INTELIGÊNCIA COMPUTACIONAL

ALGORITMOS DE CLASSIFICAÇÃO

FELIPE TORRES

CLASSIFICAÇÃO

- Classificação é uma técnica de mineração de dados para classificar objetos em uma determinada estrutura de classes.

- Este tipo de algoritmo tenta predizer a qual classe um determinado objeto pertence;

- Existem três tipos de técnicas de classificações clássicas: Árvores de Classificação, Árvores de Regressão e CART (C&RT – Classification and Regression Tree).

EXEMPLOS DE CLASSIFICAÇÃO

- Classificar tumores em benignos e malignos.
- Classificar pacientes em sintomáticos e assintomáticos.
- Classificar pacientes em resistentes ou não a um determinado tratamento ou droga.
- Classificar artigos em bons e ruins.

MODELOS DE CLASSIFICAÇÃO

- Construção do modelo
  - Com base no dataset de treinamento, modelo (regra, árvore de decisão, formula matemática) é construído.
  - Aprendizado supervisionado (atributo classe);

- Uso do modelo
  - O modelo é usado para classificar instâncias (não vistas) do conjunto de teste, estimando a acurácia.
  - A acurácia é o percentual de instâncias corretamente classificadas.

EXEMPLO MODELO DE REGRESSÃO LINEAR

- Um modelo de regressão linear simples é usado para o caso de regressão com uma variável explicativa.

- Este modelo é utilizado para identificar a função que explica o comportamento de variáveis linear.

$$Y_i = \alpha + \beta X_i + \epsilon_i$$

Y_i: Variável explicada (dependente); é o valor que se quer atingir;
α: É uma constante, que representa a interceptação da reta com o eixo vertical;
β: É outra constante, que representa o declive(coeficiente angular)da reta;
X_i: Variável explicativa (independente), representa o fator explicativo na equação;
ε: Variável que inclui todos os factores residuais mais os possíveis erros de medição.

TIPOS DE APRENDIZADO

- Lazy learning (Aprendizado preguiçoso)
  - Simplesmente guarde todos os dados do treino e aguarde um novo teste.

- Eager learning (Aprendizado ancioso)
  - Com base em um dataset de treinamento, construa um modelo de classificação antes de receber uma instância de teste.

LAZY VS EAGER

- Lazy gasta menos tempo com o treino e mais tempo com a predição.

- Acurácia
  - Métodos lazy efetivamente usam um espaço de hipóteses mais rico já que usa várias funções lineares locais para formar sua aproximação global da função alvo

  - Eager: precisa se comprometer com uma única hipótese que cobre todo o espaço de instâncias

ÁRVORES DE DECISÃO

- Árvore em que nós internos (não-folha) são testes em atributos e cada ramo é um resultado do teste e cada nó terminal (folha) é uma classe.

- Testes podem ser binários ou multi-valorados.

- Dada uma instância de teste, seus atributos são testados a partir da raíz até encontrar um nó folha

- Pode ser convertida para regras de classificação.

ÁRVORES DE DECISÃO

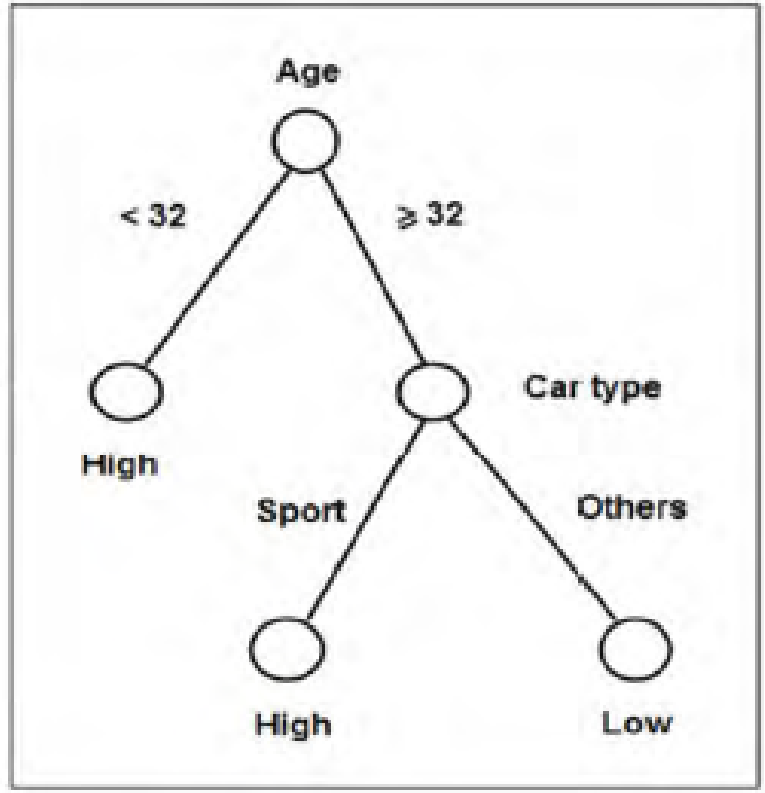
- Árvore em que nós internos (não-folha) são testes em atributos e cada ramo é um resultado do teste e cada nó terminal (folha) é uma classe.

- Testes podem ser binários ou multi-valorados.

