Machine Learning

Aula 06

Aprendizagem Supervisionada: Árvores de Decisão/Regressão

Evandro J.R. Silva

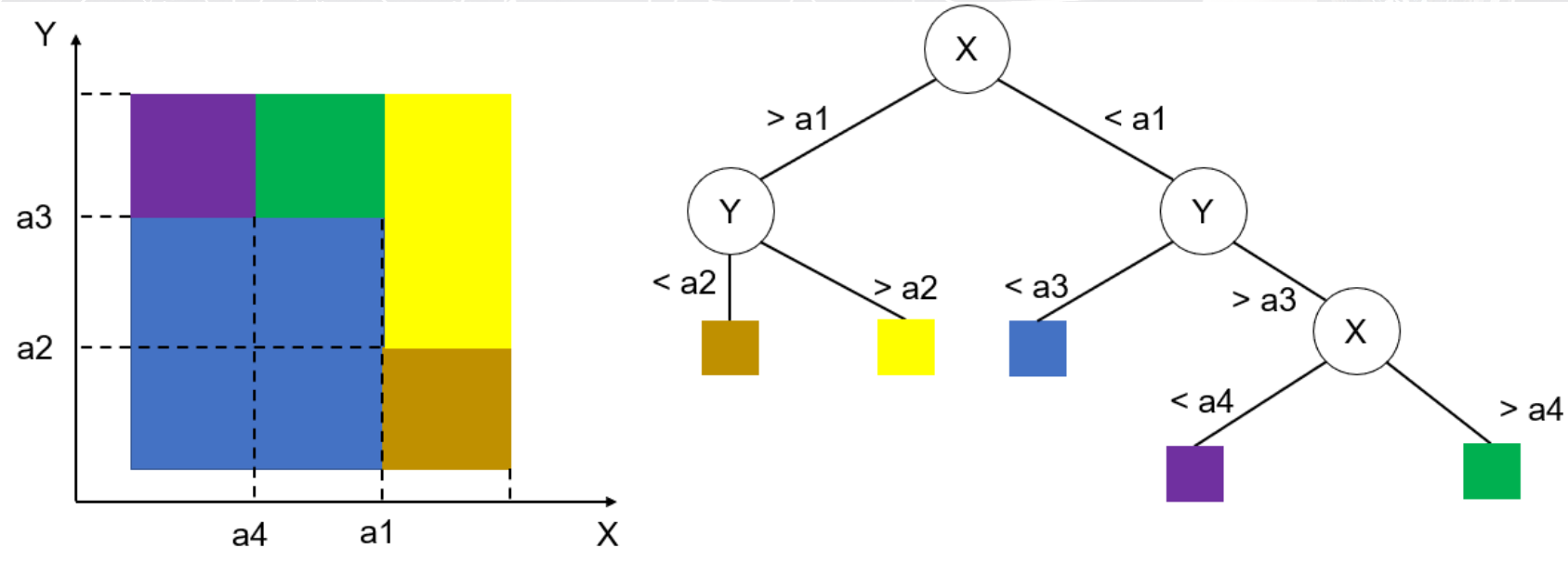
Uninassau Teresina

Sumário

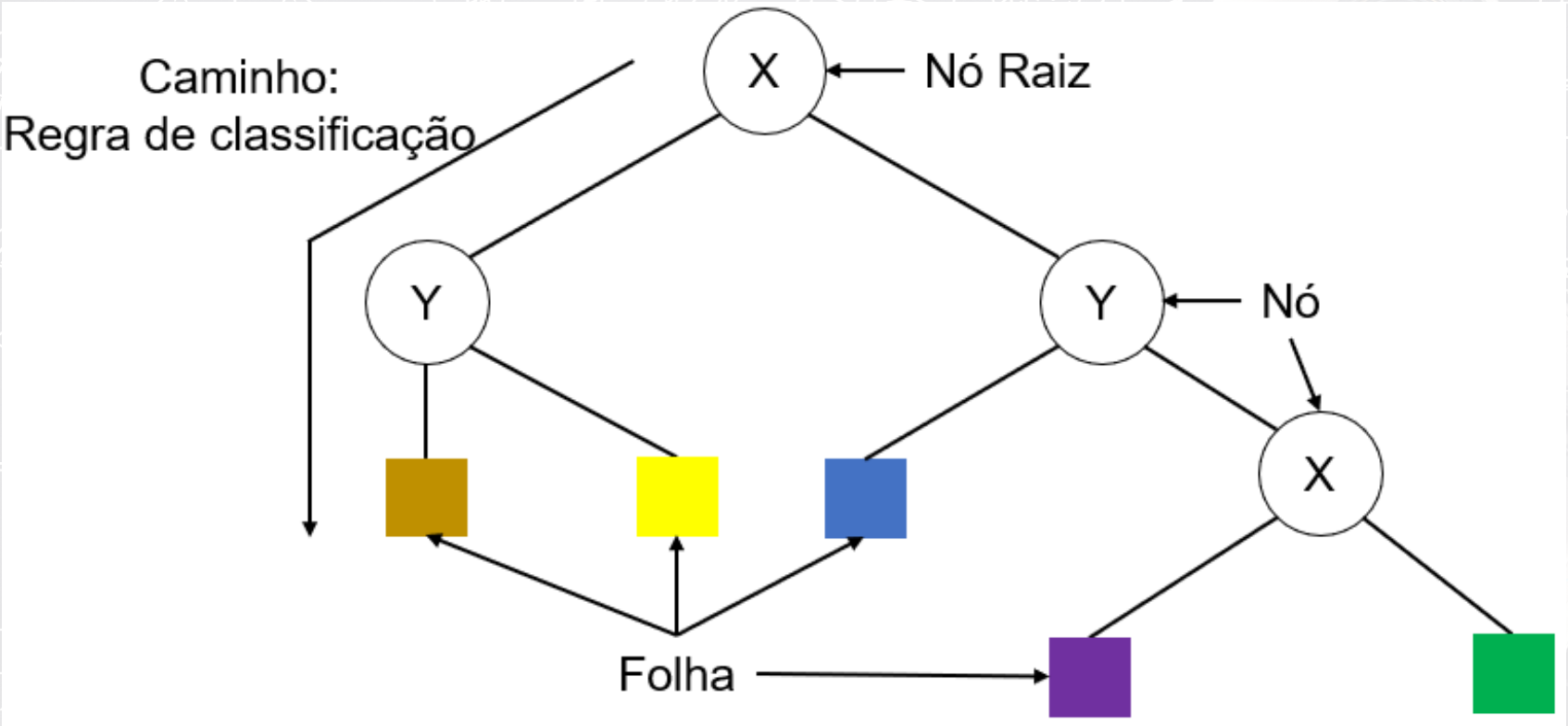
Introdução

- Árvores de Decisão são **métodos com base em procura**, sistemas orientados a conhecimento que **objetivam a criação de estruturas simbólicas que sejam compreensíveis por humanos**, além de usarem a estratégia de "dividir para conquistar" no intuito de resolver problemas de decisão.
- Além disso, **o problema complexo é dividido em problemas mais simples**, em que a mesma estratégia é usada, **as soluções dos subproblemas são combinadas na forma de uma árvore**.

