Aprendizagem de Dados Simbólicos e/ou Numéricos

Francisco de A.T. de Carvalho

Dados usuais 1/2

- $\Omega = \{1, \ldots, n\}$
  - conjunto de indivíduos (população ou universo, amostra)
- As propriedades de cada indivíduo são descritas por variáveis:
  - $Y = \{Y_1, \ldots, Y_p\}$: conjunto de p variáveis
- O_j : Domínio (conjunto dos possíveis valores) da variável de Y_j

Dados usuais 2/2

- $Y_i : \Omega \rightarrow O_i$
  $k \rightarrow x_{kj} = Y_j(k)$
- $\forall\, k \in \Omega,\ x_{kj} = Y_j(k)$
- Matriz (ou tabela) de dados:

$$X = (x_{kj})$$

Tipos de Variáveis 1/12

- Y_j é quantitativa (métrica, numérica):
  - se O_j (domínio) é idêntico ou está contido em $\Re$: $O_j \subseteq \Re$
  - Exemplos de Domínios
    - $O_j = \Re = (-\infty, \infty)$
    - $O_j = \Re_+ = [0, \infty)$
    - $O_j = [a,b] = \{x \in \Re \mid a \leq x \leq b\}$ onde $-\infty < a < b < \infty$

Tipos de Variáveis 2/12

- Y_j é quantitativa contínua
  - se O_j é um intervalo de $\Re$
    - Exemplos:

    a) Y_j é o peso de um adulto, com $O_j = [30, 250] \subseteq \Re$

    b) Y_j é o lucro de uma empresa em um determinado ano, com $O_j = \Re$

    c) Y_j é a carga de um navio em toneladas, com $O_j = \Re_+$

Tipos de Variáveis 3/12

- Y_j é quantitativa discreta se O_j é um conjunto finito ou infinito contável de valores de $\Re$
  - Exemplos de Domínios
    - $O_j = \{\xi_1,\ldots,\xi_M\} \subseteq \Re$
    - $O_j = \{\xi_1, \xi_2,\ldots\} \subseteq \Re$

a) Ex: Y_j :número de acidentes nas ruas de Recife na primeira semana de maio, com $O_j = \{0, 1, 2, \ldots\}$

b) Y_j :número de rodas de um veículo, com $O_j = \{2, \ldots,10\}$

Tipos de Variáveis 4/12

- Y_j é qualitativa (categórica)
  - se O_j (domínio) é finito e seus elementos são categorias sem significado numérico
- Y_j é qualitativa nominal
  - se O_j não possui estrutura interna
  - Dadas duas categorias x e y de O_j, x = y ou x ≠ y
    - Exemplo: Y_j é marca de um carro, com O_j = {Ford, Pegeout, Volkswagen}

Tipos de Variáveis 5/12

- Y_j é binária
  - $Y_j(k) = 1$ as vezes é interpretado como "o individuo k tem a propriedade j"
  - se O_j tem apenas duas alternativas as vezes codificada como 0 e 1
    a) Exemplo: Y_j é o sexo, com O_j = {masculino (M), feminino (F)}
    b) Y_j é a presença de asas, com $O_j = \{0,1\}$

Tipos de Variáveis 6/12

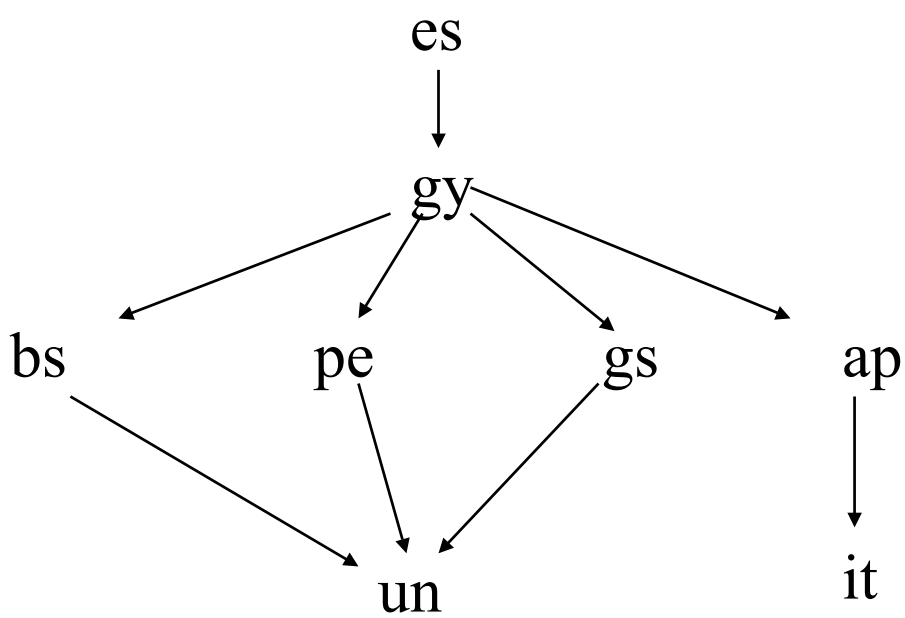
- Y_j é qualitativa ordinal
  - se existe uma ordem linear total entre as categorias de O_j
  - para a,b ∈ O_j ou a <= b ou b <= a
    - Exemplo: Y_j é a qualidade de um produto, com O_j = {insuficiente, pobre, regular,boa, excelente}

