Fundamentos de Mineração de Dados e Ciência de Dados

Técnicas de Mineração de Dados em Grande Escala (Big Data)

Marcelo de Souza Lauretto

03 de Novembro de 2025

LEMBRETE...

Marque presença!

Parte 1 - Big Data

Definição de Big Data

- **Conceito Formal**: Conjuntos de dados massivos e complexos que excedem a capacidade de processamento de ferramentas tradicionais de banco de dados.
- **Característica Principal**: Requer frameworks especiais para armazenamento e processamento distribuído.
- **Exemplos**:
  - **Redes sociais**: Análise de comportamento de milhões de usuários
  - **E-commerce**: Análise de padrões de compra em tempo real
  - **Sensores IoT**: Dados de clima, tráfego, dispositivos médicos
  - **Mercado financeiro**: Transações em tempo real nas bolsas de valores

Big Data vs Small Data

- **Small Data**:
  - Volume: GBs a TBs, armazenável em servidores únicos
  - Velocidade: Processamento em batch, atualizações periódicas
  - Variedade: Principalmente dados estruturados
  - Ferramentas: Bancos relacionais (SQL), Excel, BI tradicional
  - Exemplo: Vendas mensais de uma loja, cadastro de clientes
- **Big Data**:
  - Volume: TBs a PBs/EBs[1], requer armazenamento e processamento distribuído
  - Velocidade: Streaming em tempo real, processamento contínuo
  - Variedade: Estruturado + não-estruturado + semi-estruturado
  - Ferramentas: Hadoop, Spark, NoSQL, processamento distribuído
  - Exemplo: Análise de sentimentos em redes sociais, telemetria de veículos autônomos

[1] *PB*: Petabyte = 1000^5 bytes; *EB*: Exabyte = 1000^6 bytes

Classificação dos Tipos de Dados (1/2)

- **Dados Estruturados**:
  - Possuem esquema[2] fixo e formato tabular bem definido
  - Armazenáveis em bancos relacionais (SQL)
  - Exemplos: Tabelas de clientes, transações financeiras, planilhas Excel
  - Características: Fácil consulta, análise e agregação
- **Dados não Estruturados**:
  - Sem formato ou esquema pré-definido
  - Representam $\approx$ 80% dos dados corporativos atuais
  - Exemplos: Vídeos, imagens, PDFs, emails, posts em redes sociais
  - Armazenamento: NoSQL databases, sistemas de arquivos distribuídos

[2] Um *esquema* é a definição formal da estrutura de um banco de dados, especificando as estruturas das tabelas, chaves, relacionamentos entre entidades etc.

Classificação dos Tipos de Dados (2/2)

- **Dados Semi-Estruturados**:
  - Não possuem estrutura rígida, mas contêm metadados ou tags organizacionais
  - Permitem algum nível de análise sem esquema fixo
  - Exemplos: JSON, XML, logs de sistemas, dados de sensores com timestamp
  - Vantagem: Flexibilidade combinada com alguma estruturação
- **Importância da Classificação**:
  - Define estratégias de armazenamento adequadas
  - Determina ferramentas de análise apropriadas
  - Impacta performance e custos de processamento

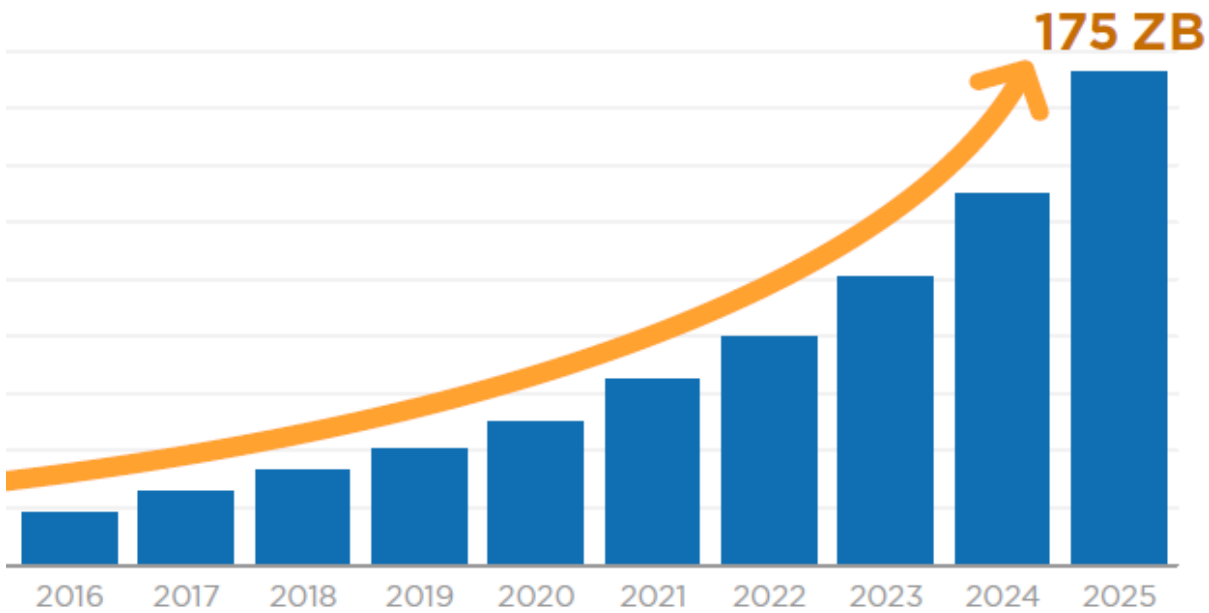
Características de Big Data - Os 5 Vs (1/3)