Introdução



Introdução

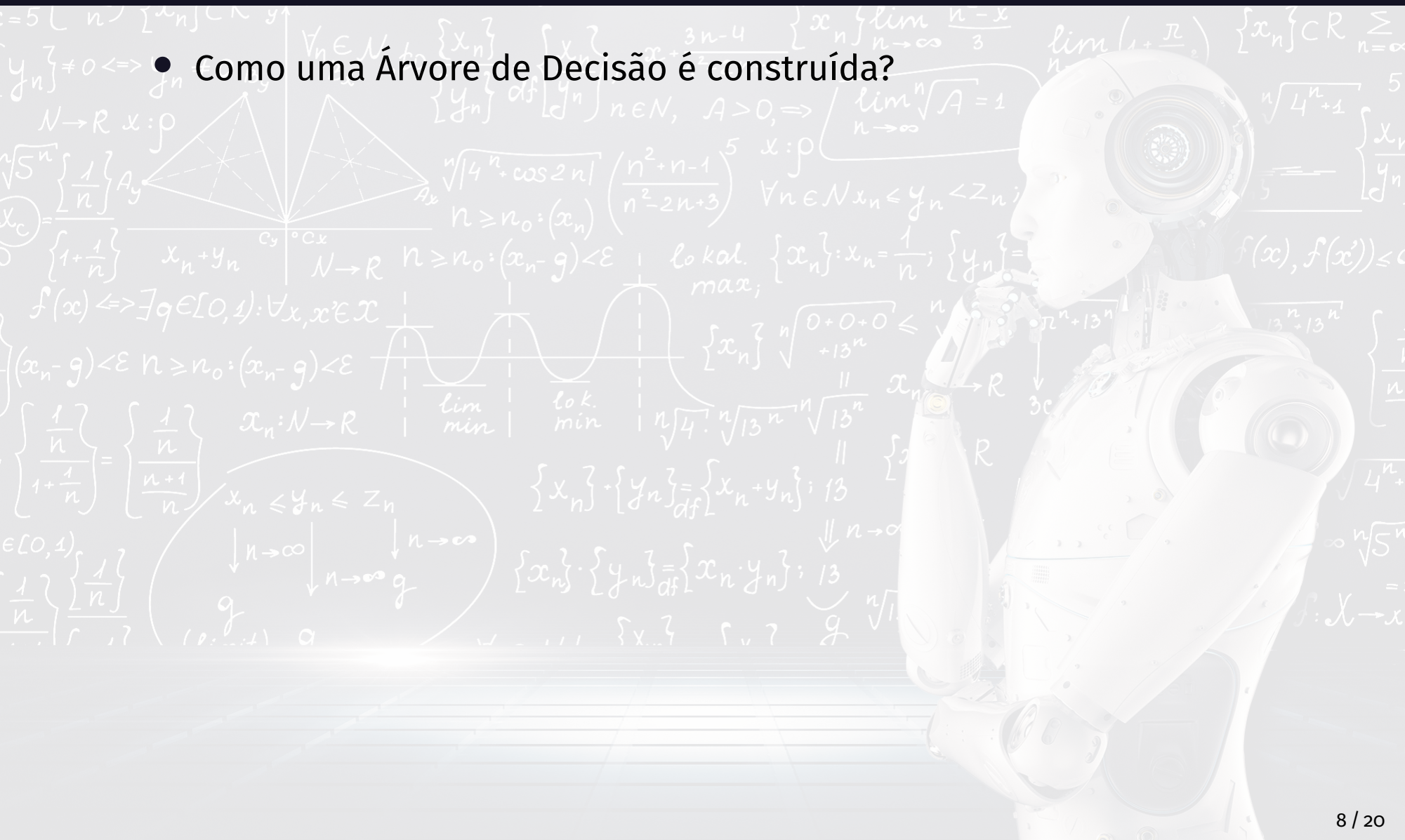


Introdução

- Sua aplicação tem o objetivo de **descobrir automaticamente regularidades implícitas** em bases de dados, **expressando-as em forma de regras**. Sua finalidade é **determinar quais campos de informação na base de dados são importantes e se relacionam com o problema**.

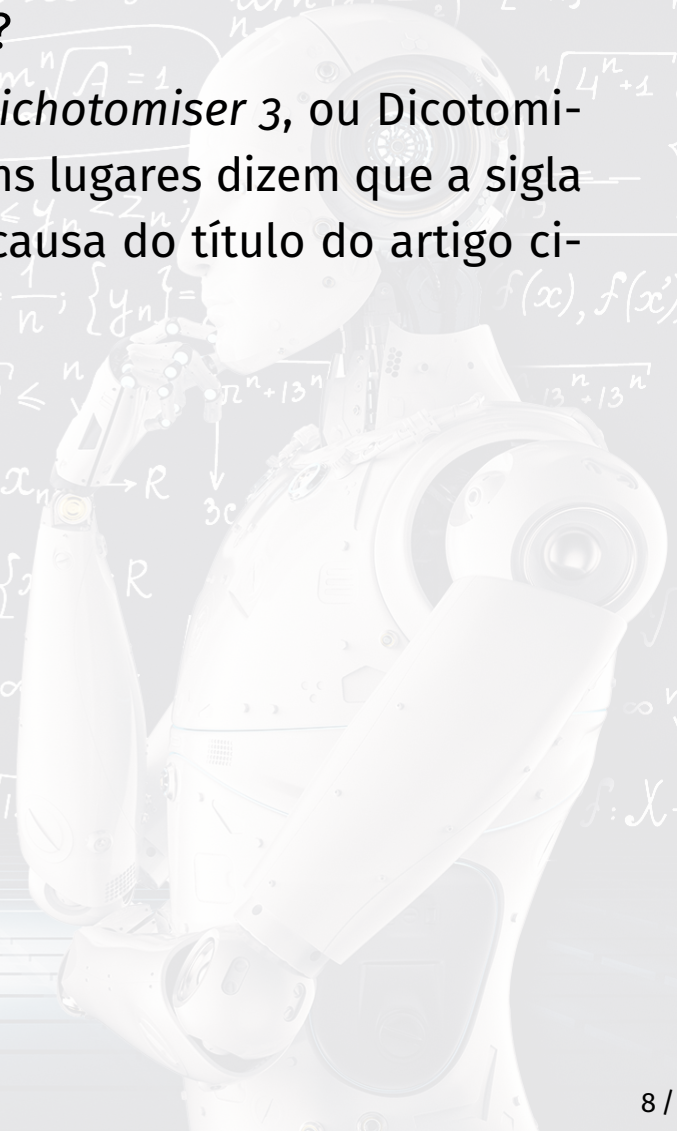
Introdução

- Como uma Árvore de Decisão é construída?



Introdução

- Como uma Árvore de Decisão é construída?
  - Algoritmo base principal: **ID3** (*Iterative Dichotomiser 3*, ou Dicotomizador Iterativo 3) por Ross Quinlan. Alguns lugares dizem que a sigla significa *Induction of Decision Tree*, por causa do título do artigo científico que o descreve.



Introdução

- Como uma Árvore de Decisão é construída?
  - Algoritmo base principal: **ID3** (*Iterative Dichotomiser 3*, ou Dicotomizador Iterativo 3) por Ross Quinlan. Alguns lugares dizem que a sigla significa *Induction of Decision Tree*, por causa do título do artigo científico que o descreve.
  - **C4.5** — evolução do ID3
    - Consegue lidar com atributos discretos e contínuos;
    - Lida também com *missing values*;
    - É capaz de lidar com atributos que possuem custos diferentes;
    - Possui a capacidade de **poda** (ou *pruning*).