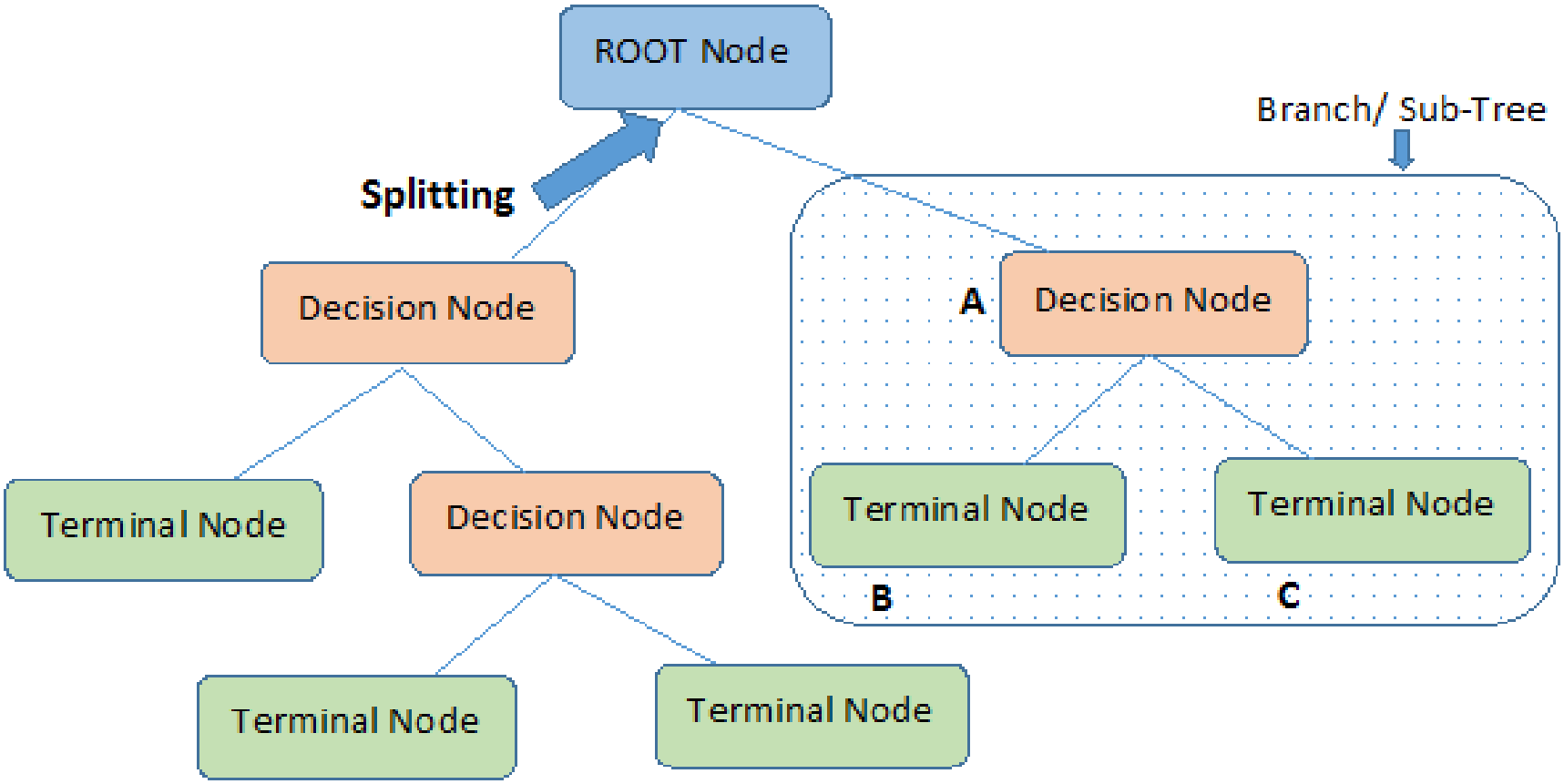
- Dada uma instância de teste, seus atributos são testados a partir da raíz até encontrar um nó folha Pode ser convertida para regras de classificação.

EXEMPLOS DE ÁRVORES DE DECISÃO

Age	Car type	Accident risk
20	Sport	High
18	Sport	High
40	Minivan	Low
50	Premium	Low
35	Compact	Low
30	Sport	High
32	Sport	High
40	Full size Van	Low
33	Mini	High
39	Convertible	Low

