

Mineração de Dados

AULA 1 Introdução

Sobre o Curso

Sobre mim

**Qual o propósito deste
curso?**

Como serão as aulas?

Sobre o Curso

Slides, Artigos, Materiais...



O que vem na sua cabeça quando falamos em “mineração de dados”?



O que é mineração de dados?

- corresponde à extração de **minerais valiosos**, como ouro e pedras preciosas, a partir de uma mina.
- embora não possam ser cultivados ou produzidos artificialmente, **existem de maneira implícita e muitas vezes desconhecida** em alguma fonte, podendo ser extraídos.
- requer **acesso à mina**, o uso de **ferramentas adequadas** de mineração, a **extração** dos minérios propriamente dita e o seu posterior **preparo** para comercialização.



O que é mineração de dados?

Dados

- 1000 milibares
- 5,1 m/s; 95°
- 30 °C
- poucas
- 1000 mts

Informação

- Pressão atmosférica = 1000 milibares
- Velocidade e direção do vento = 5,1 m/s; 95°
- Temperatura do ar = 30 °C
- Nuvens = poucas
- Visibilidade = 1000 mts

Conhecimento

- A probabilidade de chuva é baixa, portanto, posso ir à praia.

Etapas da Descoberta do Conhecimento

Mineração: Etapa Central do Processo de Descoberta de Conhecimento

Pré-processamento

Transformação

Limpeza

Seleção

Banco de Dados

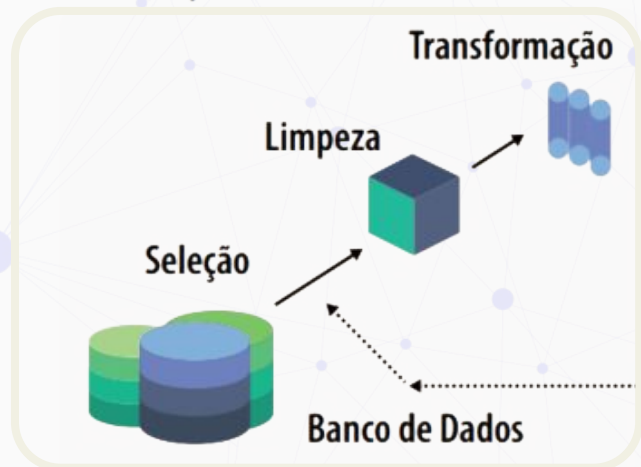
Mineração

Análise do Resultado

Conhecimento

Informação

Dado



Etapas da Descoberta do Conhecimento



Principais tarefas da mineração de dados

Análise descritiva de dados

Etapa inicial e essencial da mineração de dados.

Envolve sumarização, exploração e descrição dos dados.

Não requer algoritmos complexos de aprendizado de máquina.



Principais Técnicas Utilizadas

- Frequência: contagem de ocorrências de valores
- Tendência central: média, mediana, moda
- Dispersão: desvio padrão, variância, amplitude
- Posição relativa: quartis, percentis
- Associação: correlações entre variáveis

Principais tarefas da mineração de dados

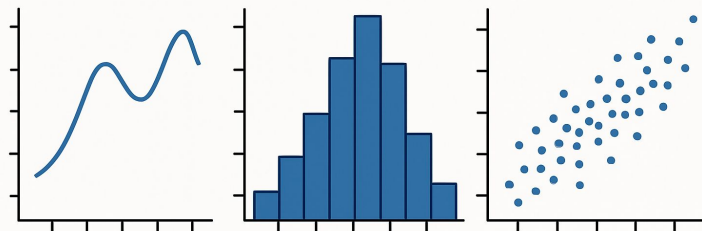
Análise descritiva de dados

Visualização de dados

- Gráficos de linha, histogramas e scatterplots
- Facilitam a interpretação visual das variáveis

Aplicações

- Análise de distribuição orçamentária por área
- Comparação de públicos por região ou tipo de evento
- Identificação de padrões de desigualdade ou concentração



Principais tarefas da mineração de dados

Predição: classificação e estimação

Construção de modelos para prever valores ou categorias

Pode prever rótulos (discretos) ou valores numéricos (contínuos)

Processo supervisionado - Usa dados com rótulos já conhecidos para treinar o modelo

Dois tipos principais de predição

- Classificação: atribui uma classe (ex: “aprovado” ou “reprovado”)
- Estimação (Regressão): estima um valor numérico (ex: R\$ 2.500,00)

Exemplos práticos

- Classificação: identificação de spams, detecção de fraudes
- Estimação: previsão de preços, estimativa de desempenho

Principais tarefas da mineração de dados

Análise de grupos (*clustering*)

Agrupamento de objetos similares entre si

Sem conhecimento prévio de categorias (dados não rotulados)