Introdução

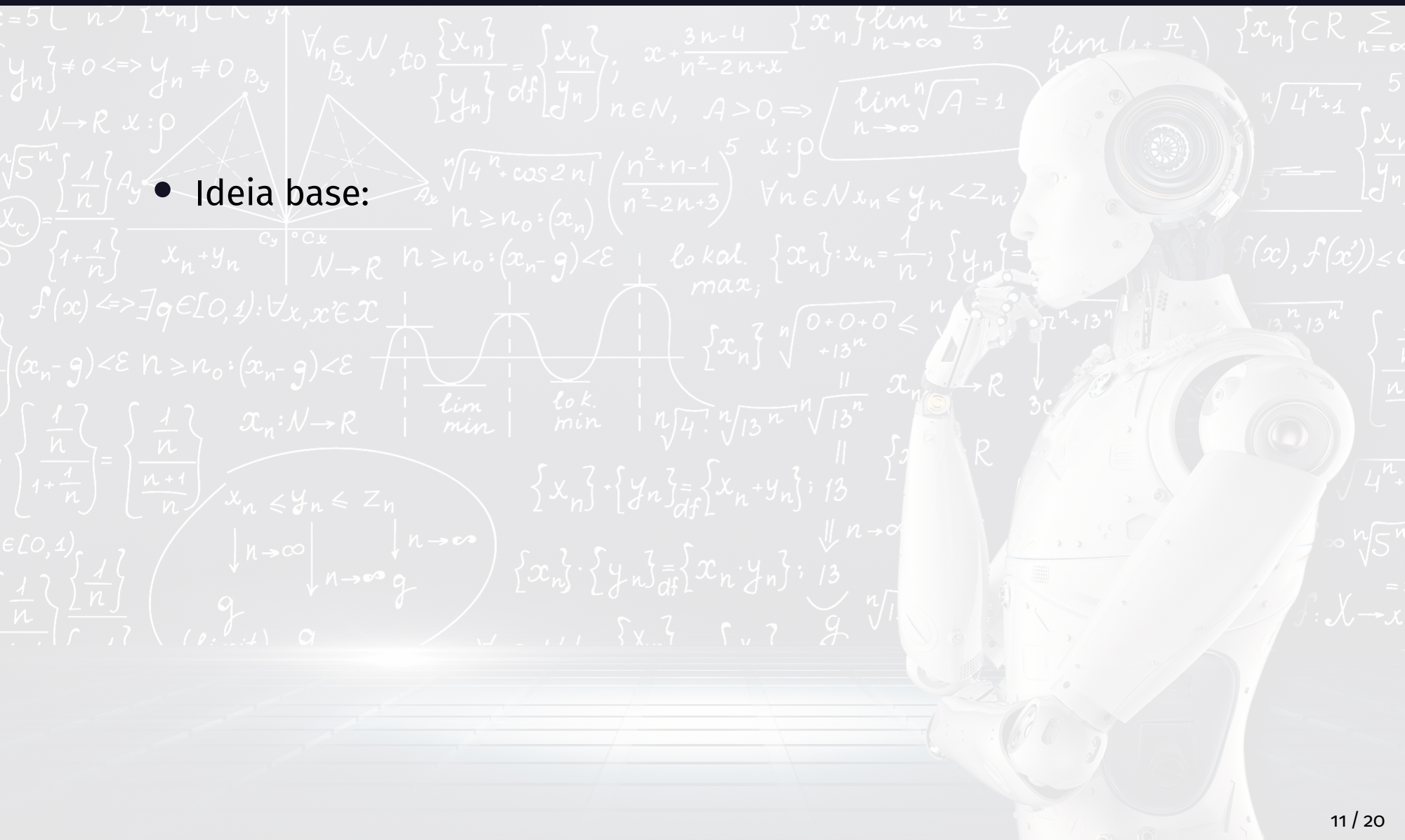
- Como uma Árvore de Decisão é construída?
  - Algoritmo base principal: **ID3** (*Iterative Dichotomiser 3*, ou Dicotomizador Iterativo 3) por Ross Quinlan. Alguns lugares dizem que a sigla significa *Induction of Decision Tree*, por causa do título do artigo científico que o descreve.
  - **C4.5** — evolução do ID3
  - **C5.0** — evolução do C4.5
    - É mais rápido;
    - É mais eficiente no uso de memória;
    - Consegue resultados similares com árvores menores;
    - Dá suporte ao **boosting** (algoritmo de comitês de classificadores — veremos sobre isso em algumas aulas); e outros.

Introdução

- O Quinlan explora comercialmente os seus algoritmos. Ou seja, para utilizá-los, precisamos pagar. Solução:
  - Implementar algoritmos de código aberto baseado nos algoritmos do Quinlan. Alguns conhecidos são o **J48**, e o **CART** (*Classification And Regression Tree*), um termo abrangente para algoritmos de Árvores de Decisão e Regressão.

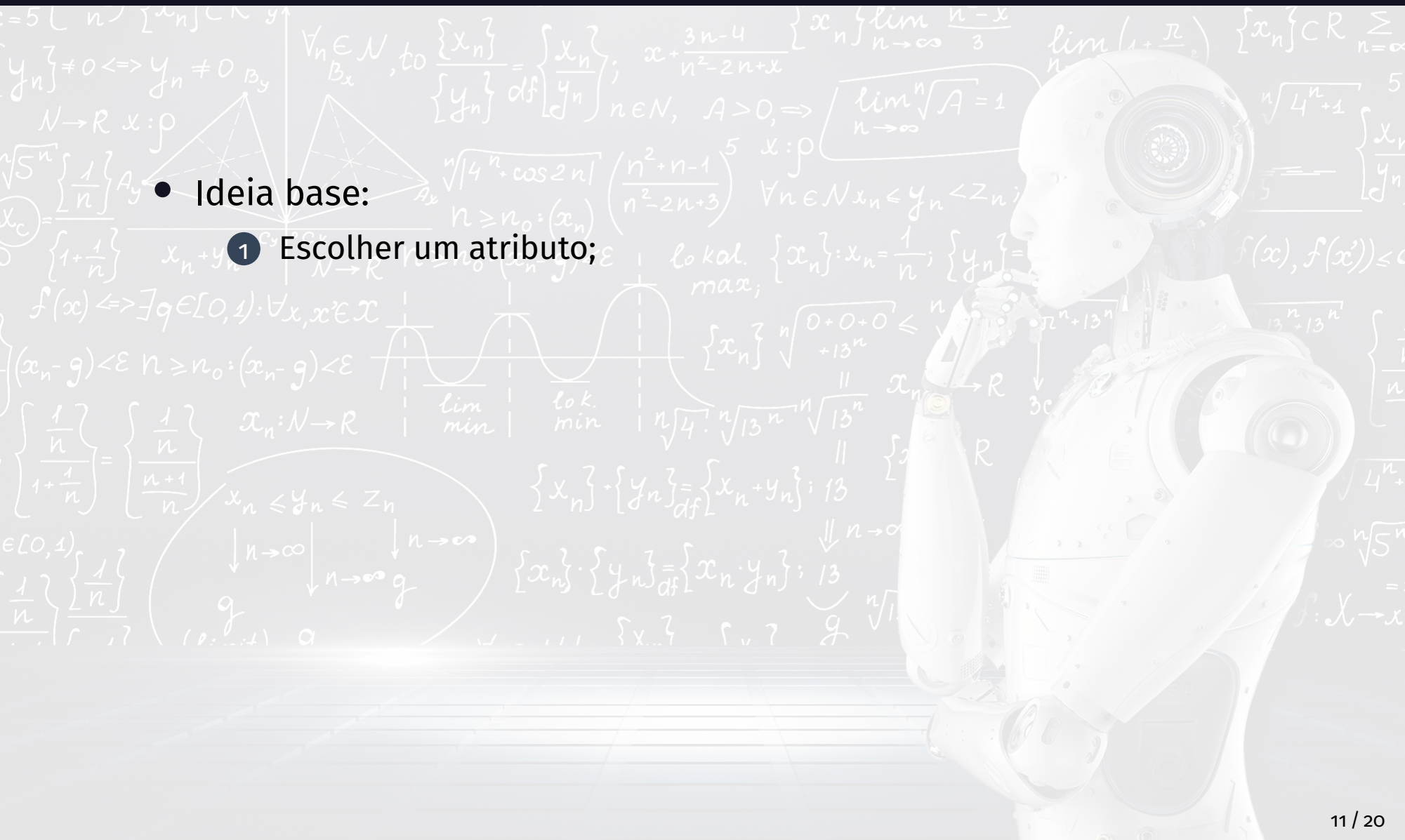
ID3

- Ideia base:



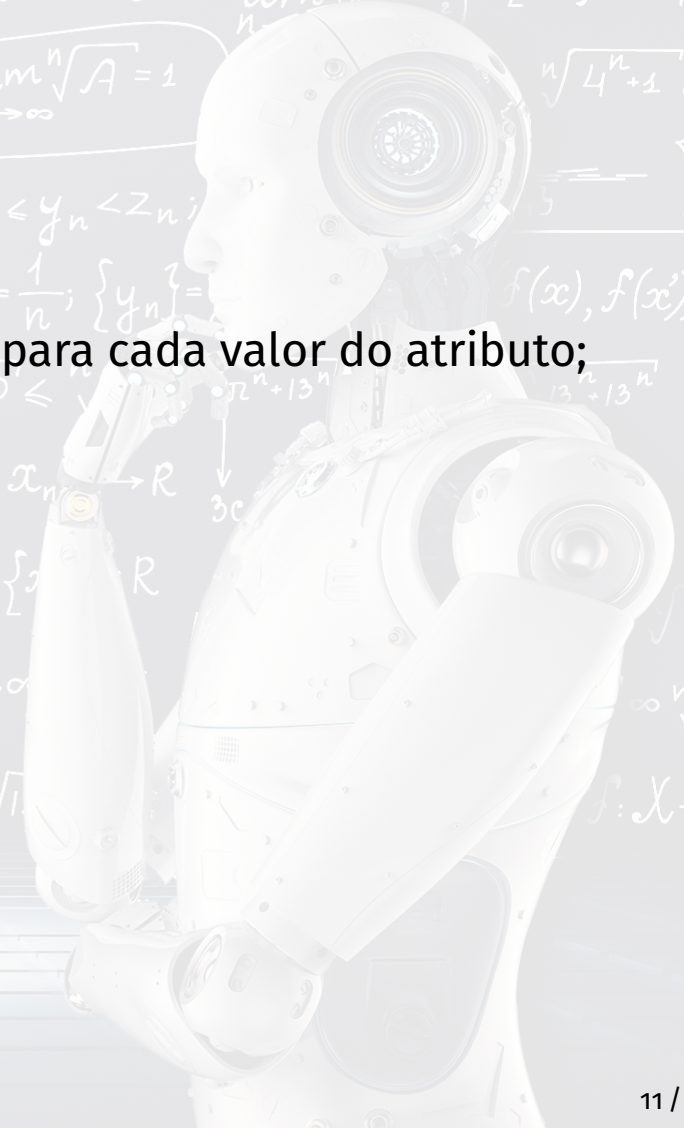
ID3

- Ideia base:
  1. Escolher um atributo;



