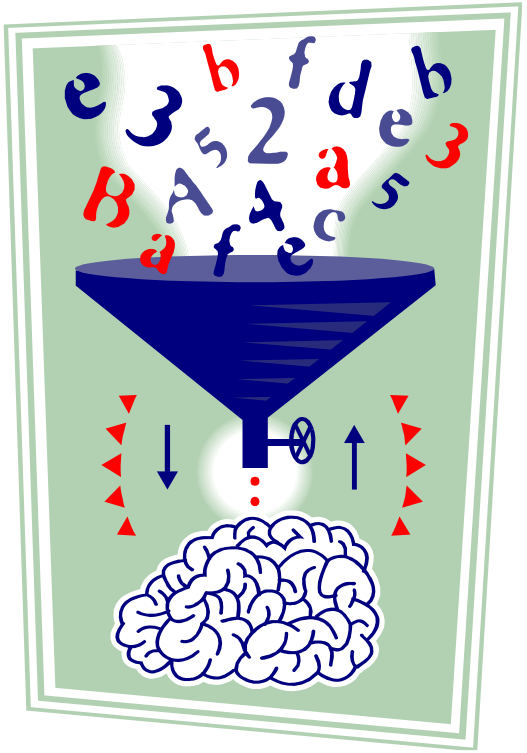
Preparação de Dados: Valores Ausentes



Eduardo R. Hruschka

Valores ausentes:

- Ocorrência comum:
  - Mau funcionamento de dispositivos de coleta de dados;
  - Dado omitido pela fonte de informação numa pesquisa;
  - Falha na digitação ou na composição da base;
- Formas de eliminação de valores ausentes:
  - Eliminar registros/atributos com valores ausentes;
    - Perda de dados pode ser considerável.
  - Preenchimento de valores (imputação)
    - Por uma constante. Ex.: Média/Moda do atributo
      - Desconsidera a relação entre atributos da base de dados
    - Por valores que tentem preservar as relações entre atributos da base de dados
      - Uso de um algoritmo de aprendizado.

Noção Intuitiva:

- Padrões de ausência (Rubin, 1977):
  - Completamente aleatória (*Missing Completely at Random – MCAR*)
  - Aleatória (*Missing at Random – MAR*)
    - Ausência de valor num atributo depende de valores de outro(s) atributo(s)
  - Não aleatória (*Missing not at Random – MNAR*)
    - Ausência de um valor num atributo relacionada a uma condição envolvendo o próprio valor do atributo

Exemplo: Pressão arterial de pacientes:
X: Medidas em Janeiro
Y: Medidas em Fevereiro:
Completo: de todos pacientes
MCAR: de pacientes escolhidos ao acaso
MAR: de pacientes com pressão < 140 em Janeiro
MNAR: registro de medidas maiores que 140

X	Y			
	Completo	MCAR	MAR	MNAR
169	148	148	148	148
126	123	-	-	-
132	149	-	-	149
160	169	-	169	169
105	138	-	-	-
116	102	-	-	-
133	150	-	-	150
109	96	-	-	-
106	148	-	-	148
176	137	-	137	-
128	155	-	-	155
131	131	-	-	-
130	101	101	-	-
145	155	-	155	155
136	140	-	-	-
146	134	-	134	-
111	129	-	-	-
97	85	85	-	-
134	124	124	-	-
153	112	-	112	-
137	122	122	-	-

Avaliando métodos de imputação – algumas considerações:

Exemplo/Atributo	a_1	a_2	a_3	a_4	Classe
e_1	k_{11}	k_{12}	k_{13}	k_{14}	A
e_2	k_{21}	?	k_{23}	?	B
e_3	k_{31}	k_{32}	k_{33}	k_{34}	A
e_4	k_{41}	k_{42}	?	k_{44}	B
e_5	k_{51}	k_{52}	k_{53}	k_{54}	A
e_6	k_{61}	k_{62}	k_{63}	k_{64}	B
e_7	?	k_{72}	k_{73}	k_{74}	A
e_8	k_{81}	k_{82}	k_{83}	k_{84}	A
e_9	k_{91}	k_{92}	k_{93}	k_{94}	B

