INTRODUÇÃO À CIÊNCIA DE DADOS

Preparação e Pré-Processamento de Dados - Parte 1

PRÉ-PROCESSAMENTO DOS DADOS

- A fase que trabalharemos agora, normalmente vem antes da análise exploratória que já vimos na aula anterior, entretanto, isso pode ser um ciclo que se renova a cada nova fonte de dados que aparece, a cada nova pergunta que se deseja responder, a cada novo atributo que se deseja experimentar no conjunto de dados.
- Alguns chamam essa fase de processamento, muitos de pré-processamento, porque consideram que o processamento é a fase onde modelos de machine learning são testados. Veremos o processamento um pouco mais adiante nessa disciplina.

PRÉ-PROCESSAMENTO DOS DADOS

- O pré-processamento é uma fase que antecede ao uso dos modelos de machine learning que veremos adiante, entretanto para que ele possa ser executado com clareza e também com assertividade é de fundamental importância que se conheça o processo completo de um projeto de DS. Não há como fazer pré-processamento se não se sabe onde se quer chegar, se não se conhece os modelos computacionais a serem testados e a que tipo de respostas se quer chegar com o projeto.
- Importante que se entenda o que é possível fazer de modo global, entretanto, é fato dizer que somente a experiência de fazer e refazer projetos é que vai tornar a fase de pré-processamento assertiva e eficaz.

PRÉ-PROCESSAMENTO DOS DADOS

- Conjuntos de dados podem apresentar diferentes características, dimensões ou formatos.

- Como vimos, dados estruturados podem ser qualitativos (nominais ou ordinais) ou quantitativos (intercalar ou racional), ou ainda podemos pensar nos dados não estruturados que podem ter origem nos e-mails, áudios, vídeos entre outra fontes!

PRÉ-PROCESSAMENTO DOS DADOS

- Os dados podem conter ruídos, imperfeições, valores incorretos ou inconsistentes, podem ser duplicados ou ausentes; os atributos podem ser independentes ou correlacionados; os conjuntos de dados podem apresentar poucos ou muitos objetos, que podem ter uma pequena ou grande quantidade de atributos.

PRÉ-PROCESSAMENTO DOS DADOS

- Técnicas de pré-processamentos tem como principal objetivo melhorar a qualidade dos dados e também procurar eliminar elementos que podem criar um falso resultado no processamento dos dados.

- Às vezes a fase de pré-processamento tem como objetivo ajustar os dados para um uso mais adequado, modelando-o para que possa ser processado.

PRÉ-PROCESSAMENTO DOS DADOS

- Por isso é tão importante conhecer os tipos de dados, as grandezas, para que seja possível identificar as necessidades de ajustes aos quais os dados precisam ser submetidos.

- Um conjunto de técnicas podem ser aplicadas e elas não têm regras, não têm sequência, não têm receita de bolo, é o olhar do cientista de dados, e sua experiência, que determinam o que precisa ser feito.

PRÉ-PROCESSAMENTO DOS DADOS

- A seguir vamos discutir técnicas já consolidadas e problemas recorrentes em conjunto de dados.
- Importante ressaltar que quando há referência ao termo "objeto", geralmente estamos nos referindo a um registro de uma tabela ou algo assim.
- O "conjunto de objetos" pode ser compreendido como um conjunto de registros, o que seria uma tabela, ou uma planilha.
- Os "atributos" são as variáveis ou, em tabelas são seus campos e em planilhas as colunas.
- Importante a compreensão dos termos utilizados na literatura.

INTEGRAÇÃO

- Dados podem ser oriundos de diversas fontes, de diversos conjuntos de dados e em determinada situação precisam ser integrados;
- Imagine que dados podem ser oriundos de uma API, com informações sobre investimento em marketing e seja preciso integrar com dados de vendas feitas em uma outra plataforma digital.
- Aspectos como: atributos correspondentes, com nomes diferentes em bases distintas; informações correspondentes em bases numéricas diferentes ou moedas (ou idiomas) diferentes.
- Em muitos casos, na integração é necessário compreender quais são os atributos necessários de cada objeto. Lembrando sempre que elevado número de atributos pode comprometer o desempenho dos algoritmos de machine learning.

ELIMINAÇÃO MANUAL DE ATRIBUTOS

- Muitas vezes, ao observar um conjunto de dados, fica claro que alguns atributos podem ser eliminados manualmente.
- Retirar os atributos pode estar relacionado, por exemplo, a anonimização de uma base (nome não é necessário).
- Em análises preditivas, quando um atributo não contribui para a estimativa de um valor, ele é irrelevante para a análise e deve ser eliminado.
- Atributos que contém o mesmo valor para todos os objetos também devem ser eliminados, por exemplo, o campo cidade em uma base que analisa dados de uma determinada cidade.

