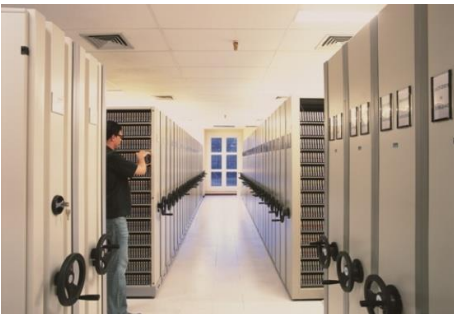

Naive Bayes

Prof. Esp. Victor Venites



SCHOOL OF AI – SÃO PAULO – AULA 9
INTRODUÇÃO A APRENDIZAGEM DE MÁQUINA

Até Aqui

Estatística –

- Centroides
- Gráficos
- Exploração de Dados

Regressão Linear –

- Álgebra Linear
- Derivadas Parciais
- Vetores e Matrizes

Árvores de Decisão –

- Classificação de Vinhos

Exemplos –

- Hands-On – 101
- Python



Introdução a Aprendizagem de Máquina



Roteiro –

- Árvores de Decisão
- **Naive Bayes**
- Support Vector Machines
- KNN
- K-means

Objetivo

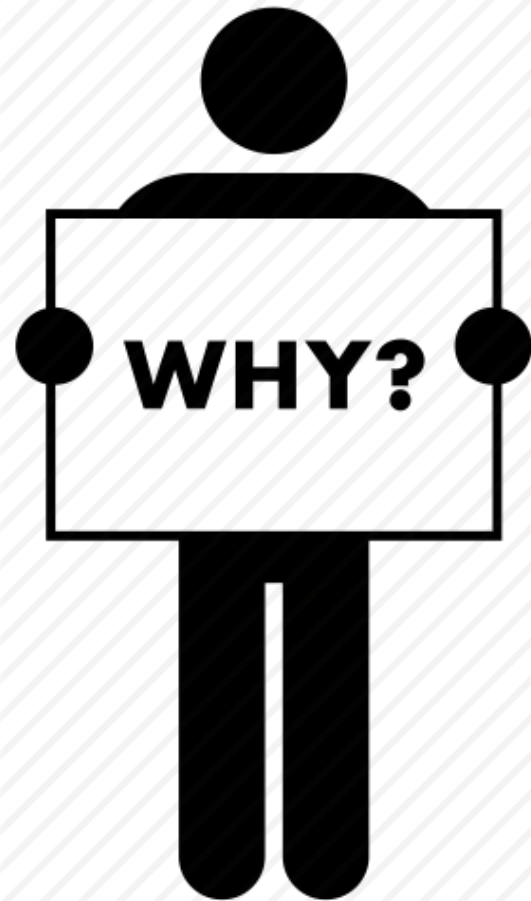
- Passar um pouco da nossa experiência
- Deixar o aluno apto para aplicar um classificador simples
- Ter noções de como fazer isso
- Partir do viés da computação
- Levantar questões... E responder a maioria!



Material: GitHub / Slides e Código

Video: Youtube - Live

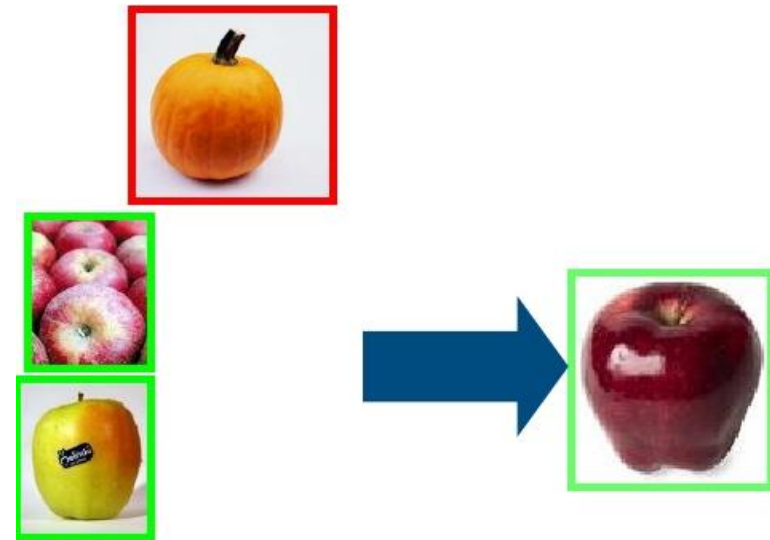
Por quê?



