Departamento de Engenharia de Teleinformática - DETI

Inteligência Computacional Aplicada
(Introdução ao Reconhecimento de Padrões)

Prof. Dr. Guilherme de Alencar Barreto

Depto. Engenharia de Teleinformática (*DETI/UFC*)

URL: www.deti.ufc.br/~guilherme

Email: guilherme@deti.ufc.br

Setembro/2007

Ementa

1. O Que é Ser Inteligente?
2. Inteligência e o Reconhecimento de Padrões (RP)
3. Conceituação Intuitiva de RP
4. Um Computador pode Reconhecer Padrões?
5. Definição Formal de RP
6. Representação na Forma de Vetor de Atributos
7. Distância entre Vetores (euclidiana, quarteirão, etc.)
8. Método do Vizinho Mais Próximo

Material Didático

1. Barreto, G. A. (2007). **Introdução ao Reconhecimento de Padrões**, Apostila para Cursos de Extensão. Disponível em http://www.deti.ufc.br/~guilherme/courses.htm

2. Bittencourt, G. A. (2006). **Inteligência Artificial: Ferramentas e Teorias**, 3a. edição, Editora da UFSC.

3. Kasabov, N. K. (1996). **Foundations of Neural Networks, Fuzzy Systems, and Knowledge Engineering**, MIT Press.

3. Repositório de Dados: http://www.ics.uci.edu/~mlearn/MLSummary.html

1. O Que Ser Inteligente?

Seria resolver um problema específico com exatidão?

(e.g. ser um mestre do xadrez ou médico especialista)

OU

Resolver problemas genéricos de modo aproximado?

(e.g. determinar a vaga adequada no estacionamento)

OU

Ter conhecimento Enciclopédico?

(ser um papai ou mamãe-sabe-tudo)

OU

Tocar um instrumento? Falar outras línguas? Jogar bola bem?

2. Inteligência e o Reconhecimento de Padrões

Seres vivos são bastante habilidosos em reconhecer padrões

- Comportamentais (fulano se comporta sempre assim!)
- Sonoros (Este barulho não é normal!)
- Táteis (Este tecido é parecido, mas a textura é diferente!)
- Visuais (Parece que vai chover hoje!)
- Olfativos (Você trocou de perfume?)
- Lógico-Matemáticos (Lembra de "Rain Main"?)

Reconhecer padrões equivale a classificar determinado objeto físico ou situação como pertencente ou não a um certo número de categorias previamente estabelecidas.

3. Conceito Intuitivo de RP -slide1

Grupo 1 (laranjas)

Grupo 2 (maças)

Este objeto

pertence a qual dos grupos anteriores?

3. Conceito Intuitivo de RP -slide2

Grupo 1 (laranjas)

Grupo 2 (maças)

Grupo 3 (tangerinas)

Este objeto

pertence a qual dos grupos anteriores?

4. Um Computador Pode Reconhecer Padrões?

- Certamente, a sua decisão é tomada com base no **grau de similaridade** entre a fruta desconhecida e as frutas conhecidas.
- Que mecanismo seu cérebro usa para realizar esta tarefa?
- Será que implementa uma **comparação** entre o objeto novo e objetos armazenados?
- Pode-se "replicar" este mecanismo em uma máquina ???
- Para comparar objetos precisamos de
  - Uma **representação** do atributos físicos das frutas.
  - Um **aprendizado** apreender o conceito laranja/maçã.
  - Uma **memória** para armazenar as frutas aprendidas.
  - Uma **regra de decisão** para classificar a nova fruta.

5. Definição Formal de RP

- Assim, para definir um Problema de RP, precisamos de:
  - Um número finito de K classes: C_1, C_2, ..., C_K
  - Um número finito de N_i objetos por classe C_i
  - Um número finito de n atributos (*features*) para representar numericamente cada objeto físico.
  - Mecanismos de memória e/ou aprendizado.
  - Uma regra de decisão para classificar novos objetos.
  - Critérios de avaliação do classificador.

6. Representação na Forma de Vetor de Atributos – slide 1

- Quais são os atributos que descrevem uma tangerina?

1. Formato (esférico/oval)?
2. Fruta cítrica?
3. Cor?
4. Casca lisa ou rugosa?
5. Cheiro ativo?

- Todos esses atributos são igualmente importantes?
- Quão difícil é a tarefa de definir os atributos de um objeto?

6. Representação na Forma de Vetor de Atributos – slide 2

- Respondendo às perguntas anteriores:

  X1. esférico.

