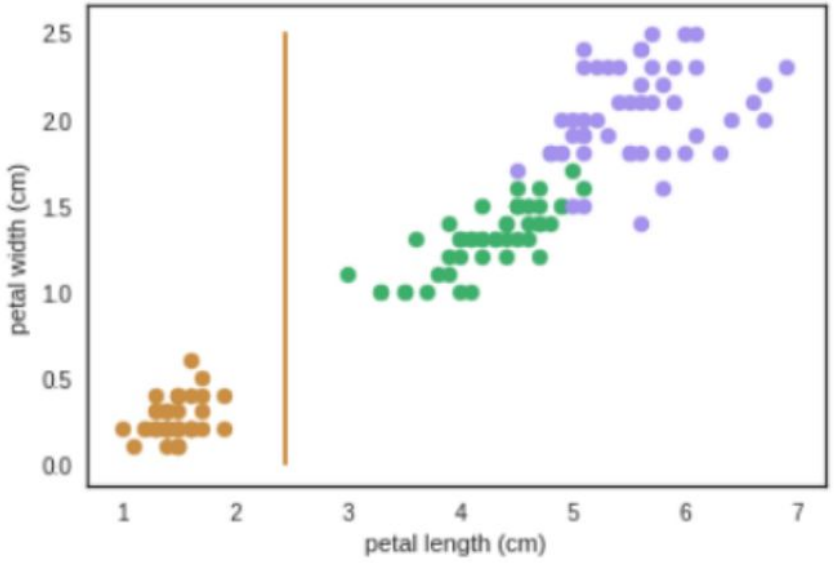
THAÍS GAUDENCIO DO RÊGO

ÁRVORE DE DECISÃO

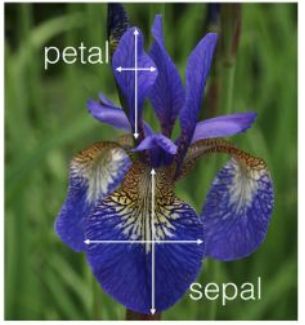
gaudenciothais@gmail.com

Exemplo - Classificação

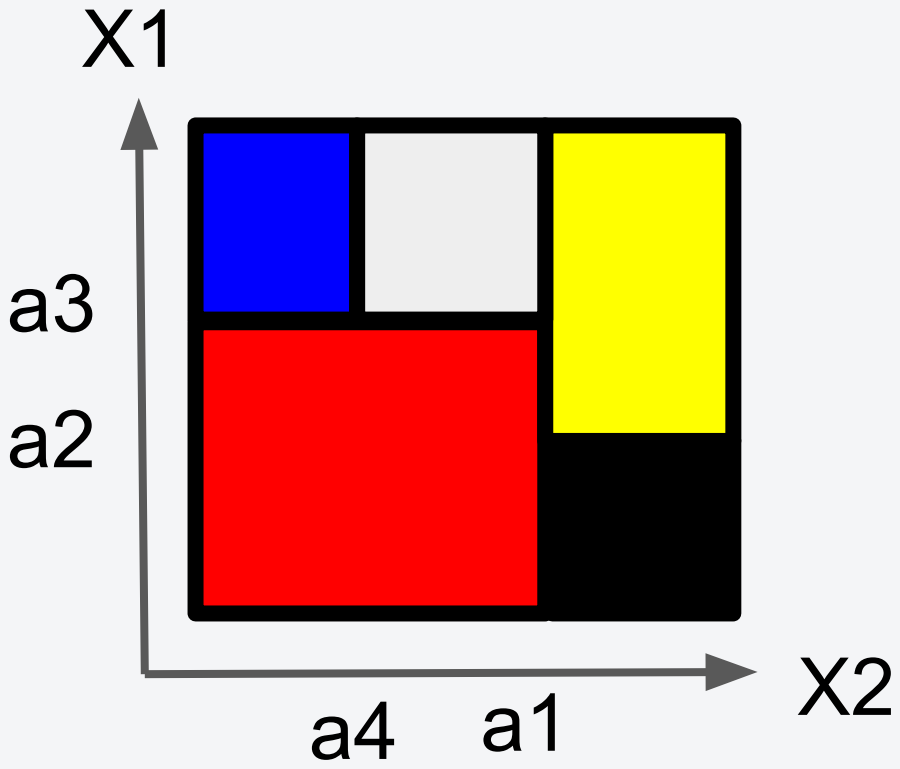
Tamanho (P)	Largura (P)	Tamanho (S)	Largura (S)	Espécie
5,1	3,5	1,4	0,2	Setosa
4,9	3,0	1,4	0,2	Setosa
7,0	3,2	4,7	1,4	Versicolor
6,4	3,2	4,5	1,5	Versicolor
6,3	3,3	6,0	2,5	Virginica
5,8	2,7	5,1	1,9	Virginica



