# Mineração de Dados em Biologia Molecular

Algoritmos de Indução de Árvores de Características

Docente: André C. P. L. F. de Carvalho
PAE: Victor Hugo Barella

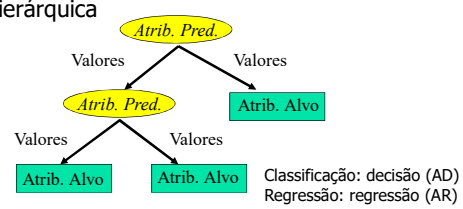
## Introdução

- Explicação das decisões pode ser importante para algumas aplicações
  - RNAs, SVMs e RNPs são caixas pretas
- Modelos interpretáveis são gerados por algoritmos de indução de:
  - Árvores de características (decisão)
  - Conjunto de regras
  - Redes Bayesianas



## Árvores de características

- Alguns algoritmos de AM particionam características (atributos) de forma hierárquica

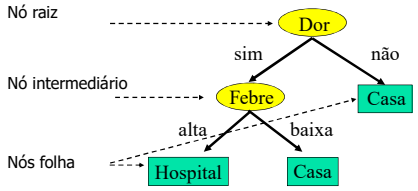



Classificação: decisão (AD)
Regressão: regressão (AR)



## Algoritmo de indução de AD

- Induzem modelos representados por ADs






## Algoritmo de indução de AD

- Existem vários, entre eles:
  - Algoritmo de Hunt
    - Um dos primeiros
    - Base de vários algoritmos atuais
  - CART
  - ID3
  - C4.5
  - VFDT



## Algoritmo de Hunt

- Seja $X_t$ o conjunto de objetos de treinamento que atingem o nó t

*Se todos os objetos de $X_t$ $\in$ a mesma classe $y_t$*
*Então t é um nó folha rotulado como $y_t$*
*Se os objetos de $X_t$ $\in$ a mais de uma classe*
*Então Selecionar um atributo preditivo teste para dividir $X_t$*
*Dividir $X_t$ em subconjuntos utilizando esse atributo*
*Aplicar algoritmo a cada subconjunto gerado*

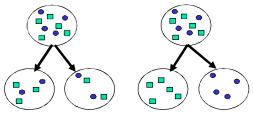
## Indução de ADs

- Geralmente usa estratégia gulosa de divisão e conquista
  - Divide progressivamente objetos baseado em um atributo preditivo de teste
    - Escolhido para otimizar algum critério
- Decisões importantes
  - Como dividir os objetos?
    - Estratégia para escolha do atributo de teste
  - Quando parar de dividir os objetos?



## Medidas para escolha de atributo

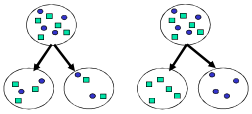
- Selecionam atributo que melhor divide os dados atuais
  - Buscam partições mais puras após divisão
    - Quanto mais homogêneas as partições, mais puras
  - Medidas de impureza





## Medidas de impureza

- Baseadas no grau de impureza dos nós filhos
  - Quando maior, pior
- Diferentes medidas geram diferentes partições
- Exemplos
  - Entropia
  - Erro de classificação
  - Gini
  - Qui-quadrado





## Medidas de impureza

$$Entropia(v) = -\sum_{i=1}^{C} p(i/v)\log_2 p(i/v)$$

$$Gini(v) = 1 - \sum_{i=1}^{C}[p(i/v)]^2$$

$$ErroClass(v) = 1 - \max_i [p(i/v)]$$

*Onde:*
*P(i/v) = fração de dados pertencente a classe i em um nó v*
*C = número de classes*
*Considera-se que $0\log_2 0 = 0$*



