Noções gerais de mineração de dados

Conceitos básicos de mineração de dados e os algoritmos baseados nos modelos preditivos e descritivos; visão geral e histórico de descoberta de conhecimento em bases de dados (KDD), sua relação com o data mining e machine learning; o processo de KDD, suas técnicas e ferramentas.

Profa. Daisy Albuquerque e prof. Fernando Cardoso Durier da Silva

1. Itens iniciais

Propósito

Compreender noções gerais de mineração de dados é relevante para a formação profissional em várias áreas, não apenas em Tecnologia de Informação, pois trata-se de um conhecimento essencial para cientistas de dados, que é uma profissão multidisciplinar.

Objetivos

- Definir a descoberta de conhecimento em bases de dados (KDD).
- Identificar as etapas do processo de KDD.
- Descrever as técnicas de mineração de dados.

Introdução

Segundo estimativas feitas pela consultoria IDC, a quantidade de dados produzida e replicada no mundo dobra a cada dois anos. Em 2013, cerca de 4,4 zettabytes ou 4,4 trilhão de gigabytes existiam no planeta. Essa projeção leva à existência, no final de 2020, de dez vezes mais dados, isto é, 44 trilhões de gigabytes.

Empresas de qualquer setor empresarial possuem a necessidade de armazenar seus dados. A questão do armazenamento dos dados pode ser considerada, de certa forma, resolvida. Contudo, o que fazer com tantos dados? Como tirar proveito deles de forma a tornar a empresa mais competitiva? Nesse contexto, surgiu a área da mineração de dados.

Neste conteúdo, vamos abordar os conceitos e técnicas de mineração de dados criadas para lidar com este volume crescente de dados que têm sido gerados pelos usuários.

Contudo, a mineração de dados é apenas uma das etapas envolvidas no conceito de descoberta de conhecimento, também conhecido por KDD.

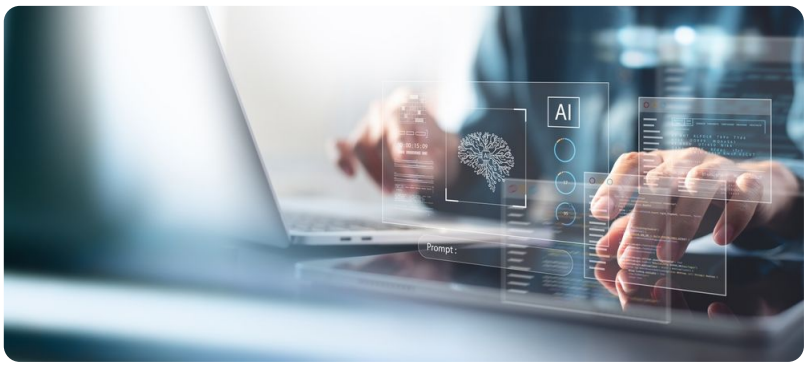
Abordaremos também o processo de KDD que permite extrair conhecimento de informações armazenadas em grandes bases de dados especializadas, a definição de mineração de dados, suas técnicas e tarefas, seus relacionamentos com a inteligência artificial e o aprendizado de máquina, além de descrever como ela contribui para o processo de descoberta do conhecimento.

Ligando os pontos

Você sabe o que é KDD? Qual a diferença entre o processo de KDD, mineração de dados e o aprendizado de máquina?

Você foi agraciado com uma bolsa de pesquisa Capes e entrará para um grupo de pesquisa numa instituição pública como estagiário bolsista. Tal instituição se ocupa de projetos de pesquisa para empresas de óleo e gás, produzindo análises e sistemas de apoio à tomada de decisão. O projeto para o qual você será alocado é voltado à criação de uma inteligência artificial capaz de dizer se um ponto de exploração é uma fonte viável de petróleo ou não.

Sua tarefa é descobrir como implementar os enigmáticos sistemas de inteligência artificial. Suponha que você ainda não sabe a diferença entre mineração de dados, aprendizado de máquina, nem IA, mas está ansioso para apreender.



Na primeira semana, o pesquisador chefe do grupo designou a base de dados para que seus assistentes explorassem e dividiu certas tarefas entre os membros do grupo. Você ficou responsável pela transformação dos dados, mas, para trabalhar, precisaria que sua colega Fátima lhe entregasse os dados pelos quais ela se responsabilizaria por selecionar e pré-processar. Depois, você deveria passar os dados transformados para Luís, que iria fazer a mineração de dados.