Para poder classificar e prever a classificação dos próximos elementos de lição

Necessidade – Exemplo

- Por exemplo, uma fruta pode ser considerada uma maçã se for vermelha, redonda e com cerca de 3 polegadas de diâmetro
- Mesmo que essas características dependam umas das outras ou da existência das outras características, todas essas propriedades contribuem independentemente para a probabilidade de que essa fruta seja uma maçã e é por isso que ela é conhecida como “Naive”

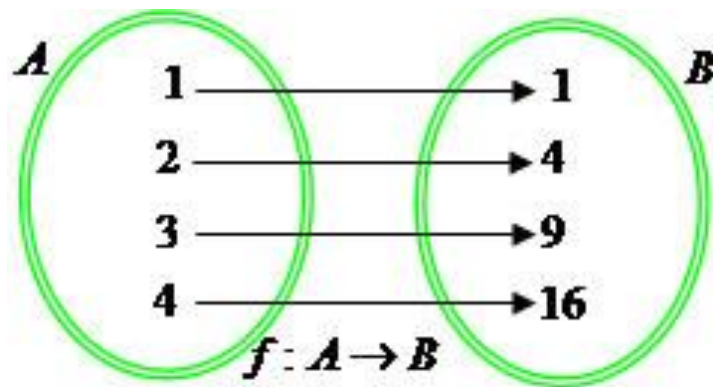


O que é Naive Bayes?

- Naive = Ingenua
- Estatística Baysiana: Teorema de Thomas Bayes(1701-1761)
- Onde com base na probabilidade de um evento ocorrer, permite qualificar e eliminar lições equivocadas(DataMining)
- Probabilidade Condicional



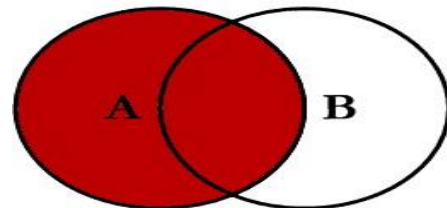
O que é Naive Bayes?



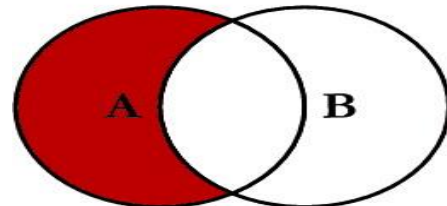
- É uma técnica de classificação baseada no Teorema de Bayes com suposição de independência entre os preditores
- Em termos simples, um classificador Naive Bayes assume que a presença de um recurso particular em uma classe não está relacionado à presença de qualquer outro recurso
- Relação entre os grupos?
- Pode haver, mas não significa causalidade.

SQL – JOIN das Tabelas

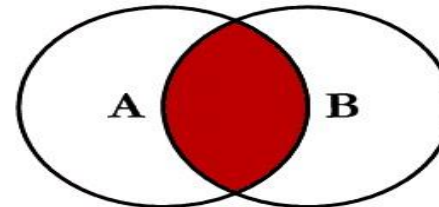
SQL JOINS



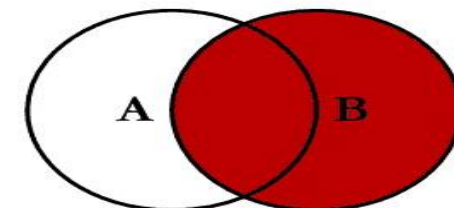
```
SELECT <select_list>
FROM TableA A
LEFT JOIN TableB B
ON A.Key = B.Key
```



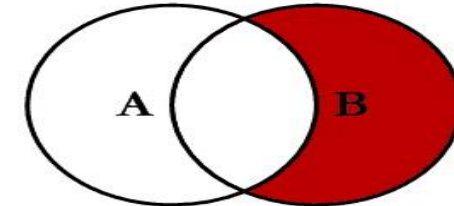
```
SELECT <select_list>
FROM TableA A
LEFT JOIN TableB B
ON A.Key = B.Key
WHERE B.Key IS NULL
```