## Exemplo

- Calcular a medida de impureza Gini para os dados abaixo:

$$Gini(v) = 1 - \sum_{i=1}^{C}[p(i/v)]^2$$

| C1 | 0 |
|---|---|
| C2 | 6 |
| Gini=? | |

| C1 | 1 |
|---|---|
| C2 | 5 |
| Gini=? | |

| C1 | 2 |
|---|---|
| C2 | 4 |
| Gini=? | |

| C1 | 3 |
|---|---|
| C2 | 3 |
| Gini=? | |



## Exemplo

$$Gini(v) = 1 - \sum_{i=1}^{C}[p(i/v)]^2$$

*P(C1) = 0/6 = 0 P(C2) = 6/6 = 1*
*Gini = 1 – P(C1)² – P(C2)² = 1 – 0 – 1 = 0*
*P(C1) = 1/6 P(C2) = 5/6*
*Gini = 1 – (1/6)² – (5/6)² = 0.278*
*P(C1) = 2/6 P(C2) = 4/6*
*Gini = 1 – (2/6)² – (4/6)² = 0.444*
*P(C1) = 3/6 P(C2) = 3/6*
*Gini = 1 – (3/6)² – (3/6)² = 0.500*

| C1 | 0 |
|---|---|
| C2 | 6 |
| Gini=0.000 | |

| C1 | 1 |
|---|---|
| C2 | 5 |
| Gini=0.278 | |

| C1 | 2 |
|---|---|
| C2 | 4 |
| Gini=0.444 | |

| C1 | 3 |
|---|---|
| C2 | 3 |
| Gini=0.500 | |

## Exercício

- Fazer os mesmos cálculos para as medidas de entropia e de erro de classificação

$$Entropia(v) = -\sum_{i=1}^{C} p(i/v)\log_2 p(i/v)$$

$$ErroClass(v) = 1 - \max_i [p(i/v)]$$

| C1 | 0 |
|---|---|
| C2 | 6 |
| E=? | |

| C1 | 1 |
|---|---|
| C2 | 5 |
| E=? | |

| C1 | 2 |
|---|---|
| C2 | 4 |
| E=? | |

| C1 | 3 |
|---|---|
| C2 | 3 |
| E=? | |

| C1 | 0 |
|---|---|
| C2 | 6 |
| Class=? | |

| C1 | 1 |
|---|---|
| C2 | 5 |
| Class=? | |

| C1 | 2 |
|---|---|
| C2 | 4 |
| Class=? | |

| C1 | 3 |
|---|---|
| C2 | 3 |
| Class=? | |



## Critério de parada

- Diversas alternativas:
  - Os objetos do nó atual têm a mesma classe
  - Os objetos do nó atual têm valores iguais para os atributos de entrada, mas classes diferentes
  - O número de objetos do nó é menor que um dada quantidade
  - Todos os atributos preditivos já foram incluídos no caminho atual



## Exemplo

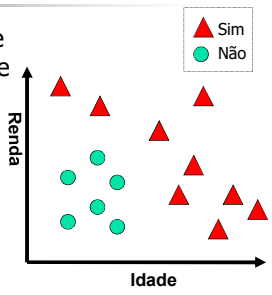
- Sejam os dados abaixo referentes a solicitações de crédito bancário
  - Construir uma árvore de decisão que classifica aplicação para cartão de crédito

| Idade | Renda | Classe |
|---|---|---|
| 20 | 2000 | Sim |
| 30 | 5200 | Não |
| 60 | 5000 | Sim |
| 40 | 6000 | Não |
| ... | | |



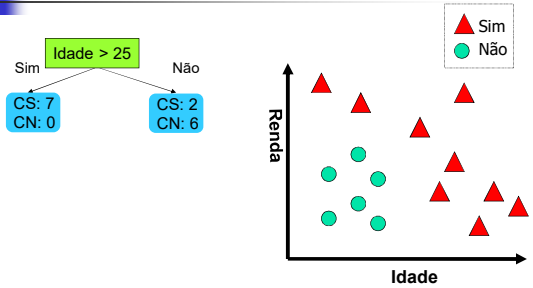
## Busca no espaço de hipóteses

- Construir uma AD que classifica solicitante de cartão de crédito
  - Aprova (Sim)
  - Não aprova (Não)
- Atributos preditivos
  - Idade
  - Renda



