

Aprendizado de Máquina

Inteligência Artificial

Msc Jose Netto
jose.netto@anhembi.br

Agenda

Aprendizado de máquina

Paradigmas

Processo de análise de dados

Exemplo ID3

Exemplo

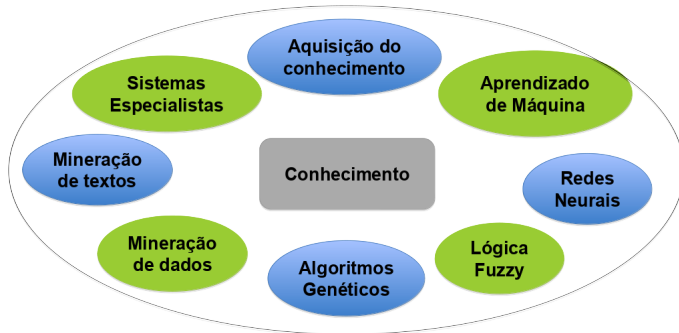


Aprendizado de Máquina

Aprendizado de máquina

Introdução

- **Técnicas chave em Inteligência Artificial (IA)**



Conceitos envolvidos

- **Armazenar conhecimento**
 - Representação → criar um modelo → algoritmo parametrizado
- **Aplicar conhecimento para resolver problemas**
 - Raciocínio (mecanismo de inferência)
 - aplicação do modelo → inferir novos dados
- **Adquirir novos conhecimentos**
 - Aprender com dados que são apresentados.

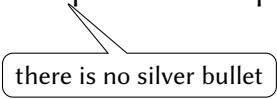
Definição

- **O que é Aprendizado de Máquina?**
 - técnicas computacionais capazes de adquirir conhecimento de forma automática.
 - capazes de tomar decisões baseado em experiências acumuladas ...
 - por meio de soluções bem sucedida de problemas anteriores.

Aprendizado de Máquina

- **Característica do aprendizado de máquina**

- É uma ferramenta poderosa para aquisição automática de conhecimento,
- contudo, não existe um único algoritmo perfeito para todos os problemas.



there is no silver bullet

Formas de aquisição de conhecimento

- **Duas formas diferentes de aquisição de conhecimento**
 - Dedução
 - Parte de um raciocínio geral para o específico.
 - Indução
 - Parte de um raciocínio específico para o geral.

Formas de aquisição de conhecimento

- **Dedução:**
 - Humanos usam raciocínio dedutivo para deduzir nova informação ...
 - a partir de informação relacionada logicamente.

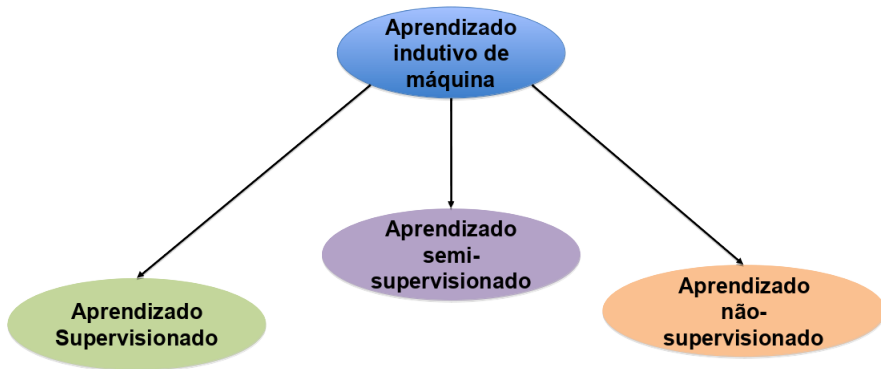
Formas de aquisição de conhecimento

- **Indução:**

- é uma forma de inferência lógica.
- permite obter conclusões a partir de um conjunto de dados.
- aprendizado acontece efetuando-se inferência indutiva sobre os dados apresentados.

Aprendizado de Máquina

- Hierarquia do aprendizado



Tipos de aprendizado

- **Aprendizado não supervisionado**

- visa-se determinar grupos ou classes por meio das informações dos dados.
- Não existe classe/rótulo pré-definida para nenhum dos dados;
- pode ser utilizado sobre um único conjunto de dados.

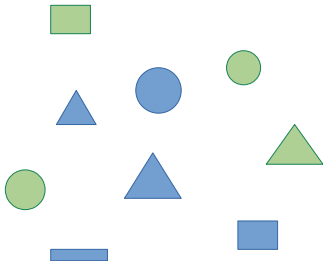
Tarefas

- **Agrupamento:**

- o indutor analisa os dados buscando padrões e/ou características em comum ...
- que sejam indícios para formação de grupos ou clusters
- Ex: identificar grupos de pessoas com padrões similares de compras.

Agrupamento

- Analise a imagem abaixo



- Como poderíamos agrupar os elementos para formar exatamente 2 grupos?
- E se fossem 3 grupos? Como ficaria o agrupamento?
- e 4 grupos?

Tarefas

- **Redução de dimensionalidade:**
 - reduz o número de atributos (colunas) dos dados.
 - avalia-se a qualidade de cada atributo e seleciona-se os melhores.
 - Ex: remoção de dados desnecessários da base de dados para melhorar performance.

