UNIVERSIDADE EDUARDO MONDLANE

FACULDADE DE ENGENHARIA

DEPARTAMENTO DE ENGENHARIA ELECTROTÉCNICA

Inteligência Artificial

Classificação em Mineração de Dados

Docentes: Eng Roxan Cadir
Eng Ruben Manhiça

Maputo, 2 de abril de 2024

Conteúdo da Aula

1. Tipos de Problemas em Mineração de dados
2. Classificação;
3. Algoritmo ID3;
4. Teoria da Informação
5. Calculo da Entropia do ID3

Tipos de Problemas em Mineração de Dados

TAREFA	DESCRIÇÃO	EXEMPLOS
Classificação	Constrói um modelo de algum tipo que possa ser aplicado a dados não classificados a fim de categorizá -los em classes	?? Classificar pedidos de crédito ?? Esclarecer pedidos de seguros fraudulentos ?? Identificar a melhor forma de tratamento de um paciente
Estimativa (ou Regressão)	Usada para definir um valor para alguma variável contínua desconhecida	?? Estimar o número de filhos ou a renda total de uma família ?? Estimar o valor em tempo de vida de um cliente ?? Estimar a probabilidade de que um paciente morrerá baseando-se nos resultados de diagnósticos médicos ?? Prever a demanda de um consumidor para um novo produto
Associação	Usada para determinar quais itens tendem a co-ocorrerem (serem adquiridos juntos) em uma mesma transação	?? Determinar quais os produtos costumam ser colocados juntos em um carrinho de supermercado
Segmentação (ou *Clustering*)	Processo de partição de uma população heterogênea em vários subgrupos ou grupos mais homogêneos	?? Agrupar clientes por região do país ?? Agrupar clientes com comportamento de compra similar ?? Agrupar seções de usuários Web para prever comportamento futuro de usuário
Sumarização	Envolve métodos para encontrar uma descrição compacta para um subconjunto de dados	?? Tabular o significado e desvios padrão para todos os itens de dados ?? Derivar regras de síntese

Classificação: Introdução

Classificação: é uma das técnicas mais utilizadas na mineração, por exemplo são comuns as tarefas de classificação de clientes em baixo, médio ou alto risco de empréstimo bancário.

“ Classificar um objecto é determinar com que grupo de entidades, já classificadas anteriormente, esse objecto apresenta mais semelhanças ”

Classificação: Definição

Definição: É o processo pelo qual examinamos as propriedades (aspectos, estrutura) de um objeto (dados) e atribuí-lo a uma das **classes predefinidas**.

É aprender o mapeamento de uma função dos objetos em uma das **classes predefinidas**.

Classificação: objetivos

Objetivo: analisar os dados de entrada (treino) e desenvolver uma **descrição** ou **modelo** para cada classe utilizando as estruturas presentes nos objetos.

Usar o relacionamento descoberto para prever a classe (o valor do atributo meta) de um registro com classe desconhecida.

Classificação: tarefas

Tarefa: descobrir um **relacionamento** entre os **atributos previsores** e o **atributo meta**, usando registros cuja **classe é conhecida**, para se **construir um modelo** de algum tipo que possa ser aplicado aos objectos não classificados para classificá-los.

Árvores de Decisão

É um **método de aprendizagem supervisionado** que constrói árvores de classificação a partir de exemplos.

Algoritmos : ID3, C4.5, (Quinlan)