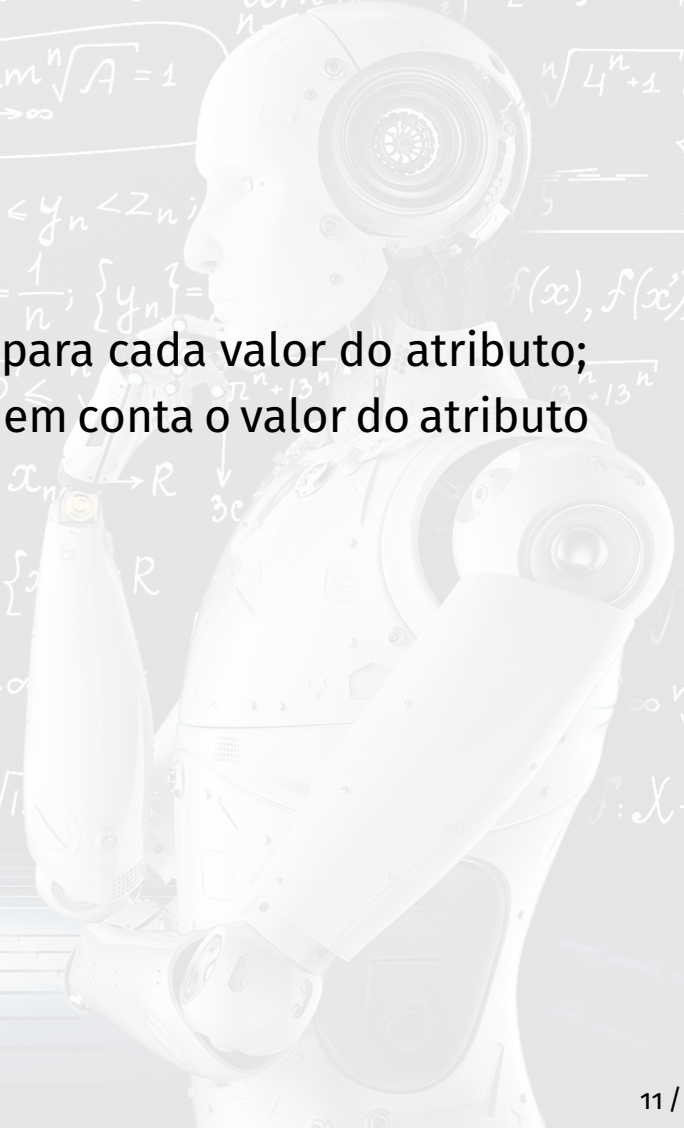
ID3

- Ideia base:
  1. Escolher um atributo;
  2. Estender a árvore adicionando um ramo para cada valor do atributo;



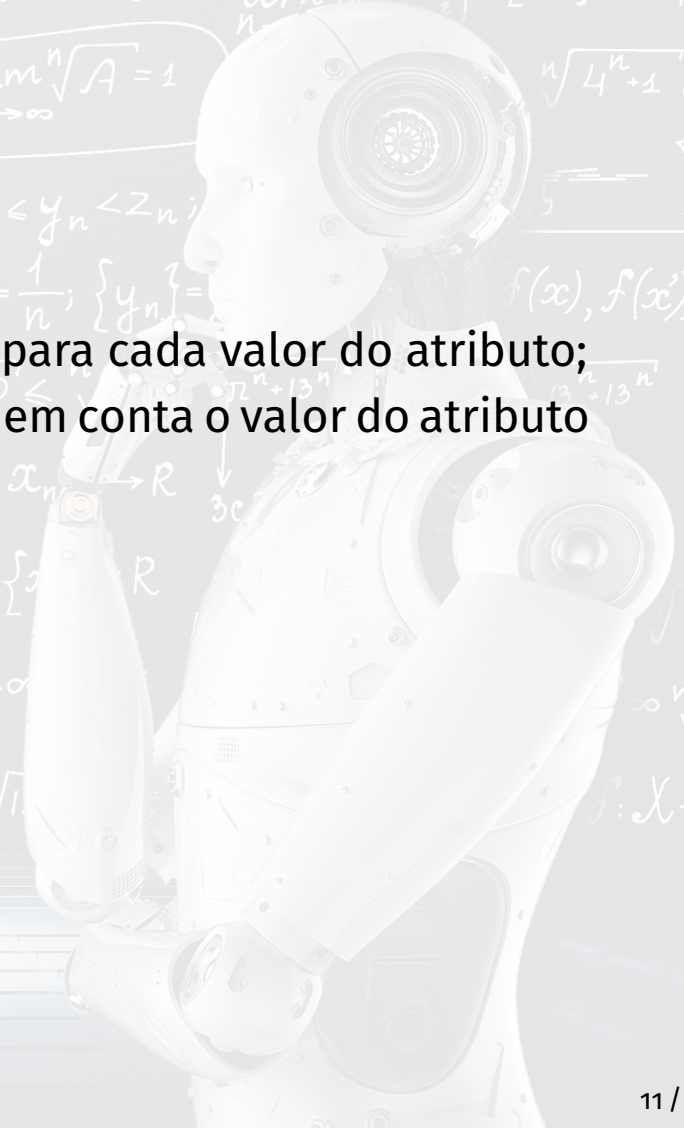
ID3

- Ideia base:
  1. Escolher um atributo;
  2. Estender a árvore adicionando um ramo para cada valor do atributo;
  3. Passar os exemplos para as folhas (tendo em conta o valor do atributo escolhido);