Tipos de aprendizado

- **Aprendizado Supervisionado**

- Necessidade de um conjunto de treinamento e um conjunto de teste.
- Os dados do conjunto de treinamento são acompanhados por “rótulos”
- os rótulos indicam a classe a que eles pertencem.
- Novos dados são classificados com base no conjunto de treinamento.

Aprendizado Supervisionado

- **Objetivo**

- Construir um classificador (indutor);
- o indutor deve ser capaz de determinar a classe de novos dados ...
- através de dados previamente rotulados.

Tarefas

- **Classificação:**

- Determinar rótulos com valores discretos.
- Ex. Detecção de spam, identificação de transação falsa
- Ex. Identificação automática de revisão de produto como boa ou má revisão.

Tarefas

- **Regressão:**

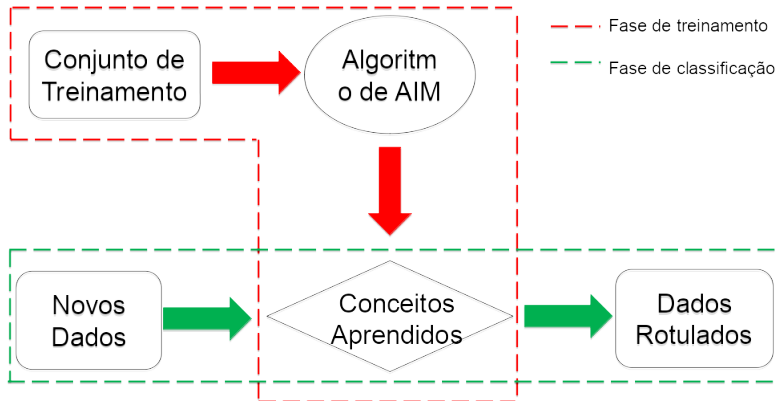
- Determinar rótulos com valores contínuos.
- Ex. Prever o faturamento do próximo período baseado nas vendas anteriores
- Ex. Prever valores faltantes em uma base de dados.

Aprendizado Supervisionado

- **Fase de treinamento e classificação**

- Basicamente as técnicas de aprendizado supervisionados possuem duas fases:
- Fase de treinamento
 - o padrão dos dados são aprendidos através de um conjunto de treinamento;
 - sendo que, os dados já possuem rótulos.
- Fase de classificação
 - o classificador treinado é utilizado para prever novos dados

Aprendizado de Máquina





Aprendizado de Máquina

Paradigmas

Paradigmas

- **Simbólico:**

- Aprende construindo representações simbólicas do problema através de exemplos.
- Constrói um modelo interpretável.
- Técnicas: Árvores de decisão, regras semânticas

Paradigmas

- **Memorização (Instance-Based ou Lazy Learning)**
 - Aprende através da utilização de uma “memória” (conjunto de dados)
 - Técnicas: kNN

Paradigmas

- **Conexionista**
 - Cria conexões entre elementos para aprender
 - Técnicas: Redes neurais artificiais

Paradigmas

- **Genético**

- Baseia-se em técnicas evolucionárias de aprendizagem
- analogia à teoria de Darwin e ao comportamento da natureza
- Técnicas: Algoritmos evolutivos, Programação genética

Paradigmas

- **Estatístico**
 - Utiliza-se de conceitos da estatística para construir um modelo de aprendizagem
 - Técnicas: Aprendizado Bayesiano, regressões



Aprendizado de Máquina

Processo de análise de dados

Definições

- **Modelo de classificação/agrupamento**
 - programa parametrizado através dos dados
 - visa extrair um bom classificador, por exemplo ...
 - a partir de um conjunto de dados rotulados.

Definições

- **Classe (Rótulo):**
 - atributo especial no qual se pretende aprender a fazer previsões a respeito.
- **Bias:**
 - qualquer preferência de uma hipótese sobre a outra.

Definições

- **Modo de aprendizado:**
 - Não incremental
 - todo conjunto de treinamento presente no aprendizado desde o início.
 - Incremental
 - quando novos dados de treinamento são adicionados.

Base de dados

- **Exemplo de base de dados**

- Flor iris
- Identificar a espécie da planta (setosa, versicolor ou virgínica)
- Utilizar o comprimento e largura da sépala e da pétala



Base de dados

- **Existem diversas definições para os conjuntos de dados.**
 - Considere o conjunto de dados estruturado abaixo.

comprimento sépala	largura sépala	comprimento pétala	largura pétala	classe
5,1	3,5	1,4	0,2	Iris-setosa
4,9	3	1,4	0,2	Iris-setosa
4,7	3,2	1,3	0,2	Iris-setosa
7	3,2	4,7	1,4	Iris-versicolor
6,4	3,2	4,5	1,5	Iris-versicolor
6,9	3,1	4,9	1,5	Iris-versicolor
6,3	3,3	6	2,5	Iris-virginica
5,8	2,7	5,1	1,9	Iris-virginica
7,1	3	5,9	2,1	Iris-virginica

Base de dados

- **Linhas**

- Cada linha pode ser chamada de dado, exemplo, registro, instância, vetor de característica, dentre outros.
- Cada linha representa uma flor diferente.

comprimento sépala	largura sépala	comprimento pétala	largura pétala	classe
5,1	3,5	1,4	0,2	Iris-setosa
4,9	3	1,4	0,2	Iris-setosa
4,7	3,2	1,3	0,2	Iris-setosa
7	3,2	4,7	1,4	Iris-versicolor
6,4	3,2	4,5	1,5	Iris-versicolor
6,9	3,1	4,9	1,5	Iris-versicolor
6,3	3,3	6	2,5	Iris-virginica
5,8	2,7	5,1	1,9	Iris-virginica
7,1	3	5,9	2,1	Iris-virginica