petal length (cm) ≤ 2.45
samples = 150
value = [50, 50, 50]
class = setosa



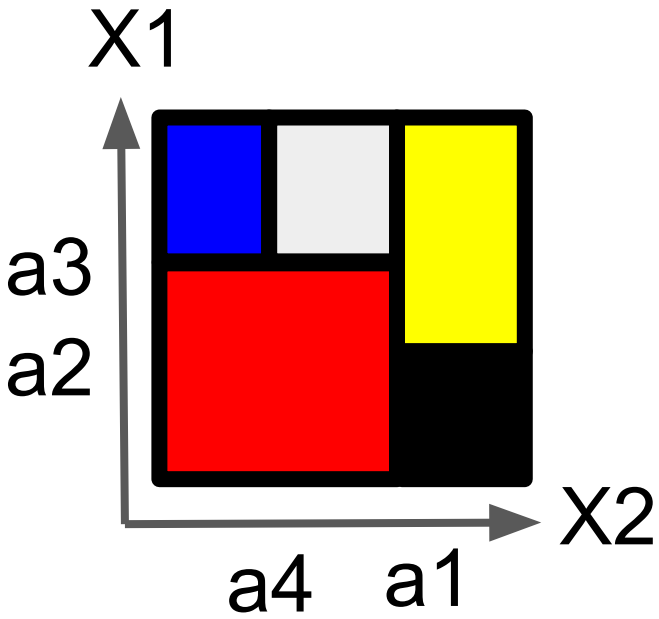
Árvore de Decisão

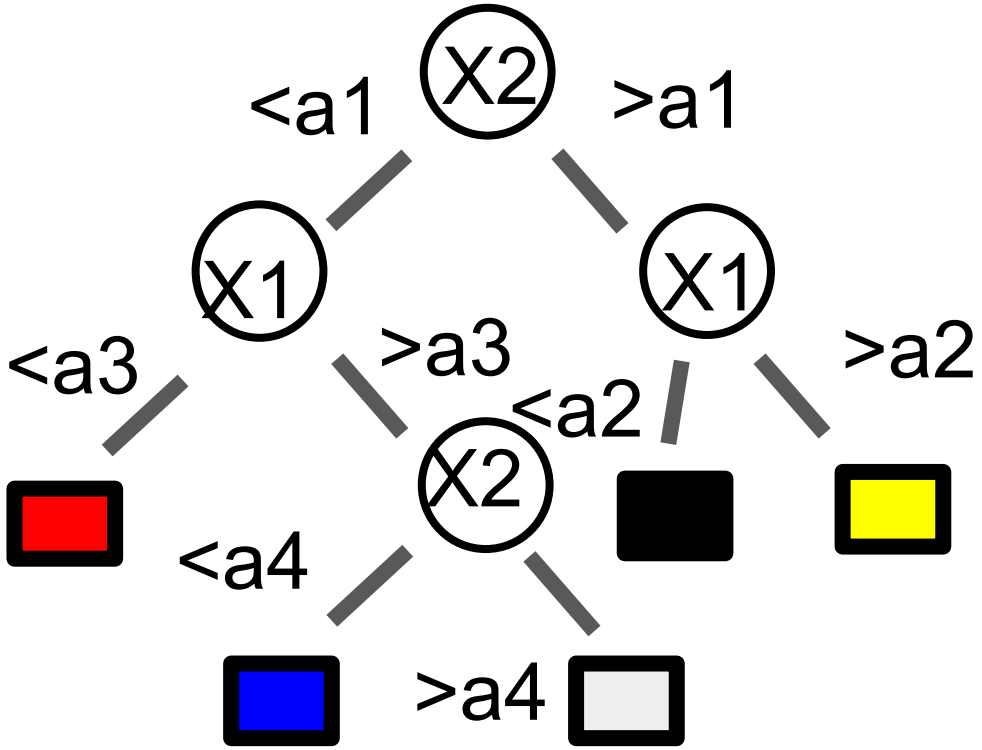


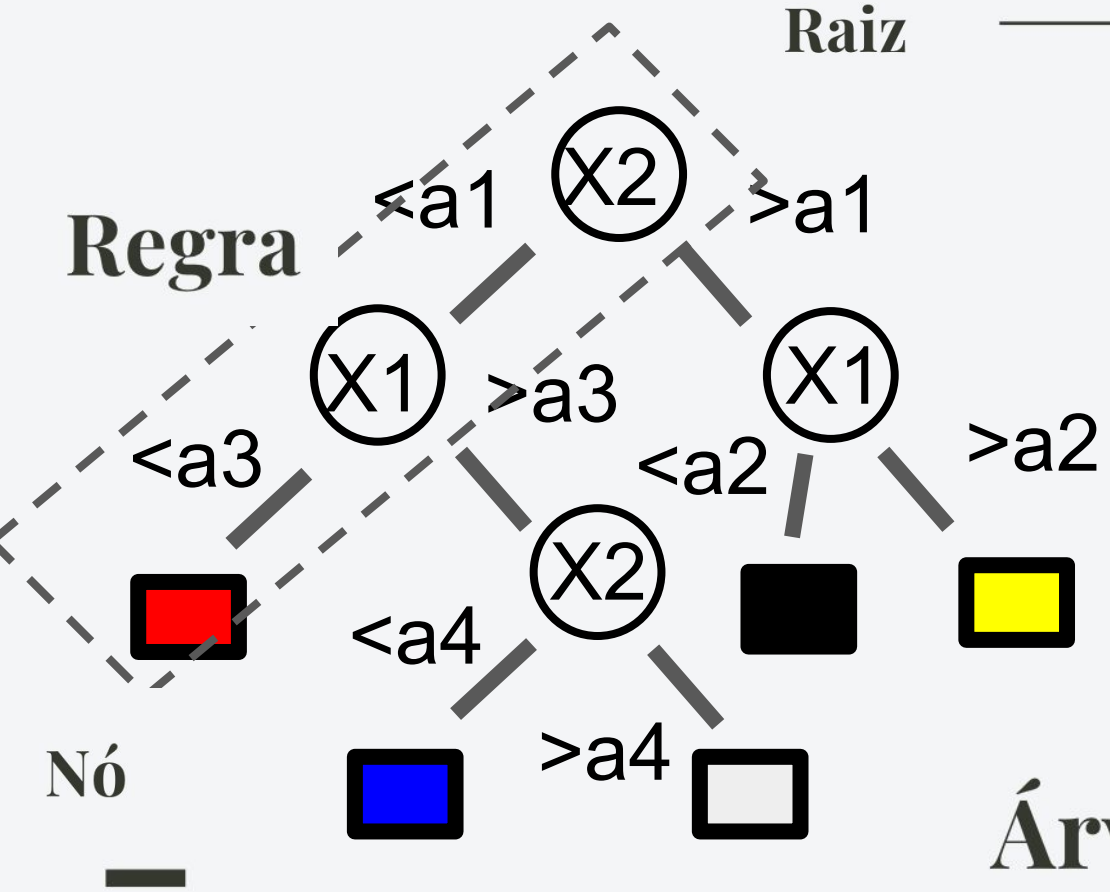
Classe Amarelo
Classe Vermelho
Classe Azul
Classe Branco
Classe Preto

Atributos (X1 e X2)
Valores de X1 e X2 - a1, a2, a3 e a4

Árvore de Decisão


X1
a3
a2
X2
a4
a1


X2
<a1
>a1
X1
X1
<a3
>a3
<a2
>a2
X2
<a4
>a4



REPRESENTAÇÃO POR ÁRVORES DE DECISÃO:

– Cada nó de decisão contem um teste num atributo.
– Cada ramo descendente corresponde a um possível valor deste atributo.
– Cada Folha está associada a uma classe.
- Cada percurso na árvore (da raiz à folha) corresponde a uma regra de classificação.

