UNINOVE
Universidade Nove de Julho
A UNINOVE É SEMPRE 10

belan.pro.br

Material de apoio

Aguardando a entrada de alunos

UNINOVE
Universidade Nove de Julho
A UNINOVE É SEMPRE 10

Árvores de Decisão

Prof. Dr. Peterson Belan
belan@uni9.pro.br

Árvores de Decisão – AD's

AD's são representações simples do conhecimento e têm sido aplicadas em sistemas de aprendizado. Elas são amplamente utilizadas em tarefas de classificação, como um meio eficiente para construir classificadores que predizem classes baseadas nos valores de atributos. Assim, podem ser utilizadas em várias aplicações como diagnósticos médicos, análise de risco em créditos, entre outros exemplos.

AD é um dos algoritmos de aprendizado mais simples e pode ser interpretada como um conjunto de regras do tipo SE...ENTÃO.

As AD's são um dos modelos mais práticos e mais usados em inferência indutiva.

Árvores de Decisão – AD's

Entre os algoritmos de AD estão: ID3 (QUINLAN, 1986), C4.5 (QUINLAN, 1993) e CART (BREIMAN et al., 1984), sendo o primeiro o mais utilizado.

A chave para o sucesso de um algoritmo AD é como gerar a árvore, ou seja, como escolher os atributos mais importantes para gerar as regras e quais regras podem ser descartadas da árvore.

O ideal é gerar a AD com base na importância dos atributos. Desta forma, o atributo considerado mais importante deve ficar na raiz da árvore.

Com isso, pode-se resolver um problema aplicando o menor número de regras.

Árvores de Decisão – AD's

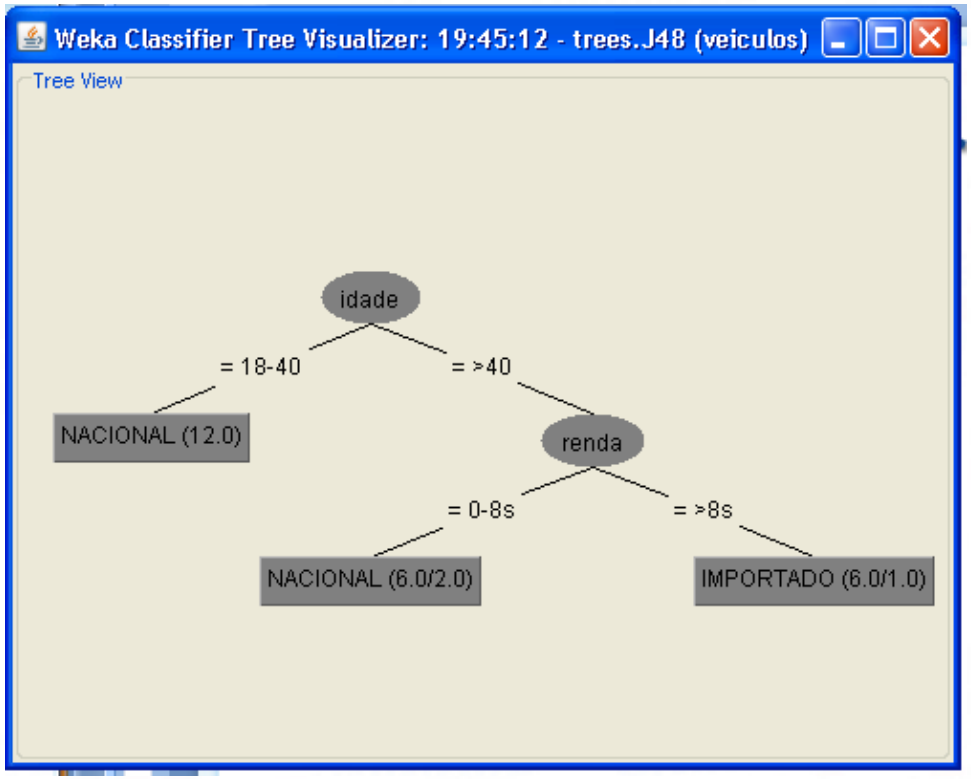
Algumas formas de escolha dos atributos:

- ✓ Aleatória: seleciona-se o atributo de forma aleatória;
- ✓ Menor quantidade de valores: seleciona o atributo com a menor quantidade de valores possíveis;
- ✓ Maior quantidade de valores: seleciona o atributo com a maior quantidade de valores possíveis;
- ✓ Ganho máximo: seleciona o atributo com maior ganho de informação esperado;
- ✓ índice Gini : seleciona o atributo com base no seu índice de dispersão estatística.

Árvores de Decisão – AD's

Exemplo: Predição da preferência de clientes em uma seguradora.

Codigo	Idade	Renda	Tipo_Veiculo
1	>40	>8s	IMPORTADO
2	18-40	0-8s	NACIONAL
3	>40	>8s	NACIONAL
4	18-40	>8s	NACIONAL
5	18-40	0-8s	NACIONAL
6	18-40	0-8s	NACIONAL
7	18-40	0-8s	NACIONAL
8	>40	>8s	IMPORTADO
9	>40	>8s	IMPORTADO
10	18-40	0-8s	NACIONAL
11	>40	0-8s	IMPORTADO
12	18-40	0-8s	NACIONAL
13	18-40	0-8s	NACIONAL
14	>40	0-8s	IMPORTADO
15	>40	>8s	IMPORTADO
16	18-40	>8s	NACIONAL
17	>40	>8s	IMPORTADO
18	>40	0-8s	NACIONAL
19	>40	0-8s	NACIONAL
20	18-40	0-8s	NACIONAL
21	18-40	0-8s	NACIONAL
22	>40	0-8s	NACIONAL
23	>40	0-8s	NACIONAL
24	18-40	0-8s	NACIONAL



Árvores de Decisão – Regras

Como já mencionado, uma árvore representa um conjunto de regras, cada uma delas começando na raiz da árvore e caminhando para baixo, em direção às folhas.

Assim, as regras que representam a AD gerada para predição da preferência de clientes da seguradora são:

SE idade = 18-40 ENTÃO tipo_veiculo = nacional

SE idade = >40 e renda = 0-8s ENTÃO tipo_veiculo = nacional

SE idade = >40 e renda = >8s ENTÃO tipo_veiculo = importado

Entendendo a construção de uma AD

Suponha que o objetivo é decidir se uma pessoa deve **Jogar Tênis**, levando em conta certos parâmetros do ambiente, como o **Aspecto** do céu, a **Temperatura**, a **Humidade** e o **Vento**. A decisão **SIM** (jogar tênis) ou **NÃO** (não ir jogar tênis) é o resultado da classificação. Para construir a AD são tidos em conta exemplos (dias) passados.

Exemplos de Treino (Conjunto de treinamento)

Dia	Aspecto	Temp.	Humidade	Vento	Jogar Ténis
D1	Sol	Quente	Elevada	Fraco	Não
D2	Sol	Quente	Elevada	Forte	Não
D3	Nuvens	Quente	Elevada	Fraco	Sim
D4	Chuva	Ameno	Elevada	Fraco	Sim
D5	Chuva	Fresco	Normal	Fraco	Sim
D6	Chuva	Fresco	Normal	Forte	Não
D7	Nuvens	Fresco	Normal	Fraco	Sim
D8	Sol	Ameno	Elevada	Fraco	Não
D9	Sol	Fresco	Normal	Fraco	Sim
D10	Chuva	Ameno	Normal	Forte	Sim
D11	Sol	Ameno	Normal	Forte	Sim
D12	Nuvens	Ameno	Elevada	Forte	Sim
D13	Nuvens	Quente	Normal	Fraco	Sim
D14	Chuva	Ameno	Elevada	Forte	Não

Algoritmo ID3

Este algoritmo segue os seguintes passos:

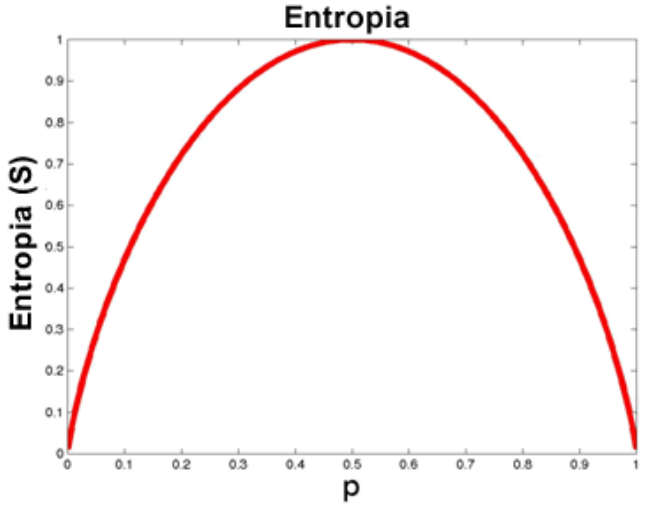
- Começar com todos os exemplos do conjunto de treinamento;
- Escolher o atributo que melhor divide os exemplos, ou seja agrupar exemplos da mesma classe ou exemplos semelhantes;
- Para o atributo escolhido, criar um nó filho para cada valor possível do atributo;
- Transportar os exemplos para cada filho levando em conta o valor do filho;
- Repetir o procedimento para cada filho não "puro". Em outras palavras, a impureza é nula se todos os exemplos pertencerem à mesma classe. Analogamente, o grau de impureza é máximo no nó se houver o mesmo número de exemplos para cada classe possível.

Pergunta importante: como saber qual atributo escolher em cada passo da construção da AD?
Para lidar com esta escolha são introduzidos dois novos conceitos: **Entropia** e **Ganho**.

Entropia

A entropia de um conjunto de treinamento pode ser definida como sendo o grau de impureza desse conjunto.
Dado um conjunto *S*, com instâncias pertencentes à classe *i*, com probabilidade p_i, temos:

$$Entropia\ (S) \equiv \sum_{i=1}^{c} - p_i \log_2 p_i$$



No nosso exemplo existem apenas duas classes: "Jogar Ténis" (positivo, ***p+***) ou "Não Jogar Ténis" (negativo, ***p-***).

Onde:
S é o conjunto de exemplos de treino;
p+ é a porção de exemplos positivos;
p- é a porção de exemplos negativos;
A entropia é dada por:

$$Entropia(S) = -p_+ \log_2 p_+ - p_- \log_2 p_-$$

S = [9+ , 5-]

$Entropia(S) =$ -9/14 * $\log_2$(9/14) + -5/14 * $\log_2$(5/14) = **0,940**

Ganho

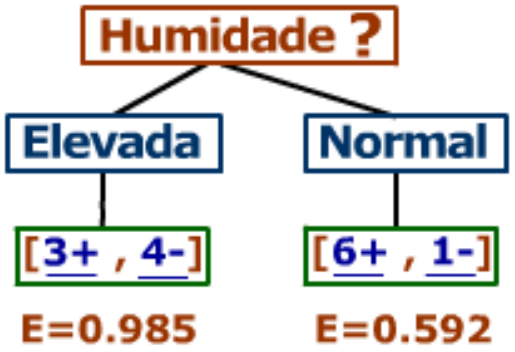
Redução esperada na entropia do conjunto *S* causada pela partição dos exemplos de acordo com o atributo A. O ganho é dado por:

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

Exemplo:

$$Ganho(S, Humidade) = 0{,}94 - \left(\frac{7}{14} * 0{,}985 + \frac{7}{14} * 0{,}592\right) = 0{,}151$$

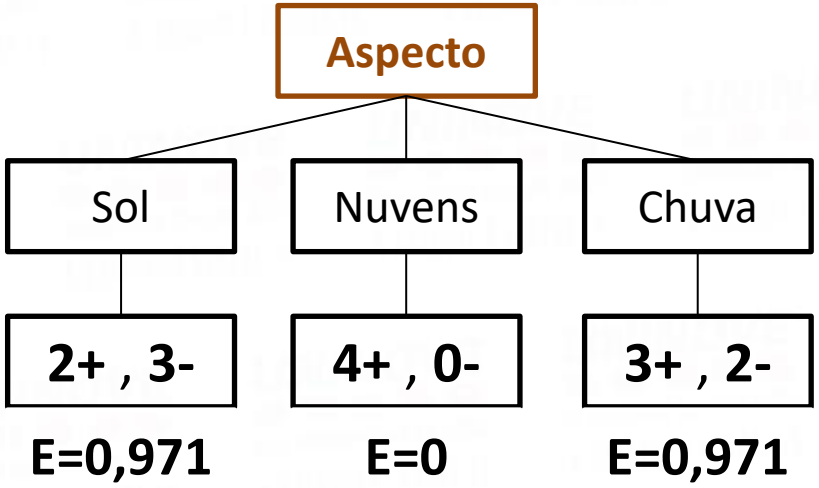
Dia	Aspecto	Temp.	Humidade	Vento	Jogar Ténis
D1	Sol	Quente	Elevada	Fraco	Não
D2	Sol	Quente	Elevada	Forte	Não
D3	Nuvens	Quente	Elevada	Fraco	Sim
D4	Chuva	Ameno	Elevada	Fraco	Sim
D5	Chuva	Fresco	Normal	Fraco	Sim
D6	Chuva	Fresco	Normal	Forte	Não
D7	Nuvens	Fresco	Normal	Fraco	Sim
D8	Sol	Ameno	Elevada	Fraco	Não
D9	Sol	Fresco	Normal	Fraco	Sim
D10	Chuva	Ameno	Normal	Forte	Sim
D11	Sol	Ameno	Normal	Forte	Sim
D12	Nuvens	Ameno	Elevada	Forte	Sim
D13	Nuvens	Quente	Normal	Fraco	Sim
D14	Chuva	Ameno	Elevada	Forte	Não

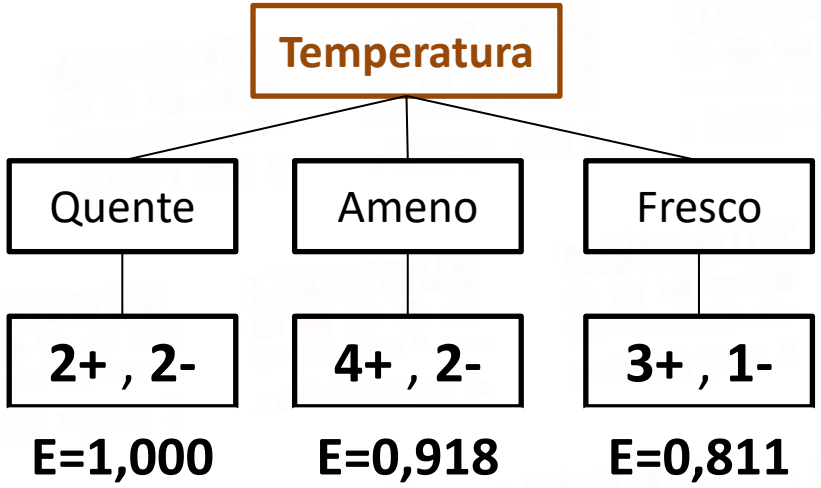


Algoritmo ID3

Primeiro passo – Definir a raiz da árvore (por qual atributo deve-se iniciar a árvore?)