Base de dados

- **Colunas**

- Cada coluna pode ser chamada de atributo ou variável.
- Cada coluna representa uma característica diferente da flor.

comprimento sépala	largura sépala	comprimento pétala	largura pétala	classe
5,1	3,5	1,4	0,2	Iris-setosa
4,9	3	1,4	0,2	Iris-setosa
4,7	3,2	1,3	0,2	Iris-setosa
7	3,2	4,7	1,4	Iris-versicolor
6,4	3,2	4,5	1,5	Iris-versicolor
6,9	3,1	4,9	1,5	Iris-versicolor
6,3	3,3	6	2,5	Iris-virginica
5,8	2,7	5,1	1,9	Iris-virginica
7,1	3	5,9	2,1	Iris-virginica

Avaliação do Modelo

- **Avaliação do aprendizado**
 - Após o processo de aprendizado ...
 - precisamos avaliar a capacidade preditiva do modelo (indutor)
 - é necessária a utilização de um novo conjunto de dados;
 - também chamado de **conjunto de teste**.

Avaliação do Modelo

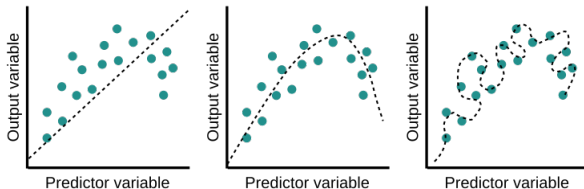
- **Conjunto de Teste**

- classificar os dados do conjunto de teste baseado no aprendizado do conjunto de treinamento
- O conjunto de teste é normalmente uma parte do conjunto original de dados ...
- que deve ser independente do conjunto de treinamento.

Avaliação do Modelo

- **Observe as três imagens**

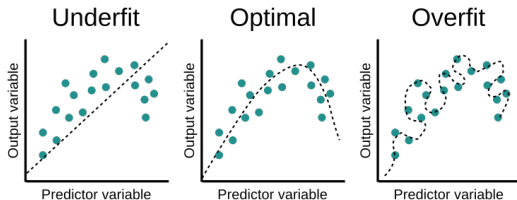
- cada uma delas representa um modelo diferente que aprendeu com os dados
- A linha pontilhada representa o modelo inferido
- Qual das três imagens melhor representa um modelo que aprendeu com os dados?



Avaliação do Modelo

- **Underfitting**

- modelo muito generalista;
- classificador não se adapta nem aos dados de treinamento.



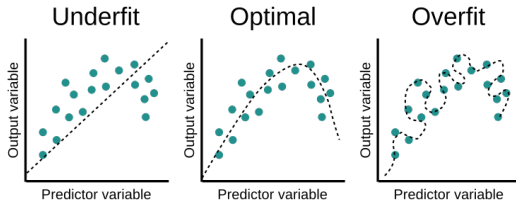
Avaliação do Modelo

- **Overfitting**

- modelo muito específico;
- classificador está muito ajustado aos dados de treinamento.

- **Problema**

- classificador é eficaz em classificar somente os dados de treinamento ...
- e inadequado para classificar outros exemplos de dados.



Avaliação do Modelo

- **Para evitar o overfitting:**
 - podemos criar um terceiro conjunto de dados;
 - também conhecido como conjunto de validação.
- **Conjunto de validação**
 - ajuda a avaliar a capacidade de generalização do classificador.
 - podemos “re-treinar” o modelo (classificador) para melhorar sua performance

Avaliação do Modelo

- **Validação cruzada (cross-validation)**
 - Técnica de validação para evitar overfitting.
 - especialmente interessante quando não temos um conjunto de teste e treinamento bem definidos.

Avaliação do Modelo

- **Ideia geral**

- Dividir os dados diversas vezes ...
- para estimar os erros de cada um dos classificadores induzidos.
- Selecionar o classificador com menor estimativa de erro.

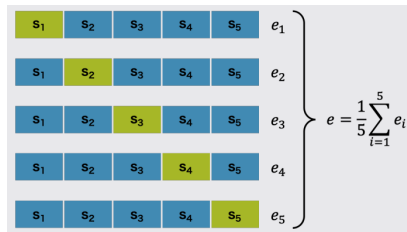
Validação Cruzada

- **Suponha uma base de dados de treinamento**
 - assuma que a base contenha 100 instâncias
 - ou seja, 100 exemplos de dados
 - podemos dividir essa base em subconjuntos de 20 instâncias cada
 - totalizando 5 subconjuntos

s_1	s_2	s_3	s_4	s_5
-------	-------	-------	-------	-------

Avaliação do Modelo

- **Exemplo de validação cruzada**
 - Vamos utilizar 4 subconjuntos para fazer treinamento ...
 - e o subconjunto restante para validar o modelo
 - Repetimos esse processo utilizando combinações diferentes



Observações

- **Analisando o modelo induzido**

- A qualidade do dados de treinamento implica na qualidade das regras do indutor.
- Não é possível descobrir algo que não esteja nos dados de treinamento.
- Seleção dos dados e dos atributos é fundamental neste processo.
- Importante o conhecimento e experiência do especialista.