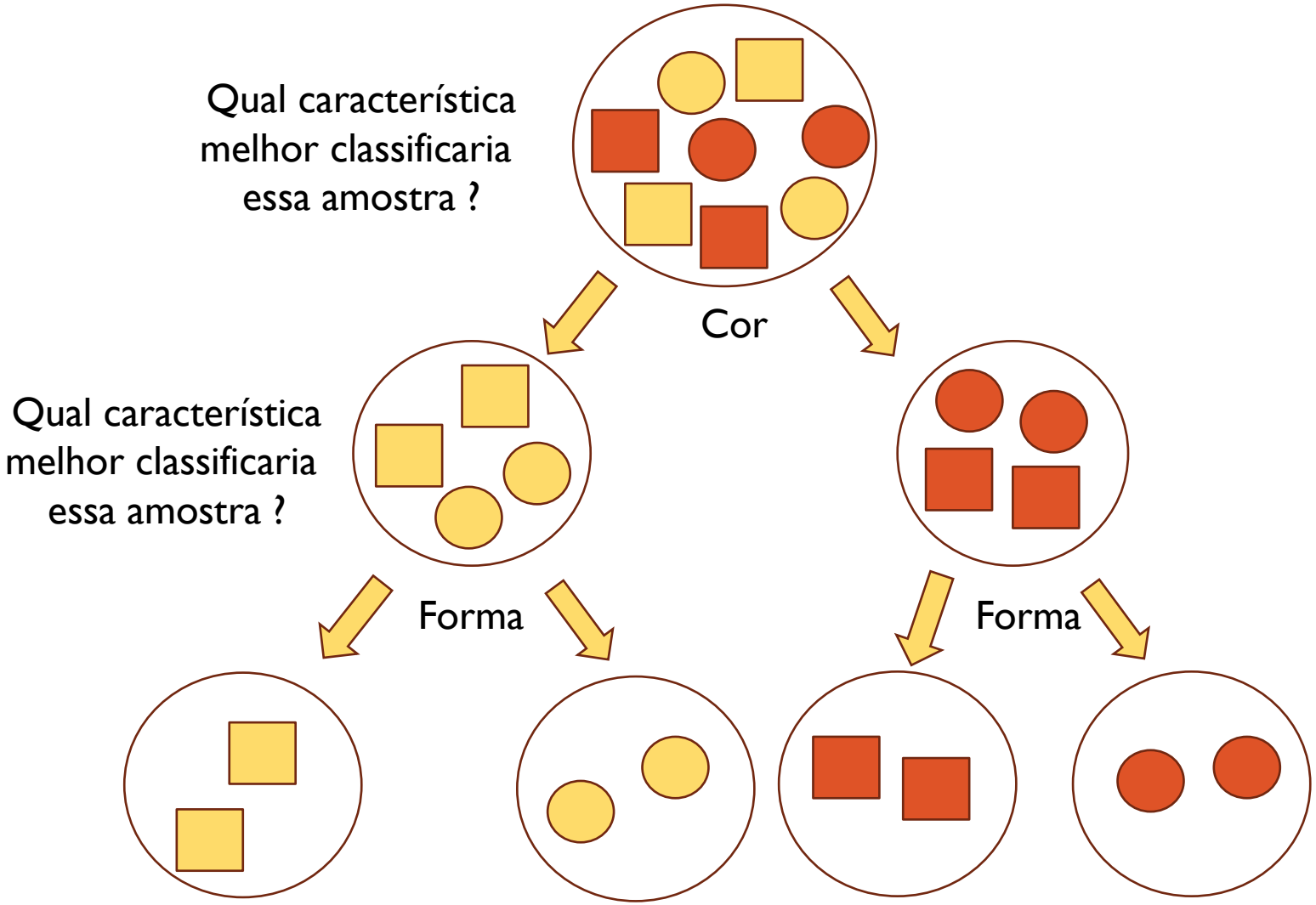


ESTRUTURA DAS ÁRVORES DE DECISÃO



Note:- A is parent node of B and C.

ALGORITMO DE ÁRVORE DE DECISÃO GENÉRICO



VANTAGENS DAS ÁRVORES DE DECISÃO

- Aprendizado e classificação são rápidos e simples.
- São de fácil interpretação.
- Podem lidar com dados multidimensionais.
- Alguns algoritmos mais utilizados são: J48, Decision Stump e o Random Tree

ALGORITMO J48

- Uma implementação do algoritmo ID3, desenvolvido pela equipe do WEKA.
- Esse algoritmo é rápido e muito acurado para amostras com variáveis contínuas em pequenos intervalos.
- Seguindo a mesma base do C4.5 o J48 possui melhorias enquanto a performance e acurácia.

ALGORITMO J48 – SELEÇÃO DE ATRIBUTOS

- Seleção de atributo por ganho de informação.
  - Entropia segundo a Teoria da Informação:
    - Medida de pureza ou impureza de um determinado conjunto de dados.
    - O quanto os dados são iguais ou diferentes entre si ?

$$entropia = \sum_{i=1}^{n} p(x_i) log2 p(x_i)$$

ALGORITMO J48 – SELEÇÃO DE ATRIBUTOS

- Seleção de atributo por ganho de informação.
  - Após medir a entropia o algoritmo pode medir o índice de ganho de informação e definir o melhor atributo.

$$Gain(S, A) = Entropia(S) - \sum_{v \in valores(A)} p(A_v) * Entropia(A_v)$$

S: Conjunto geral dos dados
A: Valores do atributo.
Av: Frequência de um determinado valor do atributo

ALGORITMO J48 – EXEMPLO DE SELEÇÃO DE ATRIBUTOS

Instancias	Expectativa	Temperatura	Humidade	Vento	Jogar tenis
1	Sol	Quente	Alta	Fraco	Não
2	Sol	Quente	Alta	Forte	Não
3	Nublado	Quente	Alta	Fraco	Sim
4	Chuva	Moderada	Alta	Fraco	Sim
5	Chuva	Fresco	Normal	Fraco	Sim
6	Chuva	Fresco	Normal	Forte	Não
7	Nublado	Fresco	Normal	Forte	Sim
8	Sol	Moderada	Alta	Fraco	Não
9	Sol	Fresco	Normal	Fraco	Sim
10	Chuva	Moderada	Normal	Fraco	Sim
11	Sol	Moderada	Normal	Forte	Sim
12	Nublado	Moderada	Alta	Forte	Sim
13	Nublado	Quente	Normal	Fraco	Sim
14	Chuva	Moderada	Alta	Forte	Não

ALGORITMO J48 – EXEMPLO DE SELEÇÃO DE ATRIBUTOS

Cálculo de ganho de informação do atributo S. Espectativa

Valores	Frequência relativa	Negativos	Positivos
Sol	5	3	2
Nublado	4	0	4
Chuva	5	2	3

Então teremos:

$$Gain(S, Expectativa) =$$

$$0{,}939 - \left(\frac{5}{14} * entropia(sol)\right) - \left(\frac{4}{14} * entropia\ (nublado)\right) - \left(\frac{5}{14} * entropia(chuva)\right)$$

$$= 0{,}245$$

ALGORITMO J48 – EXEMPLO DE SELEÇÃO DE ATRIBUTOS

Cálculo de ganho de informação do atributo S. Espectativa

Valores	Frequência relativa	Negativos	Positivos
Sol	5	3	2
Nublado	4	0	4
Chuva	5	2	3

Então teremos:

$$Entropia = -\left(\frac{9}{14} * log_2\left(\frac{9}{14}\right)\right) - \left(\frac{5}{14} * log_2\left(\frac{5}{14}\right)\right)$$

$$Entropia = -(0{,}642 * -0{,}637) - (0{,}357 * -1{,}485) = 0{,}939$$

ALGORITMO J48 – EXEMPLO DE SELEÇÃO DE ATRIBUTOS

Cálculo de ganho de informação do atributo Espectativa

Valores	Frequência relativa	Negativos	Positivos
Sol	5	3	2
Nublado	4	0	4
Chuva	5	2	3