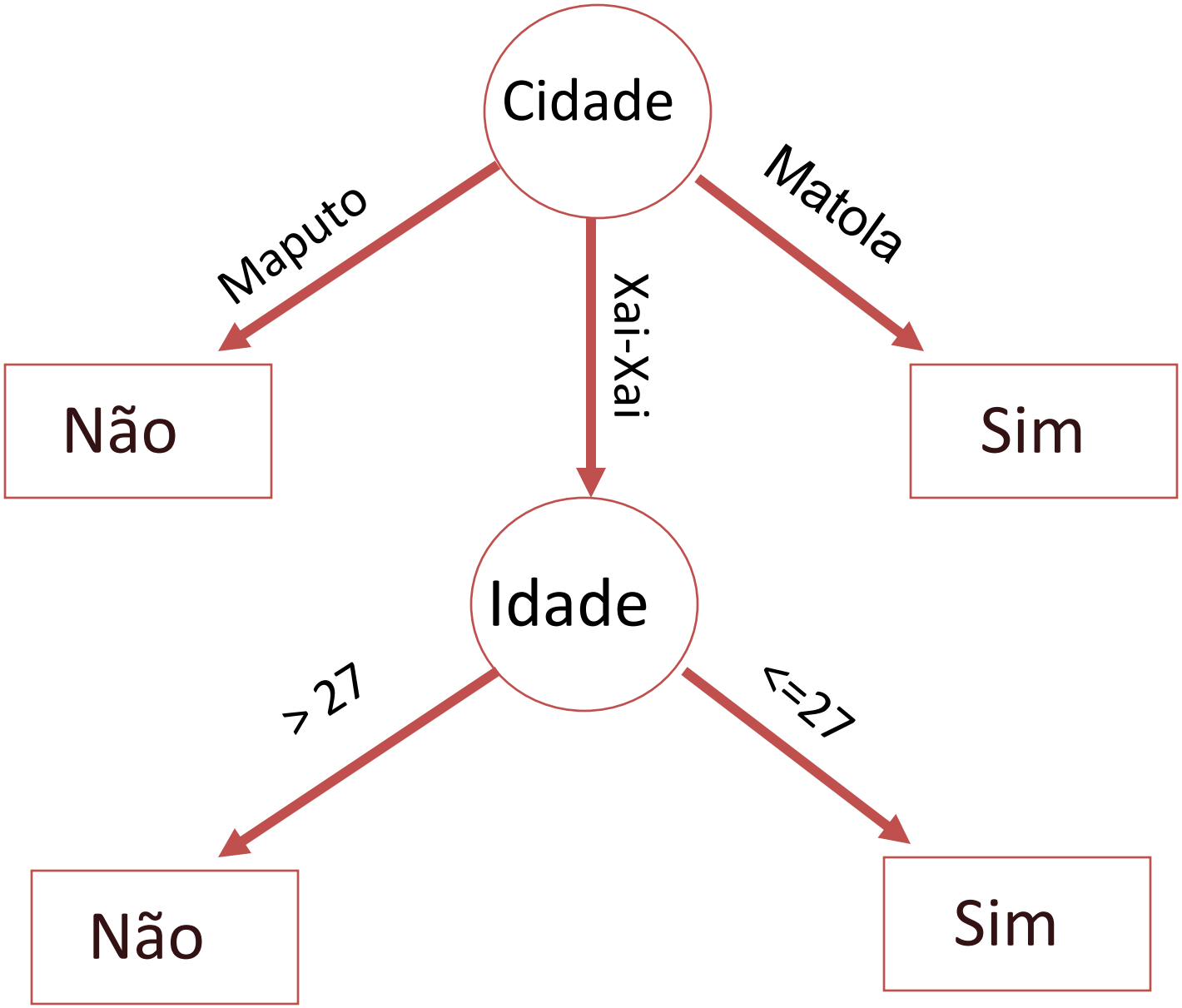
CART (Breiman)

Ross Quinlan

Leo Breiman

Árvores de Decisão: Exemplo

ID	Sexo	Cidade	Idade	Comprar ?
1	M	Xai-Xai	25	S
2	M	Matola	21	S
3	F	Xai-Xai	23	S
4	F	Matola	34	S
5	F	Xai-Xai	30	N
6	M	Maputo	21	N
7	M	Maputo	20	N
8	F	Maputo	18	N
9	F	Xai-Xai	34	N
10	M	Xai-Xai	55	N


Cidade
Maputo
Xai-Xai
Matola
Não
Sim
Idade
> 27
<=27
Não
Sim

Regras:

Se (Cidade=Maputo) Então (Decisão = Não)

Se (Cidade=Matola) Então (Decisão = Sim)

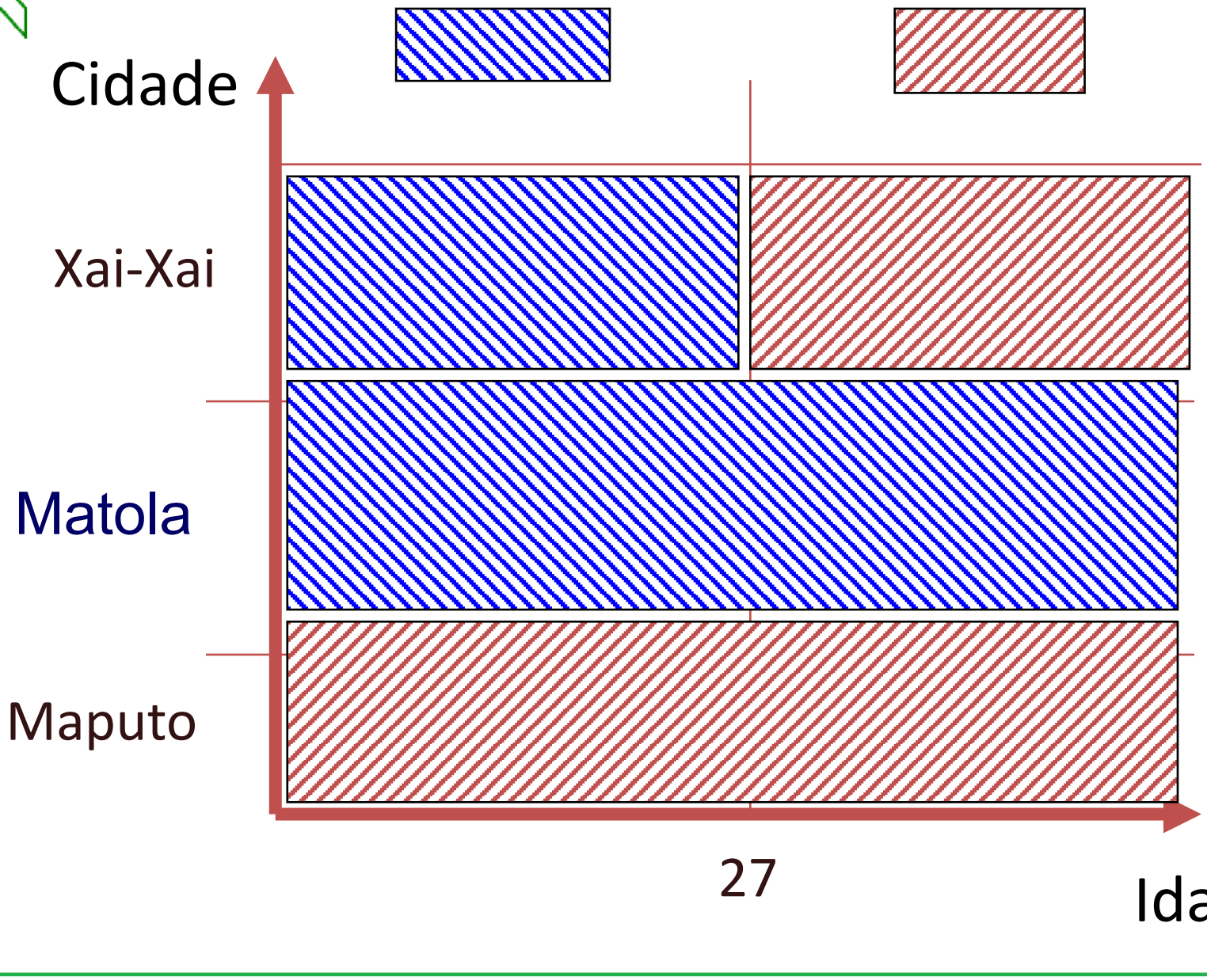
Se (Cidade=Xai-Xai e Idade <= 27) Então
(Decisão = Sim)