- Calcular a Entropia do conjunto de treinamento **(0,94)**
- Calcular o ganho de cada um dos atributos (**Aspecto**, **Temperatura**, **Humidade**, **Vento**)





$$Ganho(S, Aspecto) = 0{,}94 - \left(\frac{5}{14} * 0{,}971 + \frac{4}{14} * 0 + \frac{5}{14} * 0{,}971\right) = 0{,}247$$

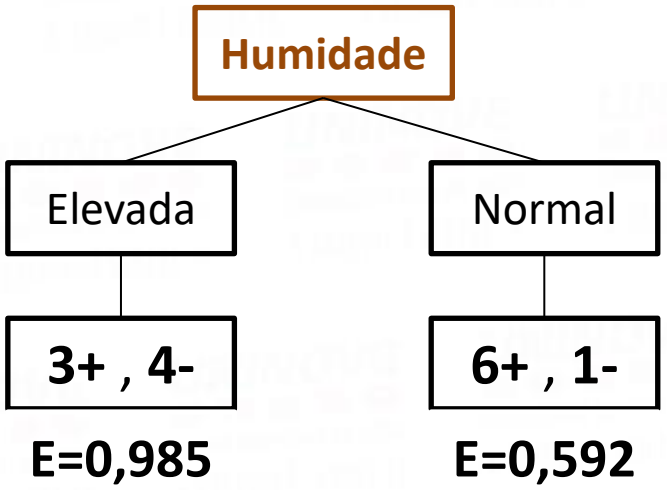
$$Ganho(S, Temperatura) = 0{,}94 - \left(\frac{4}{14} * 1{,}0 + \frac{6}{14} * 0{,}918 + \frac{4}{14} * 0{,}811\right) = 0{,}029$$

UNINOVE
Universidade Nove de Julho
A UNINOVE É SEMPRE 10

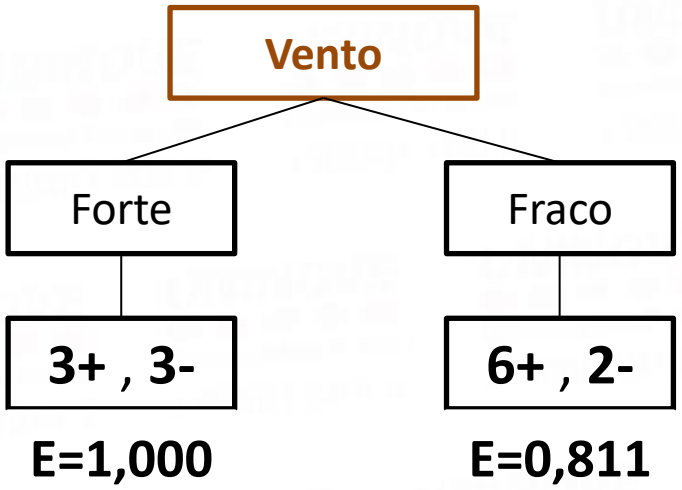
Algoritmo ID3

Primeiro passo – Definir a raiz da árvore (por qual atributo deve-se iniciar a árvore?)

- Calcular a Entropia do conjunto de treinamento **(0,94)**
- Calcular o ganho de cada um dos atributos (**Aspecto**, **Temperatura**, **Humidade**, **Vento**)



$$Ganho(S, Humidade) = 0{,}9 - \left(\frac{7}{14} * 0{,}985 + \frac{7}{14} * 0{,}592\right) = 0{,}152$$