Aprendizado não supervisionado - O modelo descobre os grupos por conta própria

Critério de agrupamento

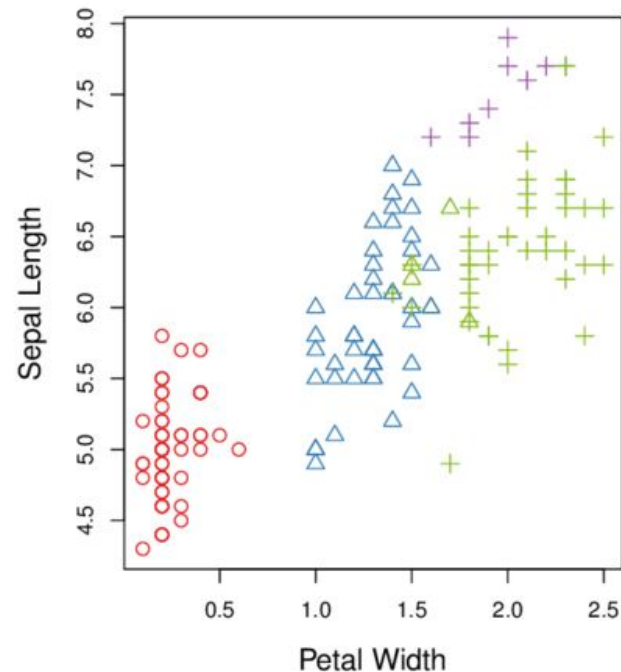
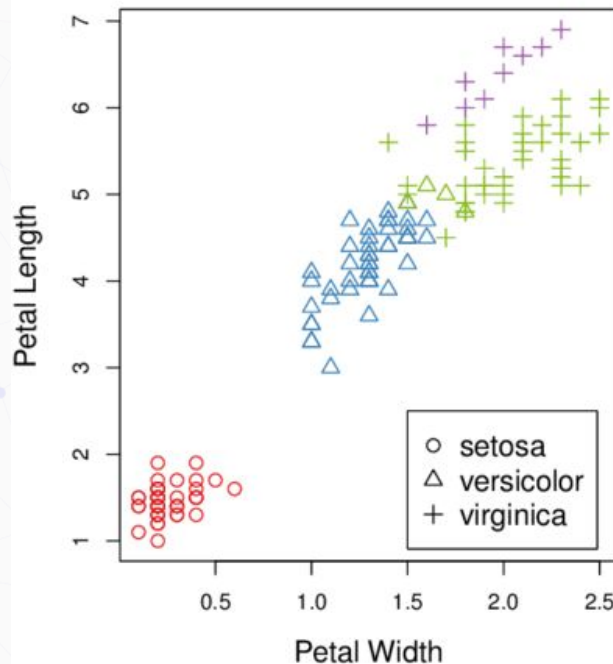
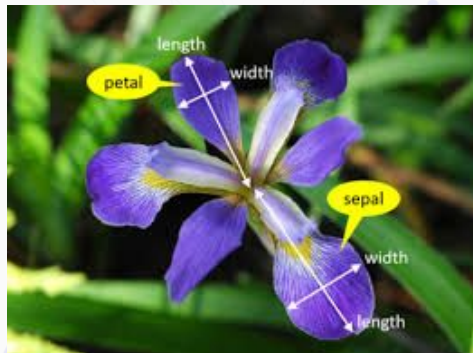
- Máxima similaridade dentro do grupo (intraclasse)
- Mínima similaridade entre os grupos (interclasse)

Aplicações

- Segmentação de público por comportamento de consumo
- Agrupamento de cidades com perfis culturais semelhantes
- Identificação de padrões em acervos museológicos ou obras digitais

Principais tarefas da mineração de dados

Análise de grupos (*clustering*)



Principais tarefas da mineração de dados

Associação

Identificação de padrões entre atributos dos dados

Foco nas relações entre variáveis, não entre objetos

Exemplo clássico: itens comprados juntos em carrinhos de supermercado

Regras de associação ($X \rightarrow Y$)

- Se X ocorre, Y também tende a ocorrer
- Exemplo: smartphone \rightarrow plano de dados
- Aplicações típicas em marketing, varejo e comportamento digital

Principais tarefas da mineração de dados

Associação

Medidas de relevância

- **Suporte (ou cobertura):** frequência da regra no total de registros
- **Confiança (ou acurácia):** proporção de vezes que a regra é verdadeira quando X ocorre

Aplicações

- Identificação de combinações frequentes em acervos (ex: temas recorrentes em obras)
- Análise de padrões de consumo cultural (ex: público que vai ao teatro também frequenta exposições)
- Sugestão automatizada de programação baseada em hábitos anteriores

Principais tarefas da mineração de dados

Detecção de anomalias

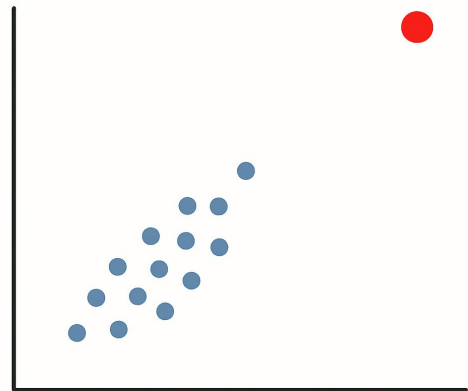
Identificação de objetos que fogem ao padrão geral dos dados