Exemplo ID3

Algoritmo ID3

- **O ID3 (Iterative Dichotomizer 3)**
 - Algoritmo desenvolvido por John R. Quinlan;
 - algoritmo baseado em árvores de decisão.

Algoritmo ID3

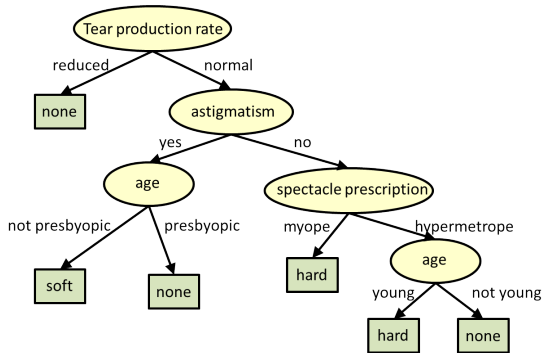
- **Árvores de decisão**

- podem ser representadas como conjuntos de regras SE ENTÃO (IF THEN).
- É um dos métodos de aprendizagem mais conhecidos.
- gera um modelo interpretável
- Aplicações:
 - Diagnóstico médico, análise de risco de crédito, mineração de dados, etc.

Algoritmo ID3

- **Características:**

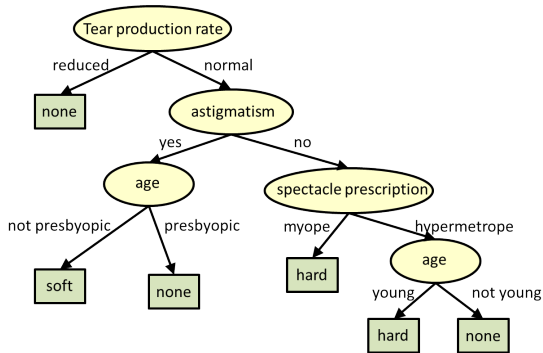
- Árvores de decisão classificam instâncias ordenando as árvores;
- começam da raiz até chegarem em alguma folha.



Algoritmo ID3

- **Características:**

- Cada **nó** da árvore especifica o **teste** de algum atributo da instância.
- Cada **ramo** parte de um nó correspondente a um dos **valores possíveis dos atributos**.



Algoritmo ID3

- **Instâncias**

- irão representar os exemplos da base de dados
- são descritas por um conjunto fixo de atributos
- cada coluna irá representar uma característica
- Ex: Temperatura e seus valores (quente, frio, etc).

Algoritmo ID3

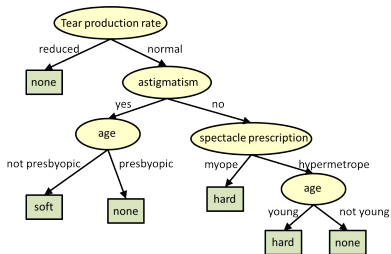
- **Classes/Rótulos**

- Elemento a ser predito/classificado.
 - tem **valores discretos de saída**
- Classificação booleana
 - (Yes ou No) para cada exemplo
- Classificação Multiclasses
 - mais de duas possibilidades (Uva, Laranja, Melancia, etc).

Algoritmo ID3

- **Inicialização do algoritmo**

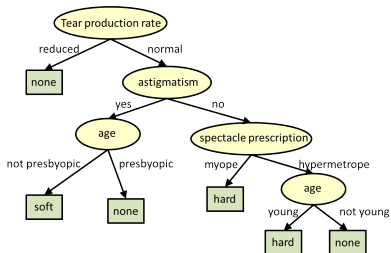
- Como determinar o atributo raiz da árvore?
 - cada atributo deve ser avaliado usando um teste estatístico
 - para determinar quão bem ele sozinho classifica os dados de treinamento.



Algoritmo ID3

- **Seleciona-se o melhor atributo**

- O **melhor atributo** é selecionado e usado como teste na **raiz** da árvore.
- Um descendente do nó raiz é então criado para cada valor possível deste atributo;
- e os dados de treinamento são ordenados para o nó descendente apropriado.



Algoritmo ID3

- **Iteração**

- O processo é repetido para cada um dos atributos restantes;
- objetivo é selecionar o melhor atributo para testar naquele ponto da árvore.
- Busca gulosa:
 - algoritmo nunca recua para reconsiderar escolhas prévias.

Algoritmo ID3

```
1 ID3 (Exemplos, atributo_classificador, Atributos)
2   cria a folha inicial da arvore Root
3   se todos os objetos são positivos em Exemplos
4     retornar uma única folhar com saída positiva
5   se todos os objetos são negativos em Exemplos
6     retornar uma única folha com a saída negativa
7   se Atributos estiver vazio
8     retornar uma unica folha com a saída mais comum na base
9   senao:
10    A = atributo de Atributos que melhor classifique Exemplos
11    Root = A
12    para cada valor de A:
13      adicionar uma nova folha a A com o valor correspondente a v
14      sendo x = resultado de Exemplos com os objetos que contenham v como atributo
15      se x gerado estiver vazio
16        adicionar uma folha a Root como valor mais comum da base
17    senao
18      ID3(x, atributo_classificador, Atributos-A)
19  retornar root
```