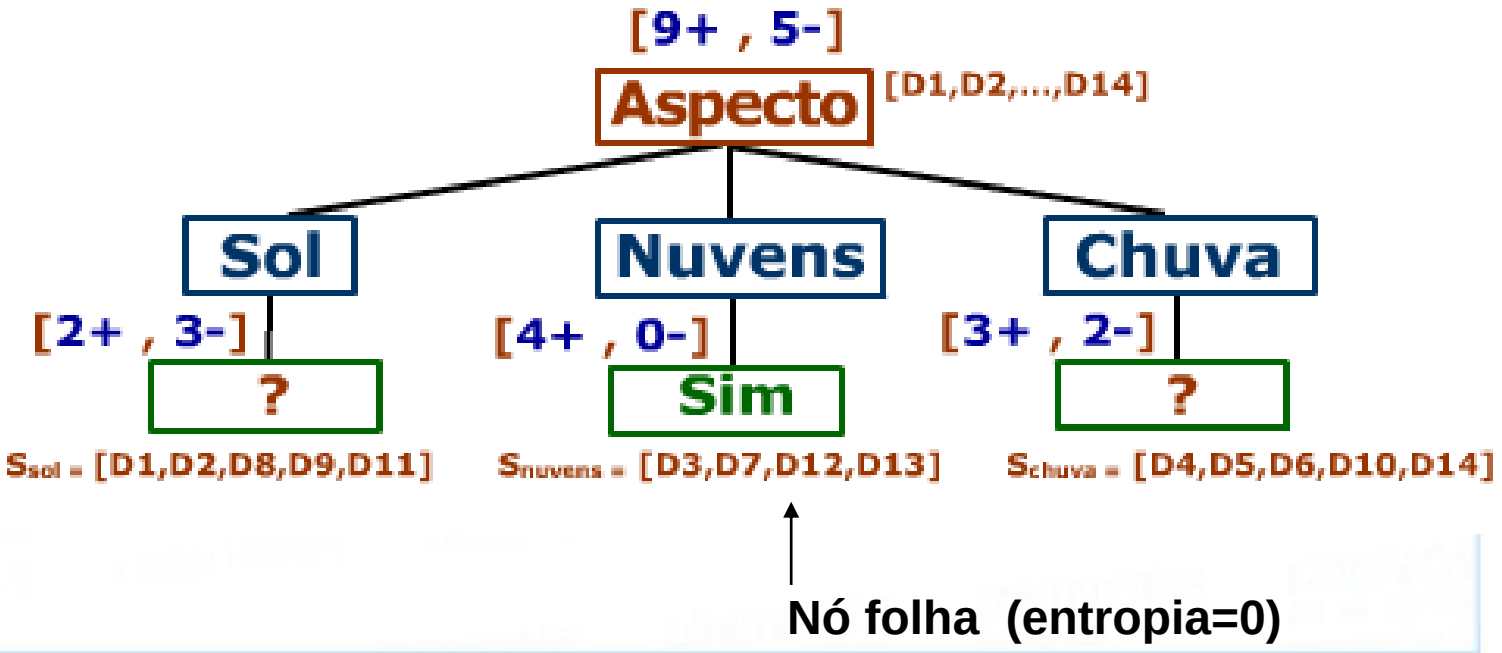
$$Ganho(S, Vento) = 0{,}9 - \left(\frac{6}{14} * 1{,}0 + \frac{8}{14} * 0{,}811\right) = 0{,}048$$

Algoritmo ID3

Primeiro passo – Definir a raiz da árvore (por qual atributo deve-se iniciar a árvore?)

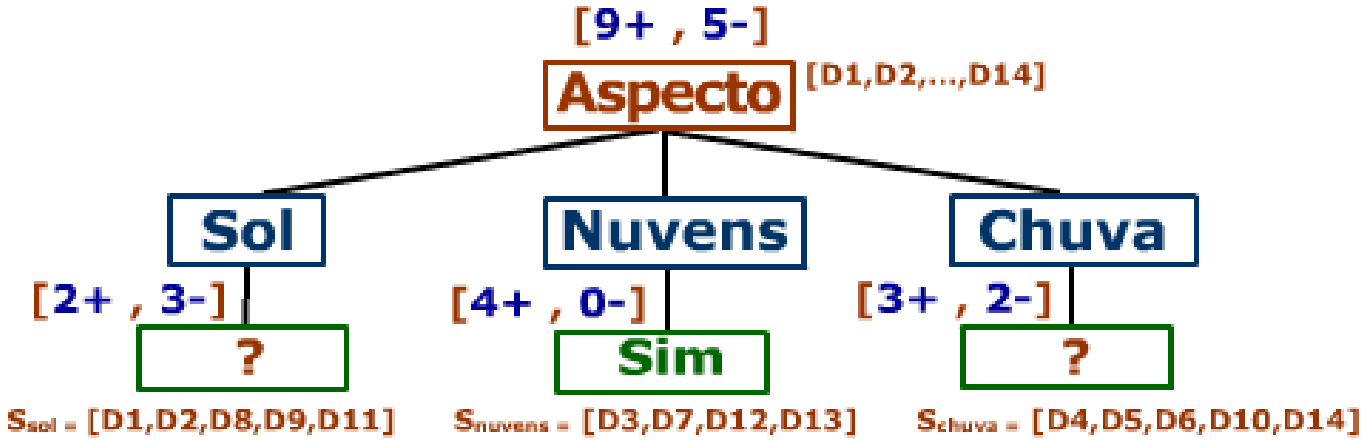
Resposta: Calculando o ganho para todos os atributos, verificamos que **Aspecto** tem maior ganho. Logo, ele será o primeiro escolhido (raiz da árvore)

$$\text{MAX}\begin{pmatrix}\text{Ganho(S,Humidade)} = 0.151\\ \text{Ganho(S,Vento)} = 0.048\\ \text{Ganho(S, Aspecto)} = 0.247\\ \text{Ganho(S,Temp)} = 0.029\end{pmatrix} = \text{Ganho(S, Aspecto)}$$



Algoritmo ID3

Próximos passos – A partir daqui, o atributo **"Aspecto"** não entra novamente na árvore. Então, cada um dos nós com entropia ≠ 0 é tomado como raiz (ou nó pai) e o processo continua levando em conta o atributo de maior ganho ainda não colocado na árvore**...**