Árvore de Decisão

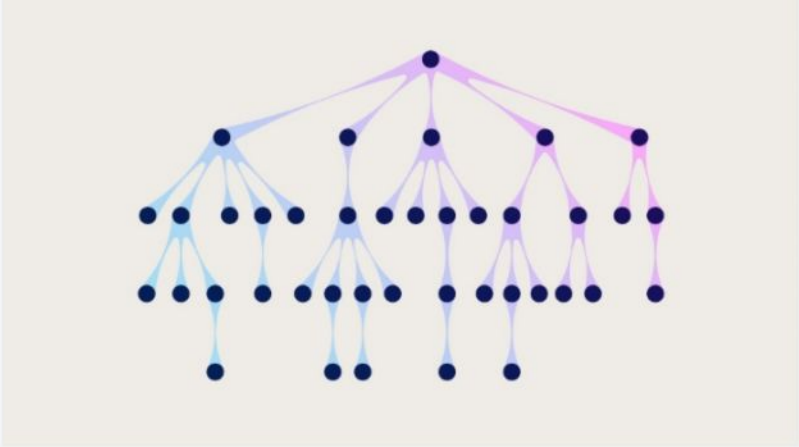
Árvore de decisão

A capacidade da árvore de decisão de criar ramificações específicas para casos particulares faz com que ela "admita exceções", tratando casos fora do padrão sem comprometer a regra geral do modelo

Instâncias (exemplos) são representadas por pares atributo-valor

Fáceis de serem implementadas e utilizadas

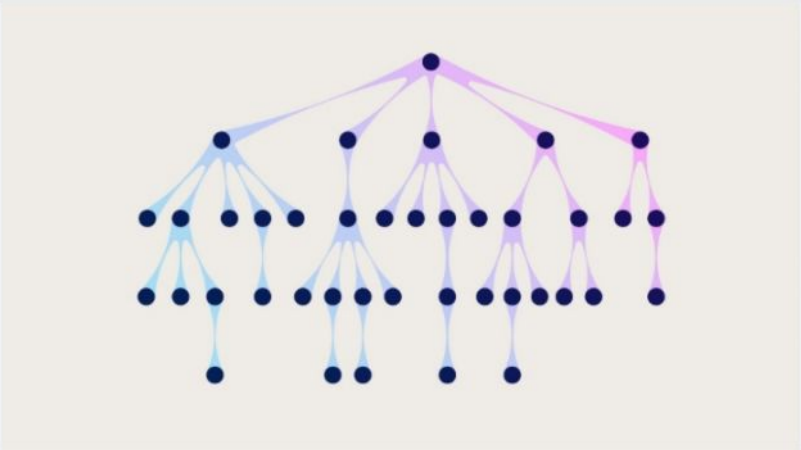
Estatística (admite exceções)



Árvore de Decisão

Uma árvore de decisão utiliza uma estratégia de dividir-para conquistar:

– Um problema complexo é decomposto em sub-problemas mais simples.
– Recursivamente a mesma estratégia é aplicada a cada sub-problema.



