Universidade Federal de Uberlândia

Faculdade de Computação

Estatística Computacional

Alunos: Pedro Henrique Resende Ribeiro **Nº:** 12011BCC004

Kemuel Santos Peres 11811BCC035

Nayara Terezinha Nunes 11911BCC006

Lista 04

Exercício 01: Considere a seguinte matriz de distâncias:

$$
\begin{array}{c|ccccc|}
 & A & B & C & D & E \\
\hline
A & 0 & 9 & 3 & 6 & 11 \\
B & 9 & 0 & 7 & 5 & 10 \\
C & 3 & 7 & 0 & 9 & 2 \\
D & 6 & 5 & 9 & 0 & 8 \\
E & 11 & 10 & 2 & 8 & 0 \\
\hline
\end{array}
$$

(a) Com base na matriz de distâncias acima, esboce o dendograma que resulta do processo de aglomeração hierárquica dessas 5 observações usando o método **complete** como a distância entre dois aglomerados.

Resolução: O método complete considera a maior distância dos dados entre dois aglomerados. Será considerado que as observa-ções possuem nomes A, B, C, D e E (ver matriz).

Pode-se notar que a menor distância existente na matriz é CE = 2 = EC (os números 0 são a distância de uma observação a ela mesma (AA, BB, etc.), ou seja, são consideradas inicialmente para formar aglomerados que contém apenas uma única observação). Dessa forma, o primeiro aglomerado a ser formado é o CE, que une as observações C e E.

Em seguida, deve-se analisar as distâncias entre A, B, D e o aglomerado CE. Como o método complete seleciona a maior distância entre dois aglomerados, as distâncias a serem escolhidas são: d(A,B) = 9, d(A,D) = 6, d(B,D) = 5, d(A,CE) = 11, d(B,CE) = 10 e d(D,CE) = 9. Dessa forma, a menor distância encontrada é d(B,D) = 5.

O próximo passo é analisar as distâncias entre A e os BD e CE. As distâncias a serem escolhidas são: d(A,BD) = 9, d(A,CE) = 11, d(BD,CE) = 10. Dessa forma, a menor distância encontrada é d(A,BD) = 9.

Por último, tem-se a distância entre os aglomerados BD e ACE. A distância entre eles, através do método complete, é d(ABD,CE) = 11. O aglomerado hierárquico é mostrado na Figura 1.

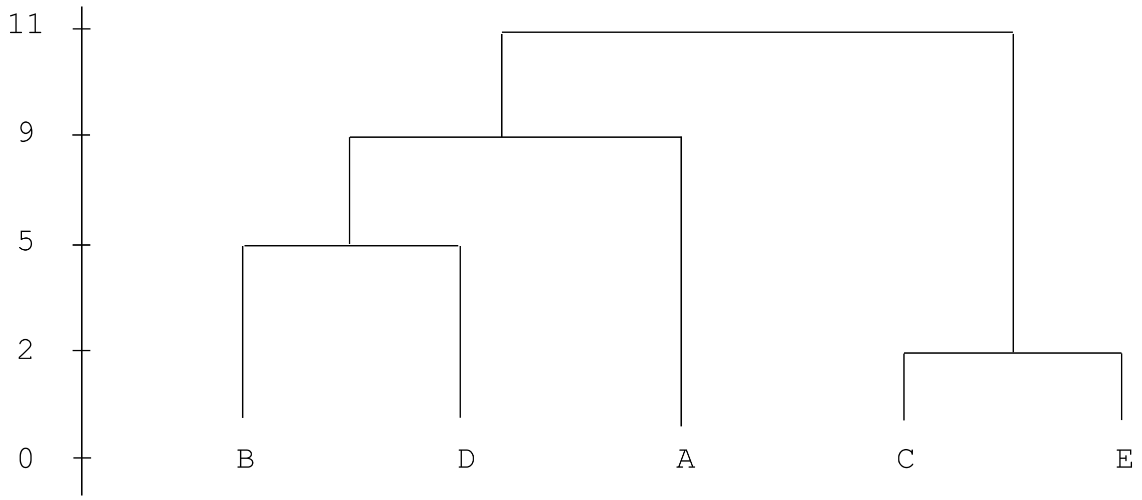

Figura 1 – Aglomerado do conjunto de dados mostrado na matriz através do método complete (não está em escala)

(b) Repita o exercício (a) utilizando o método **single** como a distância entre dois aglomerados.