Se (Cidade=Xai-Xai e Idade > 27) Então
(Decisão = Não)

Árvores de Decisão

Os métodos baseados em árvores para classificação, dividem o espaço de entrada em **regiões disjuntas** para construir uma **fronteira de decisão**.

As regiões são escolhidas baseadas em uma **optimização heurística** onde a cada passo os algoritmos selecionam a variável que provê a **melhor separação de classes** de **acordo alguma função custo.**


Cidade
Xai-Xai
Matola
Maputo
27
Idade

Algoritmo ID3

Algoritmo ID3

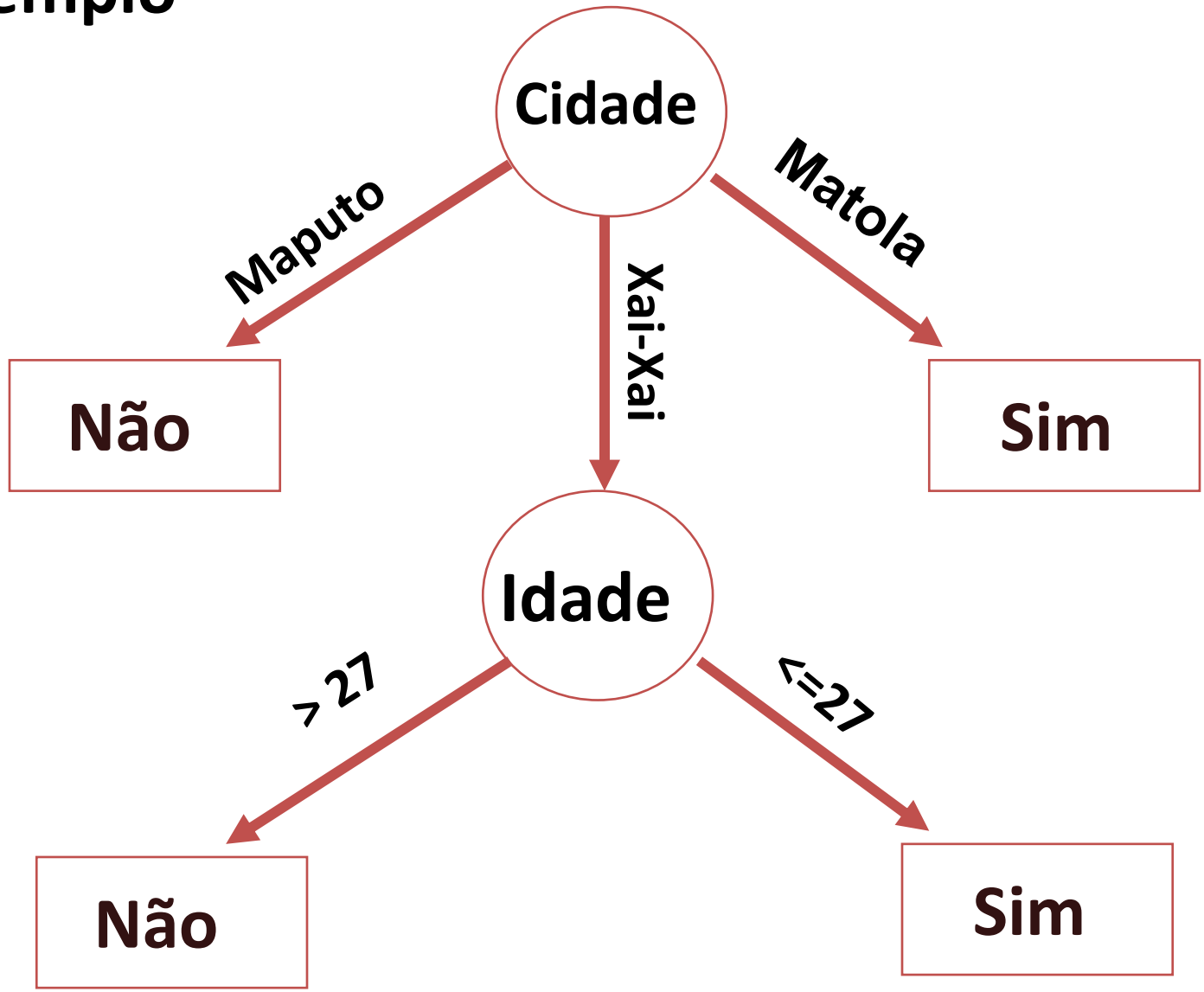
ID3, é um algoritmo simples que constrói uma **árvore de decisão** sob as seguintes premissas:

Cada **vértice** (nodo) corresponde a um **atributo**, e cada **aresta** da árvore a um **valor possível** do atributo.

Uma **folha** da árvore corresponde ao valor esperado da **decisão** segundo os dados de treino utilizados.

A **explicação** de uma determinada decisão está na **trajetória** da raiz a folha representativa desta decisão.

Exemplo



Algoritmo ID3

Cada **vértice** é **associado** ao **atributo** mais **informativo** que ainda não tenha sido considerado.

Para medir o **nível de informação** de um atributo se utiliza o conceito de **entropia da Teoria da Informação.**

Menor o valor da entropia, menor a incerteza e mais utilidade tem o atributo para a classificação.

Algoritmo ID3: Exemplo

Dia	Aspecto	Temperatura	Umidade	Vento	Decisão
1	Sol	Quente	Alta	Fraco	N
2	Sol	Quente	Alta	Forte	N
3	Nublado	Quente	Alta	Fraco	S
4	Chuva	Agradável	Alta	Fraco	S
5	Chuva	Fria	Normal	Fraco	S
6	Chuva	Fria	Normal	Forte	N
7	Nublado	Fria	Normal	Forte	S
8	Sol	Agradável	Alta	Fraco	N
9	Sol	Fria	Normal	Fraco	S
10	Chuva	Agradável	Normal	Fraco	S
11	Sol	Agradável	Normal	Forte	S
12	Nublado	Agradável	Alta	Forte	S
13	Nublado	Quente	Normal	Fraco	S
14	Chuva	Agradável	Alta	Forte	N

Algoritmo ID3: Exemplo

O número de combinações possíveis são:

Aspecto: sol, nublado, chuva

Temperatura: quente, agradável, frio

Umidade: alta, normal

Vento: fraco, forte

(3 x 3 x 2 x 2 = 36)

Algoritmo ID3

Seleção de atributos para construir a árvore de decisão.

Qual é o atributo previsor mais relevante para prever a classe a qual pertencem os dados ?

Teoria da Informação

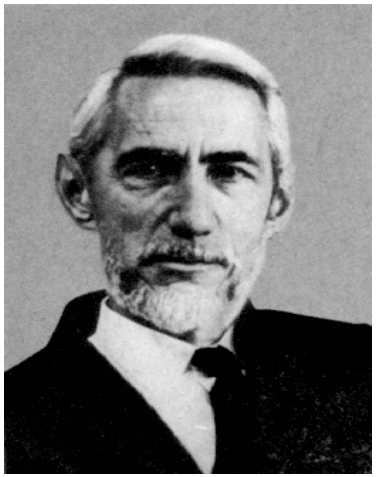
Dada uma distribuição de probabilidade

P = (p_1, p_2, . . . , p_n), a informação contida nesta distribuição, é chamada de **entropia** (função de informação de Shannon), e definida como:

$I(P) = -[p_1 * \log_2 (p_1) + . . . + p_n * \log_2 (p_n)]$

EntropiaEsperada(Decisão) = $-\sum_{i=1}^{n} p_i \log_2 (p_i)$

Observação: $\log_a N = \log N / \log a$



Claude Shannon

Teoria da Informação

Exemplos:

P = (0.5;0.5) ➔ I(P) = 1;

P = (0.67;0.33) ➔ I(P) = 0.92;

P = (1.0;0.0) ➔ I(P) = 0.0;

Observação: quanto mais uniforme é a distribuição de probabilidade, maior é a entropia e portanto maior a incerteza ou menor a informação.