  X2. sim.

  X3. alaranjado.

  X4. rugosa.

  X5. sim.

- Provavelmente NÃO!
- Horrivelmente árdua! :-(

6. Representação na Forma de Vetor de Atributos – slide 3

- O computador só entende números!
- Como transformar os atributos em números?
- Basta representar cada objeto como um vetor de atributos!

$$X = [X1 \;\; X2 \;\; \ldots \; X_n]$$

- Assim, este objeto (tangerina)

- É representado por este vetor: $X = [0 \;\; 1 \;\; 2 \;\; 1 \;\; 1]$

6. Representação na Forma de Vetor de Atributos – slide 4

- **Exercício 1**: Como os objetos laranja e maçã seriam representados como vetor de atributos?

X1 = {esférico, oval, alongado} = {0,1,2}

X2 = {não, sim} = {0,1}

X3 = {amarelo, vermelho, alaranjado,verde} = {0,1,2,3}

X4 = {lisa, rugosa} = {0,1}

X5 = {não, sim} = {0,1}

7. Distância entre Vetores - slide1

- Objeto Laranja: **X** = [0 1 2 1 0]
- Objeto Maçã: **Y** = [0 0 1 0 0]
- Objeto Tangerina: **Z** = [0 1 2 1 1]

- O vetor **Z** mais parecido com o vetor **X** ou com o vetor **Y** ???

X = [0 1 2 1 0]

Y = [0 0 1 0 0]

Z = [0 1 2 1 1]

Z = [0 1 2 1 1]

7. Distância entre Vetores - slide2

- Calcular quão "parecidos" são dois vetores equivale a calcular a distância entre eles!

- Distância *city-block* (quarteirão ou Manhattan)

$$Dq(\mathbf{X},\mathbf{Y}) = |X1 - Y1| + |X2 - Y2| + \ldots + |X_n - Y_n|$$

- Distância *euclidiana*

$$De(\mathbf{X},\mathbf{Y}) = \sqrt{(X1 - Y1)^2 + (X2 - Y2)^2 + \ldots + (X_n - Y_n)^2}$$

- Note que $Dq(\mathbf{X},\mathbf{Y}) = Dq(\mathbf{Y},\mathbf{X})$ e que $De(\mathbf{X},\mathbf{Y}) = De(\mathbf{Y},\mathbf{X})$!

7. Distância entre Vetores - slide3

- **Exercício 2**: Calcular as distâncias Dq(**X**, **Z**) e Dq(**Y**, **Z**)?
- **Exercício 3**: Calcular as distâncias De(**X**, **Z**) e De(**Y**, **Z**)?

X = [0 1 2 1 0]

Z = [0 1 2 1 1]

Y = [0 0 1 0 0]

Z = [0 1 2 1 1]

7. Distância entre Vetores - slide4

- **Resolução dos Exercícios 2 e 3**:

$\mathbf{X} = [0 \quad 1 \quad 2 \quad 1 \quad 0]$ $\qquad$ $\mathbf{Y} = [0 \quad 0 \quad 1 \quad 0 \quad 0]$

$\mathbf{Z} = [0 \quad 1 \quad 2 \quad 1 \quad 1]$ $\qquad$ $\mathbf{Z} = [0 \quad 1 \quad 2 \quad 1 \quad 1]$

$$Dq(\mathbf{X}, \mathbf{Z}) = |0-0| + |1-1| + |2-2| + |1-1| + |0-1| = 1$$

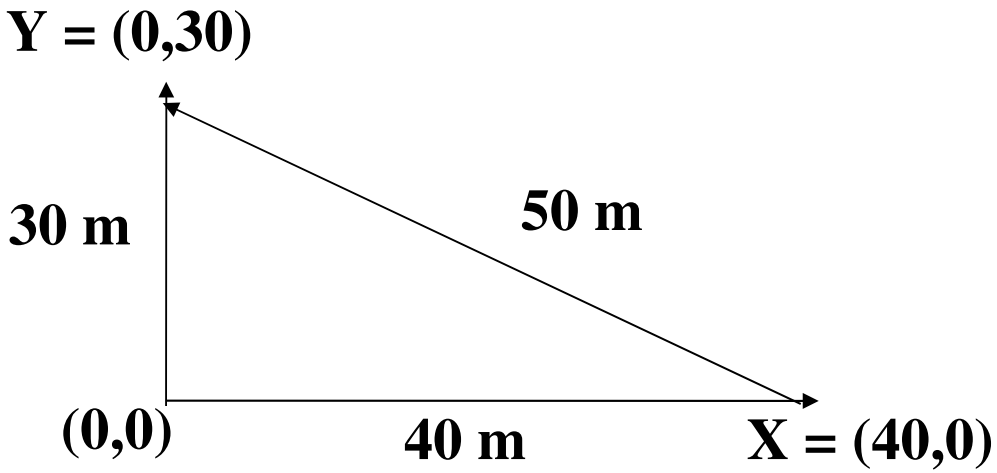
$$De(\mathbf{X}, \mathbf{Z}) = (0-0)^2 + (1-1)^2 + (2-2)^2 + (1-1)^2 + (0-1)^2 = 1$$