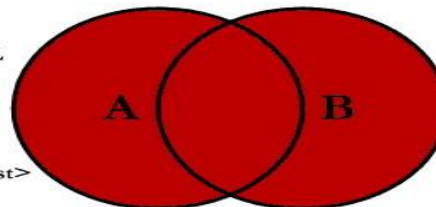
```
SELECT <select_list>
FROM TableA A
INNER JOIN TableB B
ON A.Key = B.Key
```



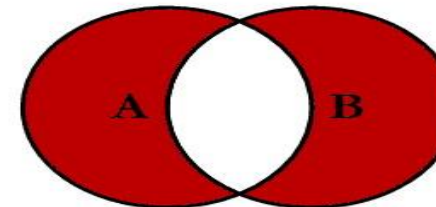
```
SELECT <select_list>
FROM TableA A
RIGHT JOIN TableB B
ON A.Key = B.Key
```



```
SELECT <select_list>
FROM TableA A
RIGHT JOIN TableB B
ON A.Key = B.Key
WHERE A.Key IS NULL
```



```
SELECT <select_list>
FROM TableA A
FULL OUTER JOIN TableB B
ON A.Key = B.Key
```

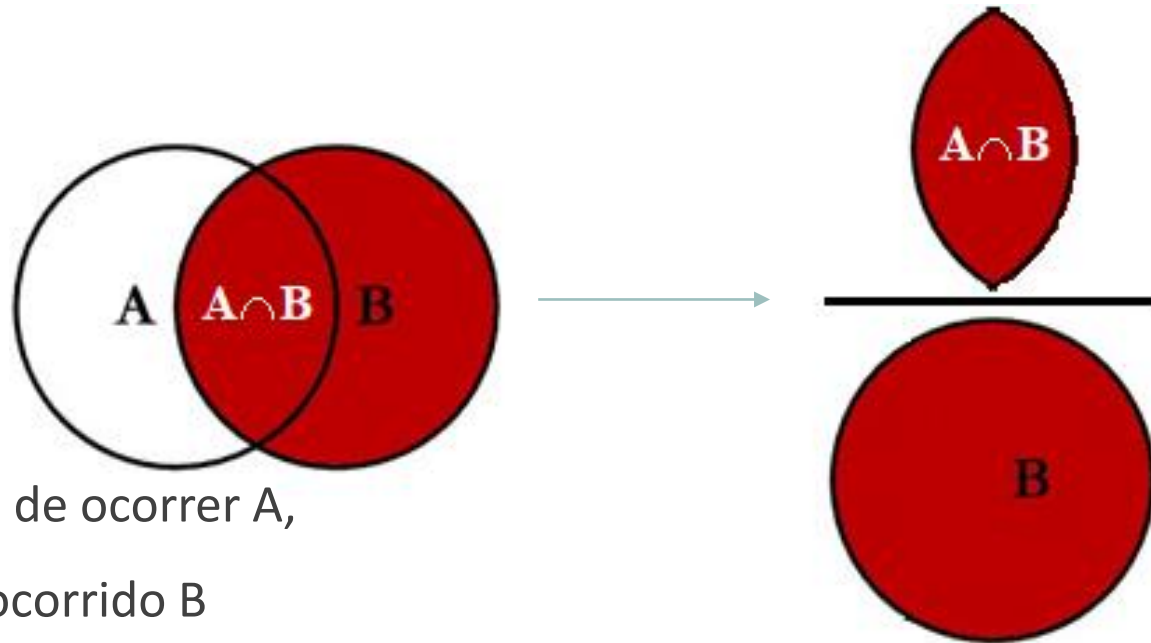


```
SELECT <select_list>
FROM TableA A
FULL OUTER JOIN TableB B
ON A.Key = B.Key
WHERE A.Key IS NULL
OR B.Key IS NULL
```

© C.L. Moffatt, 2008

O que é Naive Bayes?