Teoria da Informação

Se temos um conjunto **T** de registros, particionados em **k** classes **(C_1,. . . ,C_k),** a informação necessária para identificar a classe de um elemento de **T** é

$$\mathbf{Info(T) = I(P)}$$

onde **P** é a distribuição de probabilidade das classes **(C_1 , . . . , C_k)**

$$\mathbf{P = (C_1 / T, . . . , C_k / T)}$$

Isto é, a proporção de elementos pertencentes a classe **i.**

Algoritmo ID3: Exemplo

Dia	Aspecto	Temperatura	Umidade	Vento	Decisão
1	Sol	Quente	Alta	Fraco	N
2	Sol	Quente	Alta	Forte	N
3	Nublado	Quente	Alta	Fraco	S
4	Chuva	Agradável	Alta	Fraco	S
5	Chuva	Fria	Normal	Fraco	S
6	Chuva	Fria	Normal	Forte	N
7	Nublado	Fria	Normal	Forte	S
8	Sol	Agradável	Alta	Fraco	N
9	Sol	Fria	Normal	Fraco	S
10	Chuva	Agradável	Normal	Fraco	S
11	Sol	Agradável	Normal	Forte	S
12	Nublado	Agradável	Alta	Forte	S
13	Nublado	Quente	Normal	Fraco	S
14	Chuva	Agradável	Alta	Forte	N

Algoritmo ID3: Exemplo

Conjunto de registros **T** com 14 observações.
Duas partições **S** e **N**, com probabilidade **9/14** e **5/14**.

Info(T) = I(9/14, 5/14) =
$-(9/14 \cdot \log_2 (9/14) + 5/14 \cdot \log_2 (5/14)) =$
0.94

Teoria da Informação

Se particionamos ***T*** sobre a base dos valores do atributo **X** em conjuntos $t_1, \ldots t_n$, então a informação necessária para identificar a classe de um elemento de **T**, é:

$$\textbf{Info (X,T)} = \sum_i (t_i;T) * Info(t_i)$$

, onde t_i é o conjunto de possíveis valores do atributo **X**.

Algoritmo ID3: Exemplo

Info (Aspecto, T) = sol/T * I (sol) +
nublado/T * I (nublado) +
chuva/T * I (chuva) =

Aspecto	F_S	F_N
Sol	2/5	3/5
Nublado	4/4	0/4
Chuva	3/5	2/5

Info (Aspecto, T) =
5/14 * I (2/5, 3/5) +
4/14 * I (4/4 , 0) +
5/14 * I (3/5,2/5) = 0.693

Teoria da Informação

Definição: o ganho de informação do atributo **X**, é a diferença entre a informação necessária para identificar um elemento de **T** e a informação necessária para identificar um elemento de **T** depois que o valor de atributo **X** tenha sido considerado:

Ganho (X,T) = Info (T) - Info (X, T) ou

Ganho (X,T) = (Entropia Esperada) – (Entropia Real)

Info(T) = 0.94 (Entropia Esperada)

Info(Aspecto, T) = 0.693 (Entropia Real)

Ganho (Aspecto, T) = 0.940 - 0.693 = 0.247

Teoria da Informação

Com o objetivo de criar árvores de decisão pequenas, para identificar poucas regras, o atributo escolhido para nó da árvore é o **atributo de maior ganho.**

Algoritmo ID3

Passo 1:

Se todos os dados estão classificados em alguma das classes **então** parar ;

senão

selecionar (utilizando alguma heurística) algum atributo **A** com valores $v_1, v_2, \ldots, v_n$ e criar um nó de decisão.

Passo 2 : particionar o conjunto de dados de treino **T**, em subconjuntos $t_1, t_2, \ldots, t_n$ de acordo com os valores do atributo **A**

Passo 3 : aplicar o algoritmo recursivamente para cada conjunto de dados **ti**

Algoritmo ID3: Exemplo

Aspecto	F_S	F_N
Sol	2/5	3/5
Nublado	4/4	0/4
Chuva	3/5	2/5

Info (Sol) = 0.971

Info (Nublado) = 0.0

Info (Chuva) = 0.971

Info (Aspecto) = 0.693

(5/14 * 0.971 + 4/14 * 0.0 + 5/14 * 0.971)

Ganho (Aspecto) = 0.940 - 0.693 = 0.247

Algoritmo ID3: Exemplo

Temperatura	F_S	F_N
Quente	2/4	2/4
Agradável	4/6	2/6
Fria	3/4	1/4

Info (Quente) = 1.0
Info (Agradável) = 0.919
Info (Fria) = 0.811

Info (Temperatura) = 0.911
(4/14 * 1.0 + 6/14 * 0.919 + 4/14 * 0.811)
Ganho (Temperatura) = 0.940 - 0.911 = 0.029

Algoritmo ID3: Exemplo

Umidade	F_S	F_N
Alta	3/7	4/7
Normal	6/7	1/7

Info (Alta) = 0.984

Info (Normal) = 0.592

Info (Umidade) = 0.788

(7/14 * 0.984 + 7/14 * 0.592)

Ganho (Umidade) = 0.940 - 0.788 = 0.152

Algoritmo ID3: Exemplo

Vento	F_S	F_N
Fraco	6/8	2/8
Forte	3/6	3/6

Info (Forte) = 1.0

Info (Fraco) = 0.811

Info (Vento) = 0.892

(6/14 * 1.0 + 8/14 * 0.541)