A capacidade de discriminação de uma árvore vem da:

– Divisão do espaço definido pelos atributos em sub-espaços.
– A cada sub-espaço é associada uma classe.

Árvore de Decisão

Árvore de Decisão

A 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**
    1. **Se todos os exemplos são da mesma classe, associar essa classe à folha**
    2. **Senão repetir os passos 1 a 4**

Árvore de decisão

Atributo de entrada

Tempo	Temperatura	Umidade	Vento	Joga ou não
Sol	*85*	*85*	*Não*	*Não*
Sol	*80*	*90*	*Sim*	*Não*
Nublado	*83*	*86*	*Não*	*Sim*
Chuva	*70*	*96*	*Não*	*Sim*
Chuva	*68*	*80*	*Não*	*Sim*
Chuva	*65*	*70*	*Sim*	*Não*
Nublado	*64*	*65*	*Sim*	*Sim*
Sol	*72*	*95*	*Não*	*Não*
Sol	*69*	*70*	*Não*	*Sim*
Chuva	*75*	*80*	*Não*	*Sim*
Sol	*75*	*70*	*Sim*	*Sim*
Nublado	*72*	*90*	*Sim*	*Sim*
Nublado	*81*	*75*	*Não*	*Sim*
Chuva	*71*	*91*	*Sim*	*Não*

Atributo alvo

Árvore de decisão

Selecione um atributo

Vento

Tempo	Temperatura	Umidade	Vento	Joga ou não
Sol	85	85	Não	Não
Nublado	83	86	Não	Sim
Chuva	70	96	Não	Sim
Chuva	68	80	Não	Sim
Sol	72	95	Não	Não
Sol	69	70	Não	Sim
Chuva	75	80	Não	Sim

Tempo	Temperatura	Umidade	Vento	Joga ou não
Sol	80	90	Sim	Não
Chuva	65	70	Sim	Não
Nublado	64	65	Sim	Sim
Sol	75	70	Sim	Sim
Nublado	72	90	Sim	Sim
Chuva	71	91	Sim	Não

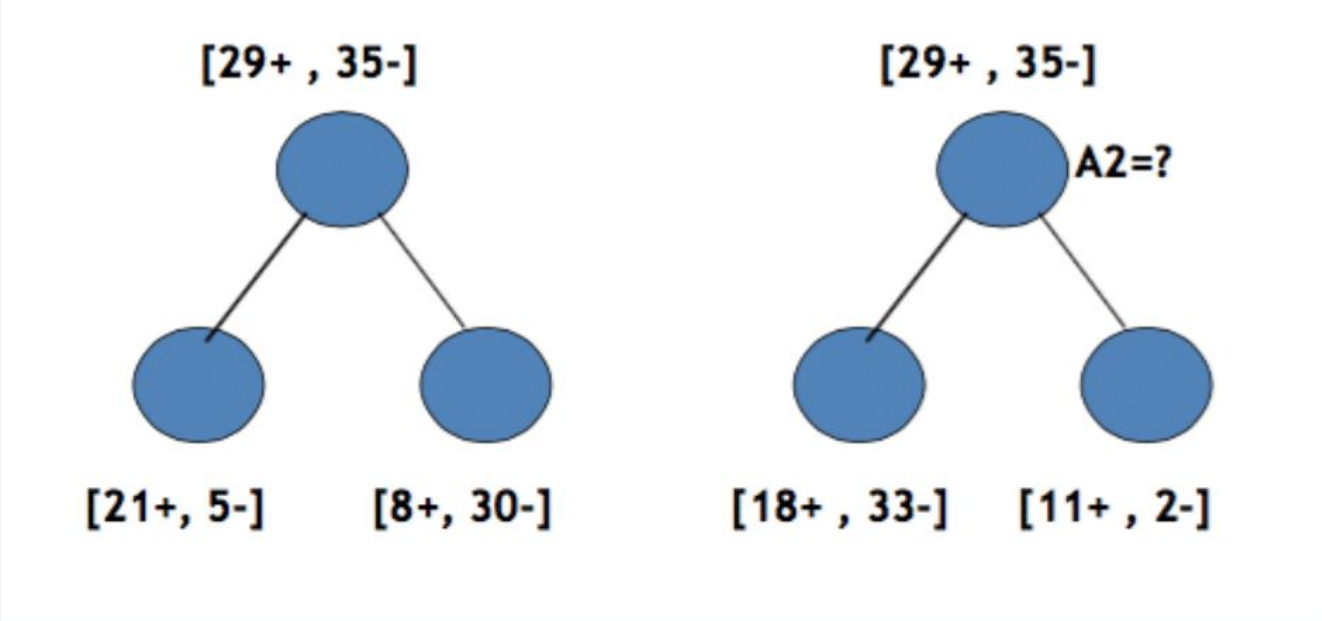
Qual é o melhor atributo?