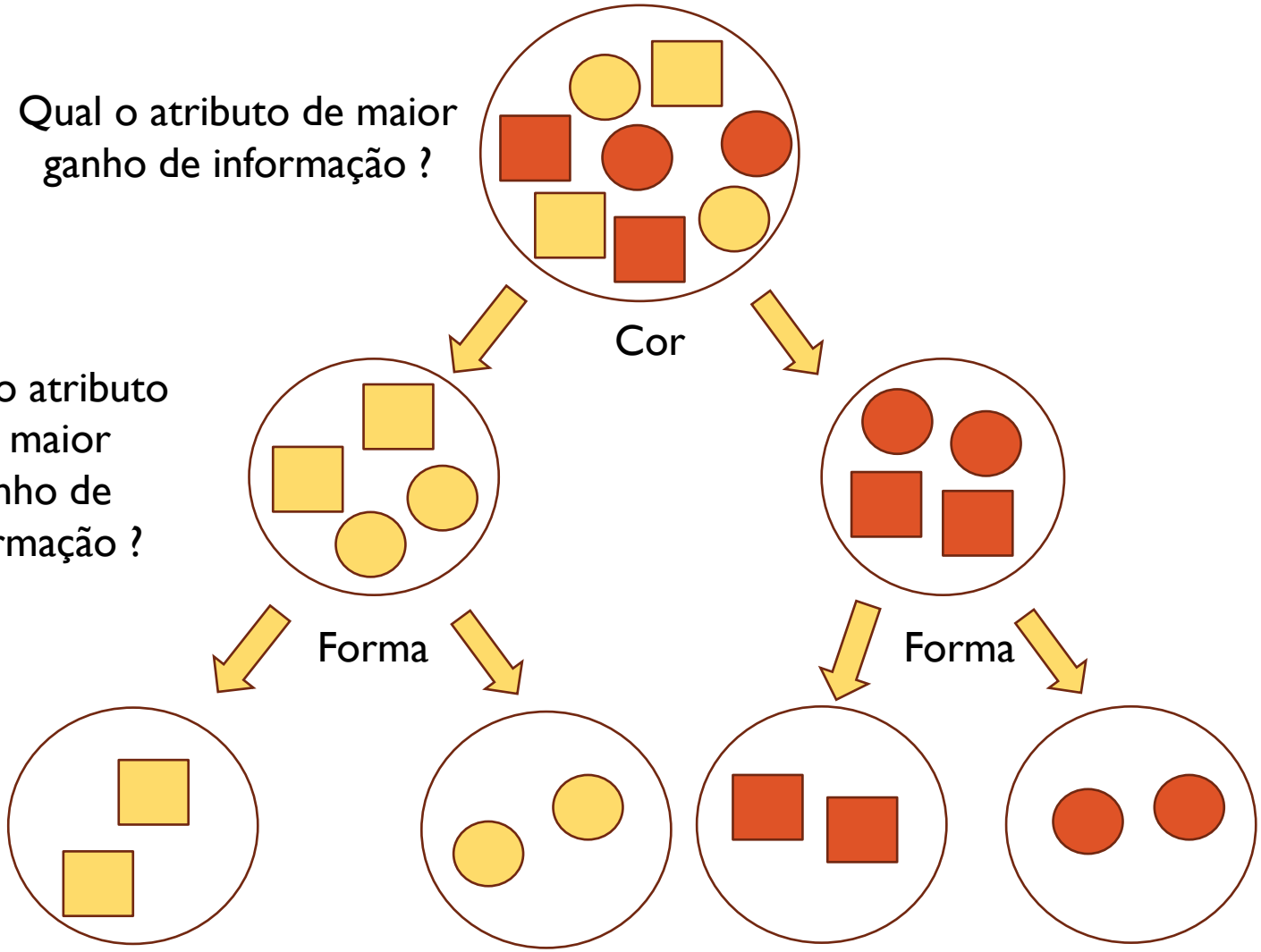
Replicando o mesmo processo teremos:

$Gain(S, Expectativa) = 0{,}245$
$Gain(S, Humidade) = 0{,}151$
$Gain(S, Vento) = 0{,}048$
$Gain(S, Temperatura) = 0{,}029$

Logo usando o J48 para esse dataset o root seria o atributo expectativa.

ALGORITMO J48

Qual o atributo de maior ganho de informação ?



Qual o atributo de maior ganho de informação ?

SELEÇÃO DE ATRIBUTOS

- IG tem viés para atributos muitos valores.

  Information Gain Ratio, J48

- O J48 Normaliza efeito de atributos multivalorados;

- Existem outras medidas de seleção de atributos:
  - Gini (CART): impuridade entre os nós
  - CHAD: teste chi-quadrado
  - C-SEP, G-statistic , Minimun Description Length (MDL).