$$Dq(\mathbf{Y}, \mathbf{Z}) = |0-0| + |0-1| + |1-2| + |0-1| + |0-1| = 4$$

$$De(\mathbf{Y}, \mathbf{Z}) = \sqrt{(0-0)^2 + (0-1)^2 + (1-2)^2 + (0-1)^2 + (0-1)^2} = 2$$

7. Distância entre Vetores - slide5

- Considere que os catetos do triângulo retângulo abaixo correspondem aos lados de um terreno.



- A distância quarteirão corresponde à distância percorrida para ir do ponto X ao ponto Y pela “calçada”, ou seja 70 m.
- A distância euclidiana é a distância percorrida atravessando o terreno pela diagonal, ou seja 50 m.

7. Distância entre Vetores - slide6

- Nos exemplos anteriores os atributos assumiram apenas valores numéricos inteiros.
- Contudo, na prática, é muito comum que alguns dos atributos assumam valores numéricos fracionários (reais).
- A título de ilustração, considere o conjunto de dados referentes à classificação de flores Iris. Este conjunto possui
  - 150 exemplos divididos em 3 classes (50 exemplos/classe).
  - Classes = {setosa, versicolor e virginica}
  - Cada exemplo é descrito por 4 atributos (unidade *cm*).

$$\mathbf{X} = [X1 \;\; X2 \;\; X3 \;\; X4] = [SL \;\; SW \;\; PL \;\; PW]$$

7. Distância entre Vetores - slide7

- X1= comprimento da sépala (SL), X2 = largura da sépala (SW)
- X3= comprimento da pétala (PL), X4 = largura da pétala (PW)

No.	SL	SW	PL	PW	Class
1	5.1	3.5	1.4	0.2	Setosa
2	4.7	3.2	1.3	0.2	Setosa
3	5.0	3.6	1.4	0.2	Setosa
4	6.5	2.8	4.6	1.5	Versicolor
5	6.3	3.3	4.7	1.6	Versicolor
6	6.6	2.9	4.6	1.3	Versicolor
7	7.1	3.0	5.9	2.1	Virginica
8	6.5	3.0	5.9	2.2	Virginica
9	6.5	3.2	5.1	2.0	Virginica
10	6.8	3.0	*5.5*	2.1	Virginica

7. Distância entre Vetores - slide8

- **Exercício 4: Calcular distâncias entre X, Y, Z usando Excel.**

$\mathbf{X}$ = [5,1 3,5 1,4 0,2] (exemplo da classe setosa)

$\mathbf{Y}$ = [6,5 2,8 4,6 1,5] (exemplo da classe versicolor)

$\mathbf{Z}$ = [7,1 3,0 5,9 2,1] (exemplo da classe virginica)

$Dq(\mathbf{X}, \mathbf{Y}) = 6{,}60$ $\qquad$ $De(\mathbf{X}, \mathbf{Y}) = 3{,}79$

$Dq(\mathbf{X}, \mathbf{Z}) = 8{,}90$ $\qquad$ $De(\mathbf{X}, \mathbf{Z}) = 5{,}30$

- No próximo slide vamos averiguar se a "representação numérica" das flores íris é consistente com a percepção visual.

7. Distância entre Vetores - slide9

- **Exercício 5: Calcular distâncias entre X, Y, Z usando "Olhômetro".** O resultado bate com o exercício anterior?