Por causa dessas dependências, você precisou pedir uma explicação a Emílio, seu chefe imediato, e ele explicou que aquilo as pessoas entendem por *data mining* ou mineração de dados, na realidade, é o processo de Descoberta de Conhecimento em Bases de Dados, KDD. Nesse processo, um volume de dados de um problema é devidamente selecionado, limpo e preparado para que um algoritmo de aprendizado de máquina minere padrões ocultos e, assim, produza uma diretriz ou regra que possa ser usada no dia a dia para automatizar ou otimizar um processo de negócio.

Luís, com os dados preparados, pode gerar uma árvore de decisão capaz de classificar um ponto de perfuração como bom ou mau candidato à fonte de petróleo. E, por fim, isso pode ser avaliado por Emílio e apresentado aos executivos da empresa de petróleo pelo pesquisador chefe.

Após a leitura do caso, é hora de aplicar seus conhecimentos! Vamos ligar esses pontos?

Questão 1

Durante esse trabalho, qual foi o principal foco e ponto em comum entre o processo de KDD, a mineração de dados (Data Mining) e o aprendizado de máquina (Machine Learning) que você identificou?

A

Modelos.

B

Dados.

C

Inteligência artificial.

D

Algoritmos.

E

Usuários.

A alternativa B está correta.

A diferença entre KDD, *data mining* e *machine learning* reside nas aplicações e instâncias ao resolver um problema. O KDD é o processo de Descoberta de Conhecimento em Bases de Dados, que contém a mineração de dados como a etapa de descoberta de padrões de forma automática por meio dos algoritmos de aprendizado de máquina. Ainda que haja essa diferença toda entre eles e uma devida hierarquização, o ponto focal e comum desses conceitos são só dados matéria-prima que o processo de KDD refina.

Questão 2

Quando você estava lidando com a classificação das tarefas de aprendizado supervisionado, qual foi o problema contraparte da classificação também do aprendizado supervisionado?

A

Regressão.

B

Agrupamento.

C

Sumarização.

D

Regras de associação.

E

Exponenciação.

A alternativa A está correta.

A classificação se ocupa do mapeamento das entradas de um conjunto de dados para a categorização/rotulamento de um objeto, enquanto a regressão se ocupa de mapear o conjunto de atributos para extrapolar o conjunto de coeficientes adequados para a predição de um valor seguindo a função matemática de regressão para um valor numérico alvo determinado.

Questão 3

Qual diferença você percebeu entre mineração de dados e aprendizado de máquina? Qual o papel da inteligência artificial nisso tudo?

Chave de resposta

A mineração de dados é a extração de padrões ocultos nos dados, de forma estatística e/ou algorítmica feita após as etapas de seleção, limpeza e preparação dos dados. O aprendizado de máquina é um ramo da IA que se ocupa de implementar de forma algorítmica os processos cognitivos humanos na máquina, como aprendizado supervisionado, não supervisionado e por reforço, por exemplo. Já a inteligência artificial é a área da computação que se ocupa de desenvolver sistemas capazes de tomar decisões automáticas de forma adaptativa e eficiente e de se comportar como humanos, utilizando, para isso, o aprendizado de máquina para implementar esses processos cognitivos, a lógica para representação do conhecimento e raciocínio, e o processo de KDD para o consumo e melhor utilização dos dados, que são o combustível de tudo isso.

Visão geral e histórico de KDD

Vejamos, a seguir, um pouco do histórico de KDD:

Década de 1960

A história da extração de conhecimento baseada em grandes volumes de dados vem se formando desde a década de 1960. Naqueles anos, o termo estatístico arqueologia de dados trazia a ideia de encontrar correlações sem uma hipótese a priori em bases de dados.

Na época, as empresas estavam migrando e direcionando seus dados para bases de dados armazenadas em computadores. Muitos especialistas armazenavam seus dados em variados recursos tecnológicos, como discos rígidos, fitas magnéticas, banco de dados etc., que forneciam um grau de segurança para a empresa.

Com a evolução da tecnologia, surgiram os Sistemas de Gerenciamento de Banco de Dados (SGBD), novos métodos de armazenamento de dados, além de outros recursos da tecnologia da informação (TI) que favoreceram a proliferação da informação.