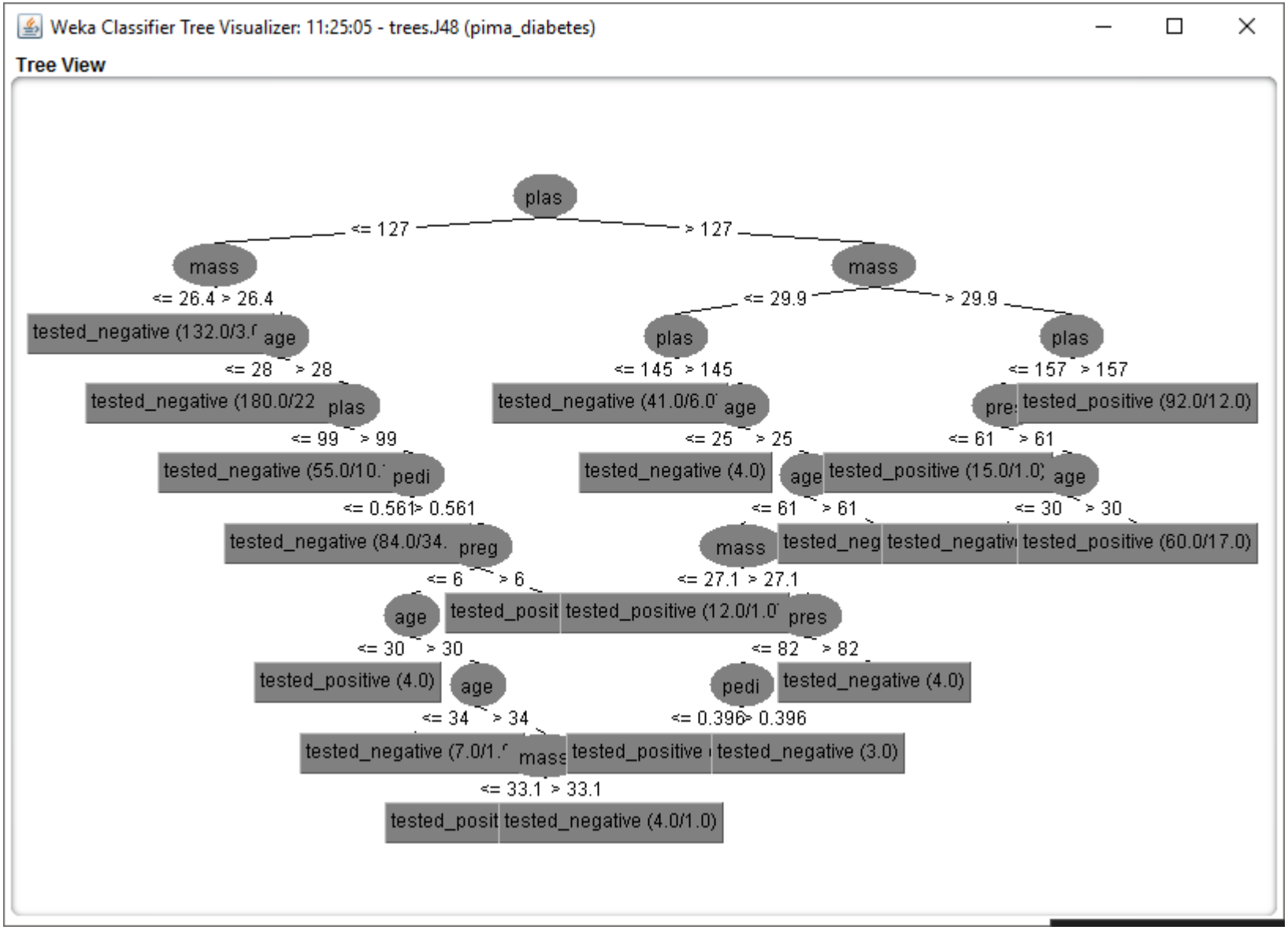
PROBLEMAS NO USO DAS ÁRVORES DE DECISÃO

- Super adaptação:
  - A árvore construída pode se super ajustar aos dados de treinamento.
  - Ramos demais. Alguns desses ramos podem refletir anomalias devido a ruídos e outliers.
  - Acurácia ruim para instâncias de teste.

COMO EVITAR A SUPER ADAPTAÇÃO

- Duas abordagens para evitar esse problema:
  - Pré-poda: Termine cedo a construção da árvore e não particione um nó se o benefício estiver abaixo de um limiar.
    - Difícil escolher um limiar apropriado.
  - Pós-poda: Remover ramos da árvore completamente construída, obtendo uma sequência de árvores progressivamente podadas.
    - Use diversos datasets para validar se a sua árvore está bem podada.

APLICAR A CLASSIFICAÇÃO USANDO O WEKA



REGRAS DE CLASSIFICAÇÃO

Training set

Age	Heart rate	Blood pressure	Heart problem
65	78	150/70	Yes
37	83	112/76	No
71	67	108/65	No

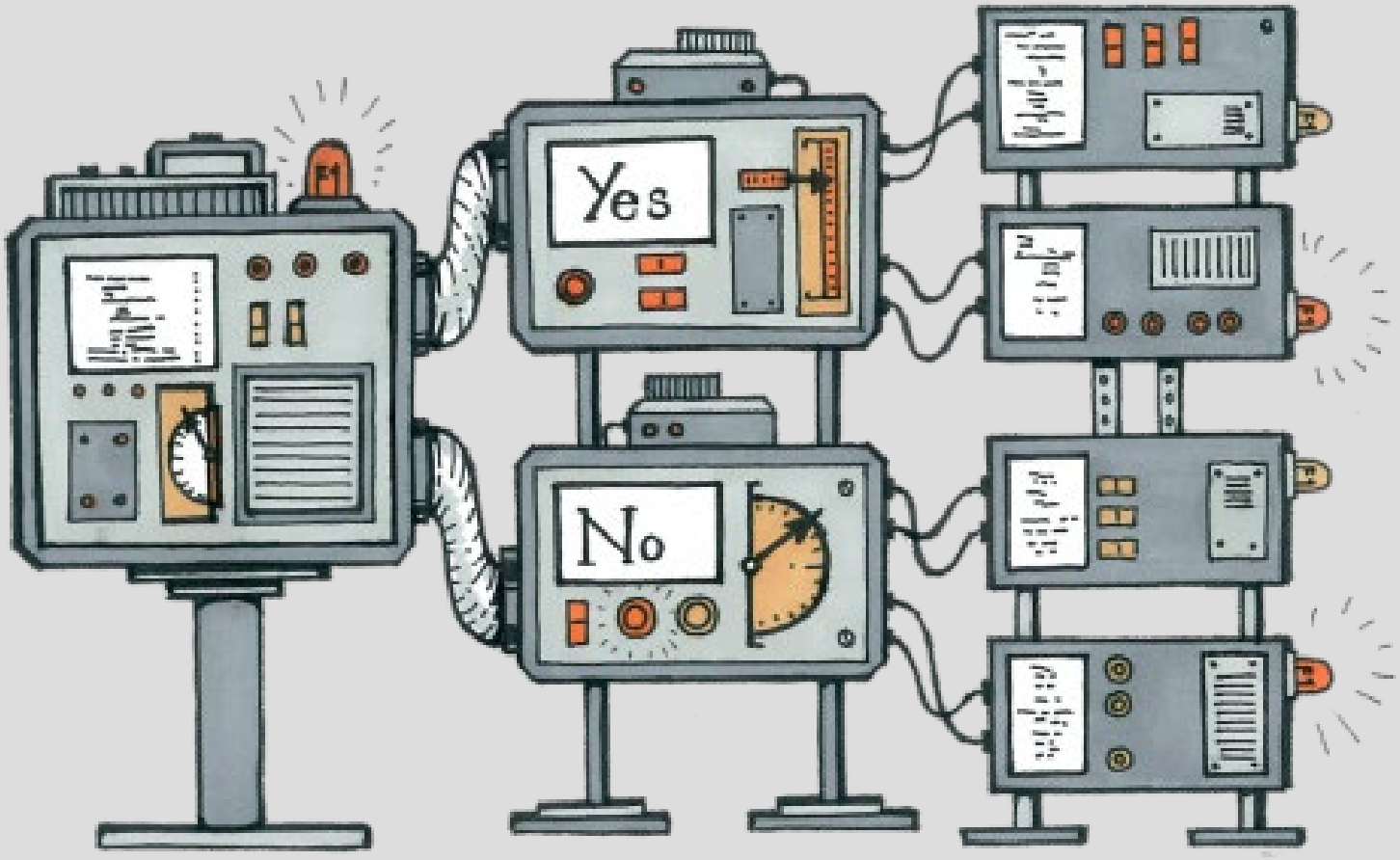
Prediction set

Age	Heart rate	Blood pressure	Heart problem
43	98	147/89	?
65	58	106/63	?
84	77	150/65	?