CRITÉRIOS PARA ESCOLHA DO ATRIBUTO

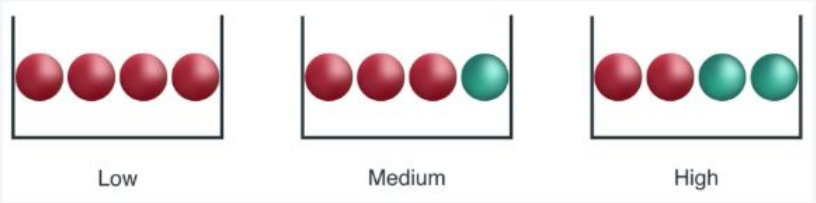
– Como medir a habilidade de um dado atributo discriminar as classes?
– Existem muitas medidas.
– Todas concordam em dois pontos:
  - Uma divisão que mantêm as proporções de classes em todas as partições é inútil.
  - Uma divisão onde em cada partição todos os exemplos são da mesma classe tem utilidade máxima

CRITÉRIOS PARA ESCOLHA DO ATRIBUTO

Qual é o melhor atributo???



Entropia



ENTROPIA CARACTERIZA A IMPUREZA DE UMA COLEÇÃO ARBITRÁRIA DE EXEMPLOS.

Entropia inicia com o trabalho de Lazare Carnot (1803)

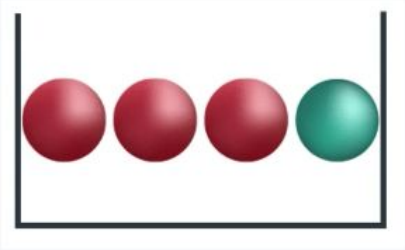
Rudolf Clausius (1850s-1860s) traz novas interpretações físicas

Claude Shannon (1948) desenvolve o conceito de Entropia em Teoria da Informação

Entropia

S É UMA AMOSTRA DOS EXEMPLOS DE TREINAMENTO

p (+) é a proporção de exemplos positivos em S
p (-) é a proporção de exemplos negativos em S



Exemplos positivos em S

Exemplos negativos em S

ENTROPIA MEDE A "IMPUREZA" DE S:

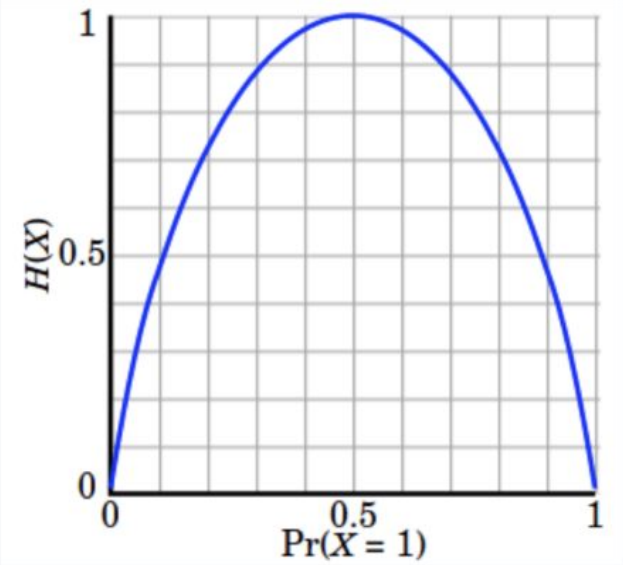
$$\text{Entropia}(S) = - p(+) \log_2 p(+) - p(-) \log_2 (-)$$

Entropia

SE P (+) É 1, O DESTINATÁRIO SABE QUE O EXEMPLO SELECIONADO SERÁ POSITIVO

Entropia é 0 (zero) (mínima)

SE P (+) É 0,5, A ENTROPIA É 1 (MÁXIMA)!!



Entropia

SUPONHA QUE S É UMA COLEÇÃO DE 14 EXEMPLOS, INCLUINDO 9 POSITIVOS E 5 NEGATIVOS

– NOTAÇÃO: [9+,5-]

A ENTROPIA DE S EM RELAÇÃO A ESTA CLASSIFICAÇÃO BOOLEANA É DADA POR:

$$Entropy([9+,5-]) = -(9/14)\log_2(9/14) - (5/14)\log_2(5/14)$$
$$= 0.940$$

Índice GINI

SIMILAR À ENTROPIA, O ÍNDICE GINI É MÁXIMO SE AS CLASSES ESTÃO PERFEITAMENTE MISTURADAS, POR EXEMPLO, NA CLASSE BINÁRIA:

$$Gini = 1 - (p_1^2 + p_2^2) = 1 - (0.5^2 + 0.5^2) = 0.5$$

Critério	Quando usar
Índice de Gini	Quando se busca uma métrica rápida e eficiente, especialmente em grandes bases de dados. Ideal para problemas com classes desbalanceadas, pois tende a favorecer a classe majoritária na divisão.
Entropia	Quando se deseja um critério mais sensível à distribuição das classes e que balanceie melhor os ramos, útil em problemas com múltiplas classes ou quando se quer evitar que um ramo fique muito dominante.