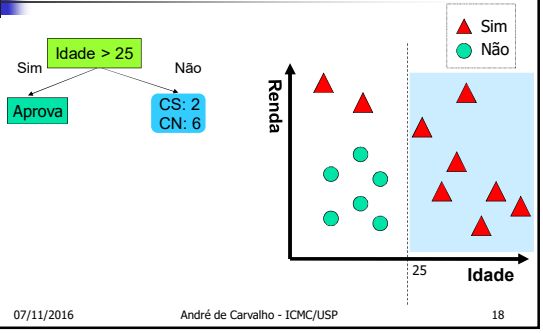



## Busca no espaço de hipóteses

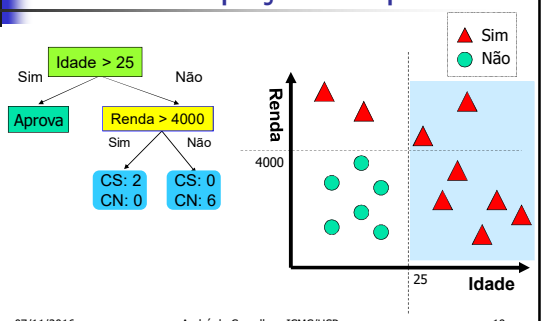





## Busca no espaço de hipóteses

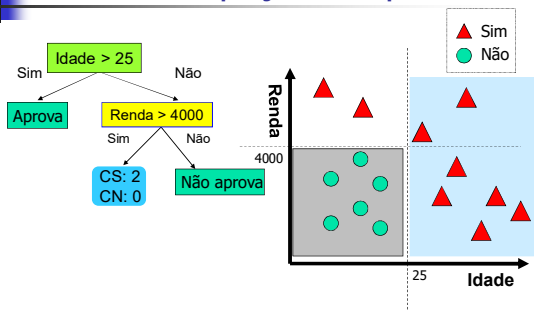


## Busca no espaço de hipóteses



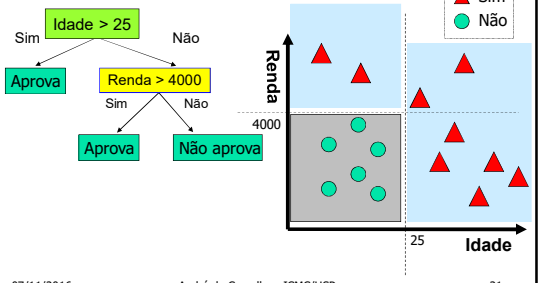



## Busca no espaço de hipóteses





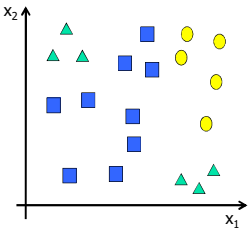

## Busca no espaço de hipóteses






## Algoritmo de indução de AD

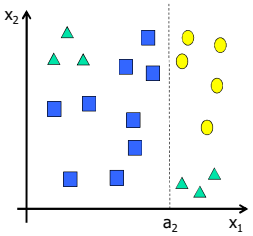
- Valores numéricos e simbólicos
- Grau de pureza é calculado para cada atributo
  - Atributo que gera menos impureza é escolhido
- Critério de parada precisa ser definido






## Algoritmo de indução de AD

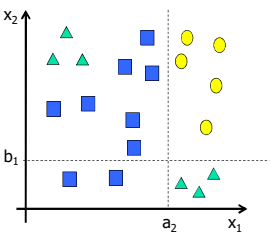
- Valores numéricos e simbólicos
- Grau de pureza é calculado para cada atributo
  - Atributo que gera menos impureza é escolhido
- Critério de parada precisa ser definido