Decompondo a base de dados:

	a_1	a_2	a_3	a_4	Classe
e_1	k_{11}	k_{12}	k_{13}	k_{14}	A
e_3	k_{31}	k_{32}	k_{33}	k_{34}	A
e_5	k_{51}	k_{52}	k_{53}	k_{54}	A
e_6	k_{61}	k_{62}	k_{63}	k_{64}	B
e_8	k_{81}	k_{82}	k_{83}	k_{84}	A
e_9	k_{91}	k_{92}	k_{93}	k_{94}	B

Registros completos (C)

	a_1	a_2	a_3	a_4	Classe
e_2	k_{21}	?	k_{23}	?	B
e_4	k_{41}	k_{42}	?	k_{44}	B
e_7	?	k_{72}	k_{73}	k_{74}	A

Registros com ausentes (?)

	a_1	a_2	a_3	a_4	Classe
e_2	k_{21}	f_{22}	k_{23}	f_{24}	B
e_4	k_{41}	k_{42}	f_{43}	k_{44}	B
e_7	f_{71}	k_{72}	k_{73}	k_{74}	A

Registros preenchidos (f)

Objetivo é inserir o menor ruído possível na base de dados. Mas como definir o que é ruído? Como medi-lo?

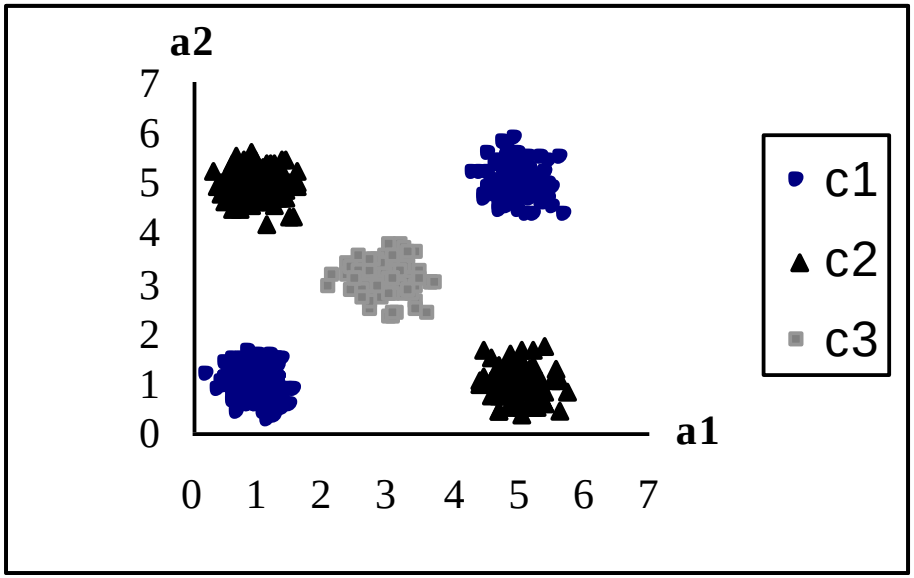
Duas abordagens:

- Predição:
  - Comparar valor imputado com valor conhecido;
  - Qual métrica poderia ser usada?
  - Viável em aplicações práticas?
  - Avaliação em dados completos diminui a informação disponível pra avaliação da ferramenta de imputação.
- Modelagem:
  - Influência na classificação, nas partições, etc.

Exemplo da influência da imputação:

- Base de dados:
  - Composta de 500 instâncias, dois atributos (a1 e a2) e atributo classe
- Instâncias de cada classe foram obtidas de distribuições Normais centradas em (1,1); (1,5); (5,1); (5,5); (3,3), e com desvio padrão igual a 0.3
- Composta de três classes:
  - c1 – 200 instâncias
  - c2 – 200 instâncias
  - c3 – 100 instâncias

Como a imputação pela média influenciará na modelagem?



Leitura sugerida para o projeto:

- Silva, J. A., Hruschka, E. R., An experimental study on the use of nearest neighbor-based imputation algorithms for classification tasks. Data & Knowledge Engineering, v. 84, p. 47-58, Elsevier, 2013.