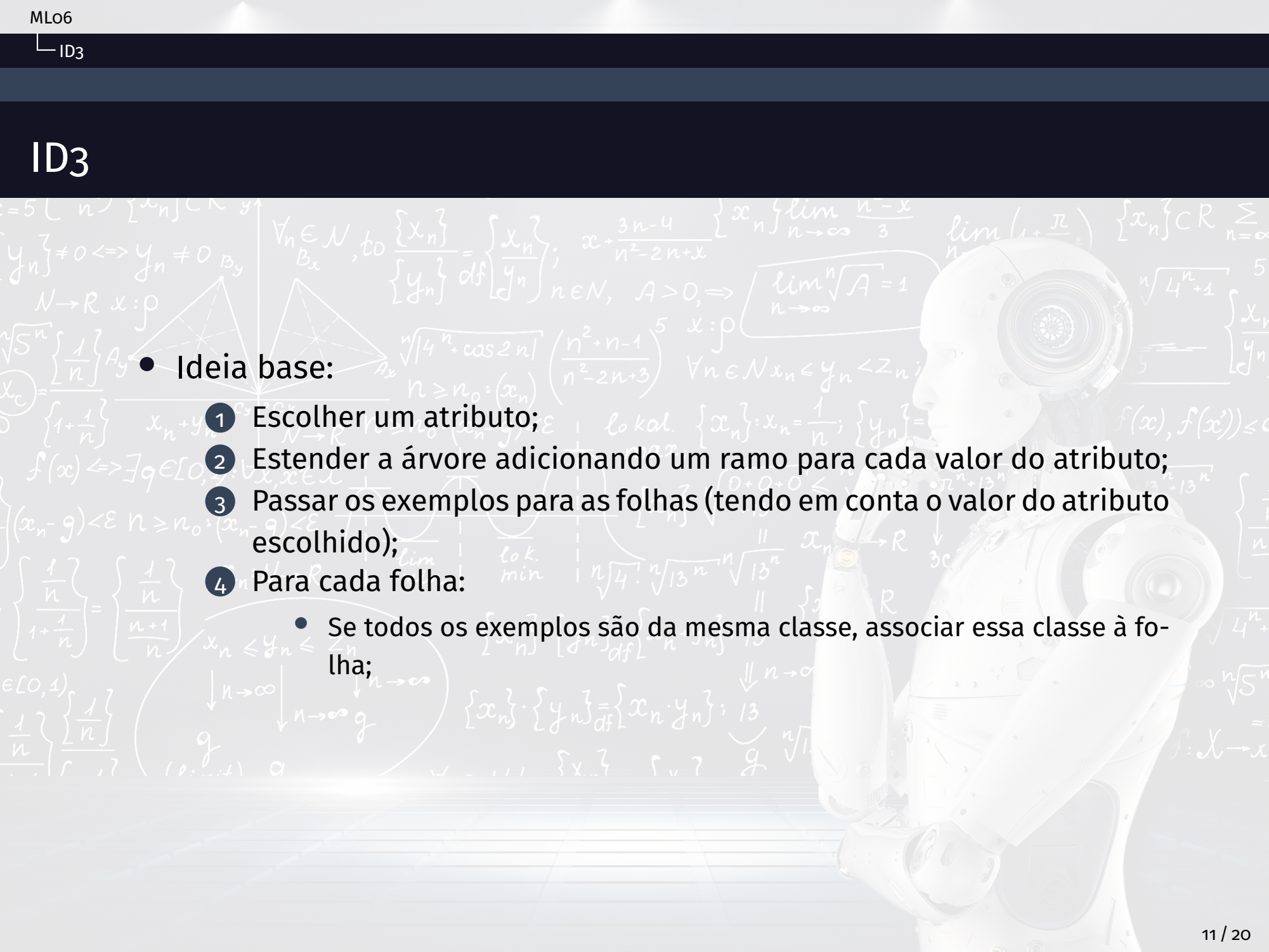
ID3

- Ideia base:
  1. Escolher um atributo;
  2. Estender a árvore adicionando um ramo para cada valor do atributo;
  3. Passar os exemplos para as folhas (tendo em conta o valor do atributo escolhido);
  4. Para cada folha:



ID3

- Ideia base:
  1. Escolher um atributo;
  2. Estender a árvore adicionando um ramo para cada valor do atributo;
  3. Passar os exemplos para as folhas (tendo em conta o valor do atributo escolhido);
  4. Para cada folha:
     - Se todos os exemplos são da mesma classe, associar essa classe à folha;

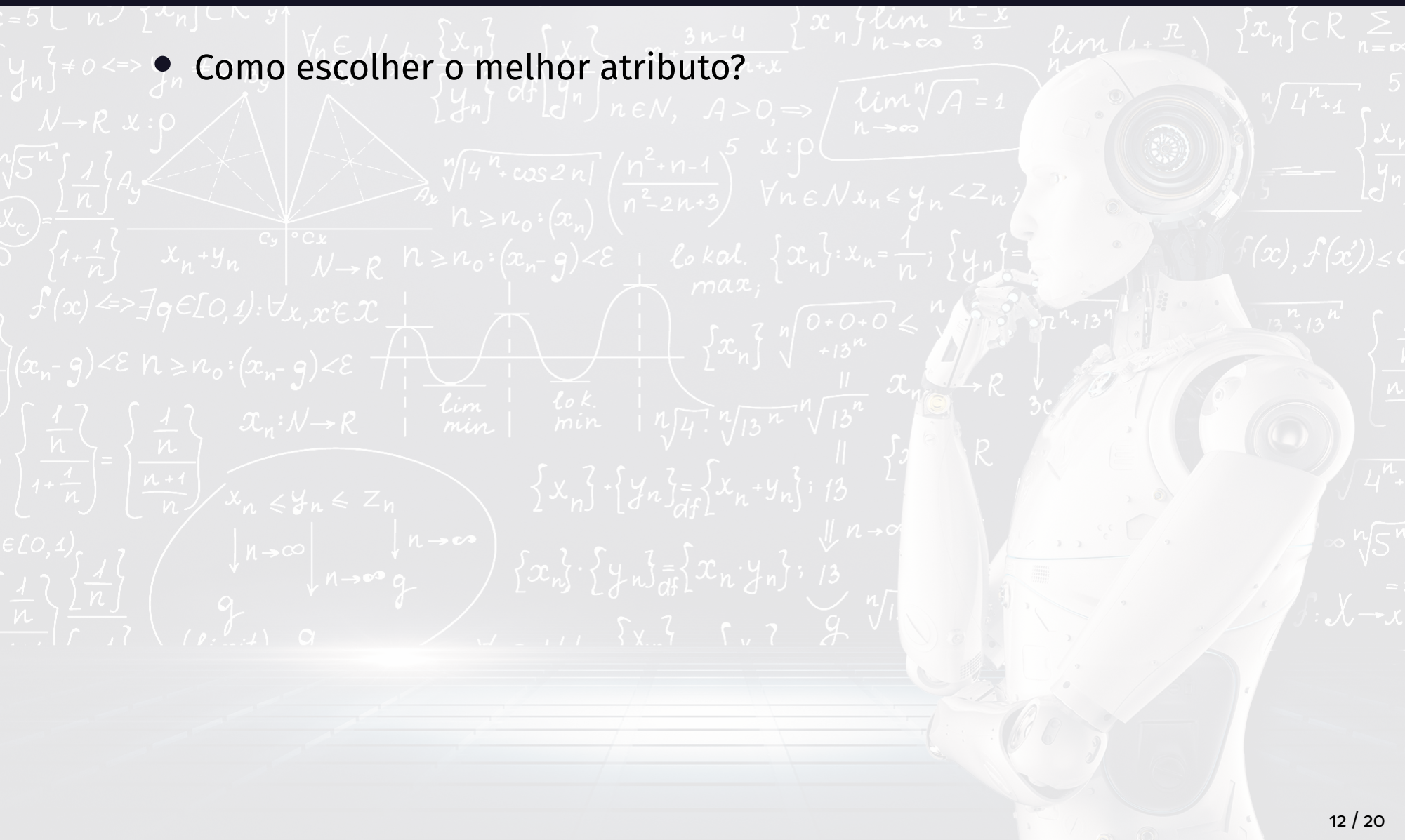


ID3

- Ideia base:
  1. Escolher um atributo;
  2. Estender a árvore adicionando um ramo para cada valor do atributo;
  3. Passar os exemplos para as folhas (tendo em conta o valor do atributo escolhido);
  4. Para cada folha:
     - Se todos os exemplos são da mesma classe, associar essa classe à folha;
     - Senão, repetir os passos 1 a 4.

ID3

- Como escolher o melhor atributo?