Resolução: O método single considera a menor distância que há entre os dados entre dois aglomerados. De forma análoga ao item (a), será considerado que as observações possuem nomes A, B, C, D e E.

Pode-se notar que a menor distância existente na matriz é CE = 2 = EC (os números 0 são a distância de uma observação a ela mesma (AA, BB, etc.), ou seja, são consideradas inicialmente para formar aglomerados que contém apenas uma única observação). Dessa forma, o primeiro aglomerado a ser formado é o CE, que une as observações C e E.

Em seguida, deve-se analisar as distâncias entre A, B, D e o aglomerado CE. Como o método single seleciona a menor distância entre dois aglomerados, as distâncias a serem escolhidas são: d(A,B) = 9, d(A,D) = 6, d(B,D) = 5, d(A,CE) = 3, d(B,CE) = 7 e d(D,CE) = 8. Dessa forma, a menor distância encontrada é d(A,CE) = 3.

O próximo passo é analisar as distâncias entre B, D e o aglomerado ACE. As distâncias a serem escolhidas são: d(B,D) = 5, d(B,ACE) = 7, d(D,ACE) = 6. Dessa forma, a menor distância encontrada é d(B,D) = 5.

Por último, tem-se a distância entre os aglomerados BD e ACE. A distância entre eles, através do método single, é d(BD,ACE) = 6. O aglomerado hierárquico é mostrado na Figura 2.

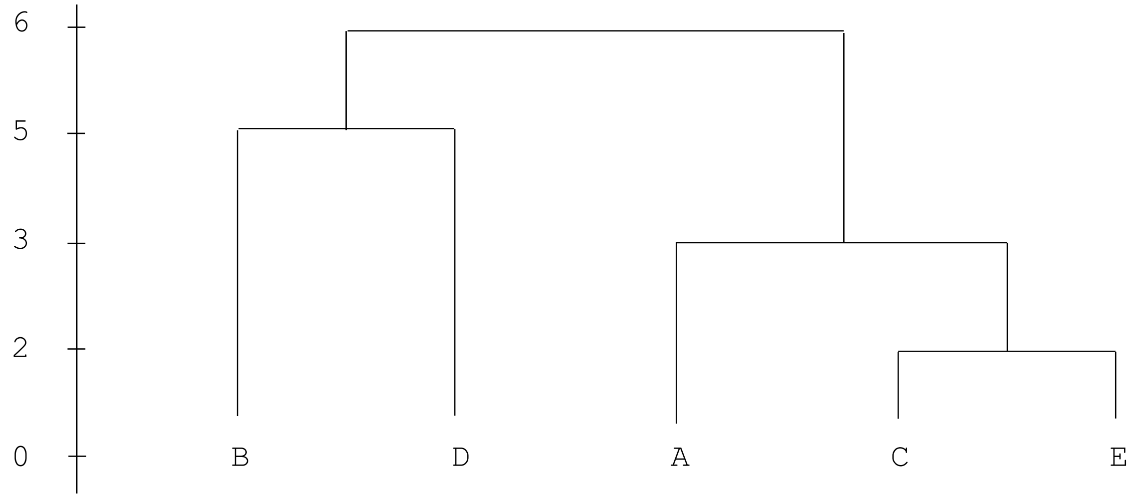

Figura 2 – Aglomerado do conjunto de dados mostrado na matriz através do método single (não está em escala)

(c) Suponha que um corte seja feito no dendograma encontrado em (a) de forma a deixar dois aglomerados. Quais observações estão em cada aglomerado?