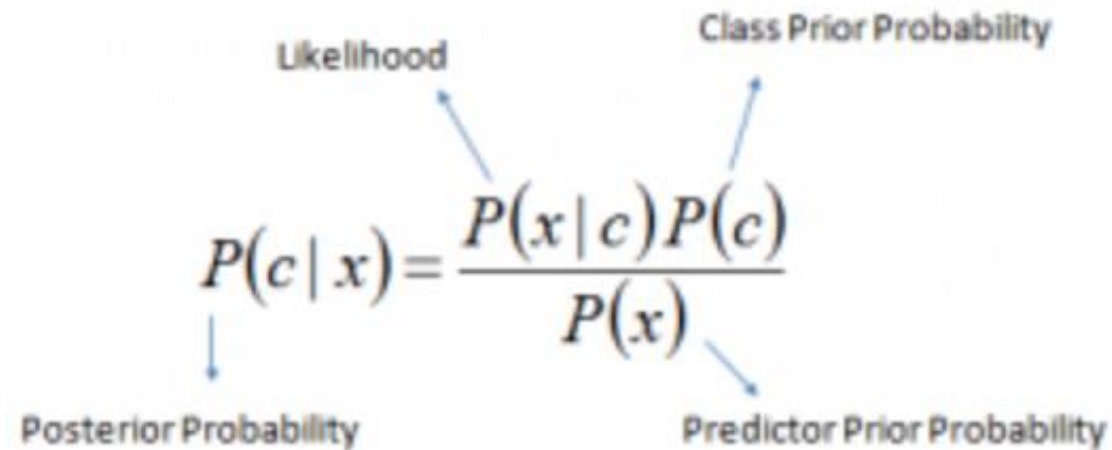
- O modelo Naive Bayes é fácil de construir e particularmente útil para conjuntos de dados muito grandes.



- Probabilidade de ocorrer A,
depois de ter ocorrido B

O que é Naive Bayes?

- O teorema de Bayes fornece uma maneira de calcular a probabilidade posterior $P(c | x)$ de $P(c)$, $P(x)$ e $P(x | c)$. Veja a equação abaixo:



The diagram shows the equation $P(c | x) = \frac{P(x | c)P(c)}{P(x)}$ with four labels and arrows: 'Likelihood' points to $P(x | c)$, 'Class Prior Probability' points to $P(c)$, 'Posterior Probability' points to $P(c | x)$, and 'Predictor Prior Probability' points to $P(x)$.

$$P(c | x) = \frac{P(x | c)P(c)}{P(x)}$$

$$P(c | X) = P(x_1 | c) \times P(x_2 | c) \times \dots \times P(x_n | c) \times P(c)$$

Como?

A partir de um conjunto de dados de treinamento de clima e a variável de segmentação correspondente "Jogar" (sugerindo possibilidades de haver jogo ou não). Agora, precisamos classificar se os jogadores jogarão ou não com base na condição climática.

Weather	Play
Sunny	No
Overcast	Yes
Rainy	Yes
Sunny	Yes
Sunny	Yes
Overcast	Yes
Rainy	No
Rainy	No
Sunny	Yes
Rainy	Yes
Sunny	No
Overcast	Yes
Overcast	Yes
Rainy	No

Frequency Table		
Weather	No	Yes
Overcast		4
Rainy	3	2
Sunny	2	3
Grand Total	5	9

Likelihood table		
Weather	No	Yes
Overcast		4
Rainy	3	2
Sunny	2	3
All	5	9
	=5/14	=9/14
	0.36	0.64

Como Funciona?



Etapa 1: converter o conjunto de dados em uma tabela de frequência;

Passo 2: Criar a tabela de “Probabilidade” (likelihood) encontrando as probabilidades para cada atributo. Tal como a probabilidade de tempo nublado (overcast) = 0,29 e a probabilidade de jogar de 0,64.

Passo 3: Agora, usamos a equação de Bayes para calcular a probabilidade posterior de cada classe.

A classe com a maior probabilidade posterior é o resultado da previsão.

Como Funciona?

Problema: Os jogadores irão jogar se o tempo estiver ensolarado. Esta afirmação está correta?