Algoritmo ID3

- **Determinando o melhor atributo:**
 - Precisamos selecionar qual atributo testar em cada nó da árvore.
 - Utilizaremos uma métrica que seja útil para selecionar o melhor atributo.
- **Qual métrica poderíamos utilizar?**

Algoritmo ID3

- **Ganho de informação:**
 - mede quão bem um atributo separa os dados de treinamento ...
 - de acordo com a classe a ser predita.
 - Para calcular o Ganho de Informação precisaremos calcular a **Entropia**

Algoritmo ID3

- **Entropia**

- Caracteriza a (im)pureza de uma coleção arbitrária de dados.
- A entropia é usada para estimar a aleatoriedade da variável a prever (classe).
- Entropia = 1
 - significa que não há tendência para nenhuma classe.
- Entropia = 0
 - significa que possui tendência máxima, ou seja, qualquer atributo classifica o conjunto de dados.

Entropia

- **Cálculo**

- Dado uma base de dados S contendo dados positivos(pos) e negativos(neg);
 - ou seja, duas possíveis classes
- a entropia de S relativa a estas duas classes é:
-

$$Entropia(S) = -(p_{pos}) \times \log_2(p_{pos}) - (p_{neg}) \times \log_2(p_{neg})$$

p_{pos} é a probabilidade do dado ser positivo em S .

p_{neg} é a probabilidade do dado ser negativo em S .

Algoritmo ID3

- **Ganho de informação** \rightarrow **Ganho** (S, A)

- redução esperada na entropia devido a partição dos dados...
- utilizando o atributo A como separador;
- ou seja, utilizando o atributo A como nó da árvore.
-

$$\text{Ganho}(S, A) = \text{Entropia}(S) - \left(\sum_{v \in \text{valores}(A)} \frac{S_v}{|S|} \times \text{Entropia}(S_v) \right)$$

Diagram illustrating the components of the formula:

- S_v : quantidade elementos em S com valor v
- $|S|$: quantidade total de elementos
- $v \in \text{valores}(A)$: para cada possível valor de A na base de dados



Aprendizado de Máquina

Exemplo

Exemplo

- Considere a base de dados a seguir

COLOR	SIZE	ACT	AGE	INFLATED
yellow	small	stretch	adult	T
yellow	small	strectch	adult	T
yellow	small	stretch	chils	F
yellow	small	dip	adult	F
yellow	small	dip	child	F
yellow	large	stretch	adult	T
yellow	large	stretch	adult	T
yellow	large	stretch	child	F
yellow	large	dip	adult	F
yellow	large	dip	child	F
purple	small	stretch	adult	T
purple	small	stretch	adult	T
purple	small	stretch	child	F
purple	small	dip	adult	F
purple	small	dip	child	F
purple	large	stretch	adult	T
purple	large	stretch	adult	T
purple	large	stretch	child	F
purple	large	dip	adult	F
purple	large	dip	child	F

Algoritmo ID3

- **Determinando a entropia da base de dados (S)**

- Base de dados

- 20 elementos no total
- 8 positivos (T)
- 12 negativos (F)

$$Entropia(S) = -(p_{pos}) \times \log_2(p_{pos}) - (p_{neg}) \times \log_2(p_{neg})$$

- Entropia da base de dados (S):

$$Entropia(S) = -\left(\frac{8}{20}\right) \times \log_2\left(\frac{8}{20}\right) - \left(\frac{12}{20}\right) \times \log_2\left(\frac{12}{20}\right)$$

$$Entropia(S) \approx 0,97$$

Algoritmo ID3

- **Ganho de informação**
 - Mede a efetividade de um atributo em classificar um conjunto de treinamento.
 - Quão bom um atributo é para classificar um conjunto de treinamento
- **Ganho de Informação de um atributo A:**
 - mede a redução na Entropia ...
 - causada pelo particionamento de dados de acordo com este atributo

Algoritmo ID3

- **Por exemplo:**

- considere atributo “age”
 - contendo os valores “adult” e “child” do conjunto de treinamento ballon.
- Agora considere que:
 - 8 dos dados positivos e 4 dos dados negativos são definidos por “age” = “adult”
 - (12 no total)
 - 8 dados negativos e nenhum positivo definidos por “age” = “child”

Algoritmo ID3

- **Ganho de informação (age)**

- O ganho de Informação ao selecionar o atributo “age” para a raiz:

- $S = [8_{pos}, 12_{neg}]$;

- $S_{adult} = [8_{pos}, 4_{neg}]$;

- $S_{child} = [0_{pos}, 8_{neg}]$

COLOR	SIZE	ACT	AGE	INFLATED
yellow	small	stretch	adult	T
yellow	small	stretch	adult	T
yellow	small	stretch	child	F
yellow	small	dip	adult	F
yellow	small	dip	child	F
yellow	large	stretch	adult	T
yellow	large	stretch	adult	T
yellow	large	stretch	child	F
yellow	large	dip	adult	F
yellow	large	dip	child	F
purple	small	stretch	adult	T
purple	small	stretch	adult	T
purple	small	stretch	child	F
purple	small	dip	adult	F
purple	small	dip	child	F
purple	large	stretch	adult	T
purple	large	stretch	adult	T
purple	large	stretch	child	F
purple	large	dip	adult	F
purple	large	dip	child	F