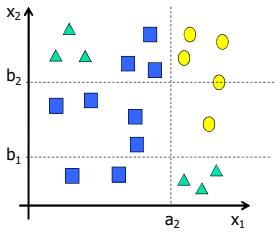

## Algoritmo de indução de AD

- Valores numéricos e simbólicos
- Grau de pureza é calculado para cada atributo
  - Atributo que gera menos impureza é escolhido
- Critério de parada precisa ser definido



## Algoritmo de indução de AD

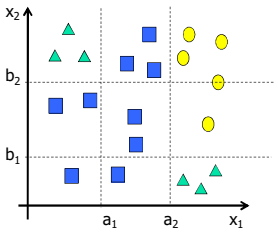
- Valores numéricos e simbólicos
- Grau de pureza é calculado para cada atributo
  - Atributo que gera menos impureza é escolhido
- Critério de parada precisa ser definido






## Algoritmo de indução de AD

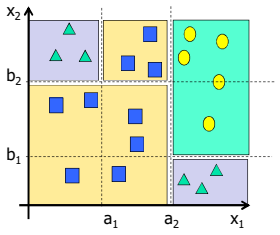
- Valores numéricos e simbólicos
- Grau de pureza é calculado para cada atributo
  - Atributo que gera menos impureza é escolhido
- Critério de parada precisa ser definido





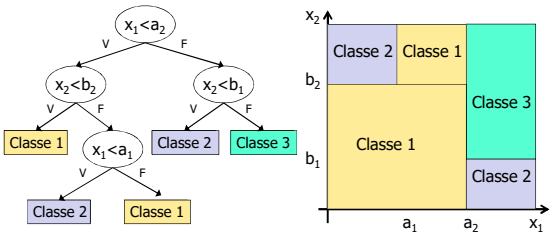

## Algoritmo de indução de AD

- Valores numéricos e simbólicos
- Grau de pureza é calculado para cada atributo
  - Atributo que gera menos impureza é escolhido
- Critério de parada precisa ser definido






## Árvore e partição do espaço de hipóteses



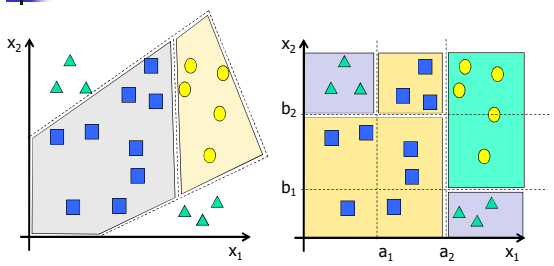



## Espaço de hipóteses

- Cada percurso da raiz a um nó folha representa uma regra de classificação
- Cada folha está associada a uma classe
  - Corresponde a um hiper-retângulo no espaço de soluções
    - Cada classe é representada por um conjunto de hiper-retângulos
    - Interseção de hiper-retângulos é um conjunto vazio
    - União de hiper-retângulos cobre todo o espaço



## RNs x AD



## Exercício

- Seja o seguinte cadastro de pacientes:

| Nome | Febre | Enjôo | Manchas | Dores | Diagnóstico |
|---|---|---|---|---|---|
| João | sim | sim | pequenas | sim | doente |
| Pedro | não | não | grandes | não | saudável |
| Maria | sim | sim | pequenas | não | saudável |
| José | sim | não | grandes | sim | doente |
| Ana | sim | não | pequenas | sim | saudável |
| Leila | não | não | grandes | sim | doente |



## Exercício

- Usando medida de entropia,
  - Induzir uma árvore de decisão capaz de distinguir:
    - Pacientes potencialmente saudáveis
    - Pacientes potencialmente doentes
  - Testar a árvore para novos casos
    - (Luis, não, não, pequenas, sim)
    - (Laura, sim, sim, grandes, sim)



## Conclusão

- Introdução
- Algoritmo de Hunt
- Medidas para selecionar divisão de atributos
- Critério de parada
- Espaço de hipóteses
- Árvores de regressão
  - Valor ou função



## Perguntas