Iris setosa (**X**)

Iris Versicolor (**Y**)

Iris Virginica (**Z**)

8. Classificadores Elementares – slide 1

Algoritmo: Vizinho Mais Próximo (*Nearest Neighbor, NN*)

1. Armazenar os exemplos em uma tabela.
2. Seja $\mathbf{X}_{new}$ um vetor cuja classe é desconhecida, ou seja:

$$\text{Classe}(\mathbf{X}_{new}) = ?$$

3. Procurar na tabela o vetor armazenado mais próximo de $\mathbf{X}_{new}$.
4. Chamar de $\mathbf{X}_{near}$ o vetor armazenado mais próximo de $\mathbf{X}_{new}$.
5. Atribuir a $\mathbf{X}_{new}$ a mesma classe de $\mathbf{X}_{near}$, ou seja:

$$\text{Classe}(\mathbf{X}_{new}) = \text{Classe}(\mathbf{X}_{near})$$

6. Se a classificação for correta incluir $\mathbf{X}_{new}$ na tabela.

8. Classificadores Elementares – slide 2

- **Observações Importantes**

1. Antes de usar o classificador, escolher *aleatoriamente* alguns poucos exemplos para "testar" o classificador.

2. Para testar o classificador "fingimos" não conhecer a classe dos exemplos selecionados no Item 1.

3. A taxa de acerto do classificador é dada por:

$$\text{Taxa de acerto} = \frac{\text{No. exemplos de teste classificados corretamente}}{\text{Número total de exemplos de teste}}$$

8. Classificadores Elementares – slide 3

- **Exercício 6 - Identificar no algoritmo do classificador NN os seguintes mecanismos:**

  Q1. Memória

  Q2. Regra de Decisão

  Q3. Aprendizado

  Respostas: Q1. Passo 1, Q2. Passo 5, Q3. Passo 6.

- **Exercício 7 - Implementar o classificador NN no Excel.**

8. Classificadores Elementares – slide 4

Algoritmo: K-Vizinhos Mais Próximos (*K-NN*)

1. Armazenar os exemplos em uma tabela.
2. Seja $\mathbf{X}_{new}$ um vetor cuja classe é desconhecida, ou seja:

$$\text{Classe}(\mathbf{X}_{new}) = ?$$

3. Encontrar na tabela os K vetores mais próximo de $\mathbf{X}_{new}$.
4. Seja C_K a classe a que pertence a maioria dos K vetores.
5. Atribuir a $\mathbf{X}_{new}$ a classe da maioria dos K vetores, ou seja:

$$\text{Classe}(\mathbf{X}_{new}) = C_K$$

6. Se a classificação for correta incluir $\mathbf{X}_{new}$ na tabela.

8. Classificadores Elementares – slide 5

- **Vantagens do Algoritmo K-NN**

  V1. Simplicidade de implementação.

  V2. Ideal para tabelas pequenas ou médias.

  V3. Não requer treinamento.

- **Limitações do Algoritmo K-NN**

  L1. Custo computacional alto para tabelas grandes.

  L2. A constante *K* é obtida por tentativa-e-erro.

- **Exercício 8 - Implementar o classificador *K*-NN no Excel.**

8. Classificadores Elementares – slide 6

- **Classificador Distância Mínima aos Centróides (DMC)**

  - Cada classe passa a ter um único vetor que a representa, chamado de **centróide**.
  - Assim, todos os exemplos de uma classe não precisam ser mais armazenados (economia de memória).
  - O centróide de uma classe é o seu vetor médio; ou seja, a média dos exemplos daquela classe.
  - Assim, um centróide de uma classe serve como um modelo que representa o objeto médio daquela classe.

8. Classificadores Elementares – slide 7

Algoritmo: Distância Mínima aos Centróides (*DMC*)

1. Armazenar apenas os centróides das classes em uma tabela.
2. Seja $\mathbf{X}_{new}$ um vetor cuja classe é desconhecida, ou seja:

$$\text{Classe}(\mathbf{X}_{new}) = ?$$

3. Encontrar na tabela o centróide mais próximo de $\mathbf{X}_{new}$.
4. Seja $\mathbf{M}_J$ o centróide mais próximo de $\mathbf{X}_{new}$.
5. Atribuir a $\mathbf{X}_{new}$ a classe do centróide mais próximo, ou seja:

$$\text{Classe}(\mathbf{X}_{new}) = \text{Classe}(\mathbf{M}_J)$$

6. Se a classificação for correta, usar $\mathbf{X}_{new}$ para recalcular $\mathbf{M}_J$.

8. Classificadores Elementares – slide 8

- **Passo 6** (cont.) **-** Atualização do centróide selecionado.

$$\mathbf{M}_J(n_J + 1) = (1\text{-}a)\mathbf{M}_J(n_J) + a\mathbf{X}_{new}$$

em que $a = 1/(n_J+1)$ é o fator de aprendizagem.

n_J é o total de exemplos usados para calcular o centróide antes da chegada de $\mathbf{X}_{new}$.