Algoritmo ID3

- Entropia do atributo Age:

$$Entropia(adult) = - \left(\frac{8}{12} \right) \times \log_2 \left(\frac{8}{12} \right) - \left(\frac{4}{12} \right) \times \log_2 \left(\frac{4}{12} \right) = 0.918$$

$$Entropia(child) = - \left(\frac{0}{8} \right) \times \log_2 \left(\frac{0}{8} \right) - \left(\frac{8}{8} \right) \times \log_2 \left(\frac{8}{8} \right) = 0$$

$$Ganho(S, A) = 0.97 - \left(\left(\frac{12}{20} \right) \times 0.918 + \left(\frac{8}{20} \right) \times 0 \right) = 0.42$$

Algoritmo ID3

- **Ganho de informação Act**

- O Ganho de Informação ao selecionar o atributo “act” para a raiz:

- $S = [8_{pos}, 12_{neg}]$

- $S_{stretch} = [8_{pos}, 4_{neg}]$

- $S_{dip} = [0_{pos}, 8_{neg}]$

COLOR	SIZE	ACT	AGE	INFLATED
yellow	small	stretch	adult	T
yellow	small	stretch	adult	T
yellow	small	stretch	child	F
yellow	small	dip	adult	F
yellow	small	dip	child	F
yellow	large	stretch	adult	T
yellow	large	stretch	adult	T
yellow	large	stretch	child	F
yellow	large	dip	adult	F
yellow	large	dip	child	F
purple	small	stretch	adult	T
purple	small	stretch	adult	T
purple	small	stretch	child	F
purple	small	dip	adult	F
purple	small	dip	child	F
purple	large	stretch	adult	T
purple	large	stretch	adult	T
purple	large	stretch	child	F
purple	large	dip	adult	F
purple	large	dip	child	F

Algoritmo ID3

- Entropia do atributo Act:

—

$$Entropia(stretch) = - \left(\frac{8}{12} \right) \times \log_2 \left(\frac{8}{12} \right) - \left(\frac{4}{12} \right) \times \log_2 \left(\frac{4}{12} \right) = 0.918$$

$$Entropia(dip) = - \left(\frac{0}{8} \right) \times \log_2 \left(\frac{0}{8} \right) - \left(\frac{8}{8} \right) \times \log_2 \left(\frac{8}{8} \right) = 0$$

$$Ganho(S, A) = 0.97 - \left(\left(\frac{12}{20} \right) \times 0.918 + \left(\frac{8}{20} \right) \times 0 \right) = 0.42$$

Algoritmo ID3

- **Ganho de informação Size**

- O Ganho de Informação ao selecionar o atributo “size” para a raiz:

- $S = [8_{pos}, 12_{neg}]$;

- $S_{small} = [4_{pos}, 6_{neg}]$;

- $S_{large} = [4_{pos}, 6_{neg}]$.

COLOR	SIZE	ACT	AGE	INFLATED
yellow	small	stretch	adult	T
yellow	small	stretch	adult	T
yellow	small	stretch	child	F
yellow	small	dip	adult	F
yellow	small	dip	child	F
yellow	large	stretch	adult	T
yellow	large	stretch	adult	T
yellow	large	stretch	child	F
yellow	large	dip	adult	F
yellow	large	dip	child	F
purple	small	stretch	adult	T
purple	small	stretch	adult	T
purple	small	stretch	child	F
purple	small	dip	adult	F
purple	small	dip	child	F
purple	large	stretch	adult	T
purple	large	stretch	adult	T
purple	large	stretch	child	F
purple	large	dip	adult	F
purple	large	dip	child	F

Algoritmo ID3

- **Entropia do atributo Size:**

$$Entropia(small) = - \left(\frac{4}{10} \right) \times \log_2 \left(\frac{4}{10} \right) - \left(\frac{6}{10} \right) \times \log_2 \left(\frac{6}{10} \right) = 0,97$$

$$Entropia(large) = - \left(\frac{4}{10} \right) \times \log_2 \left(\frac{4}{10} \right) - \left(\frac{6}{10} \right) \times \log_2 \left(\frac{6}{10} \right) = 0,97$$

$$Ganho(S, A) = 0,97 - \left(\left(\frac{10}{20} \right) \times 0,97 + \left(\frac{10}{20} \right) \times 0,97 \right) = 0$$

Algoritmo ID3

- **Ganho de informação do atributo Color**

- O Ganho de Informação ao selecionar o atributo “color” para a raiz.

- $S = [8_{pos}, 12_{neg}]$;

- $S_{yellow} = [4_{pos}, 6_{neg}]$;

- $S_{purple} = [4_{pos}, 6_{neg}]$.

COLOR	SIZE	ACT	AGE	INFLATED
yellow	small	stretch	adult	T
yellow	small	stretch	adult	T
yellow	small	stretch	child	F
yellow	small	dip	adult	F
yellow	small	dip	child	F
yellow	large	stretch	adult	T
yellow	large	stretch	adult	T
yellow	large	stretch	child	F
yellow	large	dip	adult	F
yellow	large	dip	child	F
purple	small	stretch	adult	T
purple	small	stretch	adult	T
purple	small	stretch	child	F
purple	small	dip	adult	F
purple	small	dip	child	F
purple	large	stretch	adult	T
purple	large	stretch	adult	T
purple	large	stretch	child	F
purple	large	dip	adult	F
purple	large	dip	child	F

