 Vantagens do Classificador Naive Bayes

- Fácil de implementar e computacionalmente eficiente.
- Eficaz em casos com um grande número de características.
- Desempenha bem mesmo com dados de treinamento limitados.
- Funciona bem na presença de características categóricas.
- Para características numéricas, os dados são assumidos como provenientes de distribuições normais.

0.6.4 Desvantagens do Classificador Naive Bayes

- Assume que as características são independentes, o que pode não ser verdade em dados do mundo real.
- Pode ser influenciado por atributos irrelevantes.
- Pode atribuir probabilidade zero a eventos não vistos, levando a uma generalização ruim.

0.6.5 Aplicações do Classificador Naive Bayes

- **Filtragem de Spam**: Classifica e-mails como spam ou não-spam com base em características.
- **Classificação de Textos**: Usado em análise de sentimentos, categorização de documentos e classificação de tópicos.
- **Diagnóstico Médico**: Ajuda a prever a probabilidade de uma doença com base em sintomas.
- **Avaliação de Crédito**: Avalia a credibilidade de indivíduos para aprovação de empréstimos.
- **Previsão do Tempo**: Classifica as condições climáticas com base em vários fatores.

0.6.6 Pontos Importantes

Apesar de suas suposições aparentemente simplificadas, os classificadores Naive Bayes têm funcionado muito bem em muitas situações do mundo real, como classificação de documentos e filtragem de spam. Eles requerem uma pequena quantidade de dados de treinamento para estimar os parâmetros necessários.

Os classificadores Naive Bayes podem ser extremamente rápidos em comparação com métodos mais sofisticados. A separação das distribuições de características condicionais da classe significa que cada distribuição pode ser estimada independentemente como uma distribuição unidimensional. Isso, por sua vez, ajuda a aliviar problemas decorrentes da maldição da dimensionalidade.