Tópicos Especiais em Computação I

Caracterização dos dados

Patrícia Lucas

Bacharelado em Sistemas de Informação
IFNMG - Campus Salinas

Salinas
Setembro 2021

INSTITUTO FEDERAL
Norte de Minas Gerais
Campus Salinas

Versão 1.2021

Caracterização dos dados

Referência

Capítulo 2: Análise de dados.

Inteligência Artificial: Uma abordagem de aprendizado de máquina. Katti Faceli...[et al.]. - Rio de Janeiro: LTC, 2011.

Conjuntos de dados

Análise dos dados

Quantos dados são gerados a cada dia?

500 milhões de tweets são enviados.
294 bilhões de e-mails são enviados.
4 petabytes de dados são criados no Facebook.
4 terabytes de dados são criados a partir de cada carro conectado.
65 bilhões de mensagens são enviadas no WhatsApp.
5 bilhões de pesquisas são feitas.

Em 2025, estima-se que 463 exabytes de dados serão criados a cada dia em todo o mundo - isso é o equivalente a 212.765.957 DVDs por dia!

https://www.weforum.org/agenda/2019/04/how-much-data-is-generated-each-day-cf4bddf29f/

Conjuntos de dados

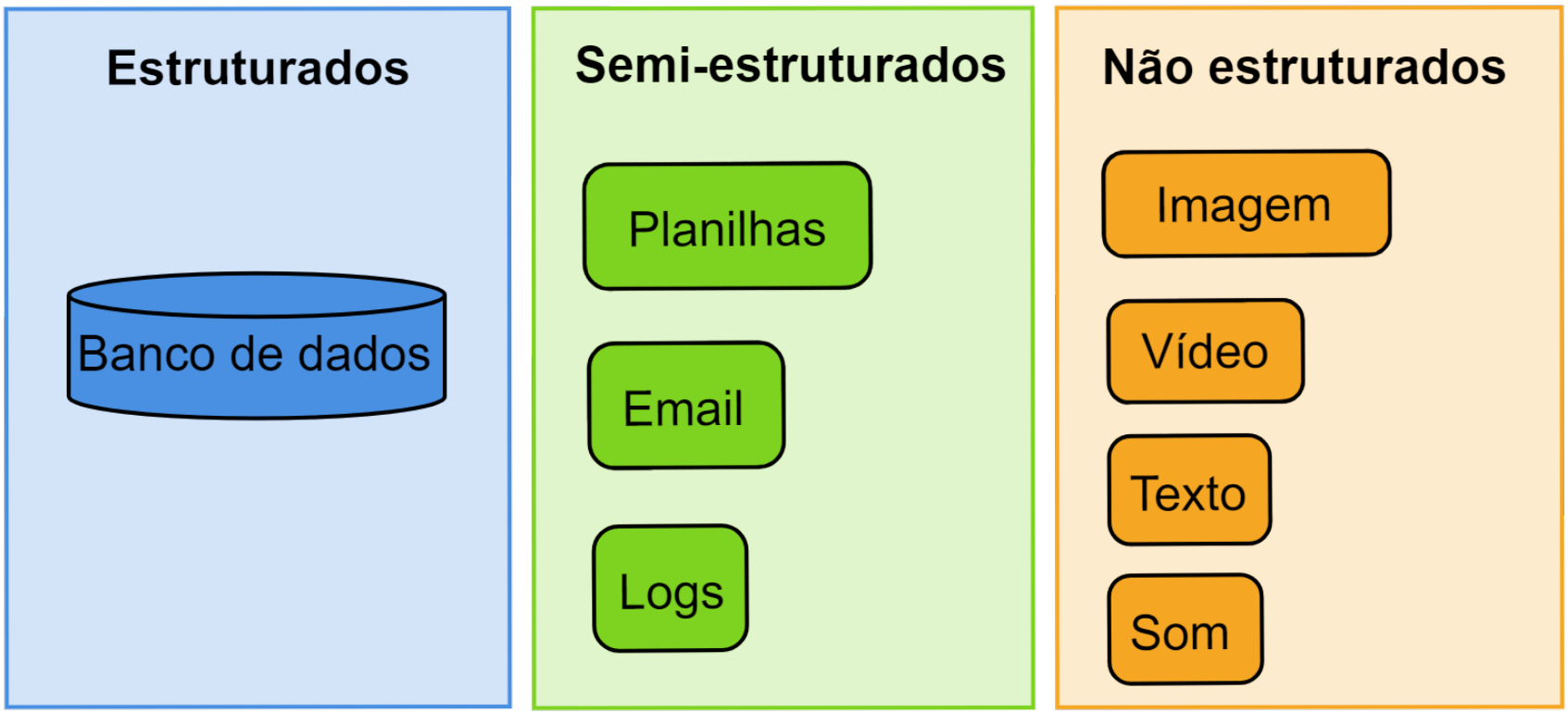
Análise dos dados

Além de serem gerados por várias fontes diferentes, esses dados também assumem formatos diferentes:

- Séries temporais
- Imagens
- Vídeos
- grafos
- texto
- páginas web...

Natureza dos dados

Caracterização dos dados



Originalidade

Caracterização dos dados

- **Primários:** dados originais sem nenhum processamento.
- **Secundários:** dados processados, agregados ou transformados.
- **Terciário:** resultado de análises sobre os dados. Ex: gráficos e tabelas resumidas.

Volume dos dados

Caracterização dos dados

- **Small data:** cabem integralmente na memória (HD) de uma máquina comum.
- **Big data:**
  - não cabem integralmente na memória.
  - necessidade de processamento distribuído.
  - Além do volume, o big data também tem as características de Variedade e Velocidade.

Geração dos dados

Caracterização dos dados