No contexto das árvores de decisão a entropia é usada para estimar a aleatoriedade da variável a prever (classe).

Dado um conjunto de exemplos, que atributo escolher para teste?

- Os valores de um atributo definem partições do conjunto de exemplos.
- O ganho de informação mede a redução da entropia causada pela partição dos exemplos de acordo com os valores do atributo.

Ganho de Informação

Ganho de Informação

A CONSTRUÇÃO DE UMA ÁRVORE DE DECISÃO É GUIADA PELO OBJETIVO DE DIMINUIR A ENTROPIA OU SEJA A ALEATORIEDADE - DIFICULDADE DE PREVISÃO - DA VARIÁVEL QUE DEFINE AS CLASSES.

Árvore de decisão

Filme	País de Origem	Grande Estrela	Gênero	Sucesso
Filme 1	*Estados Unidos*	*Sim*	*Ficção Científica*	*verdadeiro*
Filme 2	*Estados Unidos*	*Não*	*Comédia*	*falso*
Filme 3	*Estados Unidos*	*Sim*	*Comédia*	*verdadeiro*
Filme 4	*Europeu*	*Não*	*Comédia*	*verdadeiro*
Filme 5	*Europeu*	*Sim*	*Ficção Científica*	*falso*
Filme 6	*Europeu*	*Sim*	*Romance*	*falso*
Filme 7	*Do restante do mundo*	*Sim*	*Comédia*	*falso*
Filme 8	*Do restante do mundo*	*Não*	*Ficção Científica*	*falso*
Filme 9	*Europeu*	*Sim*	*Comédia*	*verdadeiro*
Filme 10	*Estados Unidos*	*Sim*	*Comédia*	*verdadeiro*

País de Origem	Sucesso
Estados Unidos	*verdadeiro*
Estados Unidos	*falso*
Estados Unidos	*verdadeiro*
Europeu	*verdadeiro*
Europeu	*falso*
Europeu	*falso*
Do restante do mundo	*falso*
Do restante do mundo	*falso*
Europeu	*verdadeiro*
Estados Unidos	*verdadeiro*

Árvore de decisão

- DOS FILMES QUE SÃO DOS EUA, 3 FORAM BEM-SUCEDIDOS E 1 NÃO FOI.
- H(EUA) = -(3/4)LOG2(3/4) – (1/4)LOG2(1/4) = 0,311 + 0,5 = 0,811
- H(EUROPEU) = 1 (METADE/METADE)
- H(DO RESTO DO MUNDO) = 0 (NENHUM BEM-SUCEDIDO)
- GANHO=1–(0,4X0,811)–(0,4X1)–(0,2X0) = 1 – 0,3244 – 0,4 – 0 = 0,2756

Grande Estrela	Sucesso
Sim	*verdadeiro*
Não	*falso*
Sim	*verdadeiro*
Não	*verdadeiro*
Sim	*falso*
Sim	*falso*
Sim	*falso*
Não	*falso*
Sim	*verdadeiro*
Sim	*verdadeiro*

Árvore de decisão

- PARA O ATRIBUTO DE "GRANDE ESTRELA":

$$H(SIM) = -(4/7)LOG2(4/7) - (3/7)LOG2(3/7) = 0,985$$

$$H(NÃO) = -(1/3)LOG2(1/3) - (2/3)LOG2(2/3) = 0,918$$

$$GANHO=1-(0,7X0,985)-(0,3X0,918) = 1 - 0,68964 - 0,275 = 0,035$$

Gênero	Sucesso
Ficção Científica	*verdadeiro*
Comédia	*falso*
Comédia	*verdadeiro*
Comédia	*verdadeiro*
Ficção Científica	*falso*
Romance	*falso*
Comédia	*falso*
Ficção Científica	*falso*
Comédia	*verdadeiro*
Comédia	*verdadeiro*

Exercício!!!

- PARA O ATRIBUTO DE “GÊNERO”:

Gênero	Sucesso
Ficção Científica	*verdadeiro*
Comédia	*falso*
Comédia	*verdadeiro*
Comédia	*verdadeiro*
Ficção Científica	*falso*
Romance	*falso*
Comédia	*falso*
Ficção Científica	*falso*
Comédia	*verdadeiro*
Comédia	*verdadeiro*