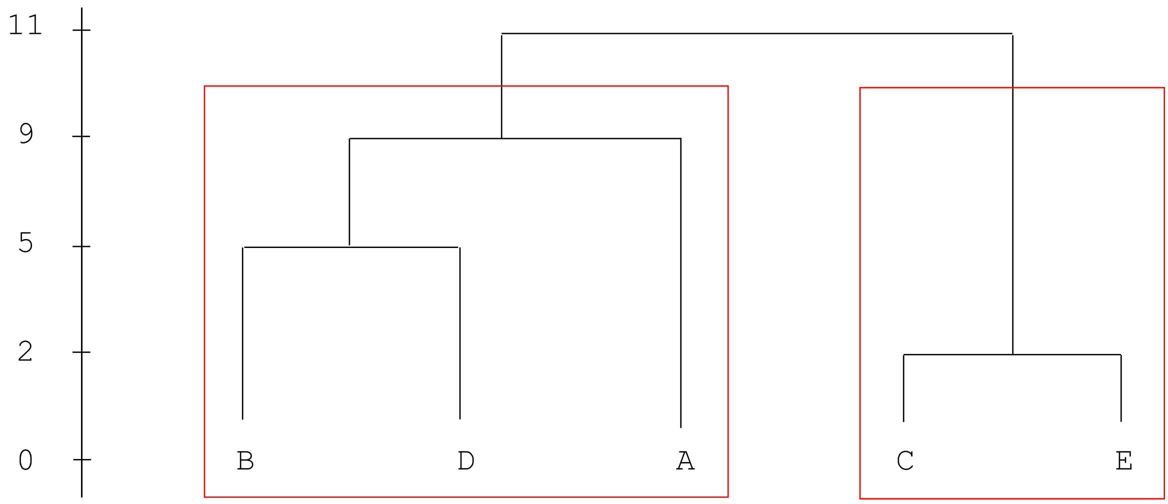

Figura 3 – Separação do dendograma de (a) em 2 aglomerados

Ao realizar o corte do dendograma encontrado em (a) para separá-lo em 2 aglomerados, o resultado obtido é um aglomerado formado por ABD e outro formado por CE na altura entre 9 e 11 unidades.

(d) Suponha que um corte seja feito no dendograma encontrado em (b) de forma a deixar dois aglomerados. Quais observações estão em cada aglomerado?

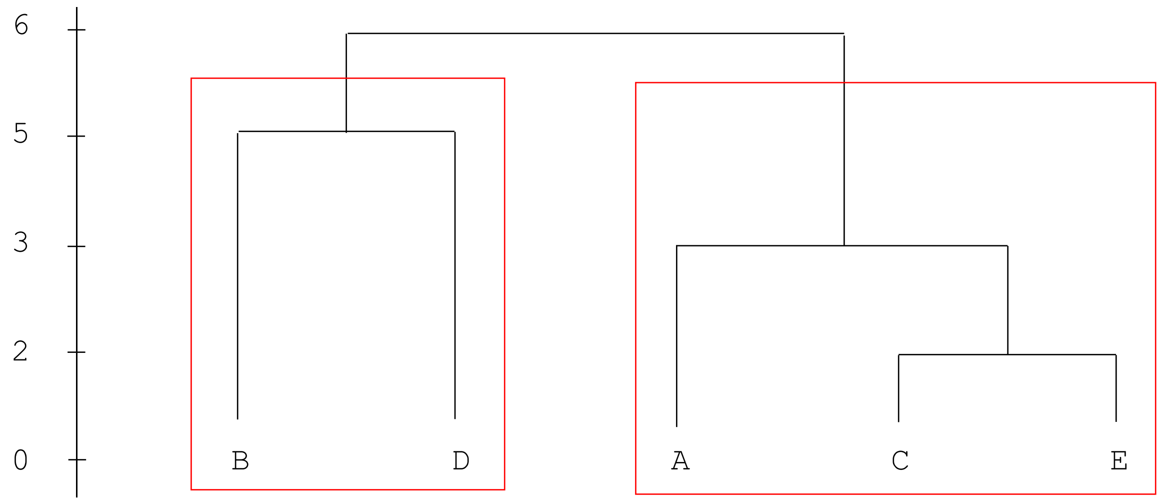

Figura 4 – Separação do dendograma de (b) em 2 aglomerados

Ao realizar o corte do dendograma encontrado em (b) para separá-lo em 2 aglomerados, o resultado obtido é um aglomerado formado por BD e outro formado por ACE na altura entre 5 e 6 unidades. Uma comparação que pode ser feita é que o método single deixou o dendograma mais compacto, ou seja, os aglomerados foram formados com distâncias menores entre si, ao contrário do método complete. Além disso, cada método produziu aglomerados diferentes, no qual a variável A é quem se diferencia no resultado.