Década de 1980

Na década de 1980, o objetivo principal era o acesso aos dados. A tecnologia disponível na época para facilitar esse acesso eram os Bancos de Dados Relacionais (RDBMS), *Structured Query Language* (SQL) e *Open Database Connectivity* (ODBC).

O surgimento dos Sistemas de Apoio à Decisão (SAD) na década de 1980 e a necessidade de reduzir o impacto das integrações entre sistemas de diversas plataformas, tanto no que se refere ao custo com a tecnologia da informação quanto ao aumento da velocidade de processamento dos sistemas de informação (SI), fizeram com que novas tecnologias fossem adotadas para esse armazenamento (INMON, 1997).

Os sistemas de informação construídos para apoiar o processo decisório geralmente armazenam seus dados em sistemas de banco de dados ou até mesmo em grandes repositórios de dados heterogêneos.

Década de 1990

Na década de 1990, surgiu o **data warehouse** — grande repositório de dados projetado para fornecer suporte à decisão — com o uso da tecnologia baseada em OLAP (*On-Line Analytical Processing*) e não mais no tradicional OLTP (*On-Line Transaction Processing*).

A tecnologia Data Warehouse permite atender sistemas de informação capazes de produzir transações de alto desempenho com objetivo de armazenar e cruzar grande volume de dados.

Um *data warehouse* consiste em um banco de dados especializado capaz de manipular um grande volume de informações obtidas a partir de bancos de dados operacionais e de fontes de dados externas à organização.

Porém, apenas parte da informação armazenada é transformada em conhecimento, isso quando não é quase totalmente esquecida nesses repositórios.

De acordo com o tipo de informação que possa ser extraído desses bancos de dados, o processo de extração pode ser considerado complexo e superar a capacidade humana de analisar essas informações e transformá-las em conhecimento (ADRIAANS; ZANTINGE, 1996).

Nesse contexto, as técnicas de análise de dados são necessárias para que os dados sejam devidamente manipulados e a extração de informações importantes ocorra com o objetivo de serem transformadas em conhecimento.

A extração de informação tem início a partir da década de 1990 com a introdução de várias técnicas de análises de dados, inclusive estatísticas. As técnicas também foram utilizadas em pesquisas científicas, cujo interesse e crescimento passou a ser evidenciado mais especificamente a partir de 1997, através de casos e ocorrências em grandes atacadistas, no mercado financeiro, governamental e industrial.

Existe uma grande motivação em usar técnicas de mineração de dados, tanto para o uso comercial como científico em diferentes áreas de estudo e mesmo em ciências aplicadas.

1

Área comercial

Na área comercial, o uso da mineração é evidenciado principalmente pelo crescimento no número de dados armazenados pelas empresas. São dados de compras e navegação pela Internet, dados de transações bancárias ou do uso de cartões de crédito. Pode-se considerar também a pressão por competição nas empresas e o barateamento e potência cada vez maior dos computadores.

Ciências

Para as ciências, são uma realidade a coleta e armazenamento de dados a altas velocidades (Gigabytes/hora) e os resultados da produção científica gerando terabytes de dados, provenientes de telescópios, sensores remotos em satélites, *microarrays* (ferramenta de análise de expressão gênica que permite investigar a expressão de centenas ou milhares de genes em uma amostra) que podem gerar dados de expressões de genes, sendo que muitas vezes as técnicas tradicionais não são apropriadas para analisar tais dados, gerando ruídos e grande dimensionalidade nos resultados produzidos.

Considerando as leis como motivadoras para o desenvolvimento da ciência da computação, temos as Leis de Moore, pelas quais:

I. A capacidade de processamento dobra a cada 18 meses, em termos de CPU, memória e cache;

II. A capacidade de armazenamento em disco dobra a cada 10 meses.

Se combinarmos as duas leis (processamento e armazenamento), produziríamos um gap cada vez mais crescente entre nossa capacidade de gerar dados e nossa habilidade de fazer uso eficiente deles.

Exemplo

A Biblioteca do Congresso (EUA), que possui atualmente cerca de 15 milhões de objetos digitais, com aproximadamente 7 petabytes em dados armazenados.

A manipulação de dados pode levar o analista a fazer descobertas em bancos de dados, previsões e até mesmo modelar um determinado tipo de cliente. Pode-se obter a previsão de vendas de um determinado produto no próximo mês e identificar as causas.