- **Volume**:
  - Uma quantidade imensa de dados é produzida diariamente: processos de negócios, sensores IoT, plataformas de redes sociais etc.
- **Velocidade**:
  - Geração em tempo real: A cada minuto:
    - 100k+ tweets, 69M+ pesquisas Google, 168M+ emails [Mohan(2024)]
  - Necessidade de processamento streaming (Spark, Flink)
  - Aplicações: Detecção de fraude, monitoramento de redes
- **Variedade**:
  - Diferentes formatos: estruturado, não-estruturado, semi-estruturado
  - Fontes diversas: textos, áudios, vídeos, sensores, transações
  - Desafio: Integração de formatos heterogêneos

Características de Big Data - Os 5 Vs (2/3)

- **Veracidade**:
  - Qualidade, confiabilidade e precisão dos dados
  - Desafio: Dados incompletos, inconsistentes ou imprecisos
  - Exemplo: Análise de sentimentos em posts com hashtags
  - Soluções: Validação, limpeza e filtragem de dados
- **Valor**:
  - Dados devem gerar valor de negócio tangível
  - Foco em insights acionáveis, não apenas em volume
  - Exemplos: Otimização de campanhas de marketing, redução de churn
  - Métrica: ROI (Return on Investment) em analytics

Características de Big Data - Os 5 Vs (3/3)



Volumes esperados de dados gerados mundialmente[3]

[Mohan(2024)]

[3] *ZB*: Zettabyte = 1000^7 bytes

Big Data Analytics: Definição

- **Definição Formal**: Processo sistemático de examinar grandes bases de dados para descobrir padrões ocultos, correlações desconhecidas, tendências de mercado e preferências de clientes.
- **Componentes Principais**:
  - **Coleta de dados**: Agregação de múltiplas fontes (IoT, redes sociais, transacionais)
  - **Pré-processamento**: Limpeza, transformação e filtragem para garantir qualidade
  - **Análise**: Aplicação de estatística, machine learning e mineração de dados
  - **Visualização**: Apresentação em dashboards, gráficos e relatórios interativos
  - **Segurança**: Proteção de dados sensíveis e conformidade com regulamentações

Importância do Big Data Analytics

- **Tomada de Decisão Melhorada**:
  - Decisões baseadas em dados ao invés de intuição
  - Identificação de padrões não visíveis em pequenos conjuntos
- **Eficiência Operacional**:
  - Otimização de processos e identificação de gargalos
  - Redução de custos através de análise preditiva de manutenção
- **Experiência do Cliente Personalizada**:
  - Análise de comportamento individual para oferecer produtos relevantes
  - Retenção de clientes através de recomendações precisas
- **Inovação e Desenvolvimento de Produtos**:
  - Identificação de novas oportunidades de mercado
  - Desenvolvimento baseado em feedback real dos usuários

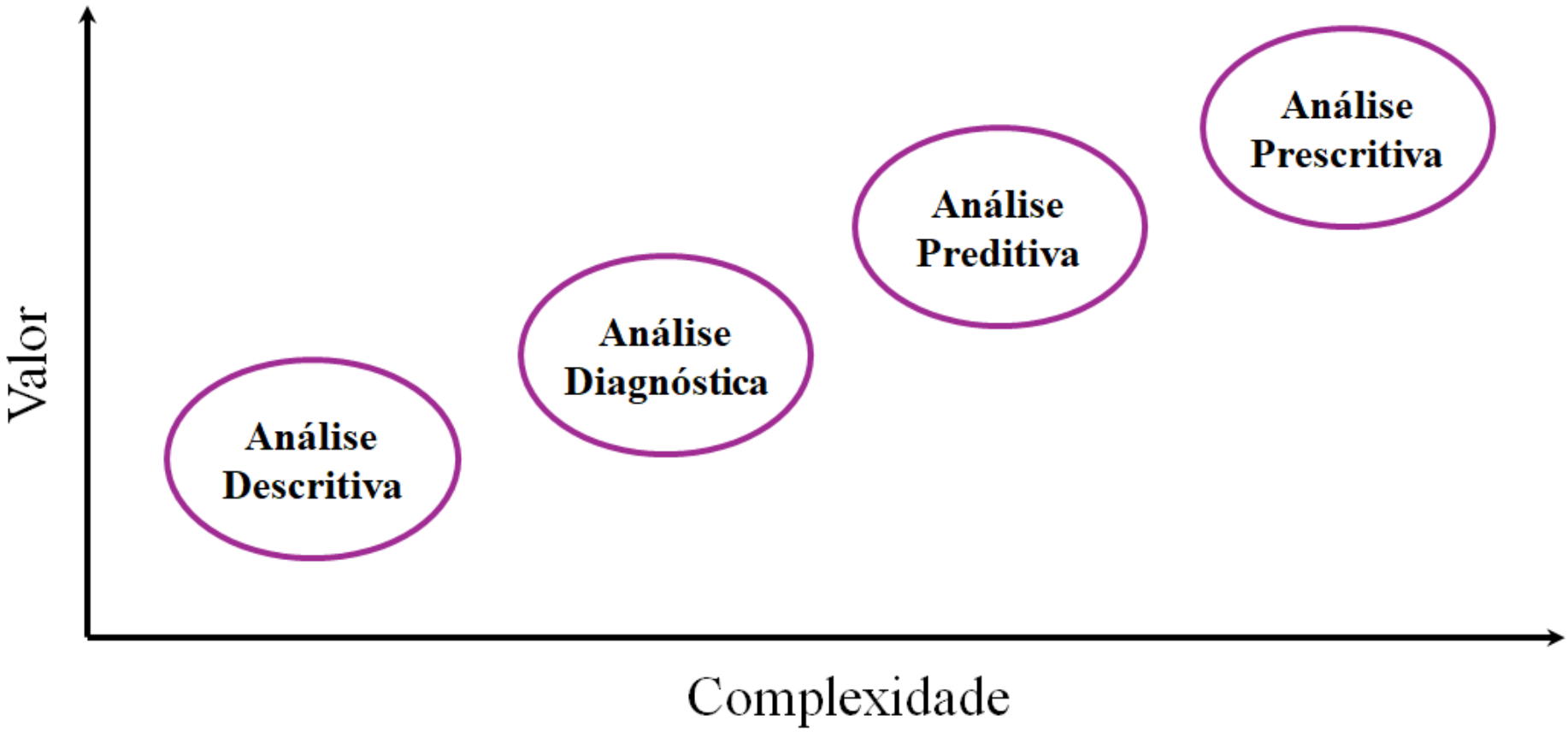
Classificação de Data Analytics (1/3)

- **Analítica Descritiva**:
  - Pergunta: "O que aconteceu?"
  - Foco: Análise de dados históricos
  - Método: Agregações, relatórios, dashboards
  - Exemplo: "Qual foi o volume de vendas em nosso aplicativo?"
- **Analítica Diagnóstica**:
  - Pergunta: "Por que aconteceu?"
  - Foco: Identificação de causas raiz
  - Método: Detecção de anomalias, descoberta de correlações simples
  - Exemplo: "Por que o volume de vendas pelo aplicativo diminuiu?"

Classificação de Data Analytics (2/3)

- **Analítica Preditiva**:
  - Pergunta: "O que pode acontecer no futuro?"
  - Foco: Previsão de eventos futuros com base nos dados disponíveis
  - Método: Aprendizado de Máquina e modelos estatísticos voltados à identificação de associações mais complexas
  - Exemplo: "Qual será o volume de vendas pelo aplicativo no próximo mês?"
- **Analítica Prescritiva**:
  - Pergunta: "O que devemos fazer?"
  - Foco: Recomendações de ações efetivas para as decisões de negócio
  - Método: Combinação de dados internos, fontes externas e Aprendizado de Máquina Interpretável
  - Exemplo: "O que devemos fazer para aumentar as vendas pelo aplicativo?"

Classificação de Data Analytics (3/3)



Complexidade $\times$ Valor das quatro classes de Data Analytics

Adaptado de [Mohan(2024)]

Desafios do Big Data (1/2)

- **Qualidade dos Dados**:
  - Dados incompletos, inconsistentes ou imprecisos
  - Impacto direto na confiabilidade dos insights
  - Solução: Processos robustos de limpeza e validação
- **Integração de Dados**:
  - Combinar fontes heterogêneas (estruturadas e não-estruturadas)
  - Diferentes formatos, esquemas e padrões
- **Segurança e Privacidade**:
  - Proteção de dados sensíveis e conformidade com LGPD/GDPR
  - Riscos de vazamento e uso indevido
  - Solução: Criptografia, controle de acesso, anonimização

Desafios do Big Data (2/2)

- **Escalabilidade e Performance**:
  - Processamento de volumes crescentes dentro de prazos aceitáveis
  - Balanceamento entre custo e performance
  - Solução: Arquiteturas distribuídas, computação em nuvem
- **Lacuna de Talentos**:
  - Escassez de profissionais qualificados (cientistas de dados, engenheiros)
  - Necessidade de habilidades multidisciplinares
  - Solução: Treinamento, parcerias acadêmicas

Parte 2 - Fluxo de Dados

Conceito: Revisão Determinístico

- **Na aula 05**: Conceito como função determinística

$$c : \mathcal{X} \to \mathcal{Y}$$

onde

  - $\mathcal{Y} = \{1, \ldots, K\}$ (Classificação)
  - $\mathcal{Y} = \mathbb{R}$ (Regressão)

- **Características**:
  - Mapeamento direto atributos $\to$ classe (ou valor)
  - Sem incerteza na predição
  - Cada X tem um único Y correspondente

Função Determinística $Y = c(X)$

Conceito: Transição para Probabilístico

- **Problema**: Mundo real tem incertezas
- **Solução**: Conceito probabilístico
- **Nova visão**:
  - X determina distribuição de Y
  - Captura variabilidade natural
  - Mais realista

Distribuição Condicional
$Y \sim P(Y|X)$

Conceito: Definição Formal

Definição Probabilística

Um **conceito** é representado pela distribuição condicional:

$$P(Y|X)$$

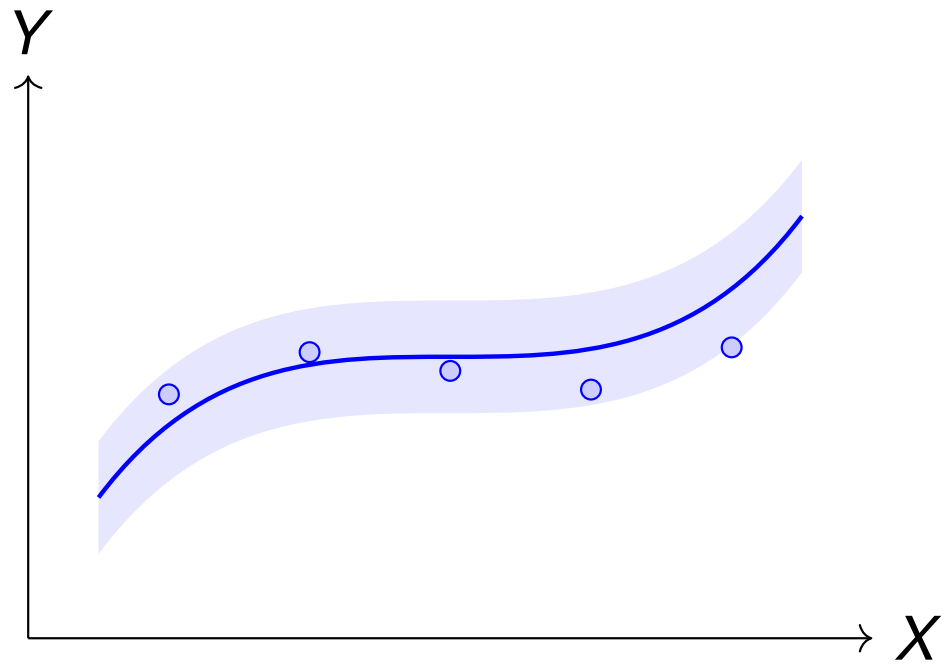
onde:

- $X \in \mathcal{X}$: vetor de atributos
- $Y \in \mathcal{Y}$: variável resposta
- $P(Y|X)$: probabilidade de Y dado X

Interpretação

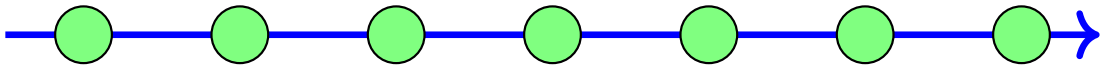
- Para cada contexto X, temos uma distribuição sobre Y
- O classificador ψ aproxima esta distribuição

Conceito: Interpretação Visual



- *Pontos azuis*: Dados observados no conjunto de treinamento
- *Linha azul*: Valor esperado $E(Y|X)$ (desconhecido) para cada valor de X
- *Área azul*: Região de incerteza de $P(Y|X)$
- *Finalidade do modelo*: buscar a melhor aproximação para $E(Y|X)$ (linha azul)

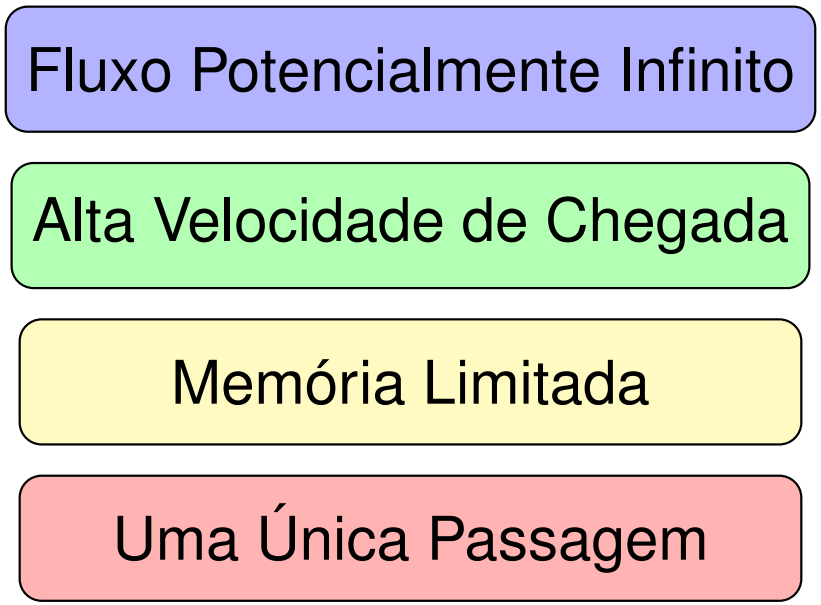
Fluxo de Dados: Definição Intuitiva



Características Principais

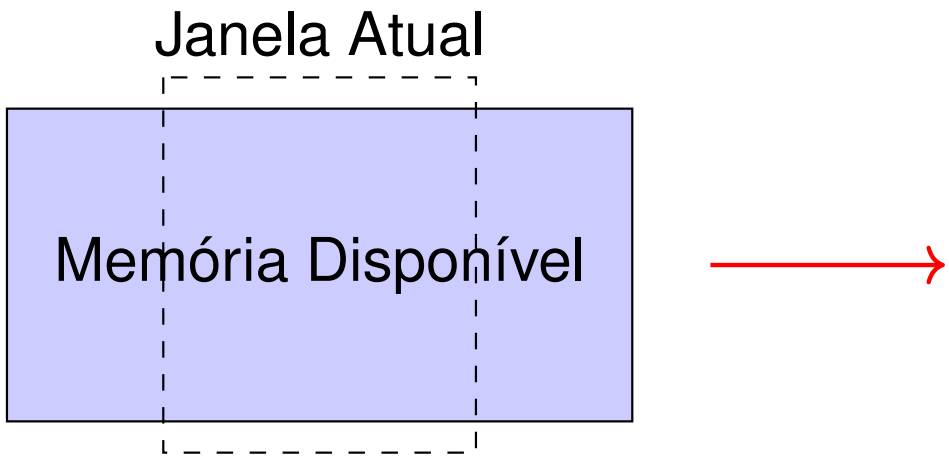
- **Sequência infinita** de dados
- Chegam **continuamente**
- Ordem temporal importante
- Não cabem na memória

Fluxo de Dados: Características Principais



- Dados processados **uma vez**
- Algoritmos precisam ser
  - **online**: processam os dados à medida que chegam
  - **adaptativos**: atualizam o modelo com os novos dados sem reprocessar tudo

Fluxo de Dados: Restrições Práticas



Consequências

- Só podemos manter uma **janela** dos dados
- Dados antigos são descartados

Fluxo de Dados: Exemplos de Aplicações

- **Aplicações Reais**
  - Redes sociais
  - Transações financeiras
  - Sensores (IoT)
  - Web logs
  - Telecomunicações

Fluxo de Dados: Desafios Técnicos

- **Processamento em Tempo Real**
  - Dados devem ser processados à medida que chegam
  - Latência mínima para decisões em tempo real
- **Limitação de Memória**
  - Dados não cabem inteiramente na memória
  - Uso de janelas deslizantes ou amostragem
- **Detecção de Mudanças**
  - Identificar quando ocorre *data shift* ou *concept drift*
  - Monitoramento contínuo da performance do modelo
- **Escalabilidade**
  - Lidar com volumes crescentes de dados
  - Manter eficiência computacional com fluxo contínuo

Fluxo de Dados: Definição Formal

- Um **fluxo de dados** é uma sequência:

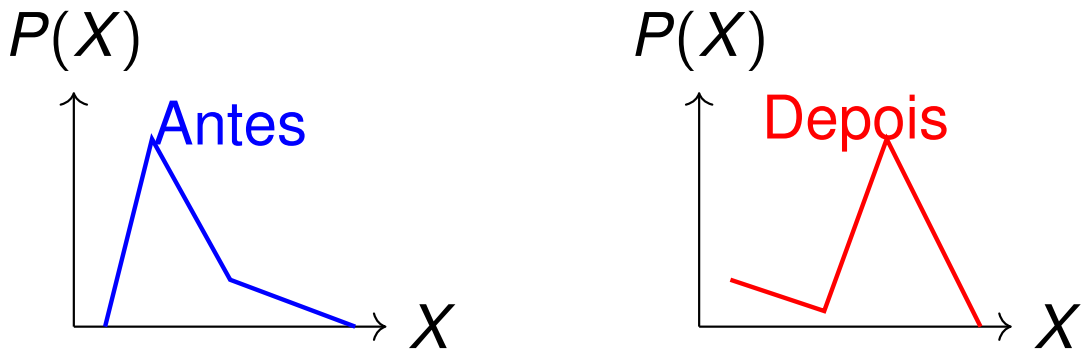
$$S = \{(x_1, y_1), (x_2, y_2), (x_3, y_3), \dots\}$$

onde cada instância (x_t, y_t) é gerada em tempo t segundo:

$$(x_t, y_t) \sim \mathcal{D}_t$$

  - $\mathcal{D}_t$: distribuição no tempo t (pode mudar)
  - Ordem temporal fundamental

Data Shift: Definição Intuitiva



Data Shift = Mudança em $P(X)$

- **Antes**: Atributos X têm uma distribuição
- **Depois**: Atributos X têm outra distribuição
- **Mesma relação**: $P(Y|X)$ não mudou
- **Problema**: Modelo vê tipos diferentes de dados

Data Shift: Exemplo

Modelo de previsão de vendas de uma loja online:

Dias Normais

- Clientes: 25-45 anos
- Renda: R$ 3.000-8.000
- Região: Sudeste

Black Friday

- Clientes: 18-60 anos
- Renda: R$ 1.500-15.000
- Região: Todo Brasil

Problema

- O modelo treinado com dados dos dias normais só reconhece os padrões de compra daqueles clientes.
- Portanto, seu desempenho será fraco quando aplicado ao público mais amplo.

Data Shift: Definição Formal

Definição Matemática

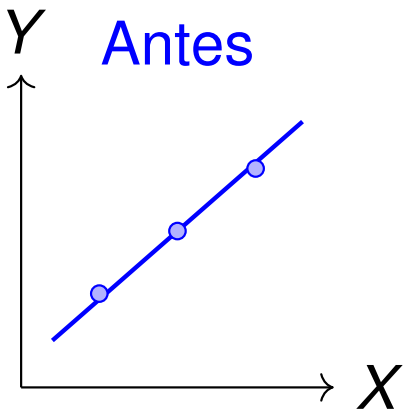
Dizemos que existe um **Data Shift** quando:

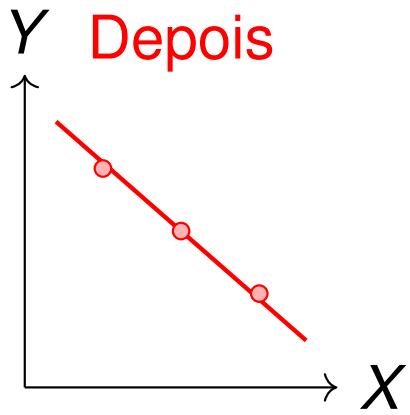
$$P_t(X) \neq P_{t+k}(X) \quad \text{para algum } k > 0$$

Interpretação

- Atributos X têm distribuições diferentes em t e $t + k$
- Relação $X \rightarrow Y$ permanece a mesma

Concept Drift: Definição Intuitiva





O que é Concept Drift?

- **Mesmos dados** X, **relação diferente** com Y
- $P(Y|X)$ mudou ao longo do tempo
- **Regras fundamentais** mudaram
- Modelo treinado no passado não funciona mais

Concept Drift: Exemplo

Mudança de Comportamento de Compras dos Jovens

Antes da Pandemia:

- Alto consumo em bares/restaurantes
- Baixo consumo em produtos home office

Durante a Pandemia:

- Baixo consumo em bares/restaurantes
- Alto consumo em produtos home office

Concept Drift: Exemplo

O que mudou?

- **População de jovens** mantida
- **Mesmos atributos**: Idade
- **Relação diferente**: Padrão de compra mudou

Características do Concept Drift

- **Antes**: $P(\text{HomeOffice}|\text{Jovem}) = \text{baixa}$
- **Depois**: $P(\text{HomeOffice}|\text{Jovem}) = \text{alta}$
- **Conceito**: "Relação idade$\rightarrow$tipo de produto" mudou

Implicação

Modelo treinado antes da pandemia preveria errado os padrões de compra durante a pandemia, mesmo mantida a mesma população!

Concept Drift: Definição Formal

Definição Matemática

Ocorre **Concept Drift** quando:

$$P_t(Y|X) \neq P_{t+k}(Y|X) \quad \text{para algum } k > 0$$

Consequência

O classificador/regressor ψ_t torna-se obsoleto!

Concept Drift: Tipos Principais

Súbito - Mudança brusca

Gradual - Transição lenta

Incremental - Fases intermediárias

Recorrente - Padrões sazonais

Concept Drift: Exemplos dos Tipos

Exemplos

- Súbito: Nova lei
- Gradual: Hábitos
- Incremental: Tecnologia
- Recorrente: Black Friday

Implicações

- Detecção automática
- Re-treino
- Adaptação

Data Shift vs Concept Drift: Comparação

Data Shift

- Muda: $P(X)$
- Mantém: $P(Y|X)$
- Problema: Dados

Concept Drift

- Muda: $P(Y|X)$
- Mantém: $P(X)$
- Problema: Modelo

Ponto de checagem

Parte 3 - Esquema Treino-teste

Limitações dos Métodos de Avaliação Tradicionais em Data Streams

- **Avaliação estática inadequada**: Em fluxos contínuos, a separação clássica entre conjuntos de treino e teste fixos não captura a natureza temporal e evolutiva dos dados.
- **Data shift / Concept drift não detectados**: Distribuições dos dados mudam ao longo do tempo, tornando modelos treinados em dados antigos gradualmente obsoletos.
- **Ineficiência no uso de dados**: Manter conjuntos de teste estáticos desperdiça informações valiosas que poderiam ser usadas para atualizar o modelo continuamente.
- **Avaliação fora do contexto real**: Não reflete o cenário de implantação onde o modelo precisa fazer predições e aprender sequencialmente.

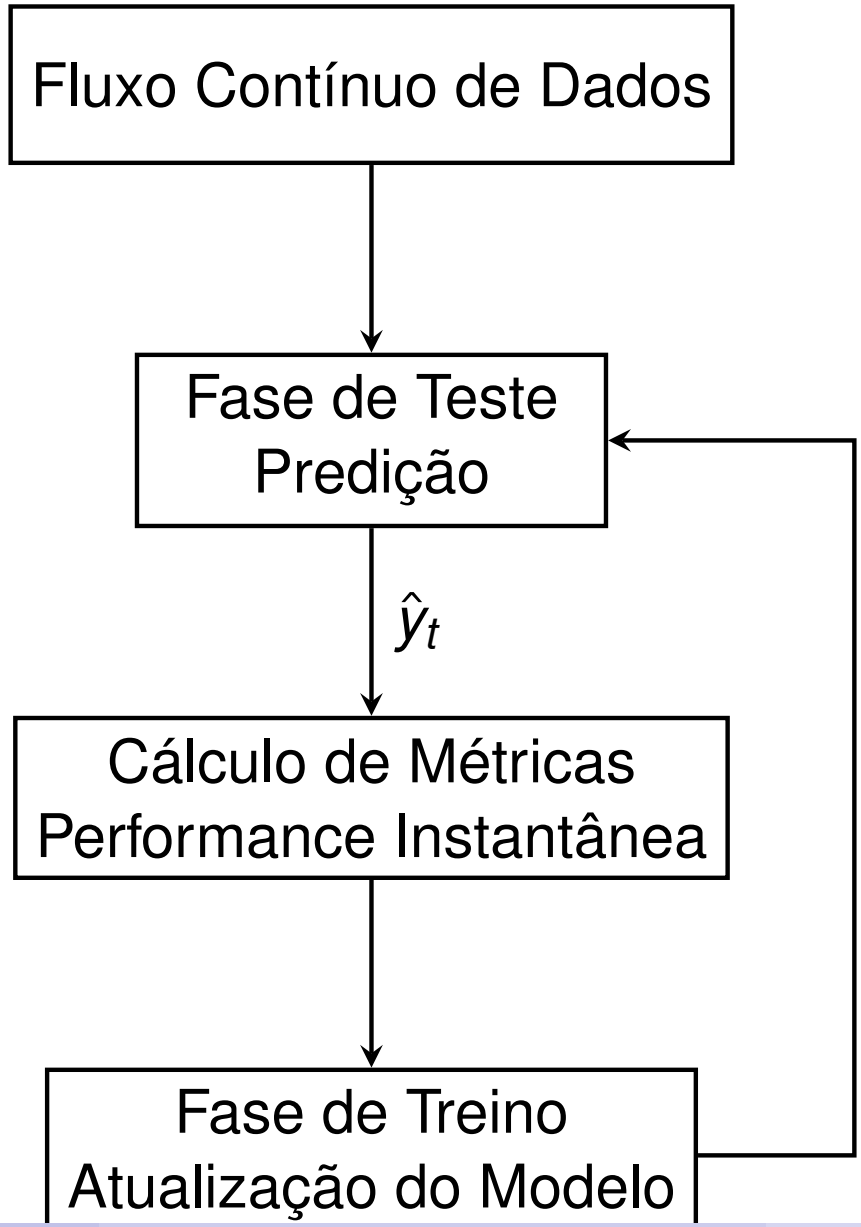
Método Prequential: Avaliação Sequencial Preditiva

Conceito Fundamental

Pre(dictive) + (se)quential = Abordagem onde cada instância ou bloco de dados é primeiro usado para teste (fazer predições) e subsequentemente para atualizar o modelo.

- **Ordem crucial**: Primeiro teste, depois treino - simulando o cenário real de implantação.
- **Avaliação contínua**: As métricas de performance são computadas incrementalmente ao longo de todo o fluxo de dados.
- **Eficiência computacional**: Processamento online que não requer armazenamento de grandes conjuntos de dados históricos.

Fluxo do Método Prequential



Algoritmo Prequencial: Versão com Blocos de Dados

Pseudocódigo para Processamento em Blocos

Entrada: Fluxo de dados S, tamanho do bloco k, modelo inicial M_0
para cada bloco $B_t = \{(x_i, y_i)\}_{i=1}^{k}$ **faça**:
  Fase de Teste:
    para cada $(x_i, y_i) \in B_t$ **faça**:
      $\hat{y}_i \leftarrow M_t.predict(x_i)$
      Calcule métrica instantânea m_i
    fim para
  Atualize as métricas acumuladas do modelo com os novos resultados $\{m_i\}_{i=1}^{k}$
  Fase de Treino:
    $M_{t+1} \leftarrow M_t.partial\_fit(B_t)$
    $t \leftarrow t + 1$
fim para

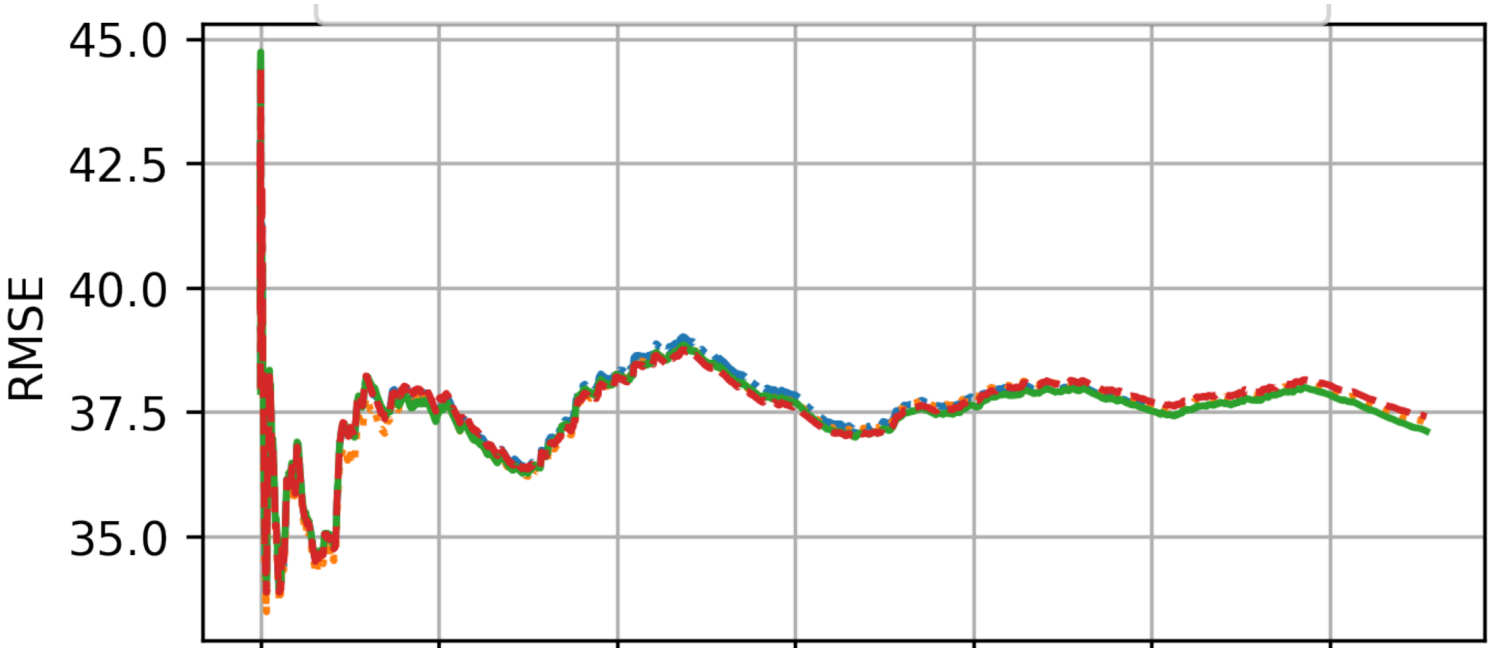
Caso particular: Quando $k = 1$, temos o processamento individual de cada nova instância.

Vantagens do Método Prequential

- Vantagens Principais
  - **Uso eficiente de dados**: Cada instância contribui tanto para avaliação quanto para aprendizado
  - **Detecção proativa de concept drift**: Mudanças na distribuição são mais rapidamente refletidas nas métricas
  - **Avaliação contínua e em tempo real**: Monitoramento constante do desempenho do modelo
  - **Simulação realista**: Reproduz o cenário de implantação onde o modelo gera predições e aprende sequencialmente

Avaliação de desempenho

- Monitoramento das métricas ao longo do tempo



[Mastelini(2023)]

Referências

Gama, J.

Knowledge Discovery from Data Streams. CRC Press, 2010.
Pré-print disponível em
http://www.liaad.up.pt/area/jgama/DataStreamsCRC.pdf

Mastelini, S. M. (2023).

Efficient online tree, rule-based and distance-based algorithms. Tese de Doutorado, Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos. doi:10.11606/T.55.2023.tde-30082023-135843.

Mohan, G.M.

Fundamentals of Big Data Analytics (Digital Notes). R20A0567, B. Tech IV Year – I Sem, Malla Reddy College of Engineering & Technology, 2024-2025.
Disponível em https://mrcet.com/downloads/digital_notes/EEE/14062023/FUNDAMENTALS%20OF%20BIG%20DATA%20ANALYTICS%20Digital%20Notes.pdf