Árvores de Decisão

As Árvores de Decisão são um dos modelos mais práticos e mais usados em inferência indutiva. Este método representa funções como árvores de decisão. Estas árvores são treinadas de acordo com um conjunto de treino (exemplos previamente classificados) e posteriormente, outros exemplos são classificados de acordo com essa mesma árvore. Para a construção destas árvores são usados algoritmos como o ID3, ASSISTANT e C4.5.

Este método de classificação pode ser facilmente compreendido através de um exemplo.

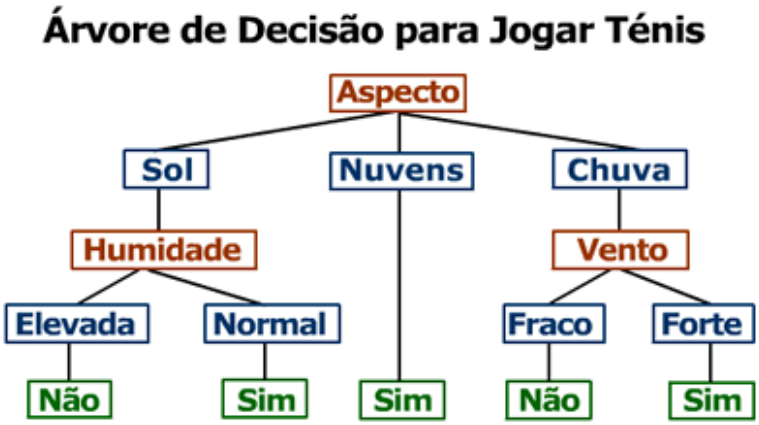
Supondo que o objectivo é decidir se vou **Jogar Ténis**. Para tal, há que ter em conta certos parâmetros do ambiente, como o **Aspecto** do Céu, a **Temperatura**, a **Humidade** e o **Vento**. Cada um destes atributos tem vários valores. Por exemplo para a temperatura pode estar **Ameno**, **Fresco** ou **Quente**. A decisão **Sim** (ir jogar ténis) ou **Não** (não ir jogar ténis) é o resultado da classificação.

Para construir a Árvore de Decisão de Jogar Ténis são tidos em conta exemplos (dias) passados.

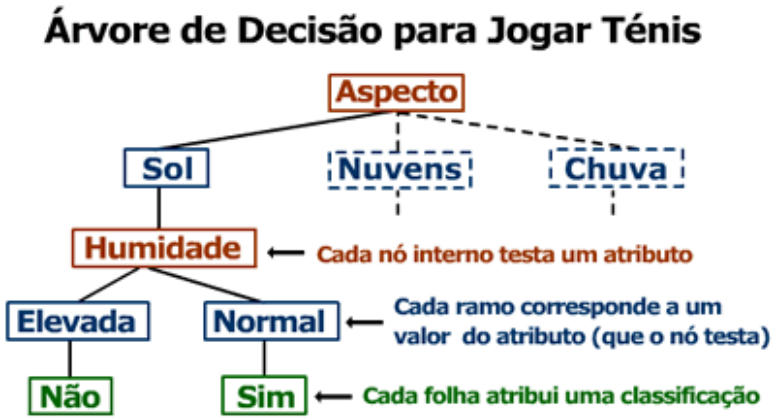
Exemplos de Treino

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

Através destes exemplos é possível construir a seguinte árvore de decisão:



A relação entre os elementos da árvore (nós e folhas) e os atributos, valores e classificações pode ser entendida na seguinte imagem:

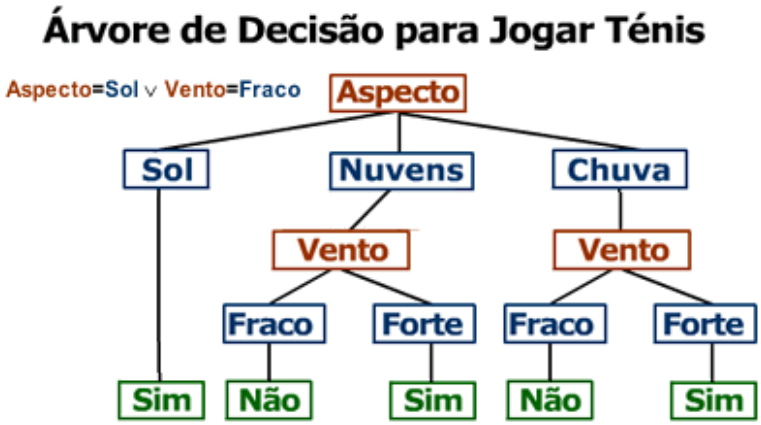


A classificação de um exemplo de acordo com a esta árvore é feita da seguinte forma:

O atributo **Aspecto** tem o valor **Sol** e a **Humidade** tem o valor **Elevada**. O exemplo é classificado com **Não**, ou seja quando esteve sol e humidade elevada não se jogou ténis. Os atributos **Temperatura** e **Vento** não são considerados, pois são desnecessário para classificar este exemplo.

Com Árvores de Decisão é possível representar a conjunção e disjunção de atributos. A árvore de decisão que representa a classificação para os dias em que o **Aspecto** é **Sol** e que o **Vento** está **Franco** encontra-se na seguinte figura.

A árvore de decisão que representa os dias em que o **Aspecto** é **Sol** ou o **Vento** está **Franco** é dada por:



Deve-se considera o uso de árvores de decisão em situações onde:

- As instâncias são descritas por pares atributo-valor;
- A função objecto (alvo) é de valor discreto;
- É pode ser necessário hipótese disjuntas;
- Os exemplos de treino poderão ter erro (*noise*);
- Faltam valores nos atributos;
- Exemplos:
  - Diagnósticos médicos;
  - Análises de risco de crédito;
  - Classificação de objectos para um manipulador de *robot* (Tan1993).

Algoritmo ID3

O algoritmo ID3 (inductive decision tree) é dos mais utilizados para a construção de árvores de decisão. Este algoritmo segue os seguintes passos:

1. Começar com todos os exemplos de treino;
2. Escolher o teste (atributo) que melhor divide os exemplos, ou seja agrupar exemplos da mesma classe ou exemplos semelhantes;
3. Para o atributo escolhido, criar um nó filho para cada valor possível do atributo;
4. Transportar os exemplos para cada filho tendo em conta o valor do filho;
5. Repetir o procedimento para cada filho não "puro". Um filho é puro quando cada atributo X tem o mesmo valor em todos os exemplos.

Coloca-se então, uma pergunta muito importante:

Como saber qual o melhor atributo a escolher?

Para lidar com esta escolha são introduzidos dois novos conceitos, a **Entropia** e o **Ganho**.

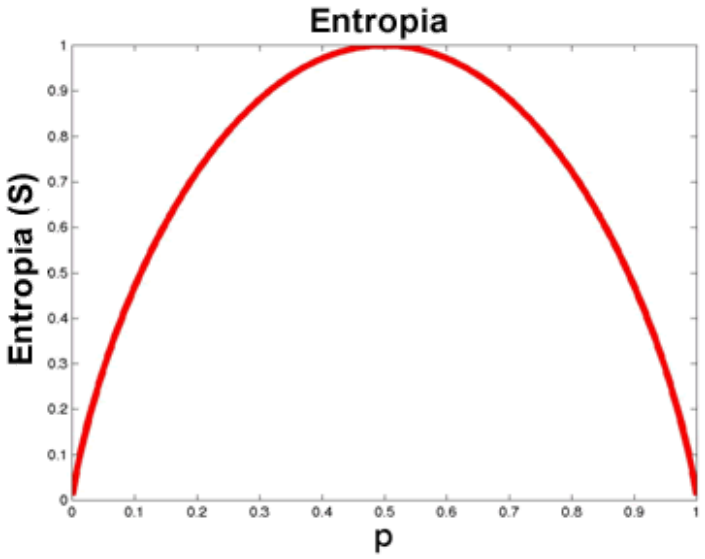
Entropia

A entropia de um conjunto pode ser definida como sendo o grau de pureza desse conjunto. Este conceito emprestado pela Teoria da Informação define a medida de "falta de informação", mais precisamente o número de bits necessários, em média, para representar a informação em falta, usando codificação óptima.