As técnicas utilizadas na Mineração de Dados propiciam a automação do processo a partir de estruturas artificialmente inteligentes. Essas estruturas envolvem técnicas necessárias para a compreensão da linguagem, percepção, raciocínio, aprendizagem e resolução de problemas.

O seu uso busca a criação de teorias e modelos com capacidade cognitiva e a implementação de sistemas computacionais baseados nesses modelos, com o objetivo de descobrir conhecimentos engendrados no banco de dados.

Assim surgiu um novo conceito conhecido como **Descoberta de Conhecimento em Base de Dados ou *Knowledge Discovery in Databases* – KDD**.

Segundo Fayyad, Piatetsky-Shapiro e Smyth (1996), o termo *Knowledge Discovery in Databases* foi cunhado no primeiro workshop de KDD, em 1989, por Piatetsky-Shapiro, com o objetivo de enfatizar que o conhecimento é o produto de uma descoberta baseada em dados.

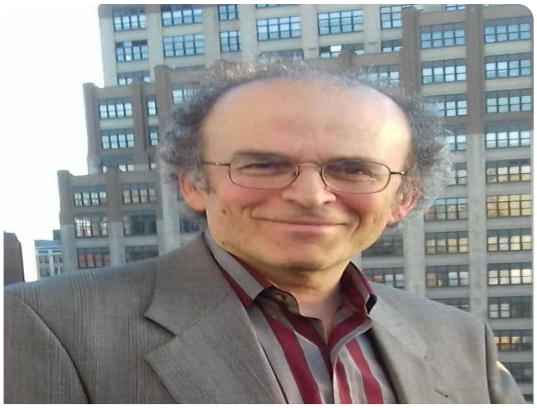


Foto de Gregory Piatetsky-Shapiro em NY, 2016.

Em seu artigo *From data mining to Knowledge Discovery in Databases*, Fayyad, Piatetsky-Shapiro e Smyth trazem como tema central a diferenciação entre o processo KDD e a Data Mining. Para os autores, o processo de KDD se refere ao conjunto de passos e/ou processos que visam a descoberta de conhecimento útil a partir dos dados, enquanto Data Mining vem a ser uma etapa desse processo de descobrimento envolvendo a fase de modelagem dos dados.

O termo **processo não trivial** remete à necessidade de execução de várias etapas, de certa forma complexas, para alcançar o objetivo de identificar padrões que sejam úteis e de fácil compreensão por meio da análise dos dados. Na visão acadêmica (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996), foram consideradas cinco etapas do processo de KDD. Em linhas gerais, vamos esboçar alguns dos seus passos básicos mais adiante, que podem ser encontrados isoladamente ou em agrupamentos nas etapas ou fases de processos de KDD usados no mercado, como o CRISP-DM (Cross Industry Standard Process for Data Mining), que possui sete fases.

Ainda aparecem duas características importantes do processo de KDD que devem ser destacadas. O KDD é: iterativo e interativo, observe que são termos distintos e seu significado é:

Iterativo

Porque prevê uma sequência de atividades em que o resultado de uma etapa (iteração) depende da outra.

Interativo

Porque o analista pode intervir nas atividades, interagindo no processo. Cada etapa do processo pode ser repetida inúmeras vezes.

Assim, KDD refere-se ao processo de extração da informação relevante ou de padrões nos dados contidos em grandes bases de dados e que sejam não triviais, implícitos, previamente desconhecidos e potencialmente úteis, objetivando a tomada de decisão (FAYYAD *et al.*, 1996).

Nesse sentido, a mineração de dados provém da análise inteligente e automática de dados para descobrir padrões ou regularidades em grandes conjuntos de dados, por meio de técnicas que envolvam métodos matemáticos, algoritmos baseados em conceitos biológicos, processos linguísticos e heurísticos, os quais fazem parte do processo de KDD responsável pela busca de conhecimentos em banco de dados (ADRIAANS; ZANTINGE, 1996; HAN; KAMBER, 2006; FAYYAD *et al.*, 1996).

Relacionamento entre KDD, Data Mining e Machine Learning

Machine learning

Esta é uma das áreas da inteligência artificial que, de maneira considerável, vem ganhando espaço no mercado, graças aos avanços dos estudos e tecnologias da internet das coisas (IoT) e de Big Data.