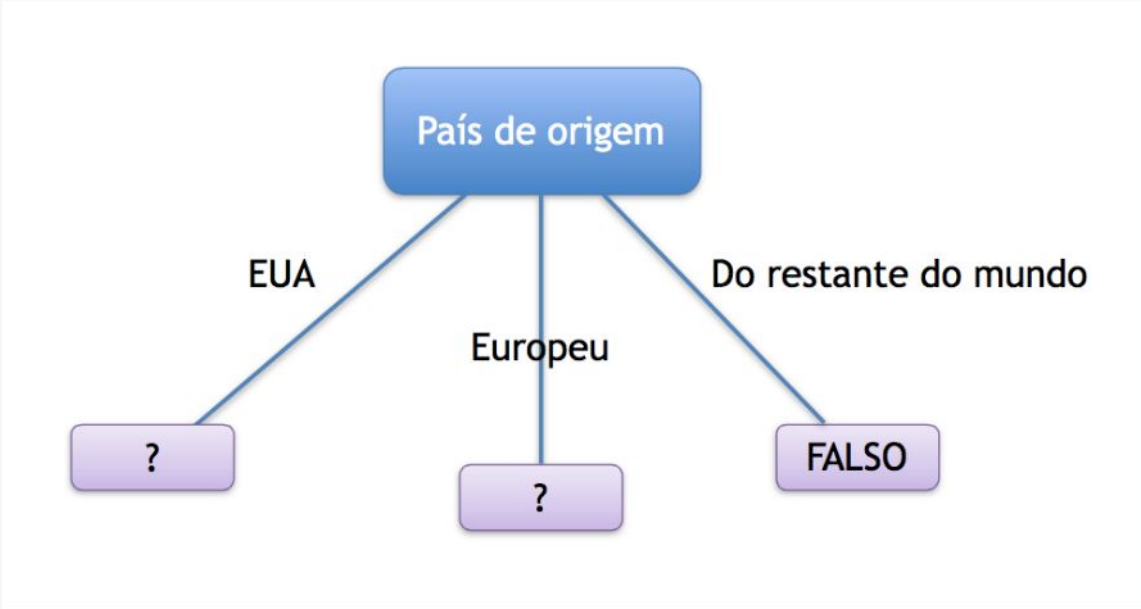
Árvore de Decisão

PARA O ATRIBUTO "GÊNERO":

H(FICÇÃO CIENTÍFICA) = 0,917
H(COMÉDIA) = 0,918296
H(ROMANCE)=0 (0 X LOG20 É 0)

GANHO=1–(0,3X0,917)–(0,6X 0,918296) – (0,1 X 0) = 1 – 0,2751 – 0,5509776 – 0 = 0,17

Árvore de Decisão



• Um teste num atributo numérico produz uma partição binária do conjunto de exemplos:
  – Exemplos onde valor_do_atributo < ponto_referência
  – Exemplos onde valor_do_atributo > ponto_referência
• Escolha do ponto de referência:
  – Ordenar os exemplos por ordem crescente dos valores do atributo numérico.

– Qualquer ponto intermediário entre dois valores diferentes e consecutivos dos valores observados no conjunto de treinamento pode ser utilizado como possível ponto de referência.

Ganho de Informação - Atributos Numéricos

Ganho de Informação - Atributos Numéricos

– É usual considerar o valor médio entre dois valores diferentes e consecutivos.

– Fayyard e Irani (1993) mostram que de todos os possíveis pontos de referência, aqueles que maximizam o ganho de informação separam dois exemplos de classes diferentes.

Temperatura	Joga
64	Sim
65	Não
68	Sim
69	Sim
70	Sim
71	Não
72	Não
72	Sim
75	Sim
75	Sim
80	Não
81	Sim
83	Sim
85	Não

Árvore de Decisão

CONSIDERE O PONTO DE REFERÊNCIA
TEMPERATURA =70,5

UM TESTE USANDO ESTE PONTO
DE REFERÊNCIA DIVIDE OS
EXEMPLOS EM DUAS CLASSES:
– EXEMPLOS ONDE TEMPERATURA < 70,5
– EXEMPLOS ONDE TEMPERATURA > 70,5

COMO MEDIR O GANHO DE
INFORMAÇÃO DESTA PARTIÇÃO?

Ganho de Informação – Atributos Numéricos

- Como medir o ganho de informação desta partição?

- Informação nas partições
  - p(sim | temperatura<70,5)=4/5
  - p(não | temperatura<70,5)=1/5
  - p(sim | temperatura>70,5)=5/9
  - p(não | temperatura>70,5)=4/9

Árvore de Decisão

- O PROBLEMA DE CONSTRUIR UMA ÁRVORE DE DECISÃO:
  – CONSISTENTE COM UM CONJUNTO DE EXEMPLOS
  – COM O MENOR NÚMERO DE NÓS

- DOIS PROBLEMAS:
  – QUE ATRIBUTO SELECIONAR PARA TESTE NUM NÓ?
  – QUANDO PARAR A DIVISÃO DOS EXEMPLOS?

Critérios de Parada

- QUANDO PARAR A DIVISÃO DOS EXEMPLOS?

– TODOS OS EXEMPLOS PERTENCEM A MESMA CLASSE.

– O NÚMERO DE EXEMPLOS É INFERIOR A UM CERTO LIMITE.

Simplificar a árvore

- DUAS POSSIBILIDADES:

– PARAR O CRESCIMENTO DA ÁRVORE MAIS CEDO (PRÉ-PRUNING).

– CONSTRUIR UMA ÁRVORE COMPLETA E PODAR A ÁRVORE (PÓS-PRUNING).