AMOSTRAGEM DE DADOS

- Muitos algoritmos de ML tem dificuldade em lidar com números grandes de objetos, levando à saturação de memória e necessidade de ampliar a escala horizontal da estrutura física.
- Quanto mais dados são utilizados, maior tende a ser a acurácia do modelo e menor a eficiência computacional.

AMOSTRAGEM DE DADOS

- Apesar de toda evolução computacional, haverá caso em que será necessário trabalhar com uma amostra dos dados, de forma que ela seja representativa o suficiente para representar o todo é menor que o conjunto de dados originais para evitar desempenhos ruins no processo.

- Um exemplo de amostra é a progressiva, que começa com uma amostra pequena e aumenta progressivamente enquanto a acurácia continuar a melhorar, até atingir um ponto que não há mais evolução.

DADOS DESBALANCEADOS

- É comum que dados de um subconjunto de uma determinada classe apareçam com frequência maior que das demais classes. Exemplo: ingressos vendidos para um show são 80% de uma área e os outros estão distribuídos entre as demais áreas.
- Esse desbalanceamento afeta muito o desempenho de alguns algoritmos de machine learning, de forma que os algoritmos favoreçam a classificação de novos dados na classe majoritária.
- Redefinir o tamanho do conjunto de dados, utilizar diferentes custos de classificação e induzir um modelo para uma classe são técnicas que podem ser utilizadas.
- Algumas situações incluem as técnicas de classificação com apenas uma classe, ou os dados são treinados separadamente por classe.

LIMPEZA DOS DADOS

- A qualidade do modelo (e dos resultados) é diretamente impactada pela qualidade dos dados.
- Dados ruidosos (que possuem erros ou valores que são diferentes do esperado), inconsistentes (que não combinam ou contradizem valores de outros do mesmo objeto), redundantes (atributos com valores repetidos no mesmo objeto) ou incompletos (com ausência de valores para parte dos dados); são motivos que impactam negativamente os resultados de uma análise.
- Essas deficiências podem ser causadas por problemas nos equipamentos que coletam os dados, na transmissão ou armazenamento, no preenchimento manual ou até em processos de integração.

DADOS INCOMPLETOS

Tabela 3.2 Conjunto de dados com atributos com valores 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

DADOS INCOMPLETOS

- A ausência de valores em alguns atributos pode ter diferentes causas:
  - o atributo não foi considerado importante (ou não era obrigatório) quando os dados foram coletados;
  - desconhecimento do valor do atributo no preenchimento dos valores do objeto;
  - distração no preenchimento;
  - inexistência de valor para o atributo em alguns registros (quantidade de partos para o gênero masculino);
  - problema com equipamento utilizado na coleta.

DADOS INCOMPLETOS

- Algumas técnicas que podem ser utilizadas:
  - eliminar objetos com valores ausentes; essa alternativa normalmente é descartada quando poucos atributos do objeto têm valores ausentes.
  - definir e preencher manualmente valores para atributos com valores ausentes;
  - usar algum método ou heurística para automaticamente definir valores para atributos com valores ausentes.
    - nesse caso é importante definir um valor onde saiba-se que era um valor ausente anteriormente;
    - utilizar média, moda ou mediana dos valores conhecidos (cuidado com isso).
    - definir um indutor baseado em outros atributos.

DADOS INCONSISTENTES

- São dados que possuem valores conflitantes em seus atributos:
  - Exemplo: Idade 3, peso 120; Volta de 5s em um circuito de Fórmula 1, com 3,5 km.
  - Outro exemplo bastante comum é o uso de escalas diferentes para fazer referência a uma mesma medida (metros e centímetros)
- Inconsistências também podem ser reconhecidas quando relações entre atributos são claramente conhecidas (valores correlacionados direta ou indiretamente).
- Algoritmos simples podem verificar existência de inconsistências, em caso de conjuntos de dados não muito grandes, dados inconsistentes podem ser removidos manualmente.

DADOS REDUNDANTES

Tabela 3.5 Conjunto de dados com objetos redundantes

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

Tabela 3.6 Conjunto de dados com atributos redundantes

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

DADOS REDUNDANTES

- Um objeto redundante é um objeto que é muito semelhante a outro no mesmo conjunto de dados.
- Também é considerado um atributo redundante quando ele pode ser deduzido a partir do valor de um ou mais atributos. Dois ou mais atributos estão correlacionados quando apresentam um perfil de variação semelhante para os diferentes objetos.
- Dados redundantes podem criar a falsa sensação de que esse perfil de objeto é mais importante que os demais, induzindo o modelo de análise.
- É importante identificar e eliminar as redundâncias, que podem ser feitas pela eliminação dos objetos semelhantes ou pela combinação dos valores dos atributos dos objetos semelhantes.

FINALIZANDO

- O pré-processamento é uma parte fundamental dos estudos de DS, quanto mais o cientista de dados tiver afinidade com esses processos, melhor serão os resultados.
- É muito importante conhecer os dados e principalmente saber como lidar com eles.
- Na próxima aula teremos demonstração de alguns processos!

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 1