$\mathbf{M}_J(n_J)$ é o centróide antes da chegada de $\mathbf{X}_{new}$.

$\mathbf{M}_J(n_J+1)$ é o novo centróide com $\mathbf{X}_{new}$ incorporado.

- **Exercício 9 - Implementar o classificador DMC no Excel.**

8. Classificadores Elementares – slide 9

- **Árvore de Decisão** (*Decision Tree*)

  classificador baseado em um conjunto regras de decisão

  SE (condição) **ENTÃO** (resultado).

  - As condições e resultados das regras são gerados a partir do particionamento recursivo do espaço de atributos.
  - As condições que compõem a parte "SE" de uma regra é chamada de **antecedentes** ou **premissas.**
  - A conclusão ou resultado que compõem a parte "ENTÃO" de uma regra é chamada de **conseqüente**.

8. Classificadores Elementares – slide 10

- **Classificador Árvore de Decisão** (continuação)

  - As condições de uma regra, em geral, envolvem intervalos para os atributos.

  - As regras são, em geral, geradas (induzidas) após uma análise cuidadosa dos dados por especialistas.

  - A seqüência correta de aplicação das regras funciona como um modelo que explica os dados armazenados.

  - A aplicação da sequência de regras vai classsificando os objetos em classes cada vez menos abrangentes.

8. Classificadores Elementares – slide 11

Algoritmo básico: Árvore de Decisão (*Decision Tree*)

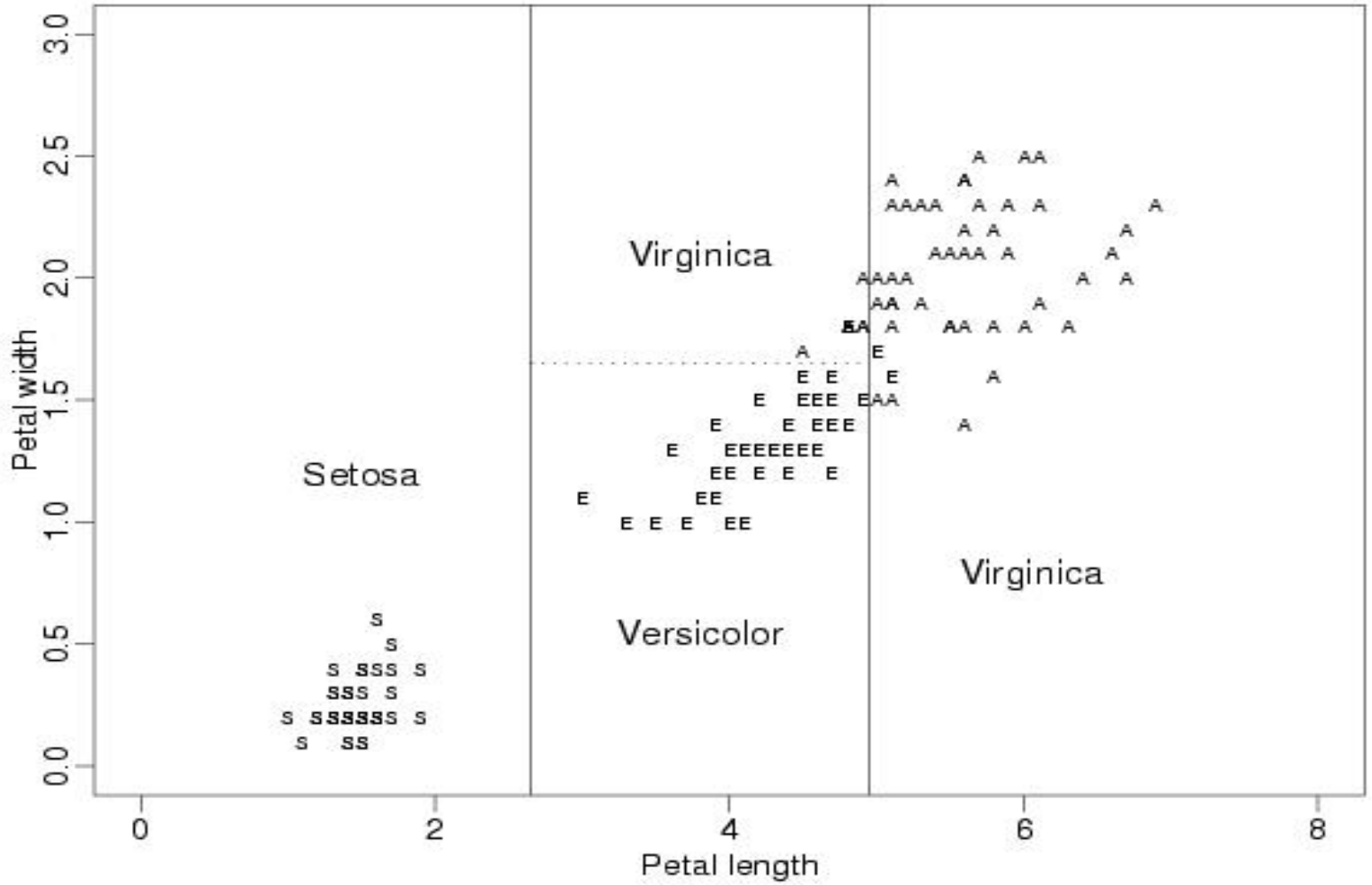
1. Fazer o gráfico de dispersão para um certo par de atributos.

2. Escolher um dos eixos e procurar pontos em que este eixo possa ser dividido por retas perpendiculares que separem bem uma classe das outras.

3. Repetir o procedimento para outros eixos, até que um critério de parada seja satisfeito.

8. Classificadores Elementares – slide 12

Exercício 8: Montar a árvore de decisão para os dados *Iris*.



8. Classificadores Elementares – slide 13

Base de Regras para o Exercício 8.

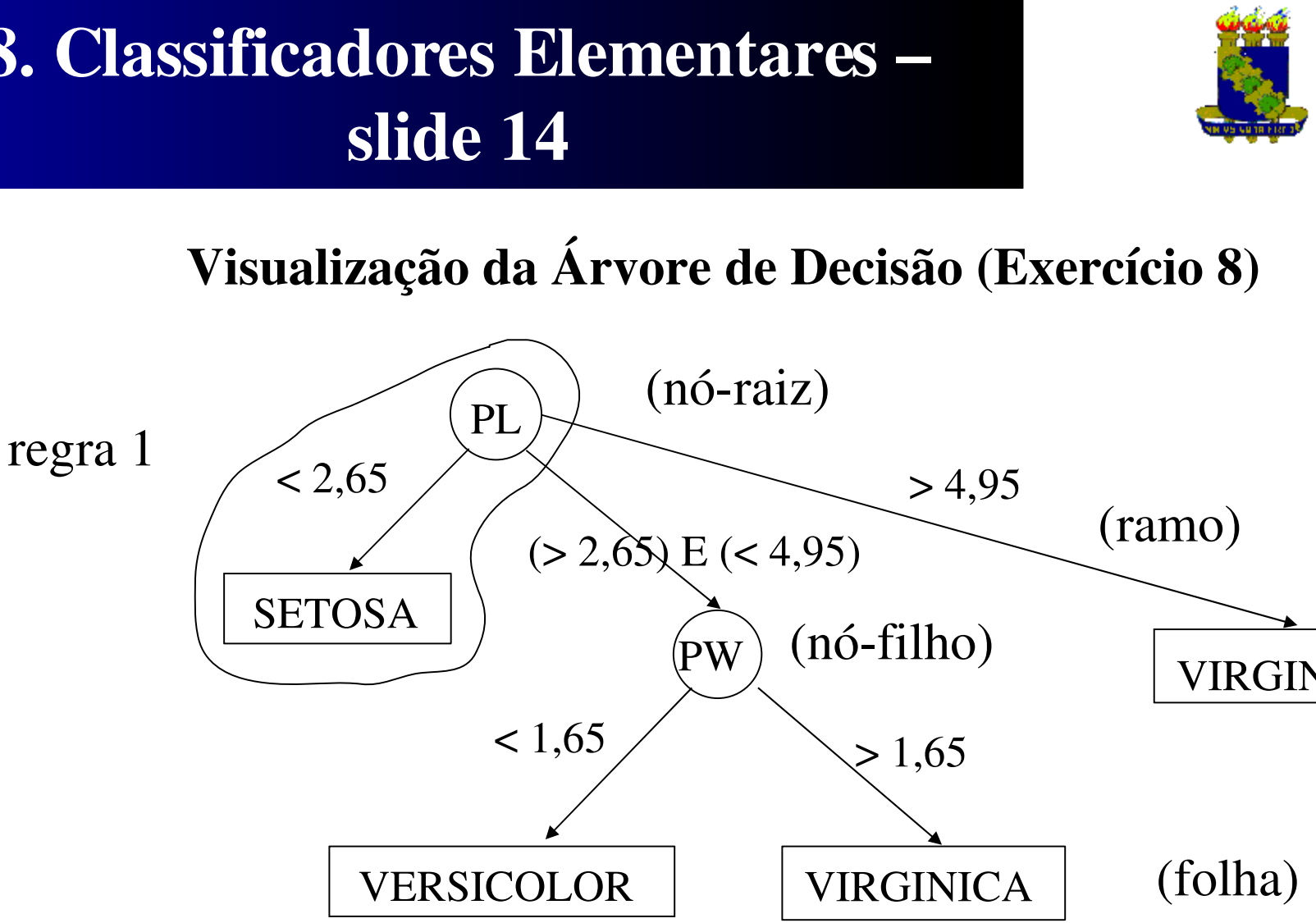
Regra 1. **SE** (PL < 2,65) **ENTÃO** classe=setosa.