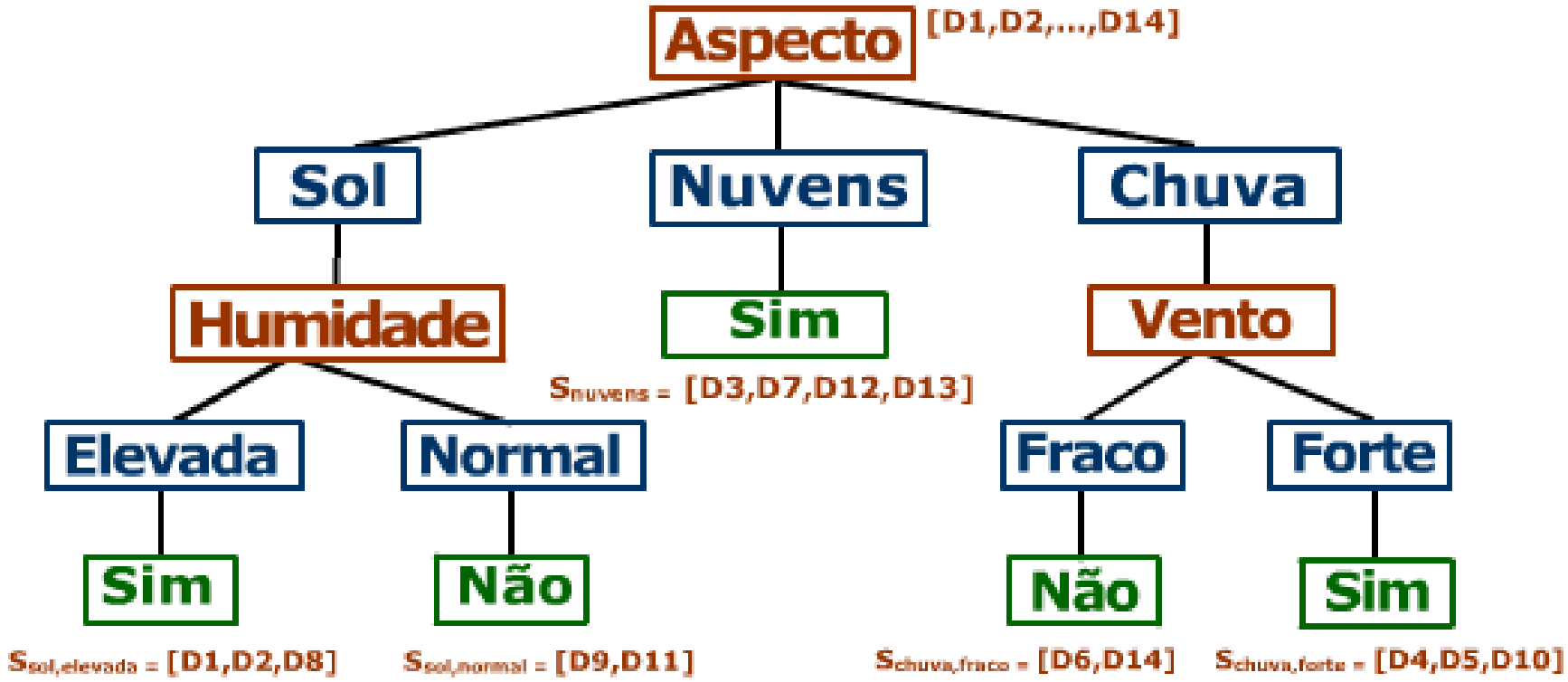
$$\text{MAX}\left(\begin{array}{l}\text{Ganho}(S_{sol}, \text{Humidade}) = 0.970-(3/5)0.0 - 2/5(0.0) = 0.970\\ \text{Gain}(S_{sol}, \text{Temp.}) = 0.970-(2/5)0.0 - 2/5(1.0)-(1/5)0.0 = 0.570\\ \text{Gain}(S_{sol}, \text{Vento}) = 0.970 = -(2/5)1.0 - 3/5(0.918) = 0.019\end{array}\right) =$$

$$= \text{Ganho}(S_{sol}, \text{Humidade})$$

Algoritmo ID3

Quando a árvore está concluída?

Quando entropia em todos os nós for nula. No exemplo da AD para decidir sobre **"Jogar Tênis"**, obtêm-se a seguinte árvore de decisão:

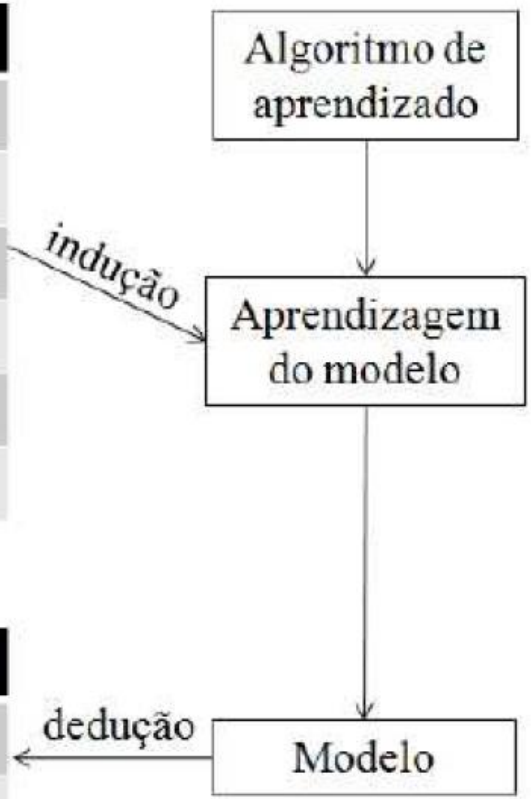


Exercício 1

Aplicação de AD para descoberta de regras visando o auxílio ao diagnóstico médico