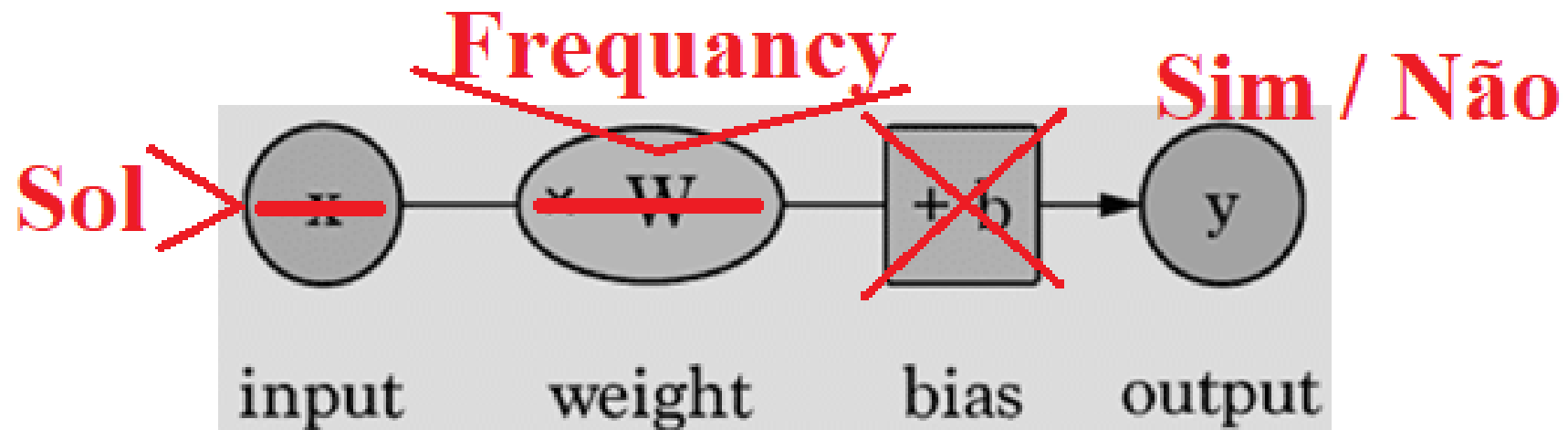
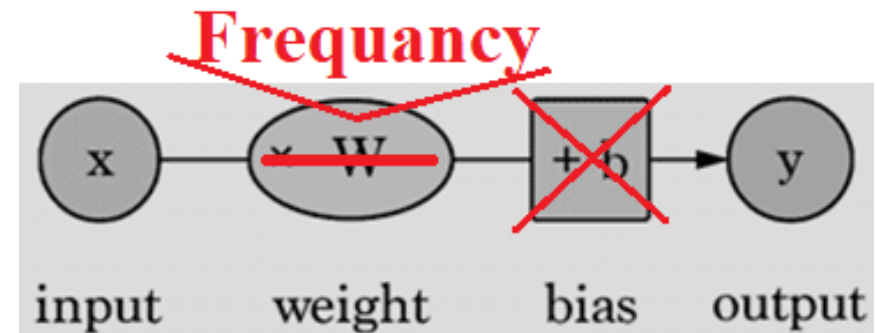
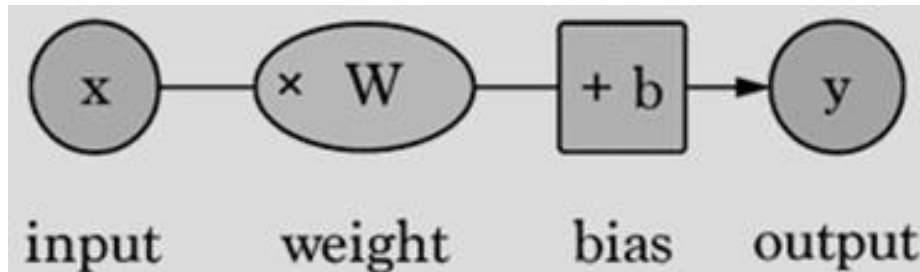
Podemos resolvê-lo usando o método discutido de probabilidade posterior, onde:

- $P(\text{Sim} \mid \text{Sol}) = P(\text{Sol} \mid \text{Sim}) * P(\text{Sim}) / P(\text{Sol})$;
- Aqui temos $P(\text{Ensolarado} \mid \text{Sim}) = 3/9 = 0.33$, $P(\text{Ensolarado}) = 5/14 = 0.36$, $P(\text{Sim}) = 9/14 = 0.64$;
- Agora, $P(\text{Sim} \mid \text{Sol}) = 0,33 * 0,64 / 0,36 = 0,60$, que tem maior probabilidade.

