INTRODUÇÃO À CIÊNCIA DE DADOS

Preparação e Pré-Processamento de Dados - Parte 2

DADOS COM RUÍDOS

- Dados com ruídos são dados que contêm objetos que, aparentemente, não pertencem à distribuição que gerou os dados analisados.
- Ruído pode ser uma variância ou erro aleatório no valor gerado de um atributo.
- Um indicador de presença de ruído é a existência de outliers, que são valores que estão além dos limites aceitáveis ou são muito diferentes dos demais valores observados para o mesmo atributo, representando, por exemplo, exceções raramente vistas.

DADOS COM RUÍDOS

Tabela 3.7 Conjunto de dados com *outlier*

Idade	Sexo	Peso	Manchas	Temp.	# Int.	Diagnóstico
28	M	79	Concentradas	38,0	2	Doente
18	F	**300**	Inexistentes	39,5	4	Doente
49	M	92	Espalhadas	38,0	2	Saudável
18	M	43	Inexistentes	38,5	8	Doente
21	F	52	Uniformes	37,6	1	Saudável
22	F	72	Inexistentes	38,0	3	Doente
19	F	87	Espalhadas	39,0	6	Doente

DADOS COM RUÍDOS

- Existem diversas técnicas de pré-processamento que podem ser aplicadas para detecção e remoção de ruídos.
  - Técnicas de encestamento, que suavizam o valor de um atributo. Primeiro os valores são ordenados, depois são divididos em faixas, e esses valores são substituídos por uma média ou mediana.
  - Técnicas de agrupamento, em que valores que não formarem grupos são considerados ruidosos ou outliers.
  - Técnicas de regressão ou classificação, que procuram determinar um valor verdadeiro para um outlier.

TRANSFORMAÇÃO DE DADOS

- Várias técnicas de machine learning estão limitadas à manipulação de valores em determinados tipos, alguns algoritmos estão restritos a valores numéricos, outros a valores qualitativos.
- Em várias situações, dependendo do modelo de machine learning a ser utilizado, será necessário converter dados qualitativos a numéricos ou vice-versa. Também há de se pensar que valores qualitativos nominais ou ordinais podem ser tratados de forma diferente.
- Exemplo: Redes Neurais Artificiais e Support Vector Machines lidam apenas com valores numéricos, portanto, quando um conjunto de dados a ser utilizado por essas técnicas apresenta atributos qualitativos, os valores precisam ser convertidos para numéricos.

TRANSFORMAÇÃO DE DADOS

- Há situações em que é necessária a transformação de valor numérico em outro valor numérico.
- Isso acontece quando os limites inferior e superior de valores dos atributos são muito diferentes, o que leva a uma grande variação de valores, ou ainda quando vários atributos estão em escalas diferentes.

TRANSFORMAÇÃO DE DADOS

- Essa transformação é realizada para evitar que um atributo predomine sobre outro.
- Normalização por amplitude (redefinindo uma nova escala de valores com limites máximo e mínimo) e por padronização (com definição de valor central e um valor de espelhamento para todos os atributos) também são utilizadas para normalizar dados numéricos.

REDUÇÃO DA DIMENSIONALIDADE

- Dimensionalidade é o tamanho horizontal (dimensão) do seu objeto, ou seja, a quantidade de atributos que um determinado objeto tem.

- Em análise de imagens, por exemplo, cada pixel representa um atributo (imagine uma imagem 1024x1024), nos estudos de genética os dados dos genes apresentam milhares de atributos também.

REDUÇÃO DA DIMENSIONALIDADE

- Em muitos algoritmos, grandes quantidades de atributos inviabilizam o processo. A redução de atributos melhora o desempenho, reduz seu custo operacional e torna os resultados mais compreensíveis.

- Duas técnicas bastante utilizadas para redução de atributos são: agregação e seleção de atributos.

REDUÇÃO DA DIMENSIONALIDADE

- As técnicas de agregação substituem atributos originais por novos atributos formados pela combinação de grupos.
- As técnicas de seleção mantêm uma parte dos atributos originais e descartam os demais atributos.
- Destaca-se novamente que mesmo com a evolução computacional e toda tecnologia que amplia o processamento, usando distribuição de processos e aumentando a escala horizontal de forma a ampliar o desempenho de algoritmos, manter atributos desnecessários em um conjunto de dados pode levar seu modelo a um custo de desempenho que seja impossível de prosseguir com a análise.