Árvores com ramos muitos compridos = Regras muito complexas - chance alta de overfitting

Árvores com ramos muito curtos = Regras muito generalistas - chance alta de underfitting

Algoritmo básico de pruning

PERCORRE A ÁRVORE EM PROFUNDIDADE PARA CADA NÓ DE DECISÃO CALCULA:

– ERRO NO NÓ
– SOMA DOS ERROS NOS NÓS DESCENDENTES

• SE O ERRO NO NÓ É MENOR OU IGUAL À SOMA DOS ERROS DOS NÓS DESCENDENTES, O NÓ É TRANSFORMADO EM FOLHA

Algoritmo básico de pruning

Qual seria o erro de classificação esperado em uma folha?

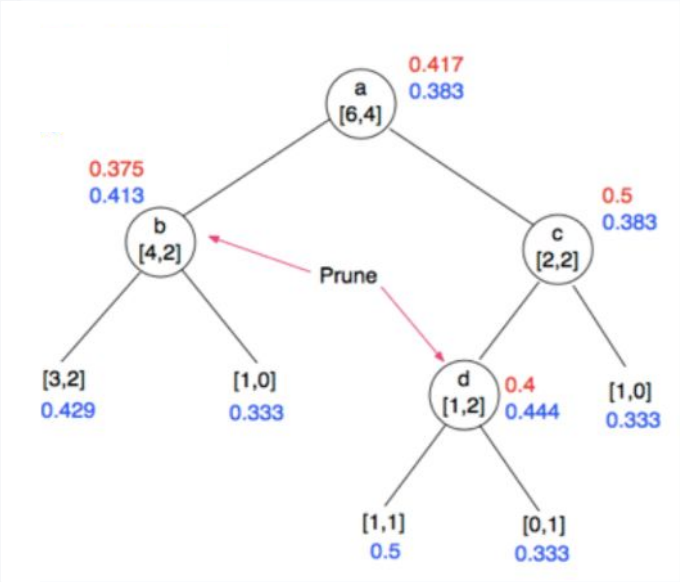
$$E(S) = (N - n + k - 1)/(N + k)$$

Este também é chamado de erro de Laplace – baseado no fato que a distribuição de probabilidades destes exemplos de serem das diferentes classes é uniforme.

S - é o conjunto de exemplos no nó
k - é o número de classes
N - exemplos em S
n - dos N exemplos em S quantos fazem parte da classe maioritária em S

PÓS-PRUNING

Poda se: o **erro estático** < **erro em relação aos nós descencentes**



$$E(S) = (N-n+k-1)/(N+k)$$

$$E(b) = (6-4+2-1)/(6+2) = 0{,}375$$

$$E(1) = (5-3+2-1)/(5+2) = 0{,}429$$

$$E(2) = (1-1+2-1)/(1+2) = 0{,}333$$

Erro em relação aos nós descentens (b) =
$(5/6)\times 0{,}429 + (1/6)\times 0{,}333 = 0{,}413$

ÁRVORE DE DECISÃO

Confiram o método Random Forest!! É uma das variações mais populares de Árvores de decisão.

ÁRVORE DE DECISÃO

O Random Forest é um algoritmo que combina várias árvores de decisão.

Funciona criando múltiplas árvores de decisão de forma aleatória, formando uma "floresta", onde cada árvore contribui para a decisão final por meio de votação.

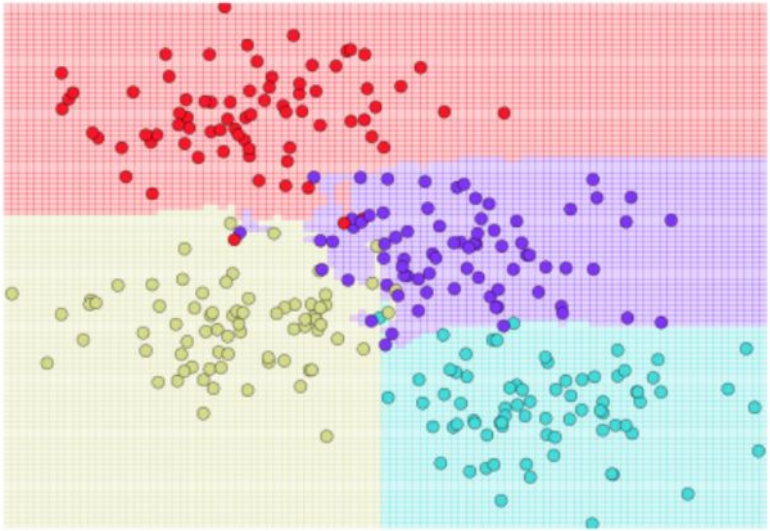
No Random Forest, cada árvore é treinada com uma amostra aleatória dos dados de treinamento, o que ajuda a reduzir a correlação entre as árvores e aumentar a diversidade do modelo.

Em problemas de regressão, a previsão final é a média dos valores previstos por todas as árvores; em problemas de classificação, o resultado mais frequente é escolhido como a previsão final.

ÁRVORE DE DECISÃO

Random Forest

```
from sklearn.ensemble import RandomForestClassifier
clf = RandomForestClassifier(n_estimators=100, random_state=0, n_jobs=-1)
visualize_tree(clf, X, y, boundaries=False);
```



Tem exercício no Jupyter!!

Slides baseados nas aulas de Teresa Ludermir - CIn (UFPE)