Algoritmo ID3

- Entropia do atributo Color:

$$Entropia(yellow) = - \left(\frac{4}{10} \right) \times \log_2 \left(\frac{4}{10} \right) - \left(\frac{6}{10} \right) \times \log_2 \left(\frac{6}{10} \right) = 0,97$$

$$Entropia(purple) = - \left(\frac{4}{10} \right) \times \log_2 \left(\frac{4}{10} \right) - \left(\frac{6}{10} \right) \times \log_2 \left(\frac{6}{10} \right) = 0,97$$

$$Ganho(S, A) = 0,97 - \left(\left(\frac{10}{20} \right) 0,97 + \left(\frac{10}{20} \right) 0,97 \right) = 0$$

Algoritmo ID3

- **Ganho de Informação:**

- Age = 0,41997309
- Color = 0,0
- Act = 0,41997309
- size = 0,0

Algoritmo ID3

- **Agora temos a raiz**

- Criamos nós filhos a partir da raiz de acordo com os possíveis valores assumidos pelo atributo “age”.

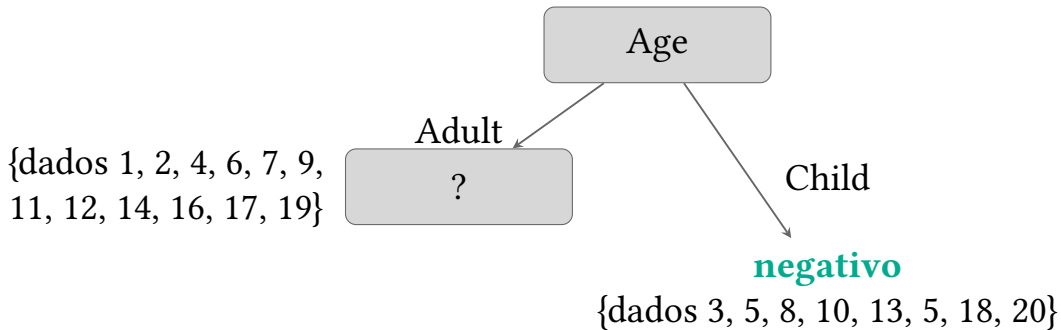
- **Demais ramos**

- Devemos proceder da mesma maneira para os demais ramos que surgem a partir da raiz.
- Em cada ramo consideramos somente os dados nele contidos.
 - Desde que haja divergência entre as classes de saída, ou seja, há classes diferentes para tal atributo.

Algoritmo ID3

- **Determinando o nó raiz**

- Um dos ramos não tem divergência entre as classes de saída;
- ou seja, entropia é igual a zero:



Nova base de dados S

estamos desconsideramos as
instâncias que contém age = child

COLOR	SIZE	ACT	AGE	INFLATED
yellow	small	stretch	adult	T
yellow	small	stretch	adult	T
yellow	small	dip	adult	F
yellow	large	stretch	adult	T
yellow	large	stretch	adult	T
yellow	large	dip	adult	F
purple	small	stretch	adult	T
purple	small	stretch	adult	T
purple	small	dip	adult	F
purple	large	stretch	adult	T
purple	large	stretch	adult	T
purple	large	dip	adult	F

Algoritmo ID3

- **Determinando a entropia do novo S**

- Assume-se a probabilidade de se pertencer a uma das duas classes (positiva ou negativa) do novo S.
- 8 positivos (T)
- 4 negativos (F)

- Logo, a Entropia desse conjunto é dada por:

–

$$Entropia(S) = - \left(\frac{8}{12} \right) \times \log_2 \left(\frac{8}{12} \right) - \left(\frac{4}{12} \right) \times \log_2 \left(\frac{4}{12} \right)$$

$$Entropia(S) \approx 0,92$$

Algoritmo ID3

- **Ganho de informação (Act)**

- O Ganho de Informação ao selecionar o atributo “act” para o nó:

- $S = [8_{pos}, 4_{neg}]$;
- $S_{stretch} = [8_{pos}, 0_{neg}]$;
- $S_{dip} = [0_{pos}, 4_{neg}]$.

COLOR	SIZE	ACT	AGE	INFLATED
yellow	small	stretch	adult	T
yellow	small	stretch	adult	T
yellow	small	dip	adult	F
yellow	large	stretch	adult	T
yellow	large	stretch	adult	T
yellow	large	dip	adult	F
purple	small	stretch	adult	T
purple	small	stretch	adult	T
purple	small	dip	adult	F
purple	large	stretch	adult	T
purple	large	stretch	adult	T
purple	large	dip	adult	F

Algoritmo ID3

- Entropia do atributo Act:

$$Entropia(stretch) = - \left(\frac{8}{8}\right) \times \log_2 \left(\frac{8}{8}\right) - \left(\frac{0}{8}\right) \times \log_2 \left(\frac{0}{8}\right) = 0$$

$$Entropia(dip) = - \left(\frac{0}{4}\right) \times \log_2 \left(\frac{0}{4}\right) - \left(\frac{4}{4}\right) \times \log_2 \left(\frac{4}{4}\right) = 0$$

$$Ganho(S, A) = 0,92 - \left(\left(\frac{8}{12}\right) \times 0 + \left(\frac{4}{12}\right) \times 0 \right) = 0,92$$

Algoritmo ID3

- **Ganho de informação (Size)**

- O Ganho de Informação ao selecionar o atributo “size” para o nó:

- $S = [8_{pos}, 4_{neg}]$;

- $S_{small} = [4_{pos}, 2_{neg}]$;

- $S_{large} = [4_{pos}, 2_{neg}]$.