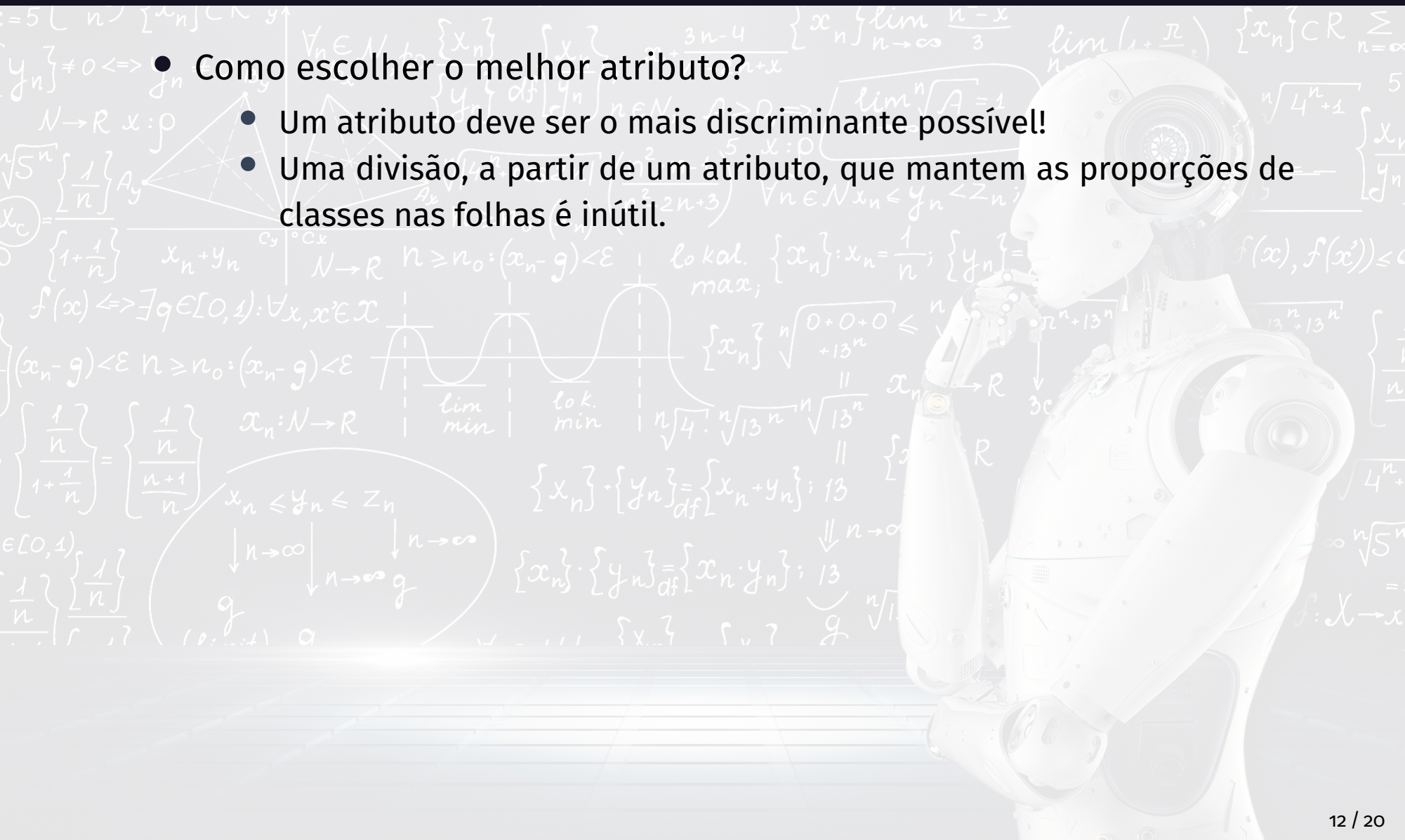


ID3

- Como escolher o melhor atributo?
  - Um atributo deve ser o mais discriminante possível!

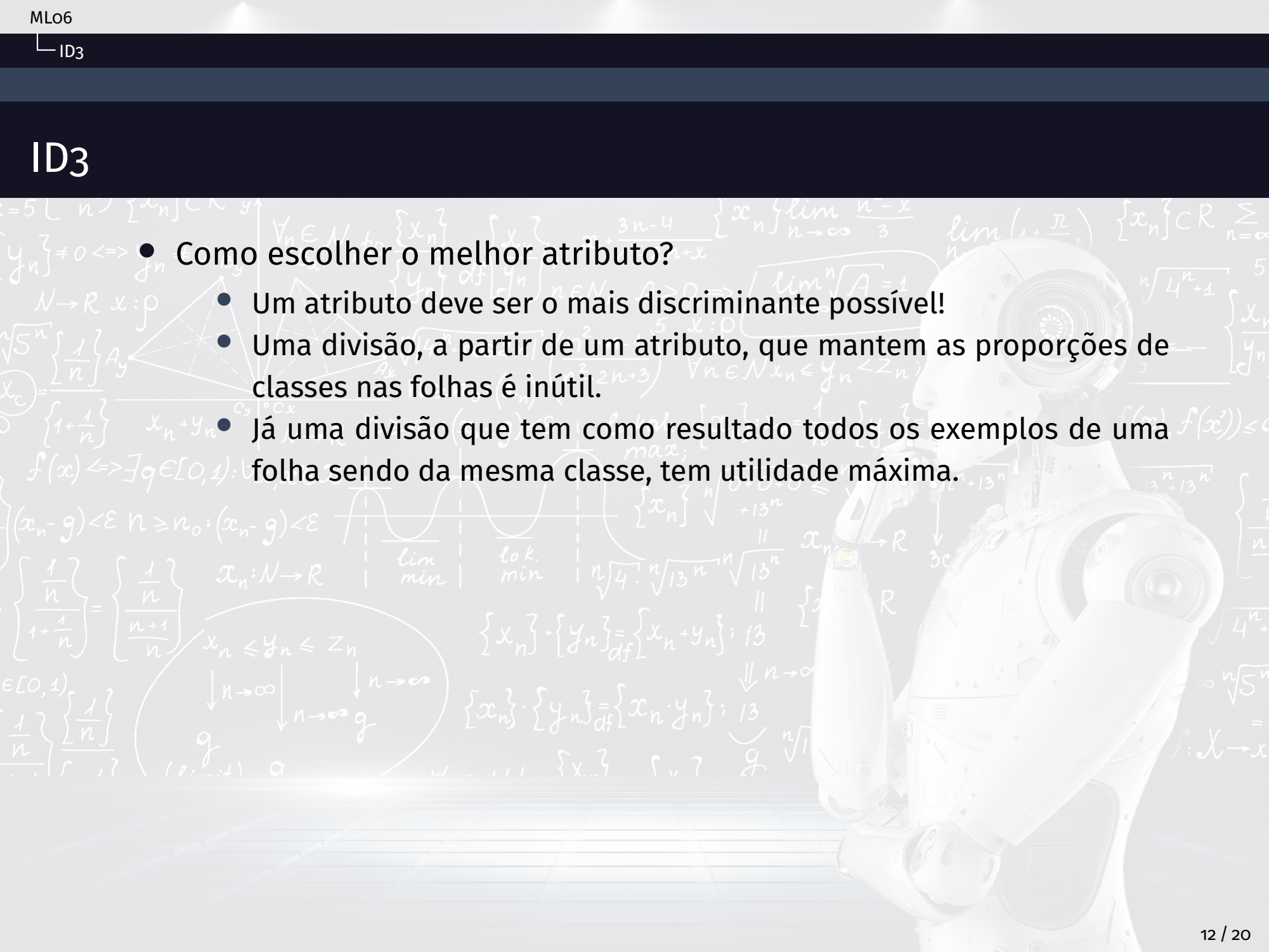
ID3

- Como escolher o melhor atributo?
  - Um atributo deve ser o mais discriminante possível!
  - Uma divisão, a partir de um atributo, que mantem as proporções de classes nas folhas é inútil.

