

# Modelos de Classificação

## Naive Bayes

---

Prof. Gustavo Willam Pereira



**INSTITUTO FEDERAL**  
Sudeste de Minas Gerais

## Naive Bayes

- O classificador Naïve Bayes se baseia no Teorema de Bayes, ou seja, qual a probabilidade de um determinado evento ocorrer dado que conhecemos algumas informações previamente.
- Abaixo segue a equação do Teorema de Bayes.

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

- Em que A e B são eventos, P(A) e P(B) são as probabilidades de ocorrência do evento A e o evento B (probabilidade priori),
- $P(A|B)$  é a probabilidade de ocorrer o evento A, se B foi observado
- $P(B|A)$  é a probabilidade de ocorrer o evento B, dado que A foi observado.

# Naive Bayes

- Vamos aplicar o Teorema de Bayes em um exemplo simples.
- Vamos imaginar que temos um problema de classificação de e-mails como spam ou não-spam.
- Vamos imaginar que de um total de 100 e-mails recebidos no dia, 20 e-mails eram spam.
- Para tentar classificar um e-mail automaticamente como spam, poderíamos verificar se uma determinada palavra específica, que está no e-mail, por exemplo, a palavra “comprar”, tem maior probabilidade de estar em mais e-mails do tipo spam.
- Então, fizemos uma análise nos e-mails que eram spam e verificamos que 15 deles tinham a palavra “comprar”.
- Por outro lado, 5 e-mails recebidos como não-spam, tinha a palavra “comprar”.

# Naive Bayes

- $P(\text{Spam}) = 20\%$
- $P(\text{Não-spam}) = 80\%$
- $P(\text{Spam}|\text{Comprar}) = ?$  (Essa é a nossa questão. Qual a probabilidade do nosso e-mail ser spam dado que tem a palavra comprar).
- $$P(\text{Spam}|\text{Comprar}) = \frac{P(\text{Comprar}|\text{Spam}) P(\text{Spam})}{P(\text{Comprar})}$$
- 
- $P(\text{Comprar}|\text{Spam}) = 15/20 = 75\%$
- $P(\text{Comprar}) = (15+5) / 100 = 20\%$
- $$P(\text{Spam}|\text{Comprar}) = \frac{75\% \cdot 20\%}{20\%} = 75\%$$

## Naive Bayes

- Então temos 75% de probabilidade de o e-mail que tem a palavra “comprar” ser um spam.
- Se fizermos o mesmo cálculo para a palavra (feature) “promoção”?
- Se o resultado encontrado for  $P(\text{Spam/Promoção}) = 90\%$ .
- Então, se assumirmos que “comprar” e “promoção” são variáveis independentes, poderíamos fazer o cálculo da probabilidade de ser spam se tiver as duas palavras como  $P(\text{Spam/Comprar}) \times P(\text{Spam/Promoção})$ .
- O algoritmo Naive Bayes faz essa suposição, que as variáveis são independentes.

## Naive Bayes

- No exemplo anterior as variáveis eram categóricas e assim foi fácil de calcular a probabilidade da variável “comprar” no conjunto de dados,  $P(\text{Comprar}) = 20\%$ .
- Mas se a variável for contínua? Veja o problema da Figura 6. Nesse exemplo queremos classificar a observação em vermelho.

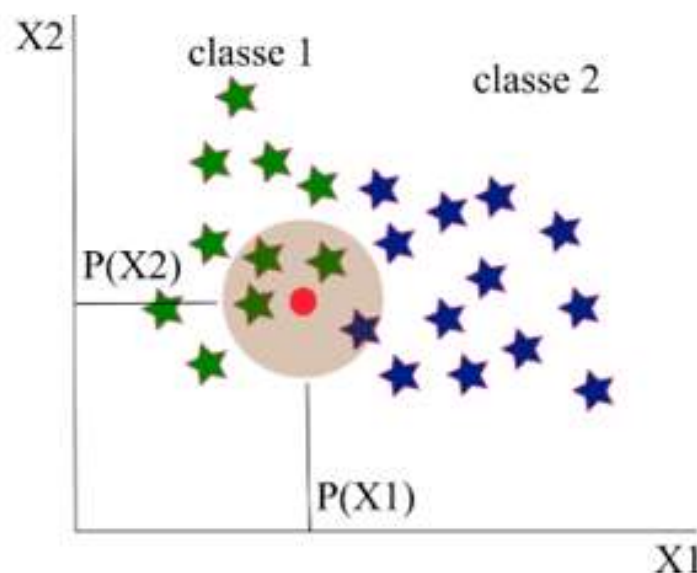


Figura 6 - Esquema de cálculo da probabilidade de cada variável contínua -  $P(X)$ .

## Naive Bayes

- Na Figura 6 existem duas classes (classe 1-verde e classe 2 - azul) e duas variáveis (X1 e X2).
- As variáveis X1 e X2 são contínuas. Nesse caso, para aplicarmos o Teorema de Bayes temos que determinar  $P(Verde|X)$ , onde X são nossas variáveis.

$$P(Verde|X) = \frac{P(X|Verde) P(Verde)}{P(X)}$$

1) Primeiro passo vamos calcular a  $P(Verde)$ .

$$P(Verde) = \frac{\text{número de instâncias verdes}}{\text{Total de instâncias}} = \frac{10}{23} = 43,5\%$$

# Naive Bayes

2) Segundo passo vamos calcular a  $P(X)$ .

Nesse caso, veja que as variáveis  $X$  são contínuas, então, definimos um círculo com centro no ponto observado, e calculamos o  $P(X)$  como:

$$P(X) = \frac{\text{número de instâncias no círculo}}{\text{Total de instâncias}} = \frac{4}{23} = 17,4\%$$



# Naive Bayes

3) Terceiro passo vamos calcular o  $P(X|Verde)$

$$P(X|Verde) = \frac{\text{número de instâncias no círculo que é verde}}{\text{Total de instâncias verde}} = \frac{3}{10} = 30,0\%$$

4) Calculamos  $P(Verde|X)$ :

$$P(Verde|X) = \frac{30\% \cdot 43,5\%}{17,4\%} = 75\%$$

Assim temos probabilidade de 75% da observação em vermelho da Figura 6 ser classe 1 (verde).



**INSTITUTO FEDERAL**  
Sudeste de Minas Gerais