Tipos de Variáveis 7/12

- Y_j é qualitativa ordinal generalizada
  - nem todo par de alternativas a, b $\in$ O_j pode ser comparado (ordem parcial)
  - o sistema de pares ordenados a < b pode ser desenhado segundo um diagrama de tipo hierarquia, reticulado, rede
    - nesse diagrama, dois níveis a, b verificam a < b se e somente se existe uma seqüência de ramos conectados que liga a à b

Tipos de Variáveis 8/12

- Exemplo: Y_j = nível educacional
  - O_j = {es = escola elementar; bs = contabilidade; gy = ginásio; ap = técnico; gs = científico; pe = pedagógico; un = universidade; it = instituto tecnológico}



Tipos de Variáveis 9/12

- Nesse exemplo:
  - Uma flecha aponta de um tipo de instituição *a* para um tipo *b* (b < a)
    - se alguém pode ser aceito por uma instituição de tipo *b* após ser graduado por uma instituição de tipo *a*

Tipos de Variáveis 10/12

- Y_j : variável taxonômica
  - com domínio O_j = {a, b,...};
  - as categorias são ordenadas em uma hierarquia
    a) Cada categoria $b \in O_j$ é um nó da hierarquia;
    b) Uma categoria *c* é descendente de *a* (*a* é ancestral de *c*) se c < a;
    c) *b* é sucessor (descendente direto) de *a* se b < a e não existe outro $d \in O_j$ tal que b < d < a

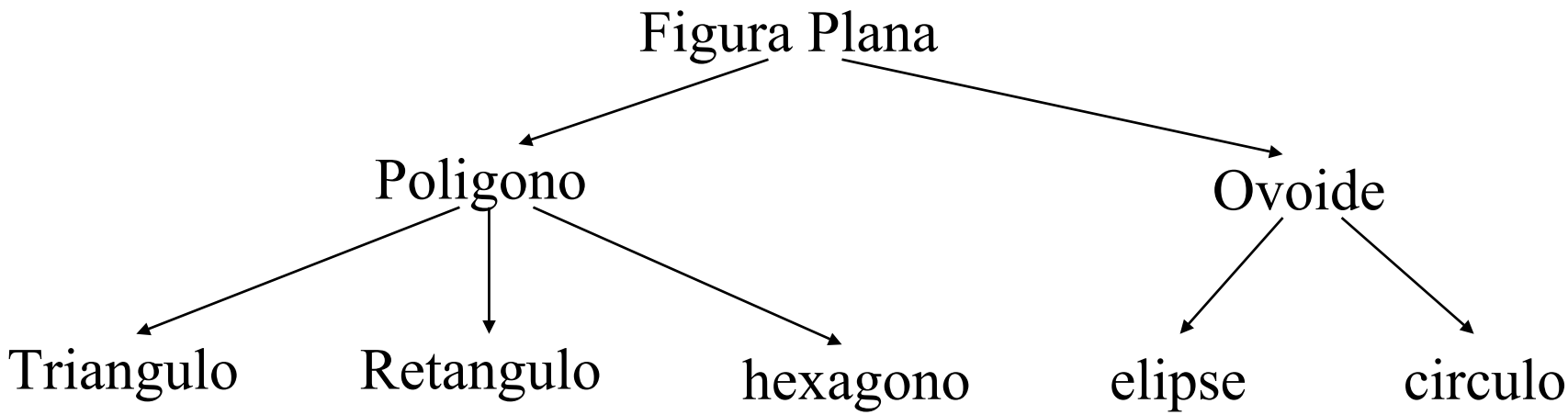
Tipos de Variáveis 11/12

d) *a* é predecessor (ancestral direto) de *b* se b < a e não existe $d \in O_j$ tal que b < d < a

e) A hierarquia contém uma única raiz

f) Uma categoria *f* que não tem sucessor é chamada de folha; os outros são nós internos.

