WORKSHOP

Digital Data Analytics

Pré-processamento de dados

Jorge da Costa Ferreira

WORKSHOP

Digital Data Analytics

TRABALHAR OS DADOS ANTES DE LHES APLICAR MODELOS INTELIGENTES CONSOME 90% DO TEMPO DE UM ANALISTA. QUE TAREFAS OCUPAM ESTA ETÁPA?

Jorge da Costa Ferreira

PRE-PROCESSAMENTO DE DADOS

FASES DE PRE-PROCESSAMENTO

Integração de dados: Integração de múltiplas bases de dados para consolidação de toda a informação relevante

Amostragem de dados: Garantir equilíbrio entre eficiência computacional e a taxa de acerto através da seleção de um número limitado de observações (bastante necessária para determinados modelos, como o k-NN).

Balanceamento de dados: Problema sentido em tarefas de classificação onde os dados a usar possuem observações consideravelmente predominantes em apenas algumas classes.

PRE-PROCESSAMENTO DE DADOS

FASES DE PRE-PROCESSAMENTO

Limpeza de dados: Tratamento de dados inconsistentes, redundantes, incompletos ruidosos ou com a presença de outliers.

Redução de dimensionalidade: Redução do número de variáveis através da obtenção de um conjunto de variáveis principais. Pode ser dividido em feature selection ou feature projection.

Transformação de dados: Alguns modelos requerem um pre-processamento com transformação de dados. Por exemplo, alguns não lidam com variáveis qualitativas e requerem a conversão para numérica, ou então necessitam de processos de normalização.

LIMPEZA DE DADOS

GERIR DADOS INCOMPLETOS OU MISSING VALUES

IDADE	GÉNERO	MANCHAS	TEMP	# INT.	DIAGNÓSTICO
-	M	-	39,5	-	Doente
18	F	Inexistente	37,0	4	Saudável
49	F	Espalhadas	38,5	8	Doente
18	-	Uniformes	38,0	1	Saudável

LIMPEZA DE DADOS

GERIR DADOS INCOMPLETOS OU MISSING VALUES

Existem variadas formas de lidar com missing values, aqui seguem alguns exemplos:

- Eliminar os objetos com dados incompletos;
- Estimar o valor com base numa medida central (média, moda ou mediana)
- Estimar o valor com base num modelo de machine learning olhando às restantes variáveis

Adicionalmente, pode-se criar um atributo que indique a ausência daquele atributo, uma vez que isso pode ser um conceito importante.

LIMPEZA DE DADOS

TRANSFORMAÇÃO DE DADOS - CRIAÇÃO DE NOVAS VARIÁVEIS

A análise de frequências pode sugerir a criação de novas variáveis.

- As categorias com baixa frequência podem ser agrupadas entre elas (e.g. outros) ou então com categorias similares

- Possibilidade de criação de novas variáveis através da junção de duas
    - E.g. profissão x gênero

LIMPEZA DE DADOS

TRANSFORMAÇÃO DE DADOS - CONVERSÃO DE VARIÁVEIS CATEGÓRICAS PARA NUMÉRICAS

Criação de variáveis dummy binárias que transformam os dados categóricos em numéricos

IDADE	GÉNERO M	GÉNERO F	MANCHAS	TEMP	# INT.	DIAGNÓSTICO
-	1	0	-	39,5	-	Doente
18	0	1	Inexistente	37,0	4	Saudável
49	0	1	Espalhadas	38,5	8	Doente
18	-	-	Uniformes	38,0	1	Saudável

LIMPEZA DE DADOS

TRANSFORMAÇÃO DE DADOS - CONVERSÃO DE VARIÁVEIS NUMÉRICAS PARA CATEG

As variáveis numéricas podem ser divididas em classes. Existem duas estratégias para a formação de classes:

- Igual frequência
- Igual amplitude

Geralmente é preferível obter classes com igual frequência do que com igual amplitude.

LIMPEZA DE DADOS

TRANSFORMAÇÃO DE DADOS - CONVERSÃO DE VARIÁVEIS NUMÉRICAS PARA CATEG

As variáveis numéricas podem ser divididas em classes.

Faixa etária	GÉNERO	MANCHAS	TEMP	# INT.	DIAGNÓSTICO
-	M	-	39,5	-	Doente
Adolescente	F	Inexistente	37,0	4	Saudável
Adulto	F	Espalhadas	38,5	8	Doente
Adolescente	-	Uniformes	38,0	1	Saudável

LIMPEZA DE DADOS

TRANSFORMAÇÃO DE DADOS - CONVERSÃO DE VARIÁVEIS NUMÉRICAS PARA CATEGÓRICAS

As variáveis numéricas podem ser divididas em classes.

Faixa etária	GÉNERO	MANCHAS	TEMP	# INT.	DIAGNÓSTICO
-	M	-	39,5	-	Doente
Adolescente	F	Inexistente	37,0	4	Saudável
Adulto	F	Espalhadas	38,5	8	Doente
Adolescente	-	Uniformes	38,0	1	Saudável

LIMPEZA DE DADOS

TRANSFORMAÇÃO DE DADOS - STANDARDIZAÇÃO

Quando as variáveis apresentam unidades e ordens de grandeza ou dispersões diferentes, alguns modelos e análises podem sair prejudicados. O processo de estandardização vem auxiliar isso mesmo.

$$x_{ij} \rightarrow x'_{ij} = \frac{x_{ij} - \bar{x}_j}{s_j}$$

Variáveis standardizadas:

- São independentes da medida utilizada
- Tem valor médio = 0 e desvio padrão = 1

LIMPEZA DE DADOS

TRANSFORMAÇÃO DE DADOS - NORMALIZAÇÃO MIN MAX

Quando as variáveis apresentam unidades e ordens de grandeza ou dispersões diferentes, alguns modelos e análises podem sair prejudicados. O processo de estandardização vem auxiliar isso mesmo.

$$x_j^{norm} = \frac{x_j - x_j^{\min}}{x_j^{\max} - x_i^{\min}}$$

Variáveis normalizadas:

- Os valores obtidos vão sempre variar entre 0 e 1

LIMPEZA DE DADOS

DETEÇÃO DE OUTLIERS

Uma definição de outliers globalmente aceite (Hawkins, 1980), classifica um outlier como um objeto que se desvia significativamente dos outros objetos.
Como tal, outliers são pontos que devem ser identificados e potencialmente removidos nas análise a fazer.

WORKSHOP

Digital Data Analytics

Pré-processamento de dados

Jorge da Costa Ferreira