TABLE 1 – TRAINING AND PREDICTION SETS FOR MEDICAL DATABASE

```
IF (Age=65 AND Heart rate>70) OR (Age>60 AND Blood pressure>140/70)
THEN Heart problem=yes
```

ALGORITMO PART (REGRAS)



ALGORITMO NAYVE BAYES

- Classificador estatístico.
  - Previsão
    - Probabilidade de uma instância pertencer a uma determinada classe.
    - Classe de maior probabilidade é escolhida.
  - Baseado no teorema de Bayes.
  - Cada instância pode aumentar ou diminuir a probabilidade da hipótese estar correta.

ALGORITMO NAYVE BAYES – TEOREMA DE BAYES

$$P(H \mid \boldsymbol{X}) = \frac{P(\boldsymbol{X}|H)P(H)}{P(\boldsymbol{X})} = P(\boldsymbol{X}|H) \times P(H)/P(\boldsymbol{X})$$

- X é uma instância (evidência) de classe desconhecida
- H é uma hipótese de que X pertence a classe C
- Classificação determina P(H|X) (probabilidade a posteriori): a probabilidade da hipótese H verdadeira se for observada uma instância particular X
- P(H) (probabilidade a priori): a probabilidade inicial de uma instância qualquer ser da classe C
- P(X): probabilidade da instância X ser observada
- P(X|H) (likelihood): probabilidade de observar a amostra X, dado que a hipótese é verdadeira

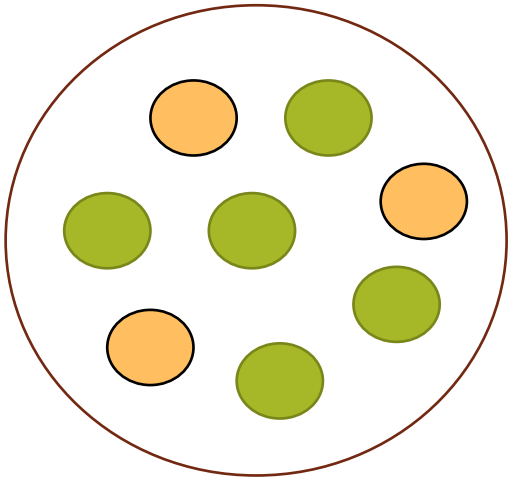
ALGORITMO NAYVE BAYES – EXEMPLOS

- Meningite causa rigidez na nuca em 50% dos casos
- Probabilidade a priori de um paciente ter menigite é 1/50,000
- Probabilidade de um paciente ter rigidez na nuca é 1/20

Se um paciente tem rigidez na nuca, qual a probabilidade de ter meningite?

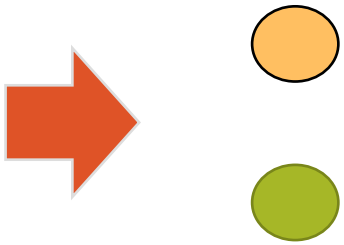
$$P(M \mid \boldsymbol{S}) = \frac{P(\boldsymbol{S}|M)P(M)}{P(\boldsymbol{S})} = \frac{0.5 \times 1/50000}{1/20} = 0.0002$$

ALGORITMO NAYVE BAYES GENÉRICO



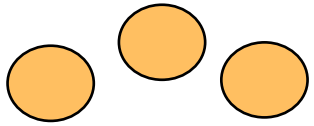
Dataset de treinamento

$$P(H \mid \boldsymbol{X}) = \frac{P(\boldsymbol{X}|H)P(H)}{P(\boldsymbol{X})}$$

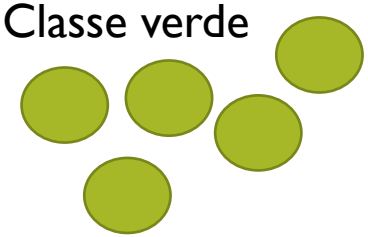


Analise de dados e definição das classes

Classe laranja



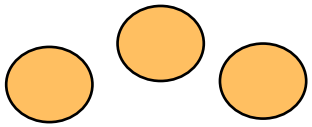
Classe verde



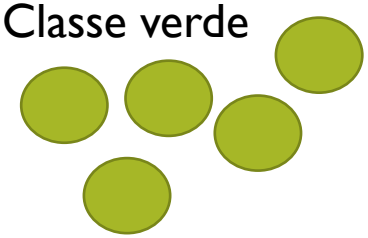
Classifica as instâncias nas classes definidas

ALGORITMO NAYVE BAYES GENÉRICO

Classe laranja



Classe verde

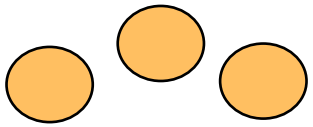


Classes definidas pelo dataset de treinamento

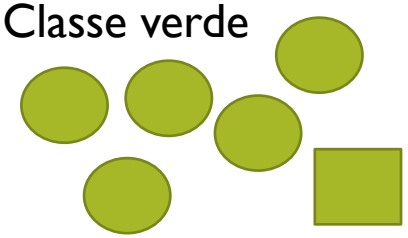
$$P(H \mid \boldsymbol{X}) = \frac{P(\boldsymbol{X}|H)P(H)}{P(\boldsymbol{X})}$$

Classificar uma nova instância

Classe laranja



Classe verde



A nova instância é classificada com base nos dados de treinamento e o modelo é reavaliado.

ALGORITMO NAYVE BAYES – MENSURAR PARÂMETROS

- Gaussian Naive Bayes
  - Indicado para quando os valores associados aquela classe são atributos contínuos.
- Multinomial naive Bayes
  - Quando os eventos avaliados gera um modelo multinominal.
- Bernoulli naive Bayes
  - Quando o modelo é composto por variáveis binárias ou boleanas.
- Semi-supervised parameter estimation
  - Quando o modelo é treinado por um dataset, previamente classificado.