- **Naturais:**
  - **Observacionais:** coleta de dados sem intervenção no processo observado.
  - **Intervencionais:** resultados de experimentos com intervenções nos valores.
- **Sintéticos:** dados gerados por um algoritmo ou um modelo generativo.

Caracterização dos dados

Introdução

Os conjuntos de dados são formados por **objetos**, que por sua vez possuem **atributos**.

atributos = atributos de entradas = vetor de características = atributos preditivos

Cada **objeto** corresponde a uma ocorrência dos dados.

Cada **atributo** está associado a uma propriedade do objeto.

Caracterização dos dados

Introdução

Os conjuntos de dados são formados por **objetos**, que por sua vez possuem **atributos**.

atributos = atributos de entradas = vetor de características = atributos preditivos

Cada **objeto** corresponde a uma ocorrência dos dados.

Cada **atributo** está associado a uma propriedade do objeto.

Caracterização dos dados

Análise dos dados

Exemplo de conjunto de dados de pacientes de um hospital:

Id.	Nome	Idade	Sexo	Peso	Manchas	Temp.	# Int.	Est.	Diagnóstico
4201	João	28	M	79	Concentradas	38,0	2	SP	Doente
3217	Maria	18	F	67	Inexistentes	39,5	4	MG	Doente
4039	Luiz	49	M	92	Espalhadas	38,0	2	RS	Saudável
1920	José	18	M	43	Inexistentes	38,5	8	MG	Doente
4340	Cláudia	21	F	52	Uniformes	37,6	1	PE	Saudável
2301	Ana	22	F	72	Inexistentes	38,0	3	RJ	Doente
1322	Marta	19	F	87	Espalhadas	39,0	6	AM	Doente
3027	Paulo	34	M	67	Uniformes	38,4	2	GO	Saudável

- Cada linha representa 1 paciente.

- Cada paciente possui um vetor de características com Nome, Idade, Sexo, Peso, Manchas, Temperatura, Número de internações e Estado onde reside.

- Cada paciente também possui um atributo de saída (rótulo), que representa o fenômeno de interesse sobre o qual se deseja fazer previsões.

- Dependendo do problema, podem existir mais de um atributo de saída ou não existir nenhum.