Ganho (Vento) = 0.940 - 0.892 = 0.048

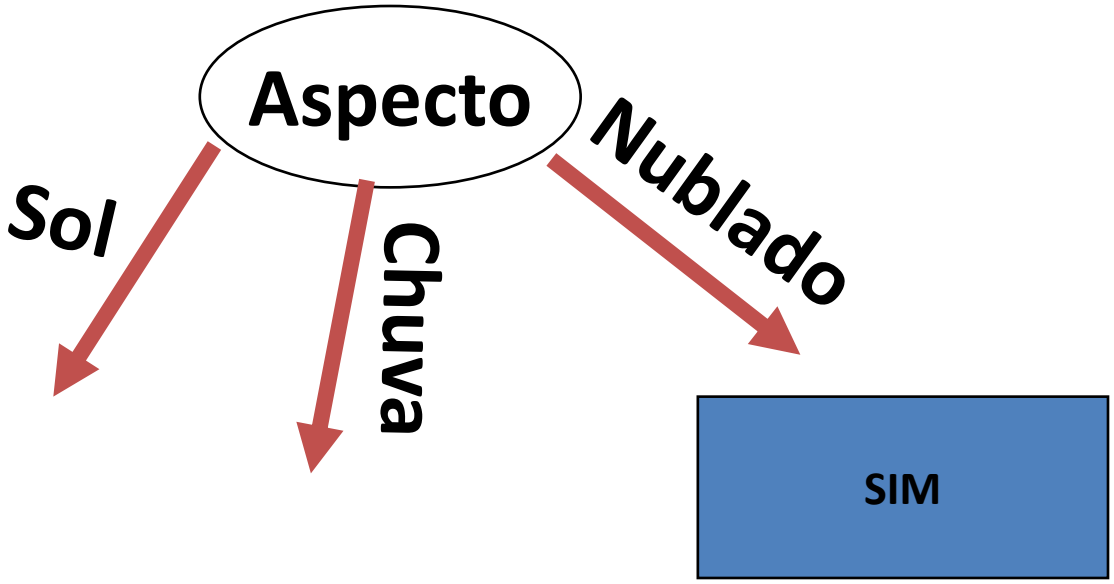
Algoritmo ID3: Exemplo

Ganho (Aspecto) = 0.247

Ganho (Temperatura) = 0.028

Ganho (Umidade) = 0.152

Ganho (Vento) = 0.048



Algoritmo ID3: Exemplo

Dia	Aspecto	Temperatura	Umidade	Vento	Decisão
1	Sol	Quente	Alta	Fraco	N
2	Sol	Quente	Alta	Forte	N
3	Nublado	Quente	Alta	Fraco	S
4	Chuva	Agradável	Alta	Fraco	S
5	Chuva	Fria	Normal	Fraco	S
6	Chuva	Fria	Normal	Forte	N
7	Nublado	Fria	Normal	Forte	S
8	Sol	Agradável	Alta	Fraco	N
9	Sol	Fria	Normal	Fraco	S
10	Chuva	Agradável	Normal	Fraco	S
11	Sol	Agradável	Normal	Forte	S
12	Nublado	Agradável	Alta	Forte	S
13	Nublado	Quente	Normal	Fraco	S
14	Chuva	Agradável	Alta	Forte	N

Algoritmo ID3 : Exemplo

Escolhemos Aspecto = Sol

Info (T) = I (2/5; 3/5) = 0.971

Temperatura

Info (Quente) = I (0/2, 2/2) = 0.0
Info (Agradável) = I (1/2, 1/2) = 1.0
Info (Fria) = I (1/1, 0/1) = 0.0

Info(Temperatura) = 0.4

(2/5 * 0.0 + 2/5 * 1.0 + 1/5 * 0.0)

Ganho (Temperatura) = 0.971 - 0.4 = 0.571

Algoritmo ID3 : Exemplo

Umidade

Info (Alta) = I (0/3, 3/3) = 0.0

Info (Normal) = I (2/2, 0/2) = 0.0

Info(Umidade) = 0.0

(3/5 * 0.0 + 2/5 * 0.0 = 0.0)