VANTAGENS E DESVANTAGENS DO NAYVE BAYES

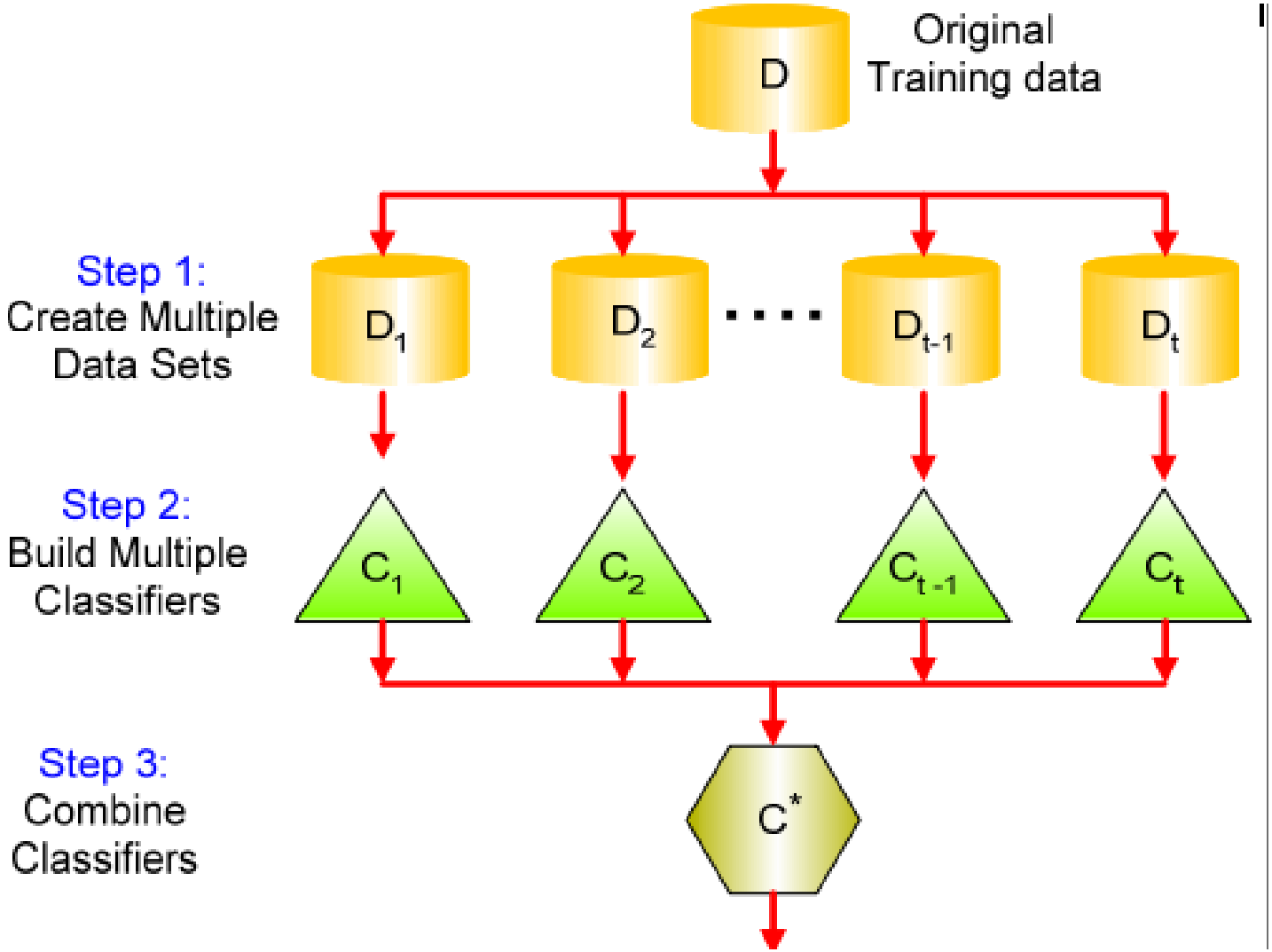
- Vantagens
  - Fácil implementação
  - Bons resultados em muitos casos

- Desvantagens
  - Pressuposição: independência dos atributos perdendo assim a acurácia.
  - Na prática sempre existe uma dependência: Ex:. Hospital: pacientes: prontuários, casos: doenças.
  - A dependência entre os atributos não podem ser mensuradas no modelo de Bayes.

ABORDAGENS USANDO MÚLTIPLOS CLASSIFICADORES

- Classificadores diferentes podem encontrar resultados diferentes para parte das instâncias.
- Ao invés de escolher um único classificador, em alguns casos são utilizados um conjunto de algoritmos.
- A classificação nesses casos virá da combinação dos resultados de todos os classificadores.

ABORDAGENS USANDO MÚLTIPLOS CLASSIFICADORES



ABORDAGENS USANDO MÚLTIPLOS CLASSIFICADORES

- Classificadores diferentes podem encontrar resultados diferentes para parte das instâncias.
- Ao invés de escolher um único classificador, em alguns casos são utilizados um conjunto de algoritmos.
- A classificação nesses casos virá da combinação dos resultados de todos os classificadores.

ABORDAGENS USANDO MÚLTIPLOS CLASSIFICADORES

- Combine uma série de k modelos de classificação, M1,M2,..., Mk, com o objetivo de criar um modelo melhor M*

- Abordagens mais comuns
  - Bagging: classificação da maioria
  - Boosting: votos ponderados de classificação

ABORDAGENS USANDO MÚLTIPLOS CLASSIFICADORES

- Bagging (bootstrap aggregating)
  - Treinamento
    - Dado um conjunto D com d instâncias, a cada iteração I, um conjunto conjunto Di com d instância instância é amostrado amostrado com reposição reposição de D (bootstrap)
    - Um modelo de classificador Mi é aprendido de cada Di
  - Classificação
    - Cada classificador Mi retorna sua predição de classe
    - O classificador agregador M* conta os votos e designa a classe com a maioria dos votos

ABORDAGENS USANDO MÚLTIPLOS CLASSIFICADORES

- Bagging (bootstrap aggregating)
  - Variações
    - Tamanho do subconjunto
    - Amostragem sem reposição
    - Amostragem de atributos e não de instâncias
    - Modelos com algoritmos diferentes

ABORDAGENS USANDO MÚLTIPLOS CLASSIFICADORES

- Boosting
  - Os classificadores posteriores focam em exemplos que foram classificados errado pelos classificadores anteriores
  - Pondere as predições dos classificadores pelo seus erros
  - Pode ser melhor que bagging, mas pode super adaptar aos exemplos difíceis (classificados errado)

ABORDAGENS USANDO MÚLTIPLOS CLASSIFICADORES

- Boosting
  - Os classificadores posteriores focam em exemplos que foram classificados errado pelos classificadores anteriores
  - Pondere as predições dos classificadores pelo seus erros
  - Pode ser melhor que bagging, mas pode super adaptar aos exemplos difíceis (classificados errado)

