

# Métodos Simples

---

Carlos Diego Rodrigues

19 de outubro de 2021

Universidade Federal do Ceará

# Métodos simples: como construir um método de aprendizado?

- O que precisamos para fazer um classificador?
- Métodos baseados na tabela de entrada:
  - ZeroR
  - OneR
  - Naive Bayes
  - Árvore de decisão

# O que precisamos para fazer um método de aprendizado?

- É um algoritmo: deve ser um conjunto bem estruturado de passos.
- Para ser útil deve permitir:
  - Atributos qualitativos e quantitativos
  - Valores faltantes
  - Ser robusto, suportar ruídos
- Vamos ver alguns esquemas básicos capazes de fazer isto, baseados na tabela de entrada do problema.

# Tabela de entrada


Atributos				Classe
Outlook	Temp.	Humidity	Windy	Play Golf
Rainy	Hot	High	False	No
Rainy	Hot	High	True	No
Overcast	Hot	High	False	Yes
Sunny	Mild	High	False	Yes
Sunny	Cool	Normal	False	Yes
Sunny	Cool	Normal	True	No
Overcast	Cool	Normal	True	Yes
Rainy	Mild	High	False	No
Rainy	Cool	Normal	False	Yes
Sunny	Mild	Normal	False	Yes
Rainy	Mild	Normal	True	Yes
Overcast	Mild	High	True	Yes
Overcast	Hot	Normal	False	Yes
Sunny	Mild	High	True	No

- É o apelido do método sem regras: sem contribuições dos atributos.
- É o mais simples método de classificação.
- Ignora todos os atributos e avalia apenas a classe.
- Consiste em escolher apenas uma resposta para qualquer entrada.
- A partir dos dados de treinamento escolhe a resposta com a maior frequência.

- É o apelido do método com apenas uma regra.
- Simples, porém preciso.
- Para cada atributo ele vai criar uma regra preditiva para a classe.
- Escolhe o atributo cuja regra tem o menor erro.

- Para cada atributo
  - Para cada valor possível do atributo, faça uma regra da seguinte forma:
    - Conte quantas vezes cada classe aparece.
    - Escolha a classe mais frequente.
    - Crie uma regra associando o valor deste atributo à classe.
  - Calcule o erro total produzido pelas regras deste atributo
- Escolha o atributo com o menor erro total.

# Algoritmo OneR

		Play Golf	
		Yes	No
Outlook	Sunny	3	2
	Overcast	4	0
	Rainy	2	3

		Play Golf	
		Yes	No
Temp.	Hot	2	2
	Mild	4	2
	Cool	3	1

		Play Golf	
		Yes	No
Humidity	High	3	4
	Normal	6	1

		Play Golf	
		Yes	No
Windy	False	6	2
	True	3	3



- É um classificador baseado no Teorema de Bayes com suposição de independência entre os atributos.
- Fácil de construir, sem parametrização e adequado para grandes conjuntos de dados.
- Muito utilizado historicamente.
- Pode ser mais eficiente do que métodos mais sofisticados.

# Teorema de Bayes

- Provê uma maneira de calcular probabilidades a posteriori  $P(c|x)$  a partir de outras probabilidades  $P(c)$ ,  $P(x)$  e  $P(x|c)$  que podem ser encontradas na tabela.

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

- No algoritmo *OneR* nós escolhemos apenas um atributo. Já no *Naive Bayes* nós escolhemos todos os atributos.

# Exemplo com um atributo

Frequency Table		Play Golf	
		Yes	No
Outlook	Sunny	3	2
	Overcast	4	0
	Rainy	2	3



Likelihood Table		Play Golf		
		Yes	No	
Outlook	Sunny	3/9	2/5	5/14
	Overcast	4/9	0/5	4/14
	Rainy	2/9	3/5	5/14
		9/14	5/14	