Regra 2. **SE** (PL > 4,95) **ENTÃO** classe=virginica.

Regra 3. **SE** (2,65 < PL <4,95) **E** (PW < 1,65),
ENTÃO classe=versicolor.

Regra 4. **SE** (2,65 < PL <4,95) **E** (PW > 1,65),
ENTÃO classe=virginica.

8. Classificadores Elementares – slide 14



OBS: Cada percurso na árvore (da raiz à folha) corresponde a uma regra de classificação.

8. Classificadores Elementares – slide 15

- **Vantagens do Classificador Árvore de Decisão**

  V1. Simplicidade de implementação.

  V2. "Imita" o processo de raciocínio humano.

- **Limitações do Classificador Árvore de Decisão**

  L1. Cansativo quando se tem muitos atributos.

- **Exercício 10 - Implementar uma árvore de decisão no Excel.**

8. Classificadores Elementares – slide 16

- **Exemplo: Concessão de Empréstimo Bancário**

Que seqüência de perguntas (regras) deve ser construída a fim de conceder crédito a alguém?

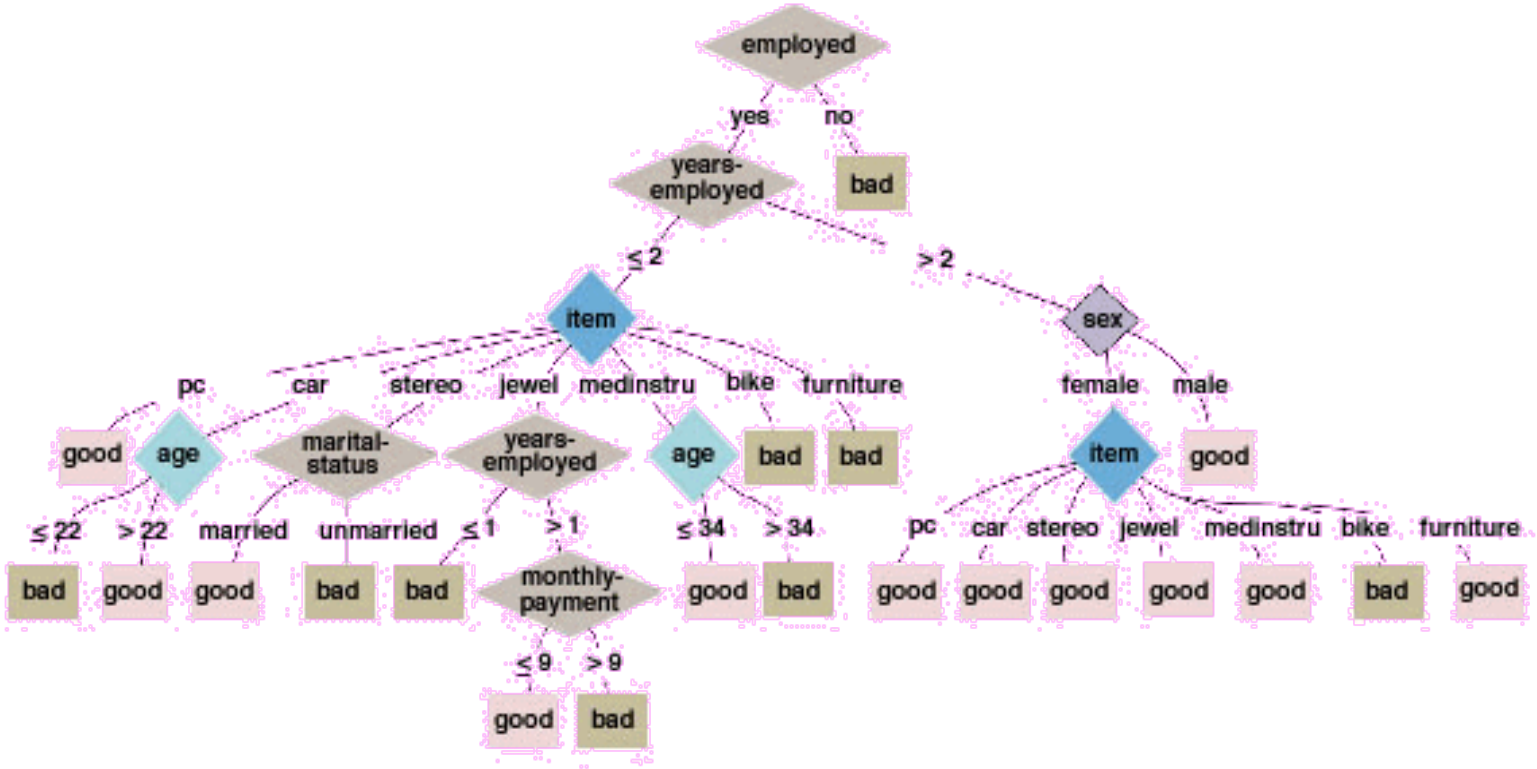
Que tal começar por:

1. O requisitante está empregado? (SIM/NÃO)
   1.1. Caso negativo, não conceder crédito.
   1.2. Caso afirmativo, fazer próxima pergunta.
2. Há quantos anos o requisitante está empregado?
   2.1. Se mais de 2 anos, qual o sexo do requisitante?
   2.21 Se menos de ou igua a 2 anos, quais suas posses?

8. Classificadores Elementares – slide 17

- **Exemplo: Concessão de Empréstimo Bancário (cont.)**

E assim sucessivamente até não haver mais dúvidas.



8. Classificadores Elementares – slide 18

Avaliação de Resultados

- **Matriz de Confusão (MC):** Tabela que mostra o resultado da classificação automática, comparando-o com o resultado real.

  > As colunas correspondem às classes reais.
  > As linhas correspondem às classes preditas.

Exemplo:

	Setosa	Versicolor	Virginica
Setosa	5	0	0
Versicolor	0	5	2
Virginica	0	2	4

8. Classificadores Elementares – slide 19

Avaliação de Resultados (cont.-1)

> Na tabela anterior:

(a) Cinco exemplos da setosa foram corretamente classificados.

(b) Dos sete exemplos da versicolor, 5 foram corretamente classificados, enquanto 2 foram classificados como virginica.

(b) Dos seis exemplos da virginica, 4 foram corretamente classificados, enquanto 2 foram classificados como versicolor.

8. Classificadores Elementares – slide 20

Avaliação de Resultados (cont.-2)

> Vantagens da Matriz de Confusão

(*i*) permite quantificar instantaneamente as taxas de acerto e erro de um classificador.

Por exemplo, usando a tabela anterior:

$$\text{taxa de acerto} = \frac{\text{Soma dos elementos na diagonal}}{\text{Soma dos elementos na tabela}}$$

$$= (5+5+4)/(5+5+4+2+2) = 14/18 = 0{,}78\ (78\%)$$

(*ii*) permite visualizar os TIPOS de erros mais cometidos.

8. Classificadores Elementares – slide 21

- **Tipos de Erros: Falso Positivo (FP) e Falso Negativo (FN)**

Considere o problema de concessão de crédito. O sistema deve confirmar se deve conceder crédito ao cliente.

(*i*) ***Erro FP***: ocorre quando um "mau cliente" é classificado pelo sistema como "bom cliente". Neste caso, o cliente é recompensado indevidamente.

(*ii*) ***Erro FN***: ocorre quando um "bom cliente" é classificado pelo sistema como "mau cliente". Neste caso, o cliente é penalizado injustamente.

8. Classificadores Elementares – slide 22

- **Matriz de confusão para classificação binária**

	Bom cliente	Mau cliente
Bom cliente	**TP**	**FP**
Mau cliente	**FN**	**TN**

TP: True Positive (positivos verdadeiros)

TN: True Negative (negativos verdadeiros)

Importante - Os diferentes tipos de erro têm impacto (custo) diferente para a empresa que concede o crédito!