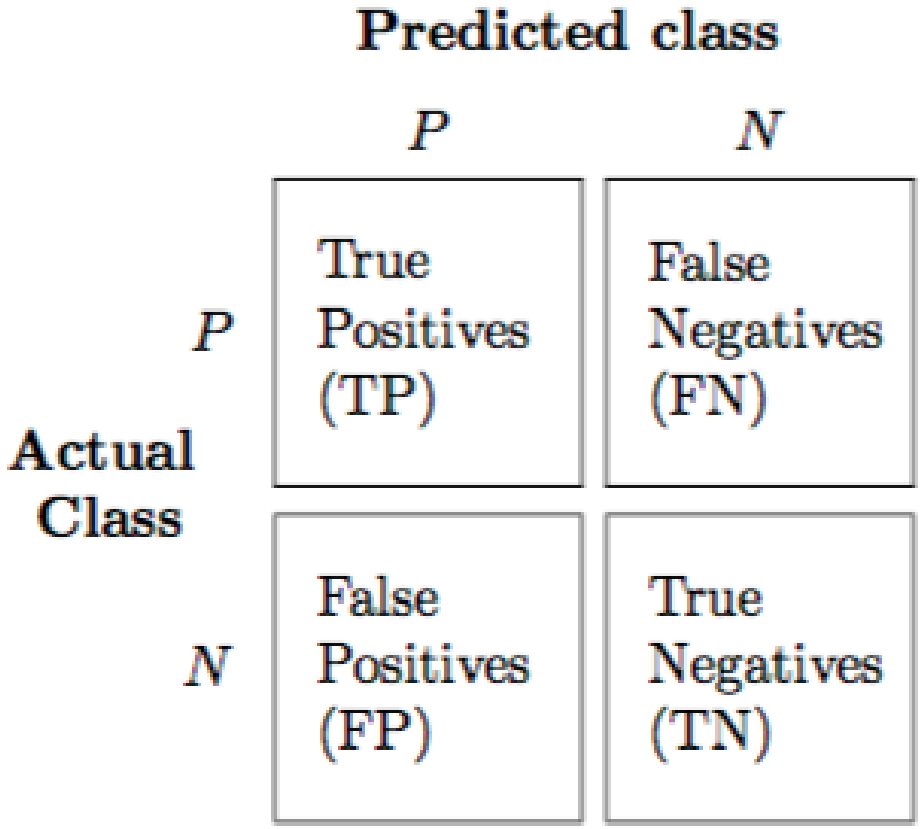
AVALIAÇÃO DOS CLASSIFICADORES

- Como avaliar um classificador ?

- Como obter uma estimativa de avaliação confiável ?

AVALIAÇÃO DOS CLASSIFICADORES

- Matriz de confusão



AVALIAÇÃO DOS CLASSIFICADORES

- Matriz de confusão

		Condition (as determined by "Gold standard")		
		Condition Positive	Condition Negative	
Test Outcome	Test Outcome Positive	**True Positive**	**False Positive** (Type I error)	Positive predictive value = $\frac{\Sigma \text{ True Positive}}{\Sigma \text{ Test Outcome Positive}}$
	Test Outcome Negative	**False Negative** (Type II error)	**True Negative**	Negative predictive value = $\frac{\Sigma \text{ True Negative}}{\Sigma \text{ Test Outcome Negative}}$
		Sensitivity = $\frac{\Sigma \text{ True Positive}}{\Sigma \text{ Condition Positive}}$	**Specificity** = $\frac{\Sigma \text{ True Negative}}{\Sigma \text{ Condition Negative}}$	

AVALIAÇÃO DOS CLASSIFICADORES

- Medidas
  - Acurácia= (TP + TN)/All
  - Error rate = 1 – accuracy = (FP + FN)/All
  - Sensibilidade ou Recall = TP/P
  - Especificidade = TN/N
  - Precisão = TP/P’

AVALIAÇÃO DOS CLASSIFICADORES -EXEMPLO

Classe Atual / Predicted class	**Cancer = yes**	**Cancer = no**	**Total**
Cancer = yes	90	210	300
Cancer = no	140	9560	9700
Total	230	9770	10000

- Precisão = 90/230 = 39.13%
- VPP = 90/300 = 30.00%

- CUIDADO COM DESBALANÇO DE CLASSES

MÉTODOS DE AVALIAÇÃO

- Holdout
- Validação cruzada (cross validation)
- Amostragem aleatória

MÉTODOS DE AVALIAÇÃO - HOLDOUT

- Particionamento aleatório em conjuntos independentes
  - Conjunto de treino (ex: 2/3) para construir modelo
  - Conjunto de teste (ex: 1/3) para estimar acurácia

- Amostragem aleatória: variação
  - Repita holdout k vezes, acurácia = média da acurácia

MÉTODOS DE AVALIAÇÃO – CROSS VALIDATION

- k-fold, valor mais popular k = 10
- Particionar os dados em k subconjuntos mutualmente exclusivos, com aproximadamente mesmo tamanho exclusivos, com aproximadamente mesmo tamanho

- Na iteração i, use Di como teste e demais como treino
  - Leave-one-out: k folds, k = quant. instâncias, para dados de pequeno tamanho
  - Stratified cross-validation: folds mantém a mesma distribuição de classes do conjunto original

MÉTODOS DE AVALIAÇÃO - HOLDOUT

- Selecionando um classificador
  - Acurácia: predizer a classe
  - Velocidade: tempo de construção do modelo (treino)
    - tempo de construção do modelo (treino)
    - tempo de aplicação do modelo (tempo de classificação/tempo)
  - Robustez: lidar com ruído e valores ausentes
  - Escalabilidade: eficiência em bases de dados em disco
  - Interpretabilidade: compreensão do modelo

DESAFIO

- Abra o dataset IRIS disponibilizado no WEKA. Explore os seguintes aspectos:

  - Descubra do que se trata esse dataset, a semântica dos atributos e os tipos deles.
  - Realize o pré-processamento caso necessário.
  - Tente classificar as instâncias usando o WEKA.
  - Qual foi o melhor classificador ?

INTELIGÊNCIA COMPUTACIONAL

ALGORITMOS DE CLASSIFICAÇÃO

FELIPE TORRES