Ganho (Umidade) = 0.971 - 0.0 = 0.971

Algoritmo ID3 : Exemplo

Vento

Info (Fraco) = I (1/3, 2/3) = 0.919

Info (Forte) = I (1/2, 1/2) = 1.0

Info(Vento) = 0.951

(3/5 * 0.919 + 2/5 *1.0) = 0.951

Ganho (Vento) = 0.971 - 0.951 = 0.020

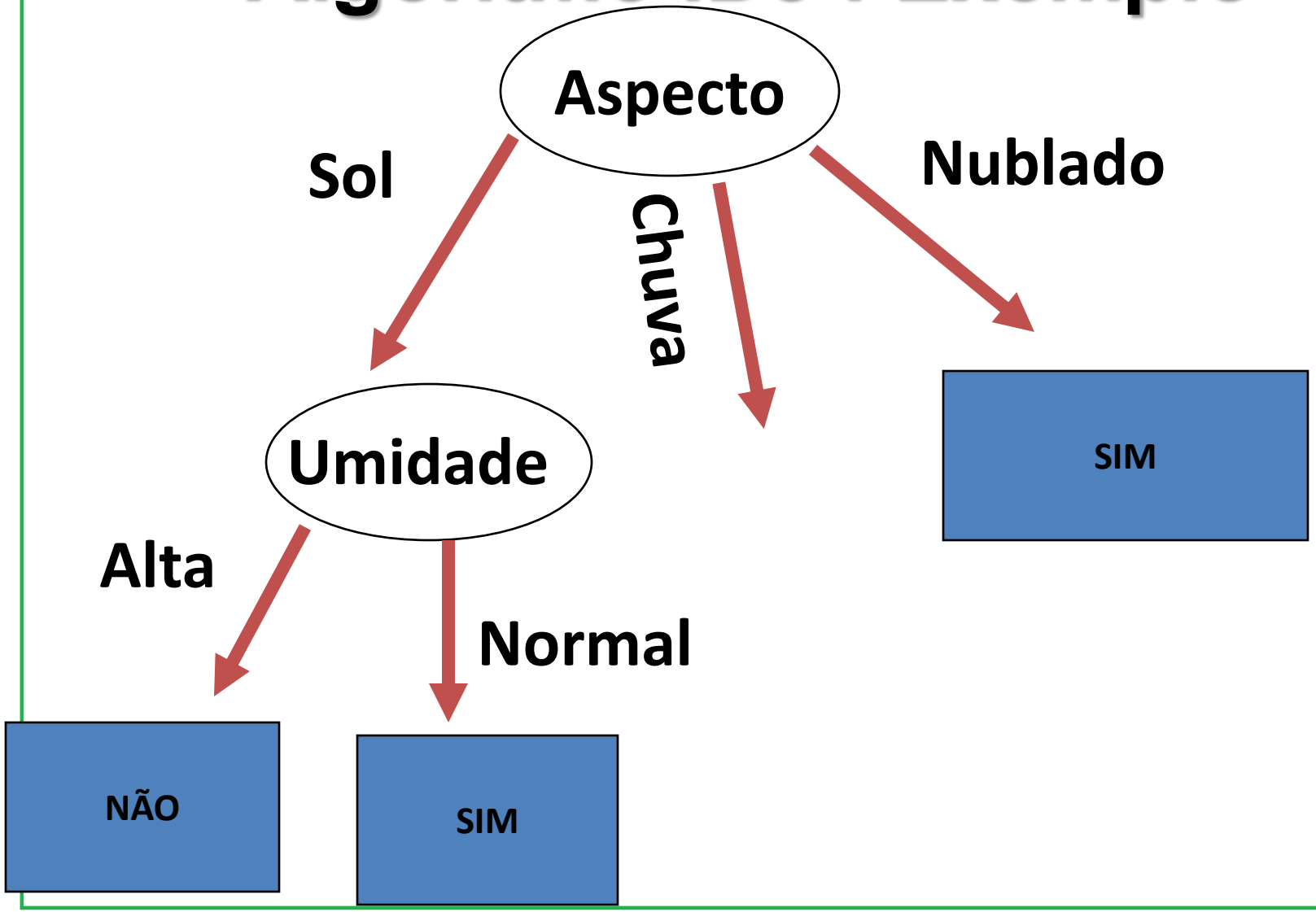
Algoritmo ID3 : Exemplo

Ganho (Temperatura) = 0.571

Ganho (Umidade) = 0.971

Ganho (Vento) = 0.020

Algoritmo ID3 : Exemplo



Algoritmo ID3 : Exemplo

Dia	Aspecto	Temperatura	Umidade	Vento	Decisão
1	Sol	Quente	Alta	Fraco	N
2	Sol	Quente	Alta	Forte	N
3	Nublado	Quente	Alta	Fraco	S
4	Chuva	Agradável	Alta	Fraco	S
5	Chuva	Fria	Normal	Fraco	S
6	Chuva	Fria	Normal	Forte	N
7	Nublado	Fria	Normal	Forte	S
8	Sol	Agradável	Alta	Fraco	N
9	Sol	Fria	Normal	Fraco	S
10	Chuva	Agradável	Normal	Fraco	S
11	Sol	Agradável	Normal	Forte	S
12	Nublado	Agradável	Alta	Forte	S
13	Nublado	Quente	Normal	Fraco	S
14	Chuva	Agradável	Alta	Forte	N

Algoritmo ID3 : Exemplo

Escolhemos Aspecto = Chuva

Info (T) = I (3/5; 2/5) = 0.971

Temperatura
Info (Quente) = I (0/0, 0/0) = 0;
Info (Agradável) = I (2/3, 1/3) = 0.919
Info (Fria) = I (1/2, 1/2) = 1.0

Info(Temperatura) = 0.951
3/5 * 0.919 + 2/5 * 1.0 = 0.951

Ganho (Temperatura) = 0.971 - 0.951 = 0.020

Algoritmo ID3 : Exemplo

Vento

Info (Fraco) = I (3/3, 0/3) = 0.0

Info (Forte) = I (0/2, 2/2) = 0.0

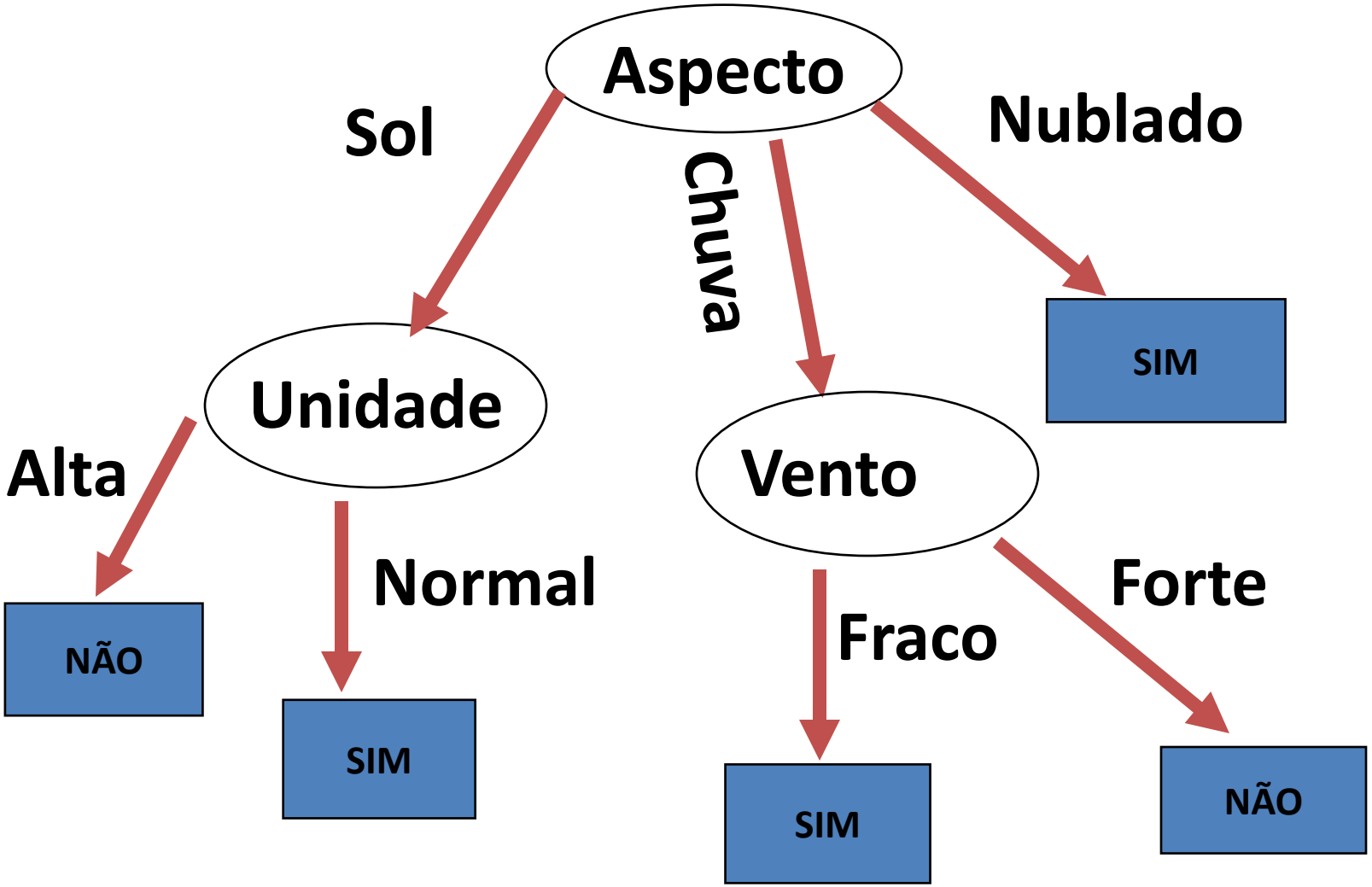
Info(Vento) = 0.0

(3/5 * 0.0 + 2/5 *0. 0)

Ganho (Vento) = 0.971- 0.0 = 0.971

Ganho (Temperatura) = 0.020

Algoritmo ID3 : Exemplo



Algoritmo ID3 : Exemplo

Se Aspecto = **Sol** e Umidade = **Alta**

Então Jogar = **Não**

Se Aspecto = **Sol** e Umidade = **Normal**

Então Jogar = **Sim**

Se Aspecto = **Chuva** e Vento = **Fraco**

Então Jogar = **Sim**

Se Aspecto = **Chuva** e Vento = **Forte**

Então Jogar = **Não**

Se Aspecto = **Nublado** Então Jogar = **Sim**

Algoritmo ID3 : Exemplo

Qual será a decisão, se o dia estiver com sol, temperatura fria, umidade alta e vento forte ?

FIM!!!

Duvidas e Questões?