Tipos de Variáveis 12/12



- Y_j: tipo da figura no plano
- O_j = {triangulo, retângulo, hexágono, elipse, circulo, polígono, ovóide, figura plana}

Vetor de Dados

- Vetor de variáveis p-dimensional

$$\mathbf{X} = \begin{pmatrix} Y_1 \\ \vdots \\ Y_p \end{pmatrix} = (Y_1, \ldots, Y_p)' \in O_1 \times \cdots \times O_p$$

- Vetor de de dados usuais

$$\mathbf{x}_k = \mathbf{X}(k) = \begin{pmatrix} x_{k1} \\ \vdots \\ x_{kp} \end{pmatrix} = (x_{k1}, \ldots, x_{kp})' \in O_1 \times \cdots \times O_p$$

Matriz de Dados

- Matriz de dados usuais n x p

$$\tilde{\mathbf{X}} = \left(x_{kj}\right)_{n\times p} = \begin{pmatrix} x_{11} & \cdots & x_{1p} \\ \vdots & x_{kj} & \vdots \\ x_{n1} & \cdots & x_{np} \end{pmatrix} = \begin{pmatrix} \mathbf{x}_1' \\ \vdots \\ \mathbf{x}_p' \end{pmatrix} = (\mathbf{y}_1, \ldots, \mathbf{y}_p)$$

Tabela de dados usuais

Indivíduo	Idade	Peso	Sexo	Altura
ω_1	25	60	F	1.65
ω_2	32	65	M	1.60
ω_3	28	58	F	1.75
:	:	:	:	:

Dependência entre Variáveis 1/6

- Diferentes tipos de dependências:
  - lógica
  - hierárquica
  - estocástica
- Dependência Lógica

Existe dependência lógica entre duas variáveis Y e Z

  - se os valores de Z dependem
    - logicamente ou funcionalmente dos valores de Y

Dependência entre Variáveis 2/6

- Exemplo
  - Y : peso de uma pessoa (Kg)

  Z : altura de uma pessoa (cm)

  r_1: se $Y \leq 55$ então $Z \leq 180$

  r_2: $Y \in [0,55] \Rightarrow Z \in [0,180]$

  r_3: se [peso ≤ 55] então [altura ≤ 180]

Dependência entre Variáveis 3/6

- Dependência Hierárquica
  - Uma variável Z depende hierarquicamente de uma variável Y
    - se o conjunto O_z de valores z para Z é especificado em dependência dos valores y $\in$ Y
- Exemplo
  - Variável mãe

Y : tipo de comercio varejista

O_Y : {loja de carros, loja de computador, ...}

Dependência entre Variáveis 4/6

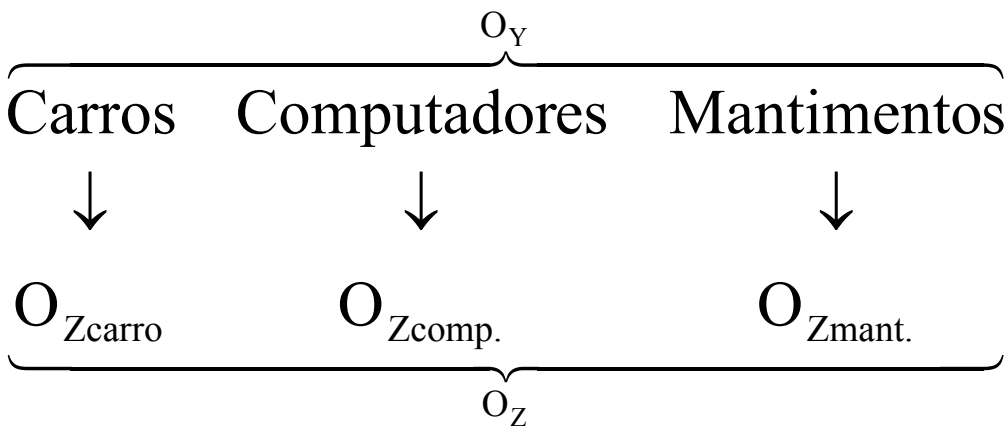
- Variável filha

Z : atacadista, cujo domínio é:

se y = loja de carros então O_{Zcarro} = {FORD, FIAT, ...}

se y = loja de comput. então O_{Zcomp} = {IBM, ...}

se y = mantimentos então O_{zmant} = {todos os agricultores locais}



Dependência entre Variáveis 5/6

- Dependência Hierárquica

Caso especial: Z não faz sentido (não aplicável) para algumas categorias y de Y

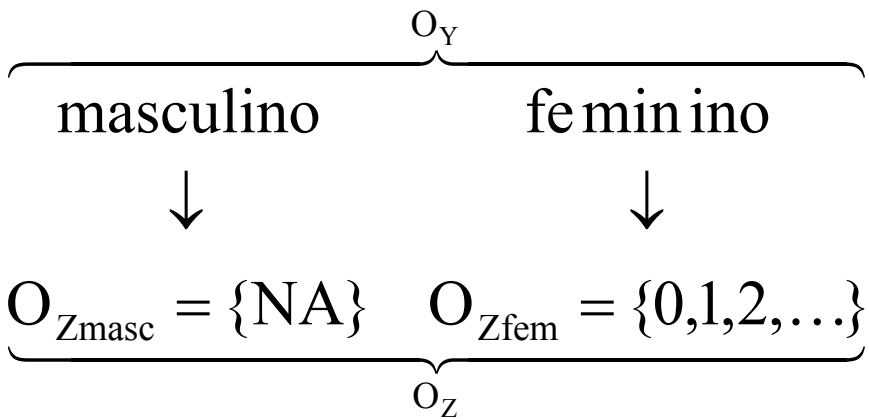
- Y : sexo

O_Y : {masculino, feminino}

- Z : número de crianças

O_{Zfem} = {0, 1, 2, ...}

O_{zmasc} = {não aplicável} = {NA}



Dependência entre Variáveis 6/6

- Dependência Estocástica

  - p variáveis aleatórias $Y_1, \ldots, Y_p$ são chamadas estocásticamente independentes se

$$P(Y_1 \in B_1, \ldots, Y_p \in B_p) = \prod_{j=1}^{p} P(Y_j \in B_j)$$

$\forall\ B_j \subset O_j,\ j = 1,$

Dados Simbólicos 1/2

- Dados simbólicos
  - informações complexas, expressas por
    - intervalos, conjuntos, histogramas, distribuições de probabilidade.
- Situações onde aparece esse tipo de dados
- Dados simbólicos para indivíduos (objetos de primeira ordem)
  - Y_j : Tempo de estudo diário
  - $Y_j(k)$ = [0,6] (em horas)
  - $Y_j(k)$ = (nada(0.5), uma(0.4),>uma(0.10))

Dados Simbólicos 2/2

- Dados simbólicos para classes de indivíduos (objetos de segunda ordem, objetos agregados)
  - Y_j : Instituições bancárias de uma cidade
    $Y_j(k)$ = {Banco do Brasil, Caixa, Itaú, Bradesco}
  - Y_j : Fração de votos por partido político e por estado
    $Y_j(k)$ = {(A, 0.5), (B, 0.2), (C, 0.3)}
  - Y_j : Níveis de cinza em uma região de uma imagem
    $Y_j(k)$ = $\Gamma(20,30)$

Tipos de Variáveis Simbólicas

- Notação.
  - Y_j : variável simbólica
  - E : conjunto de objetos
  - O_j : Domínio de Y
  - $k \in E$: objeto

Variáveis Multivaloradas

- Uma variável Y Multivalorada é uma função
  - $Y_j : E \rightarrow B = P(O_j)$
    $Y_j(k) \rightarrow U \subseteq P(O_j)$

- $P(O_i)$: conjunto de todos os subconjuntos de O_j
- Em muitas situações $U = \emptyset$ deve ser excluido

- No caso usual, $|y(k)|=1$

Variáveis Multivaloradas Categóricas

- Variáveis Multivaloradas nominais:
  - U subconjunto de valores não ordenados
    - Sexo(k) = {masculino, feminino}
- Variáveis Multivaloradas ordinais:
  - U subconjunto de valores ordenados
    - Grau de instrução(k) = {primário, secundário, superior}

Variáveis Multivaloradas Quantitativas

- Variáveis Multivaloradas quantitativas
  - Y: Numero de Acidentes Semanais nos 3 principais bairros de uma cidade
    - Y(k) = {20, 10, 15}

Variáveis Multivaloradas Quantitativas

- Variáveis Multivaloradas de Tipo Intervalo:
  - U = Y(k) = [α, β] é um intervalo de $\Re$
  - ou é um intervalo com respeito a uma determinada ordem < em O
  - Salários(k) = [200, 7000], k é uma empresa