Naive Bayes usa um método similar para prever a probabilidade de classes diferentes baseadas em vários atributos.

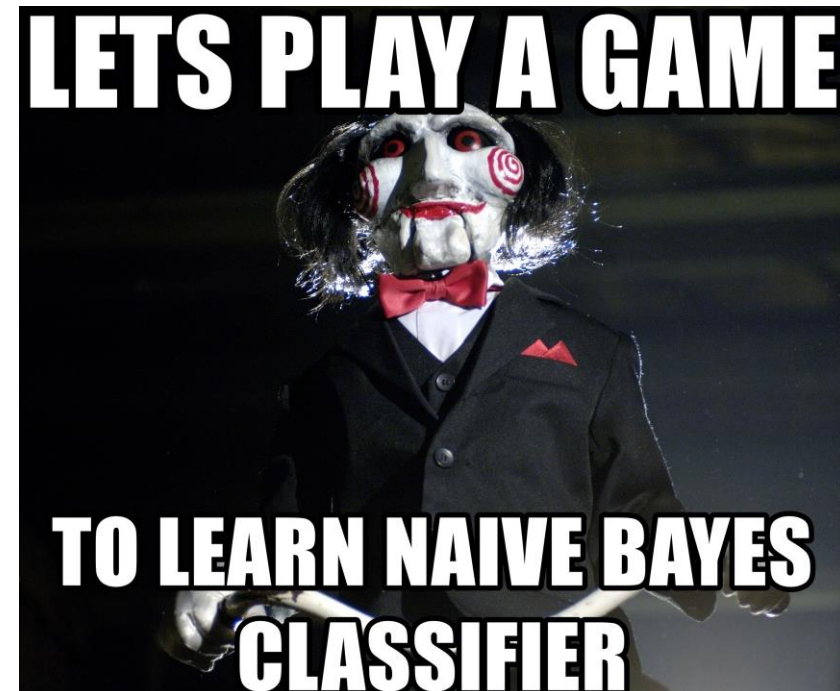
Este algoritmo é usado principalmente na classificação de texto e com problemas com várias classes.

Bayes -> Machine Learning



Passo-a-Passo

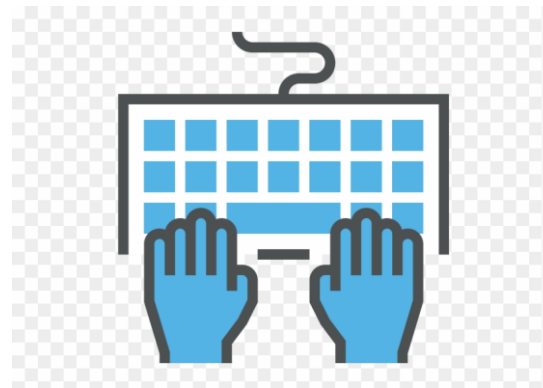
- 1 -> Importar a base
- 2 -> Visualizar
- 3 -> Separar colunas X e Y
- 4 -> Arbitrar Classificação $Y \Rightarrow 0; 1$
- 5 -> Separar linhas Treino e Teste
- 6 -> Aplica Naive Bayes
- 7 -> Calcula acurácia
- ...
- X -> Dominar o mundo



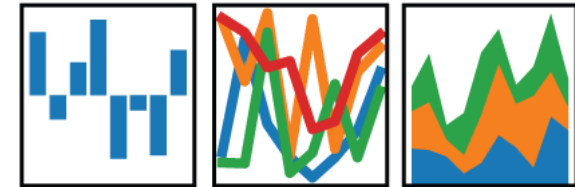
Hands-On



Jupyter Notebook
Python



pandas
 $y_{it} = \beta' x_{it} + \mu_i + \epsilon_{it}$



Prós e Contras

É fácil e rápido prever a classe do conjunto de dados de teste. Ele também funciona bem na previsão de multi classe.

Quando a suposição de independência é válida, um classificador Naive Bayes apresenta uma melhor comparação com outros modelos, como a regressão logística, e você precisa de menos dados de treinamento.