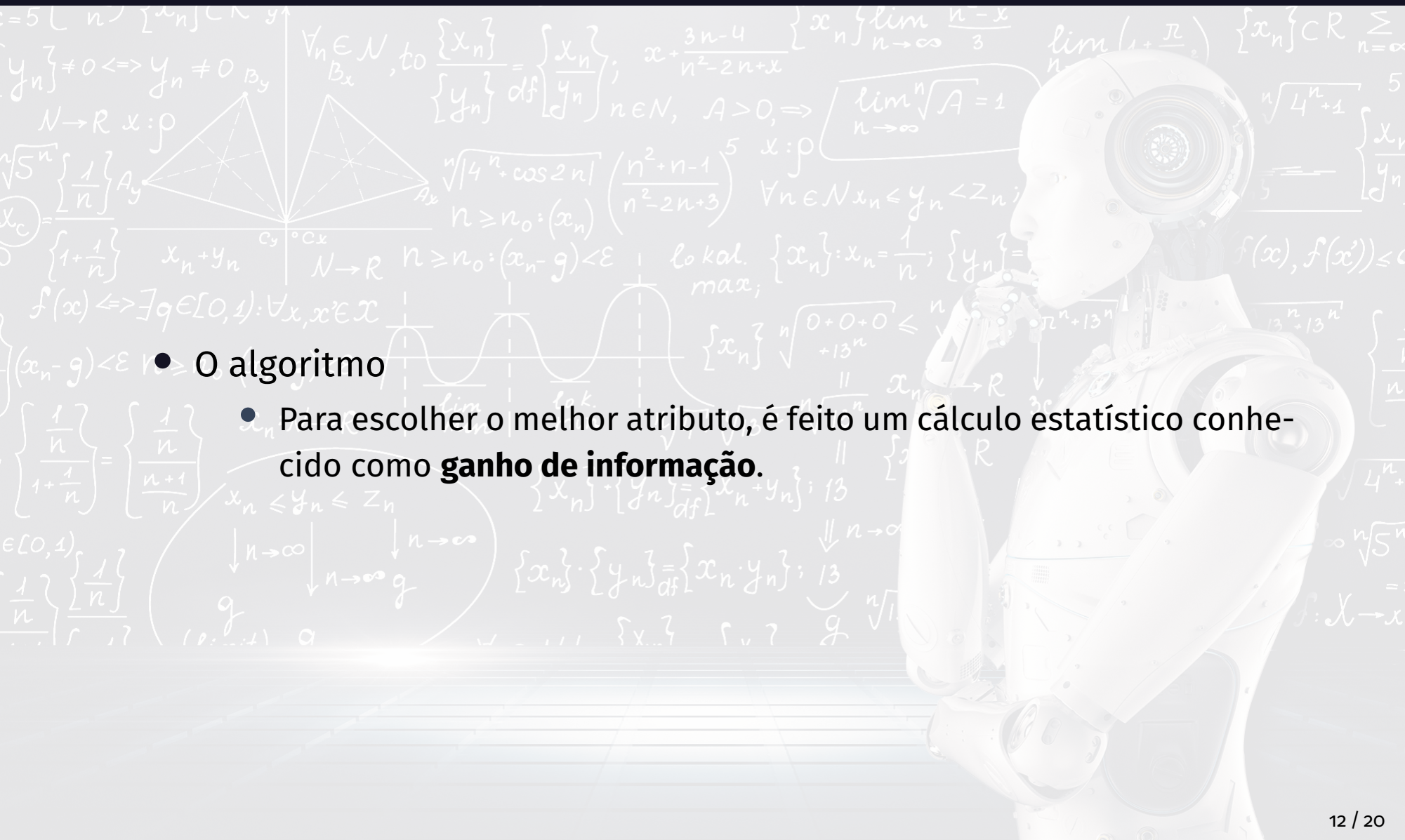


ID3

- Como escolher o melhor atributo?
  - Um atributo deve ser o mais discriminante possível!
  - Uma divisão, a partir de um atributo, que mantem as proporções de classes nas folhas é inútil.
  - Já uma divisão que tem como resultado todos os exemplos de uma folha sendo da mesma classe, tem utilidade máxima.



ID3



- O algoritmo
  - Para escolher o melhor atributo, é feito um cálculo estatístico conhecido como **ganho de informação**.

ID3

- O algoritmo
  - Para escolher o melhor atributo, é feito um cálculo estatístico conhecido como **ganho de informação**.
  - $G(S, A) \equiv Entropia(S) - \sum_{v \in Valores(A)} \frac{|S_v|}{|S|} Entropia(S_v)$

ID3

- O algoritmo
  - Para escolher o melhor atributo, é feito um cálculo estatístico conhecido como **ganho de informação**.
  - $G(S, A) \equiv Entropia(S) - \sum_{v \in Valores(A)} \frac{|S_v|}{|S|} Entropia(S_v)$
    - Conjunto de todos os possíveis valores para o atributo A;

ID3

- O algoritmo
  - Para escolher o melhor atributo, é feito um cálculo estatístico conhecido como **ganho de informação**.
  - $G(S, A) \equiv Entropia(S) - \sum_{v \in Valores(A)} \frac{|S_v|}{|S|} Entropia(S_v)$
    - Conjunto de todos os possíveis valores para o atributo A;
    - Subconjunto de S para o qual o atributo A tem valor *v*;

ID3

- O algoritmo
  - Para escolher o melhor atributo, é feito um cálculo estatístico conhecido como **ganho de informação**.
  - $G(S, A) \equiv Entropia(S) - \sum_{v \in Valores(A)} \frac{|S_v|}{|S|} Entropia(S_v)$

  - Para entender essa fórmula, vamos ver seus elementos, começando pela Entropia.

ID3

- A Entropia é uma medida que caracteriza a pureza ou impureza de um conjunto arbitrário de exemplos.
- Seja um conjunto S contendo duas classes, uma positiva ($p_{\oplus}$ = proporção de exemplos positivos) e uma negativa ($p_{\ominus}$ = proporção de exemplos negativos).

ID3

- A Entropia é uma medida que caracteriza a pureza ou impureza de um conjunto arbitrário de exemplos.
- Seja um conjunto S contendo duas classes, uma positiva ($p_\oplus$ = proporção de exemplos positivos) e uma negativa ($p_\ominus$ = proporção de exemplos negativos).
- A entropia de S é:
  $Entropia(S) \equiv -p_\oplus log_2 p_\oplus - p_\ominus log_2 p_\ominus$
  ou
  $Entropia(S) \equiv \sum_{i=1}^{c} -p_i log_2 p_i$

ID3

Exemplo

S = 14 exemplos ... [9+, 5-]

ID3

Exemplo

S = 14 exemplos ... [9+, 5-]

Entropia(S) = Entropia([9+, 5-])

ID3

Exemplo

S = 14 exemplos ... [9+, 5-]

Entropia(S) = Entropia([9+, 5-])

$= -(9/14)log_2(9/14) - (5/14)log_2(5/14) = 0,940$

ID3

- Exemplo ilustrativo

Dia	Tempo	Temperatura	Umidade	Vento	Jogar
D1	Ensolarado	Quente	Alta	Fraco	Não
D2	Ensolarado	Quente	Alta	Forte	Não
D3	Nublado	Quente	Alta	Fraco	Sim
D4	Chuvoso	Média	Alta	Fraco	Sim
D5	Chuvoso	Frio	Normal	Fraco	Sim
D6	Chuvoso	Frio	Normal	Forte	Não
D7	Nublado	Frio	Normal	Forte	Sim
D8	Ensolarado	Média	Alta	Fraco	Não
D9	Ensolarado	Frio	Normal	Fraco	Sim
D10	Chuvoso	Média	Normal	Fraco	Sim
D11	Ensolarado	Média	Normal	Forte	Sim
D12	Nublado	Média	Alta	Forte	Sim
D13	Nublado	Quente	Normal	Fraco	Sim
D14	Chuvoso	Média	Alta	Forte	Não

Tabela 1: Exemplos de treino para a decisão de jogar

ID3

- Exemplo ilustrativo

Dia	Tempo	Temperatura	Umidade	Vento	Jogar
D1	Ensolarado	Quente	Alta	Fraco	Não
D2	Ensolarado	Quente	Alta	Forte	Não
D3	Nublado	Quente	Alta	Fraco	Sim
D4	Chuvoso	Média	Alta	Fraco	Sim
D5	Chuvoso	Frio	Normal	Fraco	Sim
D6	Chuvoso	Frio	Normal	Forte	Não
D7	Nublado	Frio	Normal	Forte	Sim
D8	Ensolarado	Média	Alta	Fraco	Não
D9	Ensolarado	Frio	Normal	Fraco	Sim
D10	Chuvoso	Média	Normal	Fraco	Sim
D11	Ensolarado	Média	Normal	Forte	Sim
D12	Nublado	Média	Alta	Forte	Sim
D13	Nublado	Quente	Normal	Fraco	Sim
D14	Chuvoso	Média	Alta	Forte	Não

Tabela 1: Exemplos de treino para a decisão de jogar

ID3

- Exemplo ilustrativo

Dia	Tempo	Temperatura	Umidade	Vento	Jogar
D1	Ensolarado	Quente	Alta	Fraco	Não
D2	Ensolarado	Quente	Alta	Forte	Não
D3	Nublado	Quente	Alta	Fraco	Sim
D4	Chuvoso	Média	Alta	Fraco	Sim
D5	Chuvoso	Frio	Normal	Fraco	Sim
D6	Chuvoso	Frio	Normal	Forte	Não
D7	Nublado	Frio	Normal	Forte	Sim
D8	Ensolarado	Média	Alta	Fraco	Não
D9	Ensolarado	Frio	Normal	Fraco	Sim
D10	Chuvoso	Média	Normal	Fraco	Sim
D11	Ensolarado	Média	Normal	Forte	Sim
D12	Nublado	Média	Alta	Forte	Sim
D13	Nublado	Quente	Normal	Fraco	Sim
D14	Chuvoso	Média	Alta	Forte	Não