Exercício 02: O objetivo desse exercício é construir manualmente uma árvore de decisão para ajudar a prever quando um paciente poderá ter um ataque cardíaco. Os dados de treinamento para esse problema estão no arquivo heart.txt.

(a) Usando a impureza de Gini construa uma árvore de decisão que irá prever quando um paciente terá ou não um ataque cardíaco. Deixe indicado todos os passos da construção.

Resolução: A Tabela 1 ilustra os dados do arquivo heart.txt.

Tabela 1 – Dados do arquivo heart.txt

Patient ID	Chest pain	Sex	Smokes	Exercises	Heart attack
1	yes	yes	no	yes	yes
2	yes	yes	yes	no	yes
3	no	no	yes	no	yes
4	no	yes	no	yes	no
5	yes	no	yes	yes	yes
6	no	yes	yes	yes	no

Em primeiro lugar, deve-se calcular o índice de Gini para todas as variáveis presentes na tabela (chest pain [A], sex [B], smokes [C] e exercises [D]). A variável heart attack será denotada por [E].

Ao analisar a variável [A] (chest pain), tem-se que:

$$P_{[A] = yes} = 3/6 \text{ e } P_{[A] = no} = 3/6$$

Escolhendo o atributo [A] = yes, o conjunto resultante mostra que em 3 elementos, 3 tiveram um ataque cardíaco e 0 não tiveram. Através do índice de Gini:

$$P([A] = yes \text{ e } [E] = yes) = 3/3 = 1$$
$$P([A] = yes \text{ e } [E] = no) = 0/3 = 0$$

$$G_{[A] = yes} = 1 \times (1 - 1) + 0 \times (1 - 0) = 0$$

Escolhendo o atributo [A] = no, o conjunto resultante mostra que em 3 elemento, 1 teve um ataque cardíaco e 2 não tiveram. Através do índice de Gini:

$$P([A] = no \text{ e } [E] = yes) = 1/3$$
$$P([A] = no \text{ e } [E] = no) = 2/3$$

$$G_{[A] = no} = 1/3 \times (1 - 1/3) + 2/3 \times (1 - 2/3) = 4/9$$

A medida final para a variável [A] é dada por:

$$G_{[A]} = 3/6 \times 0 + 3/6 \times 4/9 = 0{,}22$$

Ao analisar a variável [B] (sex), tem-se que:

$$P_{[B] = yes} = 4/6 \text{ e } P_{[B] = no} = 2/6$$

Escolhendo o atributo [B] = yes, o conjunto resultante mostra que em 4 elementos, 2 tiveram um ataque cardíaco e 2 não tiveram. Através do índice de Gini:

$$P([B] = yes \text{ e } [E] = yes) = 2/4 = 1/2$$
$$P([B] = yes \text{ e } [E] = no) = 2/4 = 1/2$$

$$G_{[B] = yes} = 1/2 \times (1 - 1/2) + 1/2 \times (1 - 1/2) = 2/4 = 1/2$$

Escolhendo o atributo [B] = no, o conjunto resultante mostra que em 2 elemento, 2 tiveram um ataque cardíaco e 0 não tiveram. Através do índice de Gini:

$$P([B] = no \text{ e } [E] = yes) = 2/2 = 1$$
$$P([B] = no \text{ e } [E] = no) = 0/2 = 0$$

$$G_{[B] = no} = 1 \times (1 - 1) + 0 \times (1 - 0) = 0$$

A medida final para a variável [B] é dada por:

$$G_{[B]} = 4/6 \times 1/2 + 2/6 \times 0 = 0{,}33$$

Ao analisar a variável [C] (smokes), tem-se que:

$$P_{[C] = yes} = 4/6 \text{ e } P_{[C] = no} = 2/6$$

Escolhendo o atributo [C] = yes, o conjunto resultante mostra que em 4 elementos, 3 tiveram um ataque cardíaco e 1 não teve. Através do índice de Gini:

$$P([C] = yes \text{ e } [E] = yes) = 3/4$$
$$P([C] = yes \text{ e } [E] = no) = 1/4$$

$$G_{[C] = yes} = 3/4 \times (1 - 3/4) + 1/4 \times (1 - 1/4) = 6/16 = 3/8$$

Escolhendo o atributo [C] = no, o conjunto resultante mostra que em 2 elemento, 1 tive um ataque cardíaco e 1 não teve. Através do índice de Gini:

$$P([C] = no \text{ e } [E] = yes) = 1/2$$
$$P([C] = no \text{ e } [E] = no) = 1/2$$

$$G_{[C] = no} = 1/2 \times (1 - 1/2) + 1/2 \times (1 - 1/2) = 2/4 = 1/2$$

A medida final para a variável [C] é dada por:

$$G_{[C]} = 4/6 \times 3/8 + 2/6 \times 1/2 = 0{,}42$$

Ao analisar a variável [D] (exercises), tem-se que:

$$P_{[D] = yes} = 4/6 \text{ e } P_{[D] = no} = 2/6$$

Escolhendo o atributo [D] = yes, o conjunto resultante mostra que em 4 elementos, 2 tiveram um ataque cardíaco e 2 não tiveram. Através do índice de Gini:

$$P([D] = yes \text{ e } [E] = yes) = 2/4 = 1/2$$
$$P([D] = yes \text{ e } [E] = no) = 2/4 = 1/2$$

$$G_{[D] = yes} = 1/2 \times (1 - 1/2) + 1/2 \times (1 - 1/2) = 2/4 = 1/2$$

Escolhendo o atributo [D] = no, o conjunto resultante mostra que em 2 elemento, 2 tiveram um ataque cardíaco e 0 não tiveram. Através do índice de Gini:

$$P([D] = no \text{ e } [E] = yes) = 2/2 = 1$$
$$P([D] = no \text{ e } [E] = no) = 0/2 = 0$$

$$G_{[D] = no} = 1 \times (1 - 1) + 0 \times (1 - 0) = 0$$

A medida final para a variável [D] é dada por:

$$G_{[D]} = 4/6 \times 1/2 + 2/6 \times 0 = 0,33$$

Dentre todas as variáveis, a que apresentou o menor índice de Gini foi a [A] (chest pain) (= 0,22). Portanto, ela será a variável nó inicial da árvore de decisão. Se [A] = yes, considera-se que a pessoa teve um ataque cardíaco. Caso contrário, deve-se escolher outra variável para continuar a análise.

Em segundo lugar, tem-se que as variáveis [B] (sex) e [D] (exercises) são as que possuem os menores valores (= 0,33). Neste caso, pode-se escolher qualquer uma entre as duas. A escolhida será a variável [D]. Se [D] = no, considera-se que a pessoa teve um ataque cardíaco. Caso contrário, considera-se outra variável para continuar a análise.

Em seguida, tem-se que a variável [B] (sex) é a que possui o menor índice de Gini (= 0,33). Se [B] = no, considera-se que a pessoa teve um ataque cardíaco. Caso contrário, considera-se outra variável para continuar a análise.

Por fim, resta apenas a variável [C] (smokes), variável que apresentou o maior índice de Gini (= 0,42). Se [C] = yes, considera-se que a pessoa teve um ataque cardíaco. Caso contrário, considera-se que pessoa não teve um ataque cardíaco. Observe que neste caso já existe um erro associado a escolha (nenhum caso $G_{[C] = yes}$ e $G_{[C] = no}$ apresentou resultado igual a zero).

A Figura 5 ilustra a árvore de decisão para o conjunto de dados heart.txt.