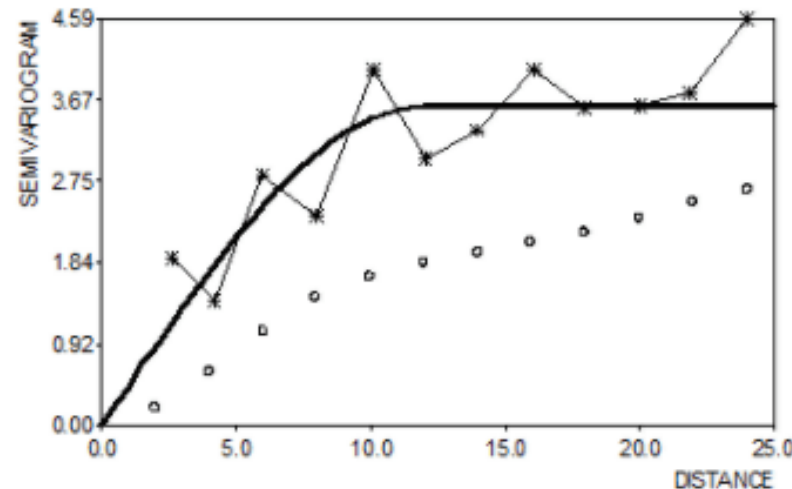
Ele funciona bem no caso de variáveis de entrada categóricas em comparação com variáveis numéricas. Para variáveis numéricas, a distribuição normal é assumida (curva de sino, que é uma forte suposição).



Prós e Contras

Se a variável categórica tiver uma categoria (no conjunto de dados de teste), que não foi observada no conjunto de dados de treinamento, o modelo atribuirá uma probabilidade 0 (zero) e não poderá fazer uma previsão.

Isso é geralmente conhecido como "Frequência Zero". Para resolver isso, podemos usar a técnica de suavização. Uma das técnicas de suavização mais simples é chamada de estimativa de Laplace.



Prós e Contras

Por outro lado, o Naive Bayes também é conhecido como mau estimador, de modo que as saídas de probabilidade do não devem ser levadas muito a sério.

Outra limitação do Naive Bayes é a hipótese de preditores independentes. Na vida real, é quase impossível obtermos um conjunto de preditores completamente independentes.

		Valor Previsto	
		Positivo	Negativo
Valor Verdadeiro	Negativo	Verdadeiros Positivos	Falsos Negativos
	Positivo	Falsos Positivos	Verdadeiros Negativos

Matriz de Confusão ajuda a medir a confiabilidade dos dados e a filtrar pelos que são verdadeiros

Revisão

Mapa Mental

Aplicações

Dúvidas

Feedback...

- O que achou da aula?
- Como foi sua experiencia?
- E os Slides? Agradáveis?



Referências Bibliográficas – Livros

Comece Pelo Porquê – Simon Sinek(2018), ISBN 978-85-431-0663-2



- **Análise Estatística com Excel Para Leigos**– Joseph Schmuller (2010), ISBN 978-85-7608-491-4

Introdução à Ciência de Dados – Fernando Amaral (2016), ISBN 978-85-7608-934-6

Referências Bibliográficas – Vídeos

Matemática Discreta –

- <https://www.youtube.com/watch?v=CPqOCi0ahss>

Matemática Discreta –

- <https://www.youtube.com/watch?v=LGt4PE7-ATI>

School of AI São Paulo –

- <https://www.youtube.com/channel/UCcQgGC19k35ayQNsspyyBhQ>



Referências Bibliográficas – Sites

MSc. José Ahirton Batista Lopes Filho



6 Easy Steps to Learn Naive Bayes Algorithm (with codes in Python and R) –

- <https://www.analyticsvidhya.com/blog/2017/09/naive-bayes-explained/>

Gaussian Naive Bayes –

- <https://medium.com/@LSchulteBraucks/gaussian-naive-bayes-19156306079b>

How Naive Bayes Algorithm Works? –

- <https://www.machinelearningplus.com/predictive-modeling/how-naive-bayes-algorithm-works-with-example-and-full-code/>

Obrigado!

Att,

Victor Venites.

LinkedIn: <http://victorvenites.com/>

E-mail: contato@victorvenites.com