COLOR	SIZE	ACT	AGE	INFLATED
yellow	small	stretch	adult	T
yellow	small	stretch	adult	T
yellow	small	dip	adult	F
yellow	large	stretch	adult	T
yellow	large	stretch	adult	T
yellow	large	dip	adult	F
purple	small	stretch	adult	T
purple	small	stretch	adult	T
purple	small	dip	adult	F
purple	large	stretch	adult	T
purple	large	stretch	adult	T
purple	large	dip	adult	F

Algoritmo ID3

- **Entropia do atributo (size):**

$$Entropia(small) = - \left(\frac{4}{6} \right) \times \log_2 \left(\frac{4}{6} \right) - \left(\frac{2}{6} \right) \times \log_2 \left(\frac{2}{6} \right) = 0,92$$

$$Entropia(large) = - \left(\frac{4}{6} \right) \times \log_2 \left(\frac{4}{6} \right) - \left(\frac{2}{6} \right) \times \log_2 \left(\frac{2}{6} \right) = 0,92$$

$$Ganho(S, A) = 0,92 - \left(\left(\frac{6}{12} \right) \times 0,92 + \left(\frac{6}{12} \right) \times 0,92 \right) = 0$$

Algoritmo ID3

- **Ganho de informação (Color)**

- O Ganho de Informação ao selecionar o atributo “color” para o nó:

- $S = [8_{pos}, 4_{neg}]$;

- $S_{yellow} = [4_{pos}, 2_{neg}]$;

- $S_{purple} = [4_{pos}, 2_{neg}]$.

COLOR	SIZE	ACT	AGE	INFLATED
yellow	small	stretch	adult	T
yellow	small	stretch	adult	T
yellow	small	dip	adult	F
yellow	large	stretch	adult	T
yellow	large	stretch	adult	T
yellow	large	dip	adult	F
purple	small	stretch	adult	T
purple	small	stretch	adult	T
purple	small	dip	adult	F
purple	large	stretch	adult	T
purple	large	stretch	adult	T
purple	large	dip	adult	F

Algoritmo ID3

- Entropia do atributo (color):

$$Entropia(yellow) = - \left(\frac{4}{6} \right) \times \log_2 \left(\frac{4}{6} \right) - \left(\frac{2}{6} \right) \times \log_2 \left(\frac{2}{6} \right) = 0,92$$

$$Entropia(purple) = - \left(\frac{4}{6} \right) \times \log_2 \left(\frac{4}{6} \right) - \left(\frac{2}{6} \right) \times \log_2 \left(\frac{2}{6} \right) = 0,92$$

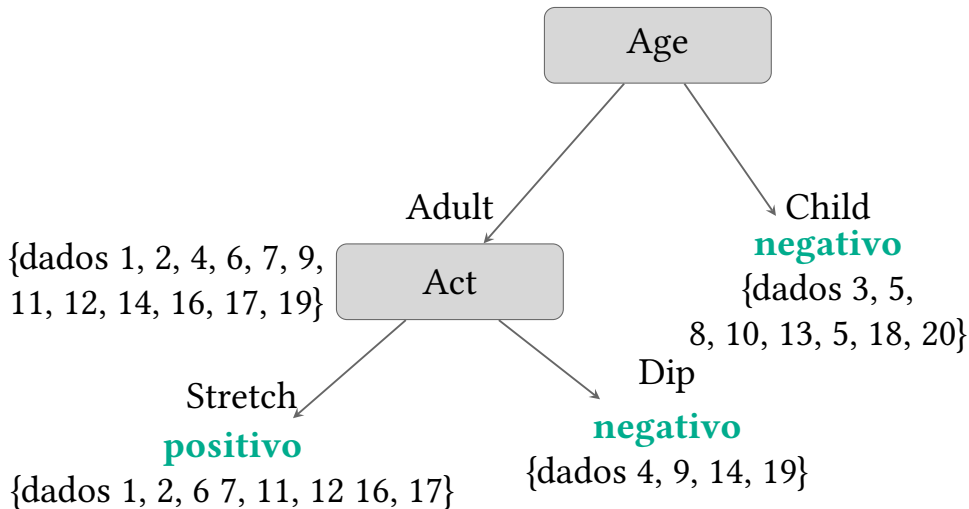
$$Ganho(S, A) = 0,92 - \left(\left(\frac{6}{12} \right) \times 0,92 + \left(\frac{6}{12} \right) \times 0,92 \right) = 0$$

Algoritmo ID3

- **Ganho de Informação:**

- Color = 0,0000000
- Act = 0,92
- size = 0,0000000

Algoritmo ID3



Algoritmo ID3

- **Observação**

- Atributos existentes incorporados acima de determinado nó ...
- não entram na avaliação de Ganho de Informação desse nó.
- Neste caso um novo nó será criado, mas o atributo “Act” não será mais avaliado.

Algoritmo ID3

- **Iterações**

- O algoritmo continua até que uma das duas condições seja satisfeita:
 - Todos atributos foram incluídos no caminho da raiz até as folhas
 - Dados de treinamento associados a um ramo apresentam o mesmo valor de saída (positivo ou negativo)

Obrigado

jose.netto@anhembi.br