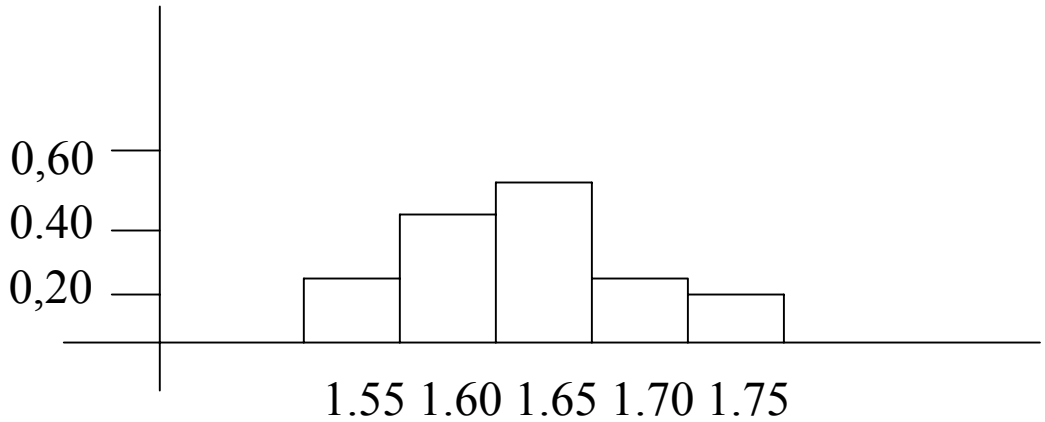
Variáveis Modais

- Uma variável Modal Y_j é uma função
  - $Y_j : E \rightarrow B = M(O_j)$
    $k \rightarrow (U(k), \pi(k))$

  onde
  - $\pi(k)$ é uma medida ou uma distribuição (de freqüências, de probabilidade, de pesos) definida no domínio O_j de Y_j
  - $U(k) \subseteq O_j$ é o suporte de π no domínio O_j
  - $M(O_j)$ é uma família de medidas não negativas definidas em O_j

Exemplo

- C = {C_1, C_2, ..., C_{10}} os 10 centros da UFPE
  - Y : altura dos funcionários no centro C_i
  - Histograma das alturas



  - Uma distribuição normal $N(168,48.4)$ com média 168 e variância 48.4

Vetor de Dados Simbólicos

- Vetor de de dados simbólicos

$$\mathbf{x}_u = \mathbf{X}(u) = \begin{pmatrix} \xi_{u1} \\ \vdots \\ \xi_{up} \end{pmatrix} = (\xi_{k1}, \ldots, \xi_{kp})' \in B_1 \times \cdots \times B_p$$

Matriz de Dados Simbólicos

- Matriz de dados simbólicos n x p

$$\underline{\mathbf{X}} = \left(\xi_{kj}\right)_{n\times p} = \begin{pmatrix} \xi_{11} & \cdots & \xi_{1p} \\ \vdots & \xi_{kj} & \vdots \\ \xi_{n1} & \cdots & \xi_{np} \end{pmatrix} = \begin{pmatrix} \xi'_1 \\ \vdots \\ \xi'_p \end{pmatrix} = (\mathbf{y}_1, \ldots, \mathbf{y}_p)$$

Tabela de Dados Simbólicos 1/2

- Conjunto de objetos
  - E = $\{a_1, a_2, a_3, a_4\}$:
    - 4 cidades da região metropolitana
- Conjunto de variáveis simbólicas
  - Y_1 : população (mínimo e Maximo nos anos 90-95)

    B_1 : intervalos de $\Re^+ = O_1$
  - Y_2 : espectro dos partidos políticos em uma cidade

    B_2 : distribuições de freqüências de
    - O_2 = {Democratas(D), Conservadores(C), Socialistas(S)}

Tabela de Dados Simbólicos 2/2

- Y_3 : grandes instituições bancárias em uma cidade

B_3 : subconjuntos de

- O_3 = {BB, Caixa, Itaú, Bradesco}

classe	População	Espectro partidário	Bancos
a_1	[80, 100]	(D 0.4 C 0.3 S0.3)	{BB, Caixa}
a_2	[100, 130]	(D 0.1 C 0.3 S0.6)	{Caixa, Itaú}
a_3	[8, 10]	(D 0.3 C 0.5 S 0.2)	{Bradesco}
a_4	[10,13]	(D 0.3 C 0.1 S 0.6)	{BB, Bradesco}