ETAPAS DE PRÉ-PROCESSAMENTO

TRATAMENTO DE DADOS AUSENTES

Cristiane Neri Nobre

Tratamento de dados ausentes

- Base de dados podem contem dados ausentes e isto apresenta dificuldades relacionadas à **qualidade dos dados**.

- Dados ausentes podem ser causadas por:
  - problemas nos equipamentos que realizam a coleta
  - a transmissão e o armazenamento dos dados
  - ou problemas no preenchimento ou na entrados dos dados por seres humanos

- Algumas técnicas de AM conseguem lidar bem com dados ausentes, mas isso não acontece com todos os métodos de AM

Tratamento de dados ausentes

Idade	Sexo	Peso	Manchas	Temp.	# Int.	Diagnóstico
—	M	79	—	38,0	—	Doente
18	F	67	Inexistentes	39,5	4	Doente
49	M	92	Espalhadas	38,0	2	Saudável
18	—	43	Inexistentes	38,5	8	Doente
21	F	52	Uniformes	37,6	1	Saudável
22	F	72	Inexistentes	38,0	3	Doente
—	F	87	Espalhadas	39,0	6	Doente
34	M	67	Uniformes	38,4	2	Saudável

Int = Número de internações

Quais as razões para dados ausentes?

- O atributo não foi considerado importante na época da coleta (campo e-mail em 1990)
- Falta de conhecimento do campo (ex: tipo sanguíneo)
- Distração
- Falta de necessidade ou obrigação de apresentar um valor para o atributo (Ex. renda)
- Inexistência de um valor para o atributo em algumas instâncias (ex: número de partos de for homem)
- Problema com equipamento ou processo utilizado para coleta, transmissão e armazenamento de dados

Tratamento de dados ausentes

Quais as alternativas para se resolver este problema de dados ausentes?

1. Eliminar as instâncias com dados ausentes
2. Definir e preencher manualmente valores para os atributos com valores ausentes
3. Utilizar algum método ou heurística para automaticamente definir valores para atributos com valores ausentes
4. Empregar algoritmos de AM que lidam internamente com valores ausentes

Tratamento de dados ausentes

1. Eliminar as instâncias com dados ausentes

Esta alternativa é geralmente empregada quando um dos atributos com valores ausentes de uma instância é o que indica a sua classe.

Esta estratégia não é recomendada quando:

- Quando poucos atributos da instância possuem valores ausentes
- Quando o número de atributos com valores ausentes varia muito entre as instâncias com esse problema
- Quando o número de instâncias que restarem for pequeno

Tratamento de dados ausentes

2. Definir e preencher manualmente valores para os atributos com valores ausentes

Esta estratégia não é recomendada quando:

- O número de instâncias ou atributos com valores ausentes for muito grande

Tratamento de dados ausentes

3. Utilizar algum método ou heurística para automaticamente definir valores para atributos com valores ausentes

Esta estratégia é uma das mais recomendadas

Como imputar estes dados ausentes?

- Moda, média dos valores dos atributos
- Empregar um indutor para estimar o valor do atributo

Tratamento de dados ausentes

Exemplo:

Idade	Sexo	Peso	Manchas	Temp.	# Int.	Diagnóstico
—	M	79	—	38,0	—	Doente
18	F	67	Inexistentes	39,5	4	Doente
49	M	92	Espalhadas	38,0	2	Saudável
18	—	43	Inexistentes	38,5	8	Doente
21	F	52	Uniformes	37,6	1	Saudável
22	F	72	Inexistentes	38,0	3	Doente
—	F	87	Espalhadas	39,0	6	Doente
34	M	67	Uniformes	38,4	2	Saudável

Após a imputação pela média/moda

Idade	Sexo	Peso	Manchas	Temp.	# Int.	Diagnóstico
27	M	79	**Inexistentes**	38,0	**4**	Doente
18	F	67	Inexistentes	39,5	4	Doente
49	M	92	Espalhadas	38,0	2	Saudável
18	**F**	43	Inexistentes	38,5	8	Doente
21	F	52	Uniformes	37,6	1	Saudável
22	F	72	Inexistentes	38,0	3	Doente
27	F	87	Espalhadas	39,0	6	Doente
34	M	67	Uniformes	38,4	2	Saudável

Tratamento de dados ausentes

4. Empregar algoritmos de AM que lidam internamente com valores ausentes

Este é o caso, por exemplo, de alguns algoritmos indutores de árvore de decisão.

Em **Python**, veja os métodos:

Missforest - pip install missingpy

```
from missingpy import MissForest
```

KNNimputer -

```
from sklearn.impute import KNNImputer
```

Resumindo:

Se você deseja eliminar dados ausentes, o que pode ser feito?

- Remoção de objetos em valores ausentes em qualquer atributo preditivo
- Remoção de objetos em valores ausentes em todos os atributo preditivos
- Remoção de objetos com valor ausente em qualquer/todos os atributos preditivos selecionados
- Remoção de atributo preditivo com valor ausente em qualquer objeto
- Remoção de atributo preditivo com valor ausente em todos os objetos
- Remoção de atributo preditivo com valor ausente em um número determinado de objetos

Mas como fazer isso em Python?

- Veja o arquivo que está no CANVAS:

- **Codifica imputa e balanceia.ipynb (com a base de dados câncer. Csv e breast-cancer.csv)**

Referências: