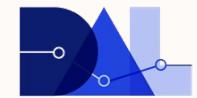


CEFET/RJ



Pré-processamento de Dados - Fundamentos

Uma jornada essencial para transformar dados brutos em insights valiosos

Eduardo Ogasawara

eduardo.ogasawara@cefet-rj.br

<https://eic.cefet-rj.br/~eogasawara>

O que é Pré-processamento de Dados?

O pré-processamento de dados é um conjunto fundamental de técnicas que prepara dados brutos para análise efetiva. Este processo multifacetado garante que seus dados estejam limpos, consistentes e prontos para mineração de dados e aprendizado de máquina.



Limpeza de Dados

Preenche valores ausentes, suaviza dados ruidosos, identifica e remove outliers, e resolve inconsistências que podem comprometer a análise.



Integração de Dados

Combina múltiplos bancos de dados, cubos de dados ou arquivos em uma fonte coerente e unificada para análise abrangente.



Redução de Dados

Aplica redução de dimensionalidade, redução de numerosidade e compressão de dados para otimizar o processamento sem perder informações essenciais.



Transformação de Dados

Realiza discretização, normalização e geração de hierarquias conceituais para preparar dados para algoritmos específicos.

- Referências:** Han, Pei & Tong (2022). Data Mining: Concepts and Techniques, 4^a ed. • García, Luengo & Herrera (2014). Data Preprocessing in Data Mining.

Qualidade de Dados: Por que Pré-processar?

A qualidade dos dados é multidimensional e crítica para o sucesso de qualquer projeto de ciência de dados. Dados de baixa qualidade levam a insights incorretos e decisões equivocadas. Compreender essas dimensões é o primeiro passo para garantir análises confiáveis.



Precisão

Os dados estão corretos ou incorretos? Quão precisos são em relação à realidade que representam?



Completude

Todos os valores necessários estão presentes? Há atributos não registrados ou indisponíveis?



Consistência

Os dados são uniformes? Existem modificações parciais ou referências pendentes entre registros?



Pontualidade

As atualizações são realizadas em tempo hábil? Os dados refletem o estado atual do sistema?



Confiabilidade

Quão confiável é a fonte dos dados? Podemos acreditar que as informações estão corretas?



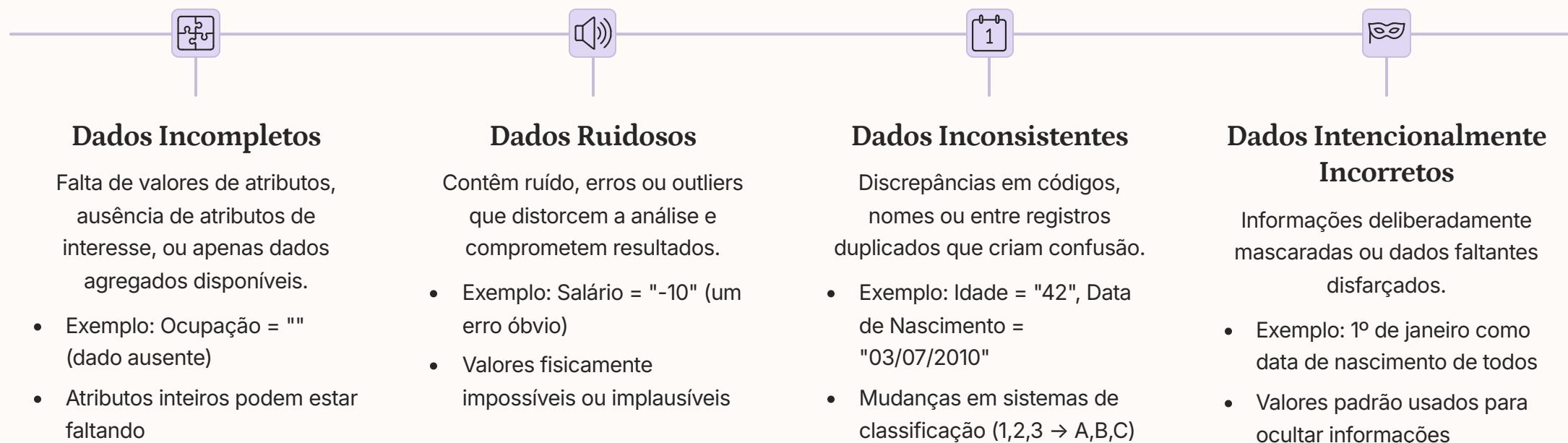
Interpretabilidade

Os dados podem ser facilmente compreendidos? A documentação e metadados são adequados?

Referências: Han, Pei & Tong (2022). Data Mining: Concepts and Techniques, 4^a ed. • García, Luengo & Herrera (2014). Data Preprocessing in Data Mining.

Limpeza de Dados: Enfrentando a Realidade

Dados do mundo real são invariavelmente "sujos" devido a falhas instrumentais, erros humanos ou de computador, e problemas de transmissão. Compreender os tipos de problemas é essencial para aplicar as técnicas corretas de limpeza.



Referência: Chu, Ilyas, Krishnan & Wang (2016). Data cleaning: Overview and emerging challenges. ACM SIGMOD.

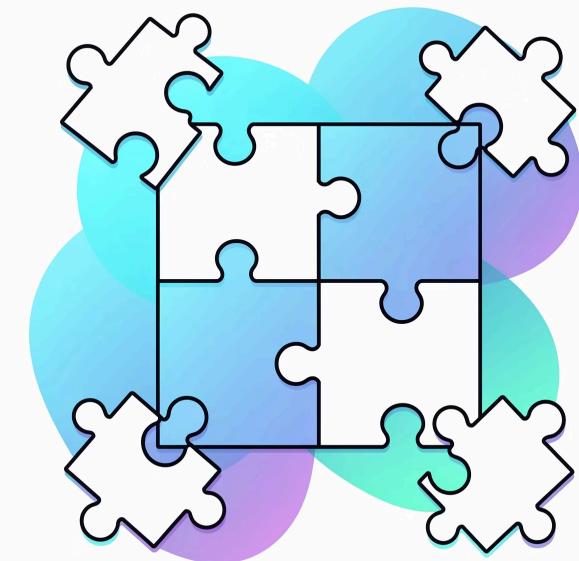
Dados Incompletos: Compreendendo o Problema

Dados nem sempre estão disponíveis, e essa realidade apresenta desafios significativos para cientistas de dados. Muitas tuplas carecem de valores registrados para diversos atributos, como a renda do cliente em dados de vendas, exigindo estratégias cuidadosas de tratamento.

Causas Principais de Dados Ausentes

- Mau funcionamento de equipamentos durante a coleta
- Dados inconsistentes com outros registros e posteriormente excluídos
- Informações não inseridas por mal-entendido ou falta de treinamento
- Certos dados não considerados importantes no momento da entrada
- Ausência de registro de histórico ou mudanças nos dados

Implicação crítica: Dados ausentes frequentemente precisam ser inferidos através de técnicas estatísticas ou de aprendizado de máquina para não comprometer a análise.



- **Referências:** Little (1988). A test of missing completely at random. JASA, 83(404). • Efron (1994). Missing data, imputation, and the bootstrap. JASA, 89(426).

Como Lidar com Dados Ausentes?

Existem múltiplas abordagens para tratar dados ausentes, cada uma com vantagens e desvantagens específicas. A escolha da técnica adequada depende da natureza dos dados, do percentual de valores ausentes e do contexto da análise.

01

Ignorar a Tupla

Geralmente feito quando o rótulo de classe está ausente em problemas de classificação. Não é efetivo quando a porcentagem de valores ausentes varia consideravelmente entre atributos.

03

Constante Global

Preencher com um valor como "desconhecido" ou criar uma nova classe. Simples, mas pode introduzir viés e criar uma categoria artificial.

05

Média por Classe

Usa a média do atributo para todas as amostras da mesma classe. Abordagem mais inteligente que considera a estrutura dos dados.

02

Preenchimento Manual

Tedioso e frequentemente inviável em conjuntos de dados grandes. Pode ser apropriado para pequenos conjuntos de dados críticos onde a precisão é fundamental.

04

Média do Atributo

Substitui valores ausentes pela média de todos os valores conhecidos do atributo. Mantém a média original, mas reduz a variância.

06

Valor Mais Provável

Baseado em inferência usando fórmulas Bayesianas ou árvores de decisão. Mais sofisticado e geralmente mais preciso.

- Referências:** Andridge & Little (2010). Hot deck imputation for survey non-response. *Int. Statistical Review*, 78(1). • Kenward & Carpenter (2007). Multiple imputation. *Statistical Methods in Medical Research*, 16(3).

Dados Ruidosos: Fontes e Características

Ruído representa erro aleatório ou variância em uma variável medida. Valores de atributos incorretos podem surgir de várias fontes, criando desafios significativos para a qualidade da análise de dados.

Principais Causas de Ruído

- Instrumentos de coleta de dados defeituosos ou mal calibrados
- Problemas durante a entrada manual de dados
- Erros durante a transmissão de dados entre sistemas
- Limitações tecnológicas dos sistemas de coleta
- Inconsistências em convenções de nomenclatura

A identificação e correção de dados ruidosos é crucial para garantir a confiabilidade dos modelos de machine learning e das análises estatísticas subsequentes.

Outros Problemas Relacionados

- Registros duplicados no banco de dados
- Dados incompletos ou parcialmente preenchidos
- Dados inconsistentes entre diferentes fontes
- Formatação inadequada ou não padronizada

 **Referência:** Zhu & Wu (2004). Class Noise vs. Attribute Noise: A Quantitative Study. Artificial Intelligence Review, 22(3).

Como Lidar com Dados Ruidosos?

Existem várias técnicas estabelecidas para identificar e tratar dados ruidosos, cada uma adequada para diferentes tipos de problemas e contextos de dados.

1

Suavização/Discretização

Primeiro ordene os dados e particione em bins de frequência igual. Depois suavize através de médias de bins, medianas de bins ou fronteiras de bins.

- Reduz variações locais mantendo tendências
- Útil para dados numéricos contínuos

2

Régressão

Suavize ajustando os dados em funções de regressão apropriadas que capturam tendências subjacentes.

- Linear, polinomial ou não-linear
- Preserva relações entre variáveis

3

Remoção de Outliers

Baseado em box-plot ou técnicas de clustering para detectar e remover valores anômalos.

- Box-plot identifica valores extremos
- Clustering isola pontos atípicos

4

Inspecção Combinada

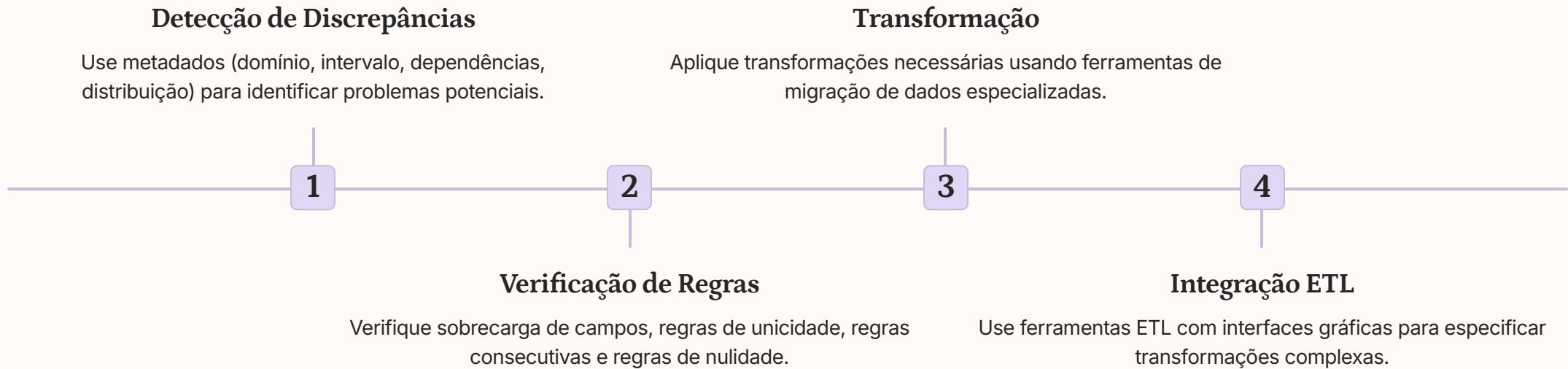
Detecção computacional de valores suspeitos seguida de verificação humana especializada.

- Combina automação com expertise
- Efetivo para casos complexos

- Referências:** Liu et al. (2002). Discretization: An enabling technique. Data Mining and Knowledge Discovery, 6(4). • Krzywinski & Altman (2015). Multiple linear regression. Nature Methods, 12(12).

Limpeza de Dados Como um Processo

A limpeza de dados não é uma tarefa pontual, mas um processo sistemático e iterativo que requer planejamento cuidadoso e ferramentas apropriadas. Este processo envolve várias etapas inter-relacionadas que garantem a qualidade dos dados.



Ferramentas e Tecnologias

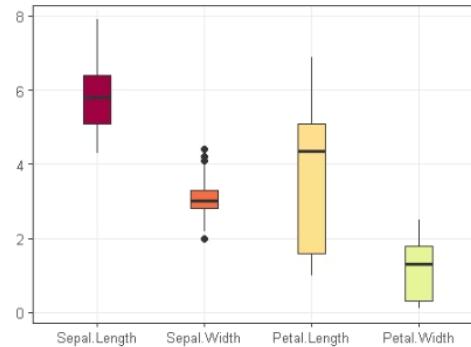
Ferramentas de Migração de Dados: Permitem que transformações sejam especificadas de forma declarativa, facilitando a conversão de formatos e estruturas.

Ferramentas ETL (Extraction/Transformation>Loading): Possibilitam que usuários especifiquem transformações através de interfaces gráficas intuitivas, reduzindo a complexidade técnica.

- Referências:** Chu et al. (2016). Data cleaning: Overview and emerging challenges. ACM SIGMOD. • Wang & Wang (2020). Time Series Data Cleaning: A Survey. IEEE Access, 8.

Remoção de Outliers Baseada em Box-Plot

O box-plot é uma ferramenta visual poderosa para identificar outliers em conjuntos de dados. Ele fornece uma representação gráfica da distribuição dos dados e destaca valores que se desviam significativamente do padrão central.



	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
16	5.7	4.4	1.5	0.4	setosa
33	5.2	4.1	1.5	0.1	setosa
34	5.5	4.2	1.4	0.2	setosa
61	5.0	2.0	3.5	1.0	versicolor

Componentes do Box-Plot

- **Q1 (1º Quartil):** 25% dos dados estão abaixo deste valor
- **Q2 (Mediana):** 50% dos dados estão abaixo deste valor
- **Q3 (3º Quartil):** 75% dos dados estão abaixo deste valor
- **IQR (Intervalo Interquartil):** $Q3 - Q1$, mede a dispersão central
- **Bigodes:** Estendem-se até $1.5 \times IQR$ além dos quartis

Esta técnica é robusta e amplamente utilizada porque é resistente a valores extremos e fornece uma visualização clara da distribuição dos dados.

Identificação de Outliers

Valores que caem além dos bigodes são considerados outliers:

- **Outliers inferiores:** Abaixo de $Q1 - 1.5 \times IQR$
- **Outliers superiores:** Acima de $Q3 + 1.5 \times IQR$

 **Referência:** McGill, Tukey & Larsen (1978). Variations of box plots. American Statistician, 32(1).

Integração de Dados: Unificando Múltiplas Fontes

A integração de dados combina informações de múltiplas fontes em um armazenamento coerente e unificado. Este processo é fundamental em ambientes empresariais modernos onde dados residem em sistemas diversos e heterogêneos.

Integração de Esquemas

Integre metadados de diferentes fontes, resolvendo diferenças semânticas.

- Exemplo: A.cust-id ≡ B.cust-#
- Mapeamento de estruturas diferentes
- Harmonização de tipos de dados

Identificação de Entidades

Identifique entidades do mundo real a partir de múltiplas fontes de dados.

- Exemplo: Bill Clinton = William Clinton
- Resolução de duplicatas
- Linkage de registros

Resolução de Conflitos

Detecte e resolva conflitos de valores de atributos para a mesma entidade.

- Representações diferentes
- Escalas distintas (métrico vs. imperial)
- Precisões variadas

A integração cuidadosa de dados de múltiplas fontes não apenas reduz redundâncias e inconsistências, mas também melhora significativamente a velocidade e qualidade da mineração de dados. Sistemas bem integrados permitem análises mais profundas e insights mais valiosos.

- **Referência:** Golshan et al. (2017). Data integration: After the teenage years. ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems.

Tratando Redundância na Integração

A redundância de dados ocorre frequentemente durante a integração de múltiplos bancos de dados e pode impactar negativamente a eficiência e qualidade da análise. Identificar e tratar redundâncias é crucial para otimizar o armazenamento e processamento.

Identificação de Objetos

O mesmo atributo ou objeto pode ter nomes diferentes em bancos de dados distintos, criando redundância implícita.

Integração Cuidadosa

Planejamento e execução cuidadosos reduzem redundâncias e melhoram a qualidade da mineração.



Dados Deriváveis

Um atributo pode ser derivado de outro, como receita anual calculada a partir de vendas mensais.

Análise de Correlação

Atributos redundantes podem ser detectados através de análise de correlação e covariância.

Benefícios da Redução de Redundância

- Economia de espaço de armazenamento
- Melhoria na velocidade de processamento
- Redução de inconsistências
- Maior qualidade dos resultados de mineração
- Simplificação da manutenção de dados

Análise de Correlação: Dados Nominais

Para dados categóricos ou nominais, a análise de correlação utiliza o teste qui-quadrado (χ^2) para determinar se dois atributos são independentes ou se existe uma relação estatisticamente significativa entre eles.

Teste Qui-Quadrado (χ^2)

Mede a discrepância entre frequências observadas e esperadas, assumindo independência entre variáveis categóricas.

$$\chi^2 = \sum \frac{(Observado - Esperado)^2}{Esperado}$$

Quanto maior o valor de χ^2 , maior a evidência contra a hipótese de independência.

Interpretação dos Resultados

Compare o valor calculado de χ^2 com valores críticos da distribuição qui-quadrado:

- Se $\chi^2 >$ valor crítico: rejeite a hipótese de independência
- Se $\chi^2 \leq$ valor crítico: não há evidência suficiente de dependência
- Considere os graus de liberdade: $(\text{linhas}-1) \times (\text{colunas}-1)$

Aplicações Práticas

Útil para identificar atributos redundantes em integração de dados e para feature selection em machine learning.

- Testar relações entre variáveis categóricas
- Detectar dependências ocultas em dados
- Validar suposições de independência em modelos



Referência: Larsen & Marx (2017). An Introduction to Mathematical Statistics and Its Applications. Pearson Education.

Cálculo do Qui-Quadrado: Um Exemplo

Vamos examinar um exemplo concreto para entender como o teste qui-quadrado funciona na prática. Considere a relação entre gostar de ficção científica e jogar xadrez.

Dados Observados

	Joga Xadrez	Não Joga
Gosta de FC	250	200
Não Gosta de FC	50	1000
Total	300	1200

Frequências Esperadas

Valores entre parênteses calculados assumindo independência:

	Joga	Não Joga
Gosta de FC	90	360
Não Gosta	210	840

01

Calcule as Frequências Esperadas

Para cada célula: $E = \frac{(total_linha) \times (total_coluna)}{total_geral}$

02

Aplique a Fórmula do χ^2

$\chi^2 = \sum \frac{(O-E)^2}{E}$ para todas as células

03

Compare com Valor Crítico

Use tabela χ^2 com $(2-1) \times (2-1) = 1$ grau de liberdade

04

Tire Conclusões

Determine se há evidência suficiente de dependência entre as variáveis

Neste exemplo, o alto valor de χ^2 indica forte evidência de que gostar de ficção científica e jogar xadrez não são independentes - existe uma associação significativa entre essas variáveis.

Análise de Correlação: Dados Numéricos

Para dados numéricos contínuos, utilizamos o coeficiente de correlação de Pearson para quantificar a força e direção da relação linear entre duas variáveis. Este é um dos métodos mais fundamentais em estatística descritiva.

Coeficiente de Correlação (r)

Mede a força e direção da relação linear entre duas variáveis numéricas.

$$r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2} \sqrt{\sum(y_i - \bar{y})^2}}$$

Interpretação do Valor

- $r = +1$: correlação positiva perfeita
- $r = 0$: sem correlação linear
- $r = -1$: correlação negativa perfeita
- $|r| > 0.7$: correlação forte
- $0.3 < |r| < 0.7$: correlação moderada
- $|r| < 0.3$: correlação fraca

Covariância

Relacionada à correlação, mas não normalizada:

$$Cov(X, Y) = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{n}$$

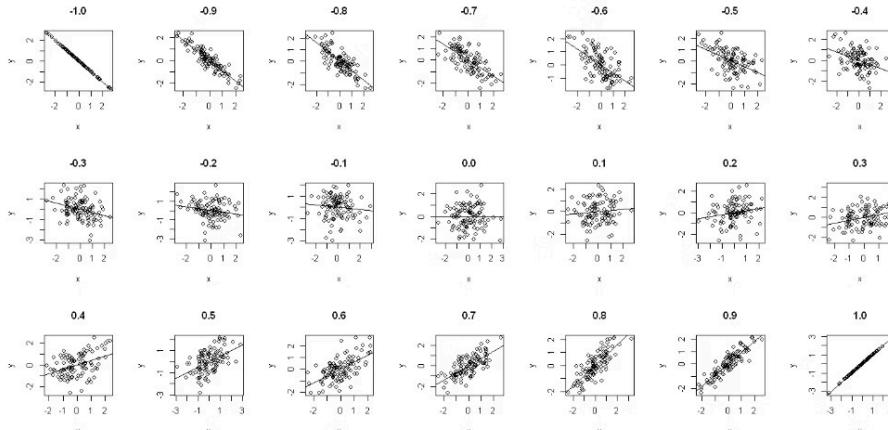
A correlação é a covariância normalizada pelos desvios padrão.

"Correlação não implica causalidade. Uma forte correlação indica associação, mas não necessariamente que uma variável causa mudanças na outra."

□ **Referência:** Larsen & Marx (2017). An Introduction to Mathematical Statistics and Its Applications. Pearson Education.

Avaliando Correlação Visualmente

Scatter plots (gráficos de dispersão) fornecem uma maneira intuitiva e poderosa de visualizar a relação entre duas variáveis numéricas. Eles permitem identificar rapidamente padrões, outliers e a natureza da correlação.



Os gráficos acima demonstram diferentes níveis de correlação, variando de -1 (correlação negativa perfeita) a +1 (correlação positiva perfeita).

Correlação Positiva Forte

Pontos formam uma linha ascendente clara. À medida que X aumenta, Y também aumenta de forma consistente.

Vantagens da Visualização

- Identifica relações não-lineares que o coeficiente de Pearson pode perder
- Revela outliers que podem distorcer a correlação
- Mostra heteroscedasticidade (variância não constante)
- Facilita comunicação de resultados para audiências não-técnicas

Sem Correlação

Pontos dispersos aleatoriamente sem padrão claro. Não há relação linear discernível entre as variáveis.

Melhores Práticas

- Sempre visualize antes de calcular correlações
- Combine análise visual com testes estatísticos
- Considere transformações para relações não-lineares
- Investigue outliers cuidadosamente antes de removê-los

A combinação de análise visual através de scatter plots e cálculo quantitativo do coeficiente de correlação fornece uma compreensão completa da relação entre variáveis, essencial para detecção de redundância e feature engineering.

Correlação Negativa Forte

Pontos formam uma linha descendente clara. À medida que X aumenta, Y diminui de forma consistente.



Pré-processamento de Dados II: Redução e Representação

Técnicas avançadas para transformar grandes volumes de dados em representações compactas e eficientes, mantendo a integridade analítica.

Estratégias de Redução de Dados

A redução de dados consiste em obter uma representação reduzida do conjunto de dados que seja muito menor em volume, mas que produza resultados analíticos idênticos ou quase idênticos aos originais.

Por que reduzir dados?

Um banco de dados ou data warehouse pode armazenar terabytes de informações. Análises complexas podem levar muito tempo para executar sobre o conjunto completo de dados, tornando a redução essencial para eficiência computacional.

Principais Estratégias

- Redução de dimensionalidade através de seleção de subconjuntos de características e criação de features
- Análise de Componentes Principais (PCA) para projeção em espaços menores
- Modelos de regressão para identificar relações essenciais
- Redução de numerosidade via amostragem, agregação e compressão

 **Referência:** M.H. ur Rehman et al., 2016, Big Data Reduction Methods: A Survey, *Data Science and Engineering*, v. 1, n. 4, p. 265–284.

Redução de Dimensionalidade

A Maldição da Dimensionalidade

Quando a dimensionalidade aumenta, os dados tornam-se cada vez mais esparsos. A densidade e a distância entre pontos perdem significado, o que é crítico para clustering e análise de outliers. As combinações possíveis de subespaços crescem exponencialmente.

01

Análise de Componentes Principais

Projeção que captura a maior variação nos dados

03

Criação de Features

Construção de novos atributos informativos

Benefícios da Redução

- Evita a maldição da dimensionalidade
- Elimina características irrelevantes e reduz ruído
- Diminui tempo e espaço necessários na mineração
- Permite visualização mais fácil dos dados

02

Técnicas Supervisionadas

Seleção de características com feedback

04

Modelos de Regressão

Identificação de relações lineares essenciais

Referência: M. Dash, H. Liu, e J. Yao, 1997, Dimensionality reduction of unsupervised data, Proceedings of the International Conference on Tools with Artificial Intelligence.

Análise de Componentes Principais (PCA)

O PCA encontra uma projeção que captura a maior quantidade de variação nos dados. Os dados originais são projetados em um espaço muito menor, resultando em redução de dimensionalidade através dos autovetores da matriz de covariância.

Dados Originais

Conjunto de dados em espaço de alta dimensão com múltiplas variáveis correlacionadas.

Transformação PCA

Identificação dos eixos de maior variância através dos componentes principais.

Projeção Reduzida

Representação compacta mantendo a informação essencial dos dados.

Os autovetores da matriz de covariância definem o novo espaço de características, ordenados por sua capacidade de explicar a variância total dos dados.

- **Referência:** H. Abdi and L.J. Williams, 2010, Principal component analysis, Wiley Interdisciplinary Reviews: Computational Statistics, v. 2, n. 4, p. 433–459.

Passos da Análise de Componentes Principais

Dados n vetores de dados de n dimensões, encontre $k \leq n$ vetores ortogonais (componentes principais) que podem ser melhor usados para representar os dados.



Normalização dos Dados

Cada atributo é ajustado para cair dentro do mesmo intervalo, garantindo que nenhuma variável domine a análise devido apenas à sua escala.



Computação dos Componentes

Calcular k vetores ortonormais (unitários) - os componentes principais. Cada vetor de entrada é uma combinação linear desses k vetores.



Ordenação por Importância

Os componentes principais são ordenados por "significância" decrescente ou força, medida pela variância que explicam.



Redução Dimensional

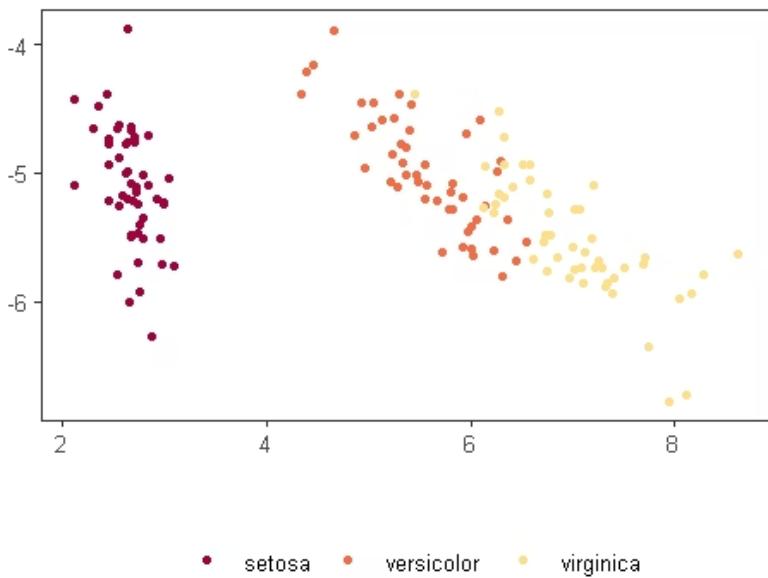
O tamanho dos dados pode ser reduzido eliminando componentes fracos (baixa variância). Usando os componentes mais fortes, é possível reconstruir uma boa aproximação dos dados originais.

Importante: O PCA funciona **apenas para dados numéricos**. Variáveis categóricas devem ser tratadas separadamente ou codificadas apropriadamente.

Referência: H. Abdi and L.J. Williams, 2010, Principal component analysis, Wiley Interdisciplinary Reviews: Computational Statistics, v. 2, n. 4, p. 433–459.

Projeção com PCA

A transformação PCA projeta dados de alta dimensão em um subespaço de menor dimensão, preservando a estrutura de variância máxima.



Dados Originais

Sepal.Length	Sepal.Width	Petal.Length	Petal.Width
5.1	3.5	1.4	0.2
4.9	3.0	1.4	0.2
4.7	3.2	1.3	0.2
4.6	3.1	1.5	0.2
5.0	3.6	1.4	0.2
5.4	3.9	1.7	0.4

Conjunto de dados bidimensional mostrando distribuição original com correlação entre variáveis.

Transformação PCA

	PC1	PC2
Sepal.Length	0.5210659	-0.37741762
Sepal.Width	-0.2693474	-0.92329566
Petal.Length	0.5804131	-0.02449161
Petal.Width	0.5648565	-0.06694199

Identificação dos eixos principais (PC1 e PC2) que capturam a máxima variância.

Dados Projetados

PC1	PC2	Species
2.640270	-5.204041	setosa
2.670730	-4.666910	setosa
2.454606	-4.773636	setosa
2.545517	-4.648463	setosa
2.561228	-5.258629	setosa
2.975946	-5.707321	setosa

Representação final no novo sistema de coordenadas definido pelos componentes principais.

A visualização acima demonstra como o PCA rotaciona o espaço de características para alinhar com as direções de máxima variância, facilitando a redução dimensional sem perda significativa de informação.

- ❑ **Referências:** A. Tharwat et al., 2017, Linear discriminant analysis: A detailed tutorial, AI Communications, v. 30, n. 2. | R-bloggers: Computing and Visualizing LDA

Seleção de Subconjunto de Atributos

Outra abordagem poderosa para reduzir a dimensionalidade dos dados, focando na identificação e remoção de atributos desnecessários.



Atributos Redundantes

Duplicam muito ou toda a informação contida em um ou mais outros atributos. Por exemplo, o preço de compra de um produto e o valor do imposto sobre vendas pago contêm informação redundante.



Atributos Irrelevantes

Não contêm informação útil para a tarefa de mineração de dados em questão. Por exemplo, o ID dos estudantes é frequentemente irrelevante para prever o GPA dos alunos.

Impacto da Remoção

- Redução do tempo de treinamento
- Melhoria na generalização do modelo
- Diminuição do overfitting
- Facilita interpretação dos resultados

Desafios

- Identificação automática de redundância
- Avaliação de relevância contextual
- Interação entre atributos
- Balanceamento entre redução e perda de informação

- Referências:** L. Carlos Molina et al., 2002, Feature selection algorithms: A survey and experimental evaluation, Proceedings - IEEE ICDM | M.A. Hall and G. Holmes, 2003, Benchmarking Attribute Selection Techniques, IEEE TKDE, v. 15, n. 6.

Busca Heurística na Seleção de Atributos

Estratégias de busca heurística são essenciais para navegar eficientemente pelo vasto espaço de possíveis subconjuntos de atributos, encontrando soluções de alta qualidade sem avaliar todas as combinações possíveis.



Seleção Forward

Inicia com conjunto vazio e adiciona atributos iterativamente, selecionando aquele que mais melhora o desempenho.

Eliminação Backward

Começa com todos os atributos e remove iterativamente aqueles que menos contribuem para o desempenho.

Busca Bidirecional

Combina estratégias forward e backward para explorar o espaço de busca de forma mais eficiente.

A escolha do método de busca depende do tamanho do conjunto de dados, do número de atributos e dos recursos computacionais disponíveis.

Critérios de Avaliação

- Acurácia do modelo resultante
- Medidas de correlação e informação mútua
- Consistência dos dados
- Distância e separabilidade entre classes

Considerações Práticas

- Complexidade computacional
- Estabilidade da seleção
- Interpretabilidade dos resultados
- Robustez a ruído e outliers

Referências: I. Kononenko and S.J. Hong, 1997, Attribute selection for modelling, Future Generation Computer Systems, v. 13, n. 2–3 | M. Dash and H. Liu, 1997, Feature selection for classification, Intelligent Data Analysis, v. 1, n. 3.

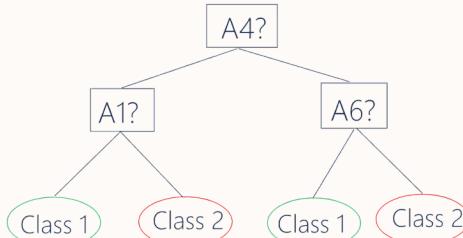
Exemplo de Indução de Árvore de Decisão para Seleção de Features

As árvores de decisão oferecem um método natural e interpretável para seleção de atributos, identificando automaticamente as características mais discriminativas através do processo de construção da árvore.

Conjunto Inicial de Atributos

{A1, A2, A3, A4, A5, A6}

Seis atributos candidatos para modelagem preditiva. O algoritmo avaliará qual subconjunto é mais informativo.



Árvore de Decisão Resultante

Durante a construção da árvore, o algoritmo seleciona automaticamente os atributos que melhor dividem os dados em cada nó, baseando-se em métricas como ganho de informação ou índice Gini.

Neste exemplo, apenas três atributos foram selecionados:

- **A4:** Atributo raiz (maior poder discriminativo)
- **A1:** Selecionado em subárvore esquerda
- **A6:** Selecionado em subárvore direita

Conjunto Reduzido: {A1, A4, A6}

Vantagens da Abordagem

- Seleção automática baseada em importância
- Captura interações entre atributos
- Interpretabilidade imediata dos resultados
- Não requer suposições sobre distribuição dos dados

Redução Alcançada

Redução de **50%** na dimensionalidade (de 6 para 3 atributos), eliminando A2, A3 e A5 que não contribuíram significativamente para a classificação.

Referência: J.R. Quinlan, 1996, Bagging, boosting, and C4.5, Proceedings of the National Conference on Artificial Intelligence, p. 725–730.

Geração de Features

Criar novos atributos que capturam informações importantes no conjunto de dados de forma mais eficaz do que os atributos originais, aumentando o poder preditivo dos modelos.

Extração de Atributos

Técnicas específicas do domínio para extrair características relevantes dos dados brutos. Por exemplo, extraer características de imagens, texto ou sinais temporais.



Mapeamento para Novo Espaço

Transformação dos dados para um espaço diferente onde padrões se tornam mais evidentes. Inclui Transformada de Fourier para análise de frequência e Transformada Wavelet para análise multi-resolução.



Construção de Atributos

Criação de novos atributos através de operações sobre os existentes, incluindo discretização de variáveis contínuas e combinações matemáticas de features.

Transformada de Fourier

Decompõe sinais temporais em componentes de frequência, revelando padrões periódicos ocultos nos dados originais. Especialmente útil para análise de séries temporais.

A geração eficaz de features requer profundo conhecimento do domínio e criatividade para identificar representações que amplificam sinais relevantes e suprimem ruído.

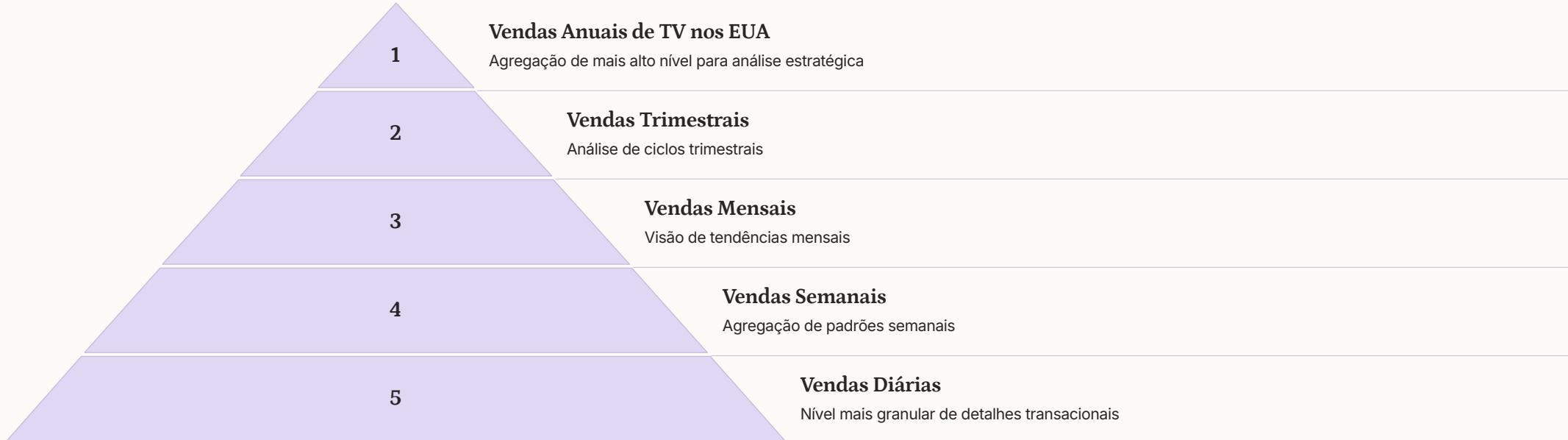
Transformada Wavelet

Oferece análise tempo-frequência com resolução variável, capturando tanto características locais quanto globais. Ideal para sinais não-estacionários.

- Referências:** S. Markovitch and D. Rosenstein, 2002, Feature generation using general constructor functions, Machine Learning, v. 49, n. 1 | I. Daubechies, 1990, The Wavelet Transform, IEEE Transactions on Information Theory, v. 36, n. 5.

Agregação de Dados

A agregação envolve sumarização e construção de cubos de dados, permitindo análises em diferentes níveis de granularidade e reduzindo substancialmente o volume de dados a ser processado.



Benefícios da Agregação

- Redução drástica no volume de dados
- Análise mais rápida em dados sumarizados
- Identificação de tendências de longo prazo
- Suporte a diferentes níveis de decisão

Operações Comuns

- SUM, AVG para métricas numéricas
- COUNT para frequências
- MIN, MAX para limites
- Agregações compostas e hierárquicas

Referência: V. Harinarayan, A. Rajaraman, and J.D. Ullman, 1996, Implementing Data Cubes Efficiently, SIGMOD Record, v. 25, n. 2, p. 205–216.

Transformação de Dados

Uma função que mapeia todo o conjunto de valores de um determinado atributo para um novo conjunto de valores substitutos, onde cada valor antigo pode ser identificado com um dos novos valores.



Construção de Atributos/Features

Novos atributos construídos a partir dos existentes através de agregações complexas e combinações matemáticas que revelam relações latentes.



Hierarquia de Conceitos

Navegação em taxonomias conceituais, movendo dados para níveis mais abstratos (por exemplo, de cidade para estado para país).



Normalização

Escalonamento dos valores para cair dentro de um intervalo menor e especificado (por exemplo, 0 a 1), eliminando viés de escala entre variáveis.



Discretização / Suavização

Conversão de variáveis contínuas em categóricas ou redução de ruído através de técnicas de suavização como médias móveis.



Mapeamento Categórico

Transformação de categorias em representações numéricas ou codificações apropriadas para algoritmos de aprendizado de máquina.

A escolha da transformação apropriada depende das características dos dados, dos requisitos do algoritmo e dos objetivos da análise.

- Referência: S. Ramírez-Gallego et al., 2017, A survey on data preprocessing for data stream mining: Current status and future directions, Neurocomputing, v. 239, p. 39–57.

Técnicas de Normalização

A normalização é crucial para garantir que diferentes atributos contribuam de forma equilibrada para a análise, eliminando vieses causados por diferenças de escala e unidades de medida.

1	Min-Max (Escalonamento) Transforma os dados para um intervalo específico, tipicamente [0, 1]. Fórmula: $v' = (v - \min) / (\max - \min)$ Preserva as relações entre valores originais e é ideal quando os limites são bem definidos.
2	Z-Score (Padronização) Transforma os dados para ter média 0 e desvio padrão 1. Fórmula: $v' = (v - \mu) / \sigma$ Útil quando os dados seguem distribuição normal e quando outliers são importantes.
3	Normalização Decimal Move o ponto decimal para normalizar valores. Fórmula: $v' = v / 10^j$, onde j é o menor inteiro tal que $\max(v') < 1$ Simples mas raramente usada em prática moderna.

Quando Usar Min-Max

- Dados com limites conhecidos
- Preservação de relações exatas
- Redes neurais e imagens
- Sem outliers severos

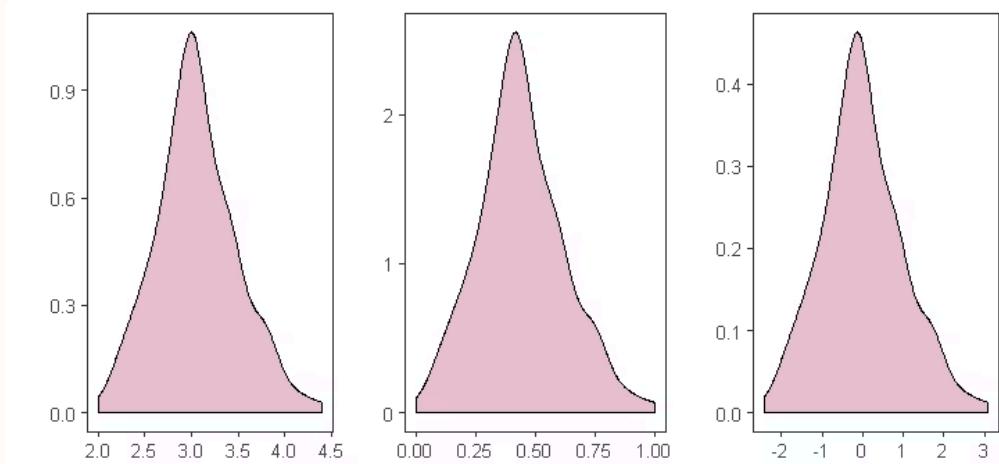
Quando Usar Z-Score

- Dados com distribuição normal
- Presença de outliers
- Limites desconhecidos
- Análise estatística

 **Referência:** E. Ogasawara et al., 2009, Neural networks cartridges for data mining on time series, Proceedings of the International Joint Conference on Neural Networks, p. 2302–2309.

Comparação Visual de Normalização

Visualização das diferentes técnicas de normalização aplicadas ao mesmo conjunto de dados, demonstrando como cada método transforma a distribuição original.



Dados Originais

Distribuição inicial com escala e dispersão natural dos valores brutos, antes de qualquer transformação.

Min-Max [0-1]

Dados comprimidos no intervalo $[0, 1]$, preservando a forma da distribuição mas alterando a escala absoluta.

Z-Score / $N(0,1)$

Dados centralizados em zero com desvio padrão unitário, criando distribuição normal padronizada.

Escolha a técnica de normalização baseada nas características dos seus dados e nos requisitos do algoritmo de aprendizado de máquina que será utilizado.

Impacto no Modelo

Algoritmos baseados em distância (K-NN, SVM) são altamente sensíveis à escala e beneficiam-se significativamente da normalização.

Preservação de Informação

Ambas as técnicas preservam as relações relativas entre os pontos de dados, apenas alterando a escala de representação.

Aplicação Prática

Em ambientes de produção, é essencial armazenar os parâmetros de normalização (\min , \max , μ , σ) para aplicar a mesma transformação a novos dados.



Pré-processamento de Dados

III: Discretização, Suavização e Amostragem

Uma exploração abrangente das técnicas fundamentais para transformar dados contínuos em formatos discretos, reduzir ruído através de métodos de suavização e selecionar amostras representativas de grandes conjuntos de dados.

Discretização e Suavização

A discretização é o processo de transferir funções contínuas, modelos, variáveis e equações para suas contrapartes discretas. Este processo é essencial quando precisamos trabalhar com algoritmos que requerem dados categóricos ou quando queremos reduzir a complexidade de dados numéricos contínuos.

A suavização é uma técnica que cria uma função de aproximação que tenta capturar padrões importantes nos dados enquanto elimina ruído ou outras estruturas de escala fina e fenômenos rápidos. Esta abordagem é crucial para revelar tendências subjacentes em dados ruidosos.

Uma parte importante da discretização e suavização é a configuração de intervalos (bins) para proceder com a aproximação. A escolha adequada dos bins pode impactar significativamente a qualidade dos resultados obtidos.

- **Referências importantes:** Kerber (1992) desenvolveu o método ChiMerge para discretização de atributos numéricos. Liu et al. (2002) demonstraram que a discretização é uma técnica habilitadora fundamental em mineração de dados.

Métodos de Binning para Suavização de Dados

Particionamento de Largura Igual

Divide o intervalo em N intervalos de tamanho igual, criando uma grade uniforme. Se A e B são os valores mais baixos e mais altos do atributo, a largura dos intervalos será: $W = (B - A)/N$.

- Abordagem mais direta e simples de implementar
- Outliers podem dominar a apresentação dos dados
- Dados assimétricos não são tratados adequadamente
- Funciona bem para distribuições uniformes

Particionamento de Profundidade Igual

Divide o intervalo em N intervalos, cada um contendo aproximadamente o mesmo número de amostras. Esta técnica garante uma distribuição mais equilibrada dos dados entre os bins.

- Excelente escalabilidade de dados
- Cada bin tem representatividade similar
- Gerenciar atributos categóricos pode ser desafiador
- Mais robusto para dados assimétricos

- Krzywinski (2016) destaca a importância do binning adequado em dados de alta resolução, especialmente em visualizações científicas onde a escolha incorreta pode obscurecer padrões importantes.

Comparação Prática dos Métodos de Binning

Para ilustrar as diferenças entre os métodos de binning, considere dados ordenados de preço (em dólares): **4, 8, 9, 15, 21, 21, 24, 25, 26, 28, 29, 34**

Particionamento de Largura Igual (3 bins)

Cálculo: $(34-4)/3 = 10$ unidades por bin

- **Bin 1 [4-13[:** 4, 8, 9
- **Bin 2 [14-23[:** 15, 21, 21
- **Bin 3 [23-34]:** 24, 25, 26, 28, 29, 34

Particionamento de Frequência Igual (3 bins)

4 valores por bin ($12 \text{ valores} \div 3 \text{ bins}$)

- **Bin 1:** 4, 8, 9, 15
- **Bin 2:** 21, 21, 24, 25
- **Bin 3:** 26, 28, 29, 34

Técnicas de Suavização Usando Frequência Igual

Suavização por Médias dos Bins

Cada valor é substituído pela média do seu bin:

- **Bin 1:** 9, 9, 9, 9 (média = 9)
- **Bin 2:** 23, 23, 23, 23 (média = 23)
- **Bin 3:** 29, 29, 29, 29 (média = 29)

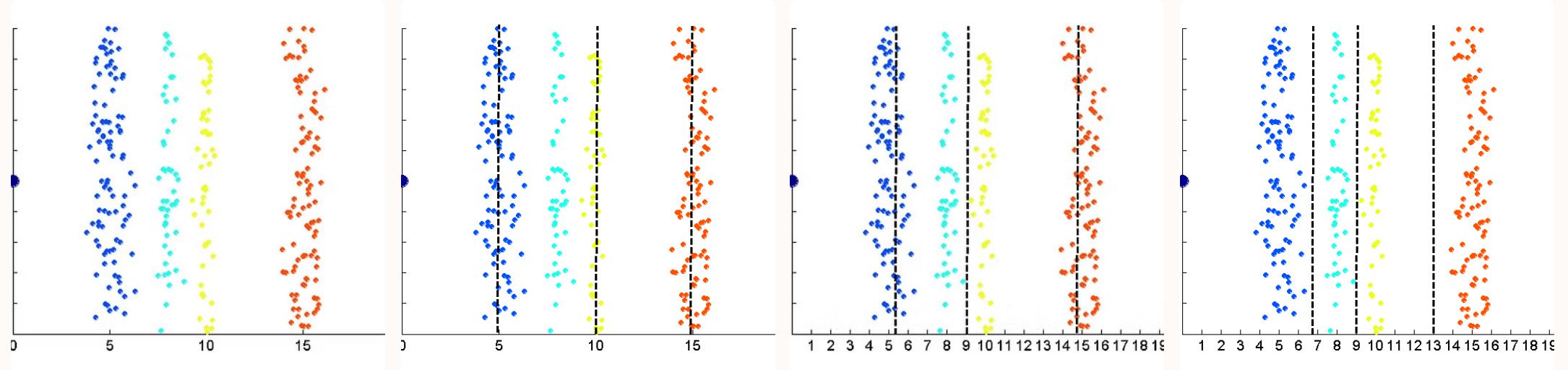
Suavização por Limites dos Bins

Cada valor é substituído pelo limite mais próximo:

- **Bin 1:** 4, 4, 4, 15
- **Bin 2:** 21, 21, 25, 25
- **Bin 3:** 26, 26, 26, 34

Influência do Binning nas Técnicas de Discretização

A escolha do número de bins e do método de binning tem impacto significativo no resultado da suavização. Observe como diferentes abordagens transformam os mesmos dados originais:



01

Dados Originais

Distribuição bruta mostrando todos os pontos de dados sem processamento

03

Binning por Frequência Igual

Suavização garantindo número similar de pontos por bin

02

Binning por Largura Igual

Suavização usando 4 bins de largura uniforme

04

Agrupamento K-means

Discretização baseada em clustering, agrupando pontos similares

Geração de Hierarquia de Conceitos

A hierarquia de conceitos organiza conceitos de forma hierárquica, geralmente associada a cada dimensão em um data warehouse. Essas hierarquias facilitam operações de drill-down e roll-up em data warehouses, permitindo visualizar dados em múltiplas granularidades.

Especificação por Especialistas

Hierarquias de conceitos podem ser especificadas por especialistas do domínio com base na semântica do domínio:

- Maior compreensão contextual
- Relacionamentos mais significativos
- Alinhamento com práticas de negócio
- Requer conhecimento especializado

Formação Automática

Hierarquias podem ser formadas automaticamente usando métodos computacionais:

- **Dados numéricos:** usando métodos de discretização demonstrados anteriormente
- **Dados categóricos:** através de agrupamento e análise de frequência
- Menos semântica, mais baseado em padrões estatísticos

Exemplos de Hierarquias de Conceitos



Hierarquia Geográfica



cidade → estado → país

Exemplo: São Paulo → SP → Brasil



Hierarquia Temporal



semana → mês → ano

Exemplo: Semana 12 → Março → 2024

Agrupamento Explícito de Dados

A especificação de uma hierarquia para um conjunto de valores pode ser feita através de agrupamento explícito de dados. Por exemplo, ao definir **{cidade, estado, país}**, estabelecemos relacionamentos claros entre diferentes níveis de granularidade geográfica.

Este tipo de estruturação permite análises mais ricas e navegação intuitiva através dos dados, desde o nível mais detalhado até visões mais agregadas. É particularmente útil em sistemas OLAP (Online Analytical Processing) onde usuários precisam alternar entre diferentes perspectivas dos dados.

Mapeamento Categórico

O mapeamento categórico é uma técnica fundamental para criar hierarquias de conceitos a partir de atributos categóricos. Este processo envolve o agrupamento sistemático de valores relacionados em categorias de nível superior.

Através do mapeamento categórico, podemos transformar dados granulares em representações mais abstratas, facilitando análises em diferentes níveis de detalhe. Esta abordagem é especialmente valiosa quando trabalhamos com grandes volumes de categorias distintas que podem ser naturalmente agrupadas.

01

Identificação de Categorias

Analizar os valores únicos do atributo categórico

02

Definição de Agrupamentos

Estabelecer categorias de nível superior baseadas em similaridade ou semântica

03

Criação do Mapeamento

Associar cada valor original à sua categoria correspondente

04

Validação da Hierarquia

Verificar se o mapeamento faz sentido no contexto do domínio

Exemplo de Mapeamento Categórico

Um exemplo prático ilustra como diferentes categorias podem ser organizadas em uma estrutura hierárquica. Observe como valores específicos são mapeados para categorias mais gerais:

▲	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
1	5.1	3.5	1.4	0.2	setosa
2	4.9	3.0	1.4	0.2	setosa
3	4.7	3.2	1.3	0.2	setosa
4	4.6	3.1	1.5	0.2	setosa
5	5.0	3.6	1.4	0.2	setosa

▲	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Speciessetosa	Speciesversicolor	Speciesvirginica
1	5.1	3.5	1.4	0.2	1	0	0
2	4.9	3.0	1.4	0.2	1	0	0
3	4.7	3.2	1.3	0.2	1	0	0
4	4.6	3.1	1.5	0.2	1	0	0
5	5.0	3.6	1.4	0.2	1	0	0

Este exemplo demonstra como valores detalhados no nível mais baixo da hierarquia são sistematicamente agrupados em categorias intermediárias, que por sua vez podem ser agrupadas em categorias ainda mais gerais. Esta estruturação multinível permite flexibilidade nas análises.

- A qualidade do mapeamento categórico depende fortemente do conhecimento do domínio. Mapas bem construídos preservam o significado semântico enquanto reduzem a complexidade dos dados.

Amostragem em Mineração de Dados

Amostragem é o processo de obter uma pequena amostra para representar todo o conjunto de dados. Esta técnica permite que algoritmos de mineração executem com complexidade potencialmente sub-linear ao tamanho dos dados, tornando análises viáveis em grandes volumes.

Princípio Fundamental

Escolher um subconjunto representativo dos dados que mantenha as características estatísticas essenciais do conjunto completo

Desafios com Dados Assimétricos

Amostragem aleatória simples pode ter desempenho muito ruim na presença de assimetria (skew) nos dados, requerendo métodos adaptativos

Métodos Adaptativos

Desenvolvimento de métodos de amostragem adaptativa, como amostragem estratificada, que consideram a distribuição dos dados

- ❑ **Observação importante:** A amostragem pode não reduzir operações de I/O em banco de dados, pois o acesso é feito página por página. A eficiência está principalmente no processamento algorítmico subsequente.

Tipos de Amostragem



Amostragem Aleatória Simples

Probabilidade igual de selecionar qualquer item do conjunto de dados. Cada elemento tem a mesma chance de ser escolhido.



Amostragem Sem Reposição

Uma vez que um objeto é selecionado, ele é removido da população. Garante que nenhum elemento seja selecionado mais de uma vez.



Amostragem Com Reposição

Um objeto selecionado não é removido da população. O mesmo elemento pode aparecer múltiplas vezes na amostra.



Amostragem Estratificada

Particiona o conjunto de dados e extrai amostras de cada partição proporcionalmente. Usada em conjunto com dados assimétricos para garantir representatividade de todos os grupos.

Amostragem: Com ou Sem Reposição

SRSWOR

(Amostragem Aleatória Simples Sem Reposição)

Neste método, cada elemento selecionado é removido do pool de candidatos. A população diminui a cada seleção, alterando as probabilidades subsequentes.

Vantagens:

- Garante diversidade na amostra
- Evita duplicações
- Melhor para amostras pequenas

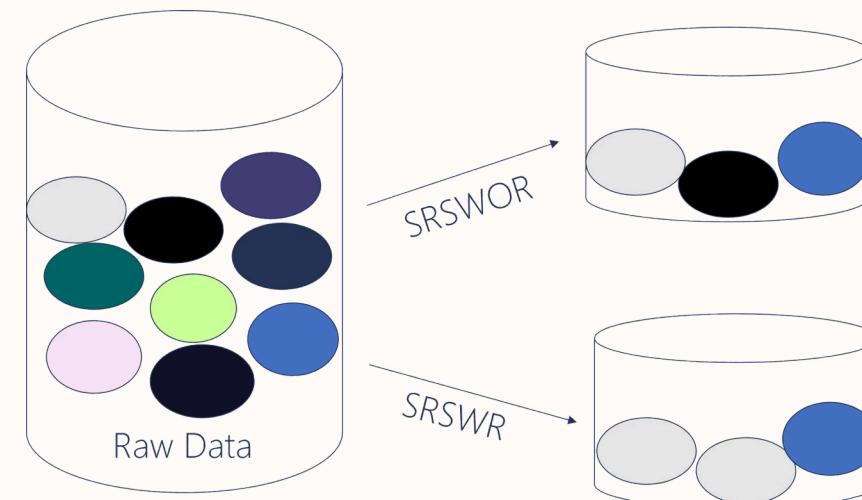
SRSWR

(Amostragem Aleatória Simples Com Reposição)

Os elementos permanecem no pool após seleção, mantendo probabilidades constantes. Cada seleção é independente das anteriores.

Vantagens:

- Simplicidade matemática
- Independência estatística
- Útil para bootstrap



Amostragem por Cluster e Estratificada

Técnicas avançadas de amostragem que consideram a estrutura natural dos dados para produzir amostras mais representativas, especialmente em conjuntos de dados complexos e heterogêneos.

Amostragem por Cluster

Divide a população em grupos (clusters) e seleciona clusters inteiros aleatoriamente. Todos os elementos dos clusters selecionados são incluídos na amostra.

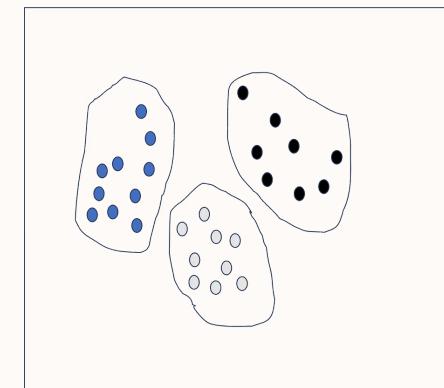
- Eficiente quando há agrupamentos naturais
- Reduz custos de coleta de dados
- Pode ter maior variância que outros métodos

Amostragem Estratificada

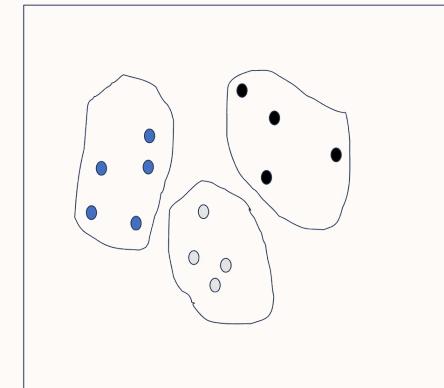
Divide a população em estratos homogêneos e amostra de cada estrato proporcionalmente à sua representação na população total.

- Garante representação de todos os subgrupos
- Reduz variância da estimativa
- Ideal para populações heterogêneas
- Requer conhecimento prévio da estrutura

Raw Data



Cluster/Stratified Sample



Exemplos Práticos de Amostragem

Visualização de diferentes abordagens de amostragem aplicadas ao mesmo conjunto de dados, demonstrando como cada método seleciona elementos de maneira distinta:

2/3

	setosa	versicolor	virginica
dataset	50	50	50
random sample	42	41	37
stratified sample	40	40	40

1/3

	setosa	versicolor	virginica
random sample	8	11	11
stratified sample	10	10	10

2/3

Taxa de Amostragem Maior

Captura aproximadamente dois terços da população original

1/3

Taxa de Amostragem Menor

Seleciona cerca de um terço dos dados disponíveis

A escolha da taxa de amostragem depende de múltiplos fatores: recursos computacionais disponíveis, variabilidade dos dados, precisão requerida e tempo de processamento. Amostras maiores geralmente produzem estimativas mais precisas, mas requerem mais recursos.

Balanceamento de Conjuntos de Dados

O Problema do Desbalanceamento de Classes

O problema de desbalanceamento de classes ocorre quando há exemplos positivos raros, mas numerosos exemplos negativos. Situações comuns incluem diagnóstico médico, detecção de fraudes, vazamento de óleo e detecção de falhas.

Métodos tradicionais assumem uma distribuição balanceada de classes e custos de erro iguais, tornando-os inadequados para dados desbalanceados. Esta realidade exige abordagens especializadas de pré-processamento.

Oversampling (Sobreamostragem)

Reamostragem de dados da classe positiva (minoritária), criando cópias ou gerando novos exemplos sintéticos para aumentar sua representação.

- Aumenta o tamanho do conjunto de dados
- Pode levar a overfitting se mal aplicado
- Técnicas como SMOTE geram exemplos sintéticos

Undersampling (Subamostragem)

Eliminação aleatória de tuplas da classe negativa (majoritária) para equilibrar a proporção entre as classes.

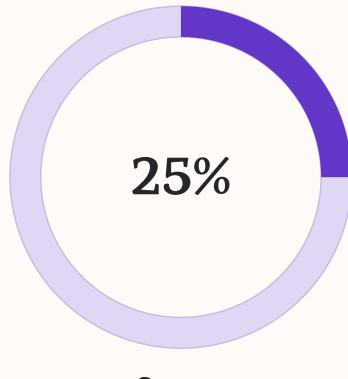
- Reduz o tamanho do conjunto de dados
- Pode perder informações importantes
- Mais rápido para treinar modelos
- Risco de perder padrões relevantes

- Referências:** Zhou e Liu (2006) desenvolveram métodos para treinar redes neurais sensíveis ao custo. Tharwat e Schenck (2020) propuseram abordagens de aprendizado ativo para dados desbalanceados.

Oversampling e Undersampling na Prática

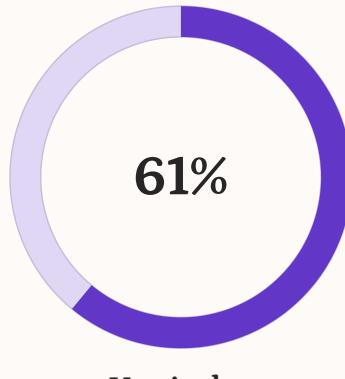
Cenário: Dataset Iris Desbalanceado

Considere que o conjunto de dados Iris apresentasse a seguinte distribuição desbalanceada:



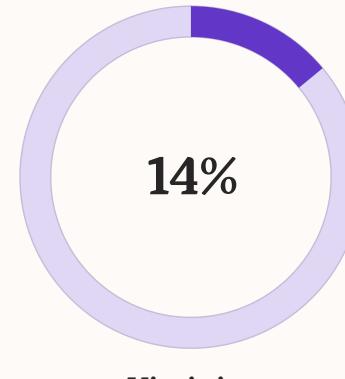
Setosa

20 observações



Versicolor

50 observações



Virginica

11 observações

Como as Técnicas Abordam o Desbalanceamento?

Abordagem de Oversampling

Aumentaria as classes minoritárias (setosa e virginica) através de:

- Replicação de exemplos existentes
- Geração de exemplos sintéticos
- Interpolação entre pontos existentes
- Resultado: ~50 observações de cada classe

Abordagem de Undersampling

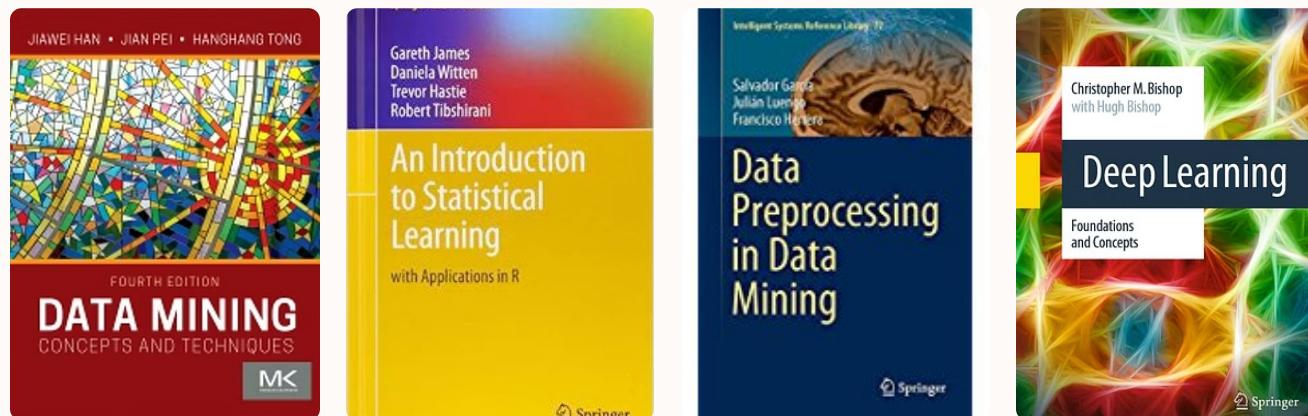
Reduziria a classe majoritária (versicolor) através de:

- Seleção aleatória de subconjunto
- Remoção de exemplos redundantes
- Preservação de exemplos informativos
- Resultado: ~15-20 observações de cada classe

A escolha entre oversampling e undersampling depende do tamanho inicial do dataset, recursos computacionais disponíveis e da importância de preservar todos os padrões presentes nos dados originais.

Referências Principais

Esta seleção de referências representa os pilares fundamentais para o estudo aprofundado de mineração de dados, cobrindo desde conceitos básicos até técnicas avançadas e aplicações contemporâneas.



1. **J. Han, J. Pei, and H. Tong**, *Data Mining: Concepts and Techniques*, 4th edition. Cambridge, MA: Morgan Kaufmann, 2022.
2. **G. M. James, D. Witten, T. Hastie, and R. Tibshirani**, *An Introduction to Statistical Learning: With Applications in R*. Springer Nature, 2021.
3. **S. Garcia, J. Luengo, and F. Herrera**, *Data Preprocessing in Data Mining*. Springer, 2014.
4. **C. M. Bishop and H. Bishop**, *Deep Learning: Foundations and Concepts*. Springer Nature, 2023.