Tabela 1: Exemplos de treino para a decisão de jogar

ID3

Ganho de Informação

$S = [9+, 5-] \longrightarrow \textit{Entropia} = 0,940$

ID3

Ganho de Informação

$S = [9+, 5-] \longrightarrow \textit{Entropia} = 0,940$

Atributo: Tempo = [Ensolarado, Nublado, Chuvoso]

ID3

Ganho de Informação

$S = [9+, 5-] \longrightarrow Entropia = 0,940$

Atributo: Tempo = [Ensolarado, Nublado, Chuvoso]

$S_{Ensolarado} = [2+, 3-]$;

ID3

Ganho de Informação

$S = [9+, 5-] \longrightarrow Entropia = 0,940$

Atributo: Tempo = [Ensolarado, Nublado, Chuvoso]

$S_{Ensolarado} = [2+, 3-]$;

$S_{Nublado} = [4+, 0-]$;

ID3

Ganho de Informação

$S = [9+, 5-] \longrightarrow Entropia = 0,940$

Atributo: Tempo = [Ensolarado, Nublado, Chuvoso]

$S_{Ensolarado} = [2+, 3-]$;

$S_{Nublado} = [4+, 0-]$;

$S_{Chuvoso} = [3+, 2-]$;

ID3

Ganho de Informação

$S = [9+, 5-] \longrightarrow Entropia = 0,940$

Atributo: Tempo = [Ensolarado, Nublado, Chuvoso]

$S_{Ensolarado} = [2+, 3-]$;

$S_{Nublado} = [4+, 0-]$;

$S_{Chuvoso} = [3+, 2-]$;

$$G(S, A) \equiv Entropia(S) - \sum_{v \in Valores(A)} \frac{|S_v|}{|S|} Entropia(S_v)$$

ID3

Ganho de Informação

$S = [9+, 5-] \longrightarrow Entropia = 0,940$

Atributo: Tempo = [Ensolarado, Nublado, Chuvoso]

$S_{Ensolarado} = [2+, 3-]$;

$S_{Nublado} = [4+, 0-]$;

$S_{Chuvoso} = [3+, 2-]$;

$$G(S, A) \equiv Entropia(S) - \sum_{v \in Valores(A)} \frac{|S_v|}{|S|} Entropia(S_v)$$

$$G(S, Tempo) \equiv 0,940 - (5/14)Entropia(S_{Ensolarado}) - (4/14)Entropia(S_{Nublado}) - (5/14)Entropia(S_{Chuvoso})$$

ID3

Ganho de Informação

$S = [9+, 5-] \longrightarrow Entropia = 0,940$

Atributo: Tempo = [Ensolarado, Nublado, Chuvoso]

$S_{Ensolarado} = [2+, 3-]$;

$S_{Nublado} = [4+, 0-]$;

$S_{Chuvoso} = [3+, 2-]$;

$$G(S, A) \equiv Entropia(S) - \sum_{v \in Valores(A)} \frac{|S_v|}{|S|} Entropia(S_v)$$

$$G(S, Tempo) \equiv 0,940 - (5/14)Entropia(S_{Ensolarado}) - (4/14)Entropia(S_{Nublado}) - (5/14)Entropia(S_{Chuvoso})$$

$$G(S, Tempo) \equiv 0,940 - (5/14)0,971 - (4/14)0 - (5/14)0,971$$

ID3

Ganho de Informação

$S = [9+, 5-] \longrightarrow Entropia = 0,940$

Atributo: Tempo = [Ensolarado, Nublado, Chuvoso]

$S_{Ensolarado} = [2+, 3-]$;

$S_{Nublado} = [4+, 0-]$;

$S_{Chuvoso} = [3+, 2-]$;

$$G(S, A) \equiv Entropia(S) - \sum_{v \in Valores(A)} \frac{|S_v|}{|S|} Entropia(S_v)$$

$$G(S, Tempo) \equiv 0,940 - (5/14)Entropia(S_{Ensolarado}) - (4/14)Entropia(S_{Nublado}) - (5/14)Entropia(S_{Chuvoso})$$

$$G(S, Tempo) \equiv 0,940 - (5/14)0,971 - (4/14)0 - (5/14)0,971$$

$$G(S, Tempo) \equiv 0,940 - 0,347 - 0 - 0,347$$

ID3

Ganho de Informação

$S = [9+, 5-] \longrightarrow Entropia = 0,940$

Atributo: Tempo = [Ensolarado, Nublado, Chuvoso]

$S_{Ensolarado} = [2+, 3-]$;

$S_{Nublado} = [4+, 0-]$;

$S_{Chuvoso} = [3+, 2-]$;

$$G(S, A) \equiv Entropia(S) - \sum_{v \in Valores(A)} \frac{|S_v|}{|S|} Entropia(S_v)$$

$$G(S, Tempo) \equiv 0,940 - (5/14)Entropia(S_{Ensolarado}) - (4/14)Entropia(S_{Nublado}) - (5/14)Entropia(S_{Chuvoso})$$

$$G(S, Tempo) \equiv 0,940 - (5/14)0,971 - (4/14)0 - (5/14)0,971$$

$$G(S, Tempo) \equiv 0,940 - 0,347 - 0 - 0,347$$

$$G(S, Tempo) \equiv 0,246$$

ID3

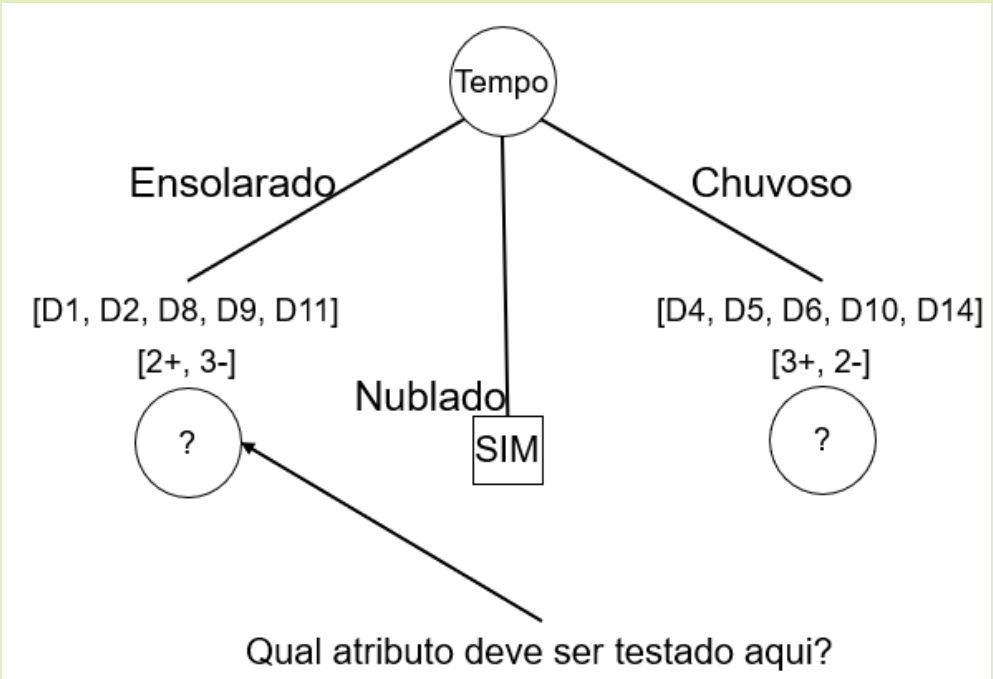
Ganho de Informação

$Ganho(S, Tempo) = 0,246$

$Ganho(S, Temperatura) = 0,029$

$Ganho(S, Umidade) = 0,151$

$Ganho(S, Vento) = 0,048$



Árvore de Decisão

Próximo Atributo

$S_{Ensolarado} = [D1, D2, D8, D9, D11]$

$$Ganho(S_{Ensolarado}, Umidade) = 0,970 - (3/5)0 - (2/5)0 = 0,970$$

$$Ganho(S_{Ensolarado}, Temperatura) = 0,970 - (2/5)0 - (2/5)1 - (1/5)0 = 0,570$$

$$Ganho(S_{Ensolarado}, Vento) = 0,970 - (2/5)1 - (3/5)0,918 = 0,019$$

FIM