Exemplo	Febre	Enjôo	Manchas	Dor	Diagnóstico
T1	sim	sim	pequenas	sim	doente
T2	não	não	grandes	não	saudável
T3	sim	sim	pequenas	não	saudável
T4	sim	não	grandes	sim	doente
T5	sim	não	pequenas	sim	saúdavel
T6	não	não	grandes	sim	doente

Exemplo	Febre	Enjôo	Manchas	Dor	Diagnóstico
N1	não	não	pequenas	sim	?
N2	sim	sim	grandes	sim	?



Exercício 2

Aplicação de AD para descoberta de regras visando o auxílio à análise de crédito (Base de Análise de Crédito do repositório da UCI - https://archive.ics.uci.edu/ml/datasets/Credit+Approval)

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1	**A1**	**A2**	**A3**	**A4**	**A5**	**A6**	**A7**	**A8**	**A9**	**A10**	**A11**	**A12**	**A13**	**A14**	**A15**	**Classe**
2	b	30.83	0	u	g	w	v	1.25	t	t	1	f	g	202	0	+
3	a	44.33	0	u	g	c	v	2.5	t	f	0	f	g	0	0	+
4	b	18.83	0	u	g	q	v	0.665	f	f	0	f	g	160	1	-
5	b	21.17	0	u	g	c	v	0.5	f	f	0	t	s	0	0	-
6	b	20	0	u	g	d	v	0.5	f	f	0	f	g	144	0	-
7	b	16.25	0	y	p	aa	v	0.25	f	f	0	f	g	60	0	-
8	b	23.17	0	u	g	cc	v	0.085	t	f	0	f	g	0	0	+
9	b	41.33	0	u	g	c	bb	15	t	f	0	f	g	0	0	+
10	b	23.08	0	u	g	k	v	1	f	t	11	f	s	0	0	-
11	b	20.42	0	?	?	?	?	0	f	f	0	f	p	?	0	-
12	b	23.17	0	?	?	?	?	0	f	f	0	f	p	?	0	+
13	a	25.58	0	?	?	?	?	0	f	f	0	f	p	?	0	+
14	b	34.58	0	?	?	?	?	0	f	f	0	f	p	?	0	-
15	b	37.58	0	?	?	?	?	0	f	f	0	f	p	?	0	+
16	a	71.58	0	?	?	?	?	0	f	f	0	f	p	?	0	+

Weka[1]

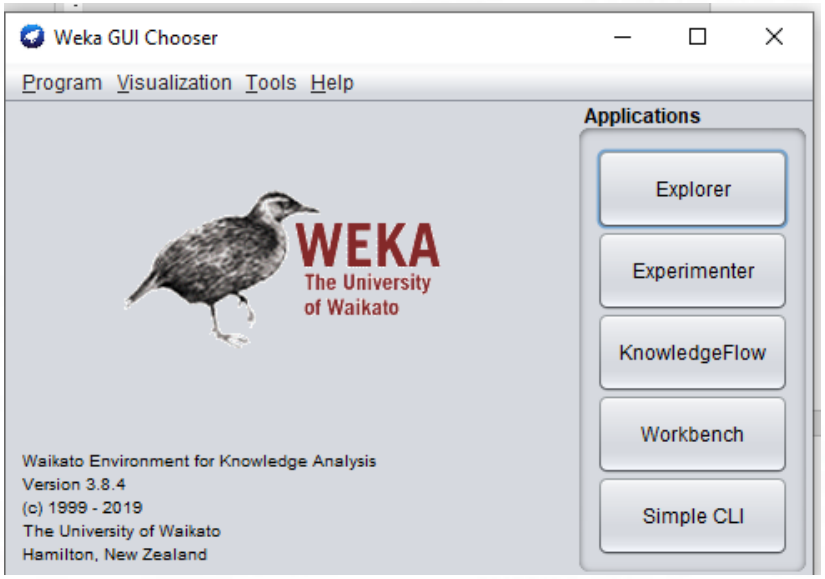
O Weka (*Waikato Environment for Knowledge Analysis*) é uma ferramenta que agrega algoritmos de inteligência artificial e de aprendizagem de máquina.

Trata-se de uma das mais populares ferramentas de mineração de dados em ambiente acadêmico.

O Weka teve seu desenvolvimento iniciado em 1993, usando Java, na Universidade de Waikato, Nova Zelândia sendo adquirido posteriormente por uma empresa no final de 2006. O Weka encontra-se licenciado ao abrigo da General Public License sendo portanto possível estudar e alterar o respectivo código fonte.

1. https://www.cs.waikato.ac.nz/ml/weka/

Weka



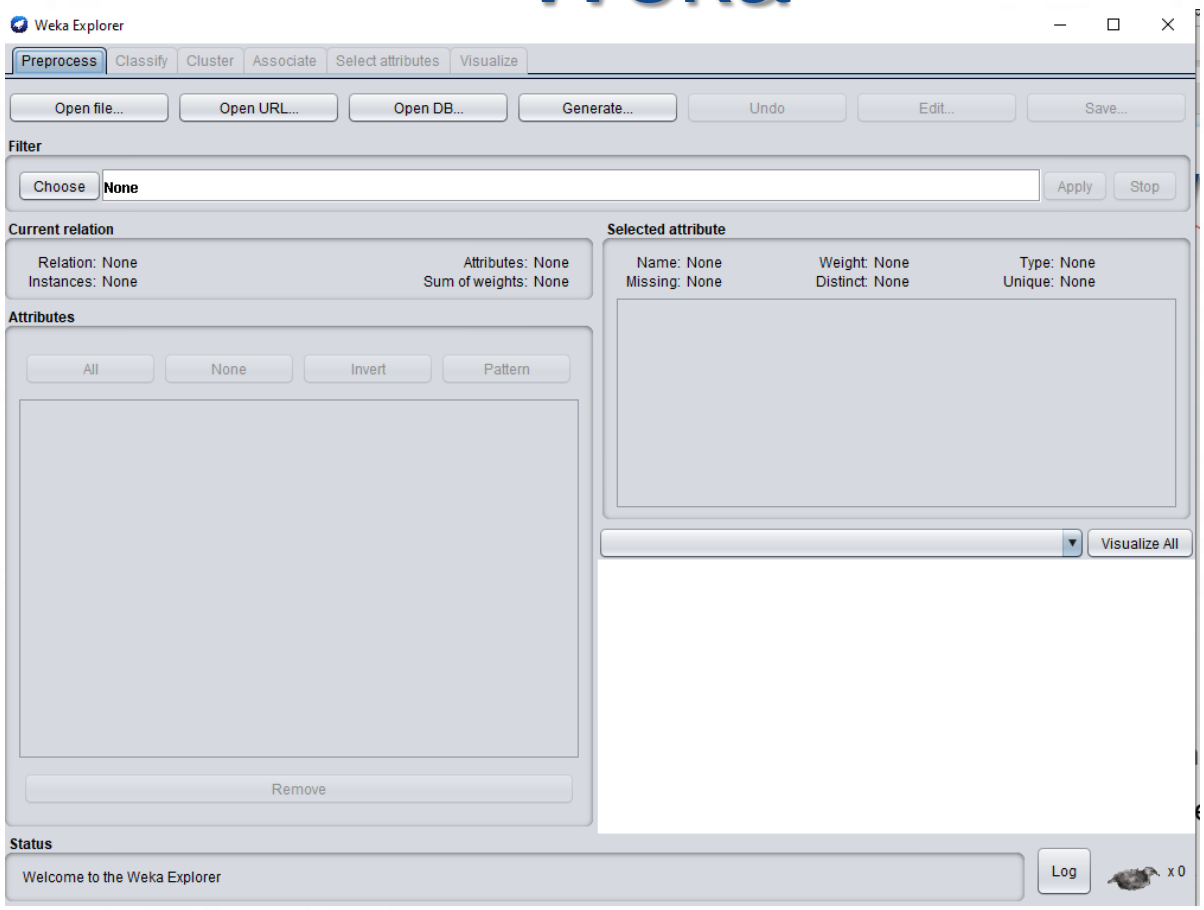
Explorer: classificadores/filtros

Experimenter: comparação de performances de algoritmos (o processamento pode levar horas, dias, semanas ou meses...)

KnowledgeFlow: Interface gráfica para o “desenho” dos métodos

Simple CLI: Interface em linha de comando

Weka



Em “Explorer” estão os algoritmos de pré-processamento, classificação, clusterização, associação, seleção de atributos e visualização de dados

Formato de arquivo no Weka (.arff)

```
@relation jogar_tenis

@attribute aspecto {chuva, nuvens, sol}
@attribute temperatura {ameno, fresco, quente}
@attribute humidade {elevada, normal}
@attribute vento {forte, fraco}
@attribute jogar {nao, sim}

@data
sol      , quente , elevada , fraco , nao
sol      , quente , elevada , forte , nao
nuvens   , quente , elevada , fraco , sim
chuva    , ameno  , elevada , fraco , sim
chuva    , fresco , normal  , fraco , sim
chuva      fresco   normal    forte   nao
nuvens   , fresco , normal  , fraco , sim
sol      , ameno  , elevada , fraco , nao
```

Referências bibliográficas

BRAMER,M. (2007). Principles of data mining. Springer, London.

MITCHELL, T. M. Machine learning. McGraw Hill: 1997.

QUINLAN, J. R. (1993). C4.5: programs for machine learning. Morgan Kaufmann

Publishers Inc., San Francisco, CA, USA.

QUINLAN, J. R. (1986). Induction of decision trees. Machine Learning, 1(1):81-106.