REDUÇÃO DA DIMENSIONALIDADE

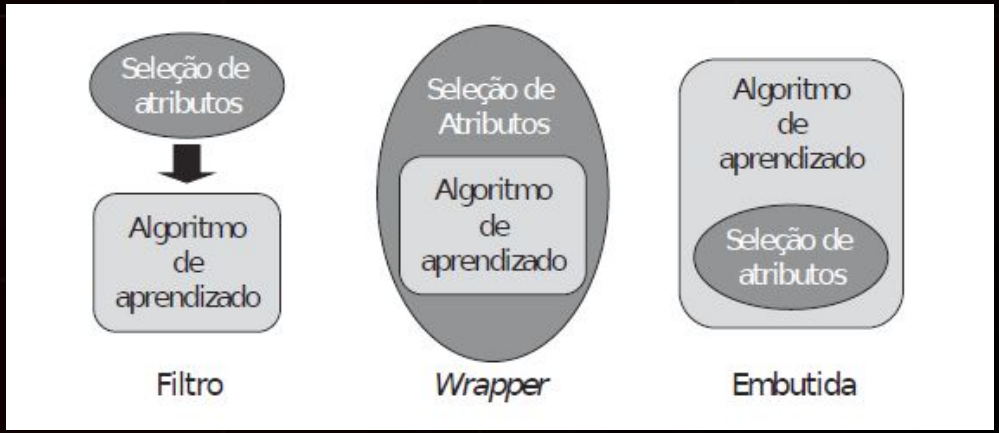
- Agregação: reduz as dimensões por combinação dos atributos.
- Uma das técnicas mais conhecidas é a de Análise de Componentes Principais (PCA). Há algoritmos de ML que reproduzem o PCA, entretanto, os grandes mestres do ML sempre dizem que essa não deveria ser uma técnica para reduzir a dimensionalidade.

REDUÇÃO DA DIMENSIONALIDADE

- O PCA descorrelaciona estatisticamente os exemplos, reduzindo a dimensionalidade do conjunto de dados original pela eliminação de redundâncias.
- Algumas áreas (biologia, finanças, medicina, entre outros) evitam agregar atributos, pois consideram os dados originais importantes para o processo de interpretação dos resultados.

REDUÇÃO DA DIMENSIONALIDADE

- Seleção de atributos: reduz a dimensão eliminando atributos.
- Não é simples identificar atributos que podem ser eliminados, principalmente quando há uma grande quantidade de valores. Relações complexas entre atributos torna tudo mais difícil.
- Algumas técnicas automáticas têm sido estudadas para avaliar a qualidade ou desempenho de um subconjunto de atributos, entre elas: a abordagem embutida, a abordagem baseada em filtro e a wrapper.



DEMONSTRAÇÃO

- Por meio de demonstração, vamos apresentar algumas soluções para o pré-processamento de dados.
- Importante ressaltar que há ainda um conjunto de possibilidades e necessidades que os dados podem ter para estarem aptos a serem enviados para a aplicação dos modelos de machine learning.
- Mesmo do ponto de vista teórico, apresentamos as principais e mais conhecidas necessidades de pré-processamento, mas há muitas outras.
- Vamos usar o Pandas para as demonstrações, mas há muitas outras possibilidades e ferramentas (procure por ETL).
- DEMONSTRAÇÃO

FINALIZANDO

- Entenda os dados, acostume-se com as técnicas, a prática e a repetição levam a melhoria do conhecimento do processo!
- Algo muito importante no processo de pré-processamento é a criatividade, em cada novo livro, em cada novo artigo, em cada nova observação é possível descobrir novas técnicas e aprender com elas.
- Alguns cientistas de dados consideram o pré-processamento uma parte muito "chata". Não tenha dúvida, dominar as técnicas é algo muito relevante para torná-lo um bom cientista de dados.

REFERÊNCIAS

- Conteúdo dessa aula é baseado no livro:

CARVALHO, André Carlos Ponce de Leon Ferreira et al. *Inteligência Artificial* - Uma Abordagem de Aprendizado de Máquina. Disponível em: Minha Biblioteca, (2nd edição). Grupo GEN, 2021.

INTRODUÇÃO À CIÊNCIA DE DADOS

Preparação e Pré-Processamento de Dados - Parte 2