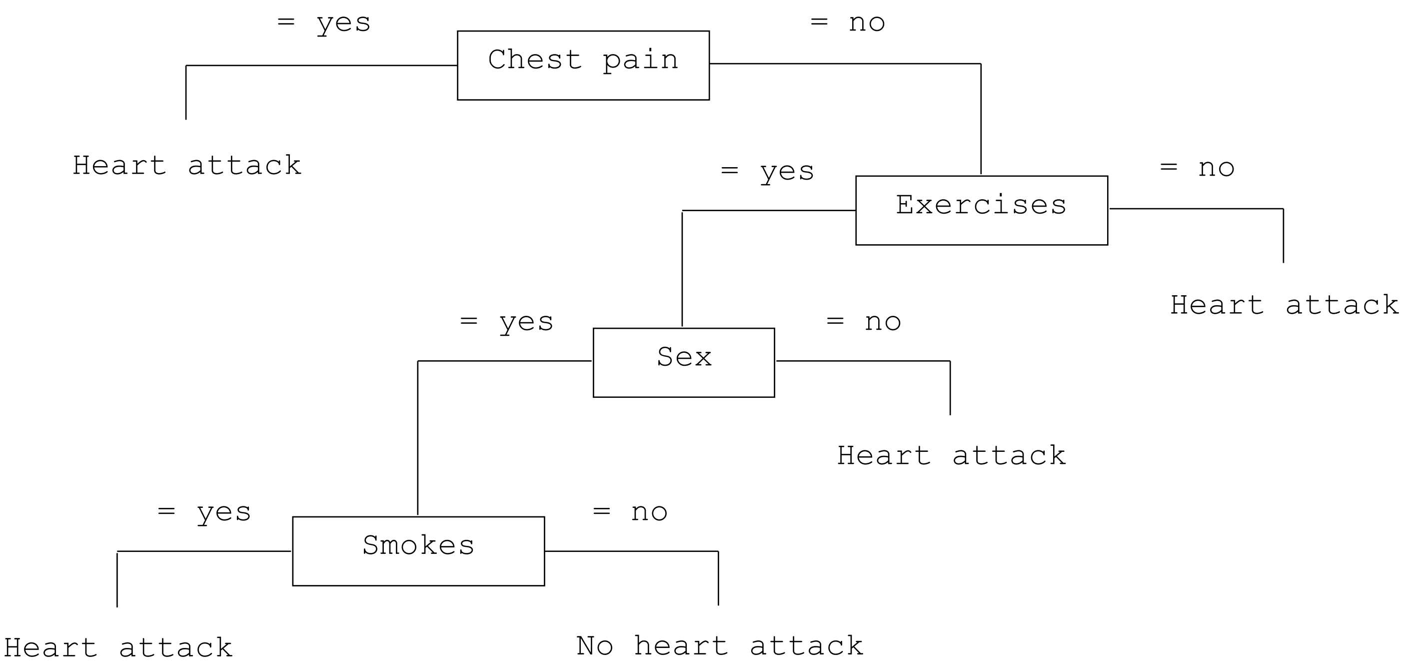

Figura 5 – Árvore de decisão para o arquivo heart.txt

(b) Traduza a árvore construída acima em um código. O código deve ser uma função cuja entrada é um vetor de tamanho 4 referentes às variáveis explanatórias do modelo (cada entrada sendo yes ou no) e cuja saída é yes (é provável que o paciente tenha um ataque cardíaco) ou no (é provável que o paciente não tenha um ataque cardíaco).

Resolução: Utilizando a árvore de decisão construída em (a), o código em R é dado por:

```r
heart_attack = function(dados){
    tamanho = length(dados)
    if(tamanho != 4){
        return(as.factor(“erro”))
    }
    if(dados[1] == “yes”){
        return(as.factor(“yes”))
    }else{
        if(dados[4] == “no”){
            return(as.factor(“yes”))
        }else{
            if(dados[2] == “no”){
                return(as.factor(“yes”))
            }else{
                if(dados[3] == “yes”){
                    return(as.factor(“yes”))
                }else{
                    return(as.factor(“no”))
                }
            }
        }
    }
}
dados <- c(“no”, “yes”, “no”, “yes”)
dados <- as.factor(dados)
decisao <- heart_attack(dados)
decisao
# Neste caso, decisao = “no”
```

Observação: a resolução das questões 3 e 4 foram feitas no R.