Anomalias também são chamadas de outliers ou valores discrepantes

Muitas vezes descartadas, mas podem ser informativas em contextos específicos

Métodos de detecção

- Modelos estatísticos baseados em distribuição de probabilidade
- Medições de distância — objetos muito distantes dos demais são considerados anômalos



Principais tarefas da mineração de dados

Detecção de anomalias

Importância prática

Útil na detecção de fraudes, erros de sistema ou eventos raros significativos

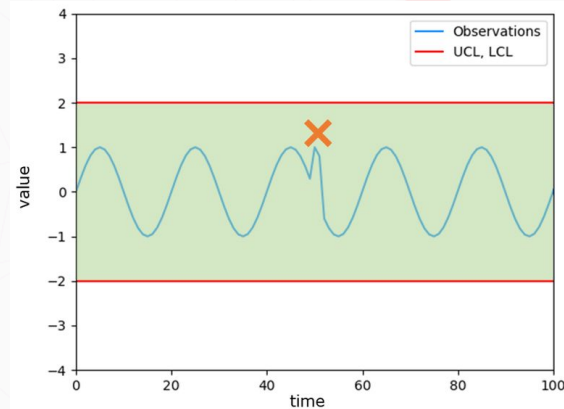
Anomalias ocorrem com frequência muito menor que os padrões normais

Aplicações

Detecção de acessos irregulares a sistemas de patrimônio digital

Identificação de obras mal catalogadas com metadados incoerentes

Análise de desvios de financiamento ou repasses anômalos



Dicas para uma análise eficaz

Estabelecer a significância da mineração

Considere a significância estatística

- Avalie a confiabilidade dos resultados
- Verifique se a base de dados foi bem preparada (amostragem, normalização)
- Correlação não necessariamente indica causalidade
- Resultados incoerentes indicam falhas no processo analítico

Considere a significância prática

- Analise se os resultados podem orientar decisões reais
- Verifique se os achados têm aplicação em políticas, gestão ou ações concretas

Dicas para uma análise eficaz

Reconhecer que as características da base de dados influenciam todos os resultados

A análise depende da base de dados utilizada

- Resultados refletem os dados disponíveis, não a realidade completa
- Bases mal preparadas podem comprometer totalmente a validade da mineração

Fatores que afetam a análise

- Quantidade de objetos (amostra pequena pode distorcer)
- Número e tipo de atributos (textos, números, categorias)
- Ausência de valores e dados incompletos
- Interdependência entre atributos (correlações invisíveis podem afetar modelos)

Dicas para uma análise eficaz

Necessidade de conhecer os dados

Análises preliminares são essenciais

- Explorar os dados antes da mineração reduz erros e interpretações equivocadas
- Permite compreender padrões, inconsistências e limitações

Técnicas recomendadas

- Medidas de tendência central (média, mediana, moda por variável)
- Análise de componentes principais (PCA) para redução de dimensionalidade
- Exploração visual com gráficos e estatísticas simples

Dicas para uma análise eficaz

Busca pela parcimônia

Modelos diferentes podem surgir de execuções distintas

- Mesma base pode gerar modelos diversos, dependendo da amostra ou do algoritmo
- Nem sempre o modelo mais complexo é o melhor

Parcimônia como critério de escolha

- Prefira modelos mais simples e interpretáveis, quando possível
- Modelos enxutos tendem a ser mais fáceis de aplicar e manter
- Complexidade desnecessária pode dificultar a adoção prática

Dicas para uma análise eficaz

Verificar os erros

Todo algoritmo de mineração deve ser avaliado

- Diagnosticar o desempenho é parte essencial do processo
- Entender os erros ajuda a melhorar a análise e os modelos

Formas de avaliação por tipo de tarefa

- Agrupamento: mede-se a qualidade dos grupos formados (ex: método de Silhueta)
- Predição: avalia-se o erro entre o valor previsto e o real (ex: RMSE, acurácia)
- Regras de associação: analisam-se suporte e confiança das regras
- Anomalias: usam-se métricas específicas para eventos raros (ex: taxa de falsos positivos)

Dicas para uma análise eficaz

Validar seus resultados

Validação é essencial para garantir a confiabilidade

- Resultados precisam ser testados e comparados para evitar conclusões precipitadas
- É parte do ciclo de melhoria contínua da análise de dados

Formas de validação

- Comparar com outras técnicas e modelos
- Avaliar a capacidade de generalização dos resultados (funciona com novos dados?)
- Combinar com outras abordagens (ex: análises estatísticas + qualitativas)
- Consultar especialistas da área para interpretar e validar os achados

Leitura deste módulo



MINERAÇÃO DE DADOS: UM OLHAR INSTIGANTE DE POSSIBILIDADES E APLICAÇÕES PARA ÓRGÃOS DA ADMINISTRAÇÃO PÚBLICA FEDERAL

Roberto Rosa da Silveira Junior
Daniel Lins Rodriguez

<https://doi.org/10.21874/rsp.v73.i3.5446>

OBRIGADO!

Até a próxima aula!



Hora da pausa! Voltamos em:

◀◀20:00-▶▶