Caracterização dos dados

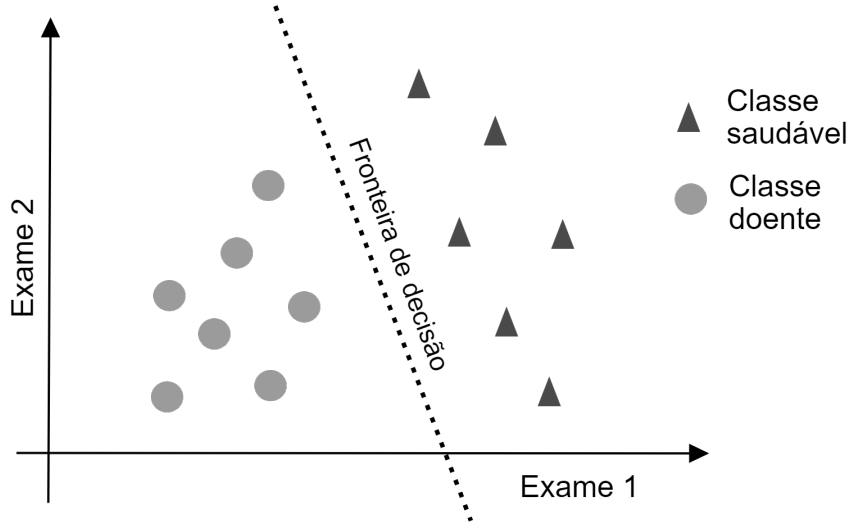
Análise dos dados

Formalmente, os dados podem ser representados por uma matriz de objetos X_n^d, em que *n* é o número de objetos e *d* é o número de atributos de entrada.

O valor *d* define a dimensionalidade dos objetos ou do espaço de objetos e podem ser vistos também como um conjunto de eixos ortogonais e os objetos, como pontos do espaço de dimensão *d*.

Já *n* defini o volume da base de dados.

Exame 2	**Exame 1**	**Diagnóstico**
1	2	Doente
1	10	Saudável



Caracterização dos dados

Análise dos dados

Quando o atributo alvo contém rótulos que identificam categorias ou classes às quais os objetos pertencem, ele é denominado classe e assume valores discretos $1, ..., k$.

Nesse caso temos um problema de classificação!

Classe majoritária: classe que possui maior número de objetos.

Classe minoritária: classe que possui menor número de objetos.

Caracterização dos dados

Análise dos dados

Se o atributo de saída contém valores numéricos contínuos, tem-se *um problema de regressão!*

Uma caso especial de problema de regressão é a previsão de séries temporais.

Caracterização dos dados

Análise dos dados

Tipos de dados: diz respeito ao grau de quantização nos dados.

- Se o atributo representa quantidades, ele é denominado quantitativo ou numérico.
  Contínuo: podem assumir um número infinito de valores.
  Exemplo: peso.
  Discreto: assumem um número finito ou infinito contável.
  Exemplo: número de filhos.

- Se o atributo representa qualidade, ele é denominado qualitativo ou categórico.
  Exemplo: tamanho(baixa, média, alta) ou sexo(masculino, feminino).

Caracterização dos dados

Análise dos dados

Escala: defini que operações podem ser realizadas sobre os valores.

- Nominais: qualitativos e não possuem relação de ordem ($=$ e $\neq$).
- Ordinais: qualitativos e possuem relação de ordem ($=$, $\neq$, $<$, $>$, $\leq$ e $\geq$).
- Intervalares: quantitativos, representados por números dentro de um intervalo.
- Racionais: quantitativos e são os que mais carregam informações, pois seus valores têm significado absoluto.

Atributo	Classificação
Id.	Nominal
Nome	Nominal
Idade	Racional
Sexo	Nominal
Peso	Racional
Manchas	Nominal
Temp.	Intervalar
#Int.	Racional
Est.	Nominal
Diagnóstico	Nominal

Caracterização dos dados

Análise dos dados

- **Completude:**
  - o quão representativos são os dados em relação ao fenômeno ou a população real?
  - O volume de dados é suficiente?
- **Integridade:** Os dados são coerentes?(falhas de coleta, transformações, armazenamento, etc.)
- **Atualidade:** quão recentes são os dados? Eles ainda são representativos?