Dado um conjunto *S*, com instâncias pertencentes à classe *i*, com probabilidade p_i, temos:

$$Entropia(S) = \sum p_i \log_2 p_i \quad (1)$$

No exemplo apenas existem duas classes de classificação, ou seja, "Jogar Ténis" (positivo, +) ou "Não Jogar Ténis" (negativo, -). Assim sendo, o valor da entropia varia de acordo com o gráfico:



Onde:

- S é o conjunto de exemplo de treino;
- p_+ é a porção de exemplos positivos;
- p_- é a porção de exemplos negativos;
- A entropia é dada pelo desdobramento da equação 1

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

Ganho

O ganho (*gain*) é define a redução na entropia. *Ganho(S,A)* significa a redução esperada na entropia de S, ordenando pelo atributo A. O ganho é dado pela seguinte equação:

$$Ganho(S,A) = Entropia(S) - \sum_{v \in values(A)} \frac{|S_v|}{|S|} \cdot Entropia(S_v)$$

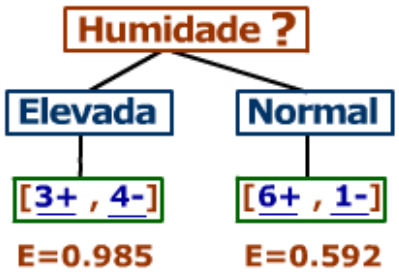
Escolha do Melhor Atributo

Para responder à pergunta anterior, "Como escolher o melhor atributo" é usado o ganho. Em cada iteração do algoritmo é escolhido o atributo que apresente uma maior ganho.

Para o primeiro passo são analisados todos os atributos, começando pela **Humidade**, por exemplo:

S = [9+ , 5-]
E = 0.940 = -9/14 $\log_2$ 9/14 – 5/14 $\log_2$ 5/14

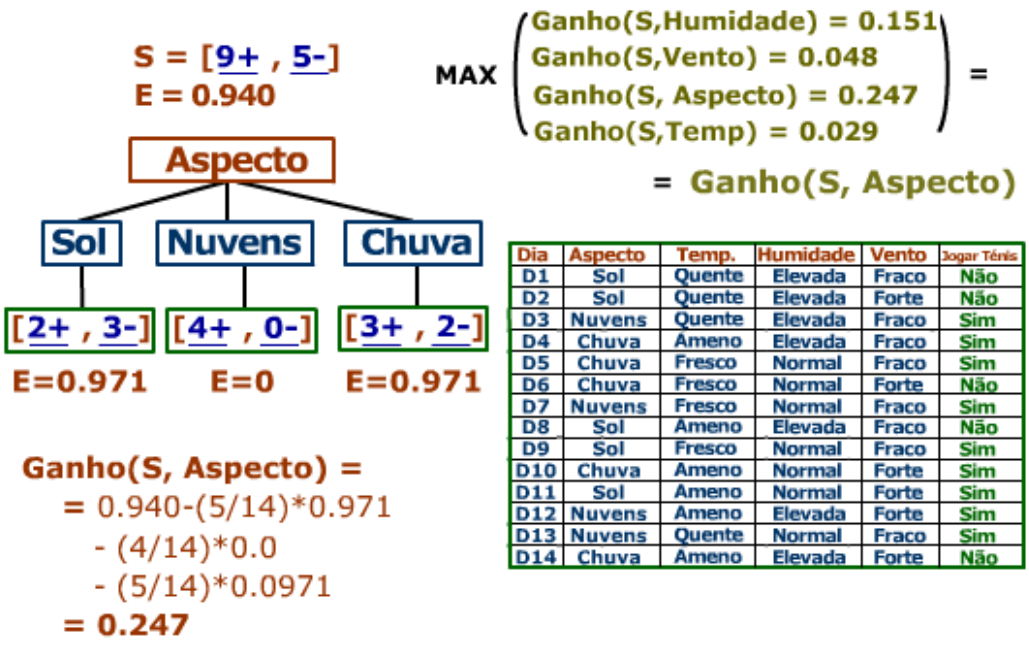
Ganho(S,Humidade) = 0.151



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

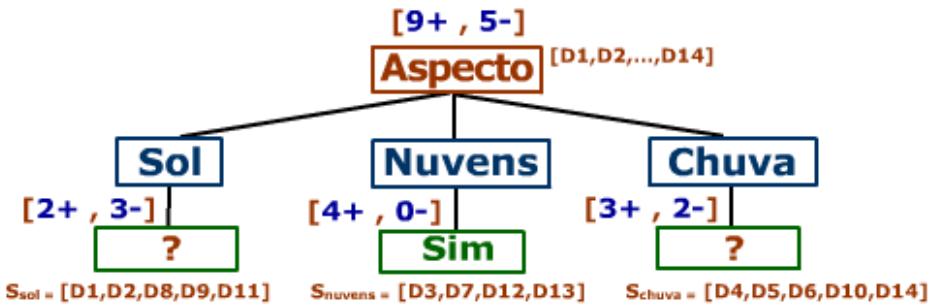
Ganho(S,Humidade) =
= 0.940-(7/14)*0.985
– (7/14)*0.592
= 0.151

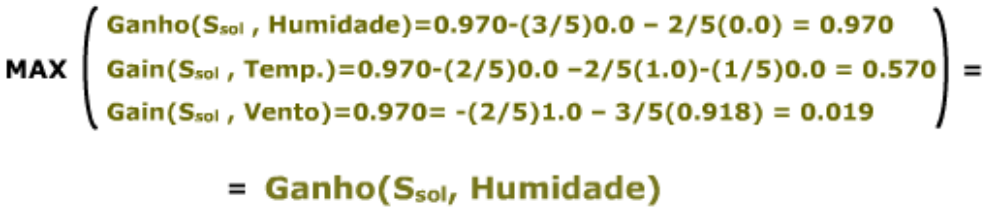
Calculando o ganho para todos os atributos, verificamos que o tem maior ganho é o **Aspecto.**



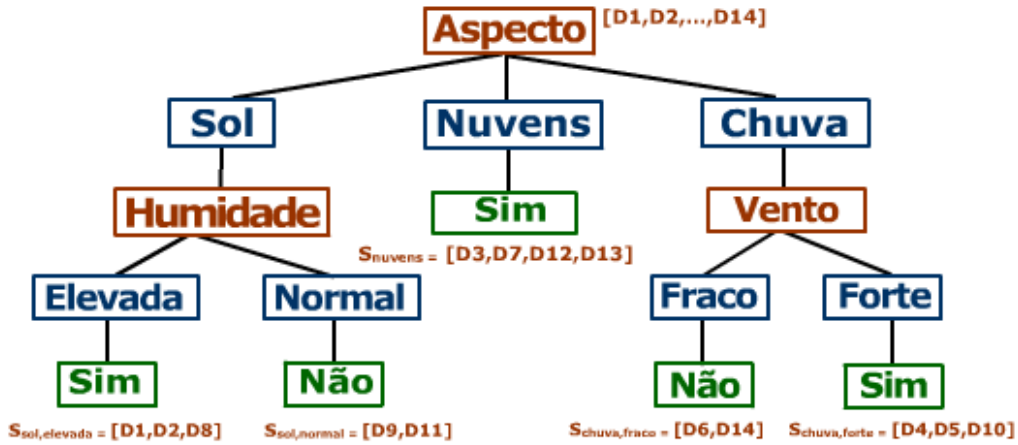
Dia	Aspecto	Temp.	Humidade	Vento	Jogar Ténis
D1	Sol	Quente	Elevada	Fraco	Não
D2	Sol	Quente	Elevada	Forte	Não
D3	Nuvens	Quente	Elevada	Fraco	Sim
D4	Chuva	Ameno	Elevada	Fraco	Sim
D5	Chuva	Fresco	Normal	Fraco	Sim
D6	Chuva	Fresco	Normal	Forte	Não
D7	Nuvens	Fresco	Normal	Fraco	Sim
D8	Sol	Ameno	Elevada	Fraco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

No próximo passo o atributo **Aspecto** já não é tido em conta:





Quando todos em todos os nós a entropia for nula, o algoritmo para e obtêm-se a seguinte árvore de decisão:



A seguinte animação auxilia a compreender o funcionamento do ID3.

No próximo exemplo é usado uma árvore de decisão para calcular o estado de activação do gene CLN2, dado os níveis de expressão de outros genes.