$$P(x | c) = P(\text{Sunny} | \text{Yes}) = 3 / 9 = 0.33$$

$$P(c) = P(\text{Yes}) = 9 / 14 = 0.64$$

$$P(x) = P(\text{Sunny}) \\ = 5 / 14 = 0.36$$

# Tabelas de frequência

		Play Golf	
		Yes	No
Outlook	Sunny	3	2
	Overcast	4	0
	Rainy	2	3



		Play Golf	
		Yes	No
Outlook	Sunny	3/9	2/5
	Overcast	4/9	0/5
	Rainy	2/9	3/5

		Play Golf	
		Yes	No
Humidity	High	3	4
	Normal	6	1



		Play Golf	
		Yes	No
Humidity	High	3/9	4/5
	Normal	6/9	1/5

		Play Golf	
		Yes	No
Temp.	Hot	2	2
	Mild	4	2
	Cool	3	1



		Play Golf	
		Yes	No
Temp.	Hot	2/9	2/5
	Mild	4/9	2/5
	Cool	3/9	1/5

		Play Golf	
		Yes	No
Windy	False	6	2
	True	3	3



		Play Golf	
		Yes	No
Windy	False	6/9	2/5
	True	3/9	3/5

## Fazendo um exemplo

- Prever o exemplo: {Rainy, Cool, High, True}

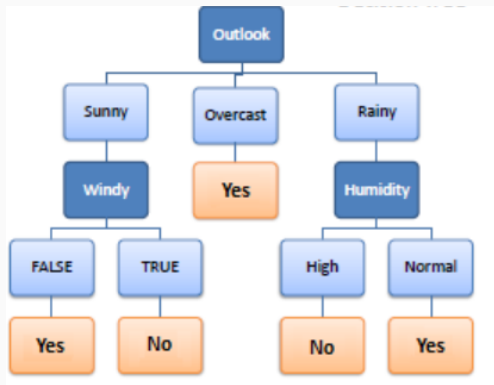
## Dois problemas

- O que fazer com os zeros na tabela?
  - O valor zero é problemático porque ele pode funcionar como um veto na natureza multiplicativa do *Naive Bayes*.
  - Basta adicionar o valor um a todas as frequências da tabela.
- O que fazer quando os dados são numéricos?
  - Dados numéricos não são imediatamente tratados pelo métodos, mas duas saídas são possíveis.
  - Construir uma distribuição de frequências, categorizar.
  - Adotar uma distribuição de probabilidade para os dados numéricos, tipicamente uma distribuição normal, calculando média e desvio padrão e então associando os valores numéricos à probabilidades à partir de uma função conhecida.

# Árvore de decisão

- Árvores de decisão criam modelos em uma estrutura de árvore: o problema original é quebrado em partes menores até encontrarmos um tamanho trivial.
- Árvores são compostas de nós de decisão e nós folhas.
- Os nós de decisão ramificam a árvore em dois ou mais galhos, onde em cada um há um conjunto menor de dados a serem considerados.
- As folhas têm idealmente uma única classe ou uma distribuição de fácil classificação.
- O nó inicial é chamado nó raiz.

## Exemplo de árvore de decisão





## Algoritmo ID3

- O algoritmo ID3 constrói uma divisão iterativa dos dados onde em cada nó de decisão ele escolhe um único a ser ramificado.
- São criados ramos para cada valor possível do atributo escolhido.
- O atributo é escolhido a partir dos conceitos de entropia e ganho de informação.
- Os atributos são utilizados conforme possam ser úteis à classificação.

# Entropia

O que é entropia?

- É uma medida de aleatoriedade da informação processada.
- É uma forma de quantificar similaridades e diferenças.

Surpresa

$$\text{surpresa}(X) = \log_2 \left( \frac{1}{P(X)} \right)$$

Entropia é o valor esperado da surpresa

$$E(X) = \sum_{i=1}^n p_i \log_2 \left( \frac{1}{p_i} \right)$$

- A entropia em uma tabela de um único atributo é definida como:

$$E(S) = \sum_{i=1}^n -p_i \log_2 p_i$$

- Em uma tabela de dois atributos define-se como:

$$E(T, X) = \sum_{x \in X} P(x) E(x)$$

- Ganho de informação é definido como:

$$G(T, X) = E(T) - E(T, X)$$

# Ganhos de informação da tabela original

		Play Golf	
		Yes	No
Outlook	Sunny	3	2
	Overcast	4	0
	Rainy	2	3
Gain = 0.247			

		Play Golf	
		Yes	No
Temp.	Hot	2	2
	Mild	4	2
	Cool	3	1
Gain = 0.029			

		Play Golf	
		Yes	No
Humidity	High	3	4
	Normal	6	1
Gain = 0.152			

		Play Golf	
		Yes	No
Windy	False	6	2
	True	3	3
Gain = 0.048			