Com esse aprendizado, os computadores podem identificar padrões entre os dados analisados e, por meio da aplicação de algoritmos especiais, serem treinados a aprenderem sozinhos, a fim de executar uma tarefa.

Softwares capazes de aprender com a experiência e informações inerentes a um grande volume de dados nos ajudam a definir o aprendizado de máquina. O objetivo é a criação de técnicas computacionais que visam o aprendizado e a construção de sistemas inteligentes com a capacidade de adquirir conhecimento de forma automática.

> Um sistema aprendiz é aquele que consegue tomar decisões com base em soluções bem-sucedidas aplicadas a problemas anteriores.

Data mining

O **data mining ou mineração de dados** é o ramo da área de banco de dados que utiliza técnicas e algoritmos para extrair informações relevantes de uma base de dados densamente povoada. Portanto, nada mais é que uma das técnicas para obtermos conhecimento em base de dados, permitindo que possamos descobrir conhecimento que esteja implícito no agrupamento de dados.

O processo mais tradicional de coleta de dados consiste, basicamente, no processamento de informação usando técnicas manuais de processamento de informação por especialistas que geram uma série de relatórios que deverão ser analisados e interpretados pelos tomadores de decisão.

Atenção

Em muitos casos, esse processo se torna impraticável devido ao volume de dados, e é nesse momento que a mineração de dados se torna uma alternativa de solução a esse problema de sobrecarga de dados.

KDD

Usama Fayyad, Gregory Piatetsky-Shapiro e Padhraic Smyth publicaram, em 1996, na revista *AI Magazine* uma definição para KDD como um processo não trivial para identificar, em bases de dados, padrões desconhecidos, válidos, potencialmente úteis e facilmente entendíveis.

Segundo eles, KDD está preocupado com o desenvolvimento de métodos e técnicas para dar sentido aos dados. No cerne do processo, está a aplicação de métodos específicos de mineração de dados para a descoberta e extração de padrões.

Sendo assim, podemos diferenciar KDD e *data mining* conforme a seguir:

KDD

Refere-se ao processo global de descoberta de conhecimentos úteis a partir de dados.

Data mining

A mineração de dados é a aplicação de algoritmos específicos para extrair padrões a partir de dados.

Em resumo, o termo KDD representa o processo de transformação dos dados de baixo nível em conhecimento de alto nível. A mineração de dados é uma das etapas desse processo e que pode ser entendida como a extração de padrões ou modelos de dados observados para avaliação e descoberta de conhecimento.

Historicamente, a noção de encontrar padrões úteis em dados tem recebido uma variedade de nomes, incluindo:

- *Data mining*;
- Extração de conhecimento;
- Descoberta de informação;
- Arqueologia de dados;
- Processamento de dados.

O termo *data mining* (mineração de dados) foi utilizado principalmente por estatísticos, analistas de dados e em sistemas de informação de gestão (SIG), além de ter ganhado popularidade na área de banco de dados.

Em virtude do intenso avanço tecnológico das últimas décadas, principalmente em relação ao grande volume de dados que vem sendo produzido e as diversas maneiras criadas para gerarmos e extrairmos informação, o desenvolvimento de algoritmos, os métodos e as aplicações conseguem tratar esse avanço de maneira eficaz e se faz cada vez mais necessário.

> Em meio a esse cenário, um dos ramos da inteligência artificial, o *machine learning* (aprendizado de máquina), tem crescido e se desenvolvido com a mineração de dados, possibilitando a criação de soluções tecnológicas inovadoras.

O aprendizado de máquina sobre esse grande volume de dados (Big Data) pode oferecer soluções e propor ferramentas metodológicas a serem aplicadas nos dados a fim de que possam ser geradas informações, que poderão ser transmitidas e aprendidas por sistemas que implementem algum nível relevante de inteligência artificial.

Desde o final do século passado, com o advento da *World Wide Web*, a sociedade produz muitos dados na Internet, o que faz existir uma gama de maneiras de processá-los, tabulá-los e inferir conhecimento sobre esses resultados e, por conseguinte, no que se refere à inovação, a criação de máquinas cada vez mais inteligentes seria uma das maneiras produtivas para que seja possível aprender com essas informações.

Segundo Carvalho e Dallagasa (2014), o processo de KDD compreende uma série de disciplinas como:

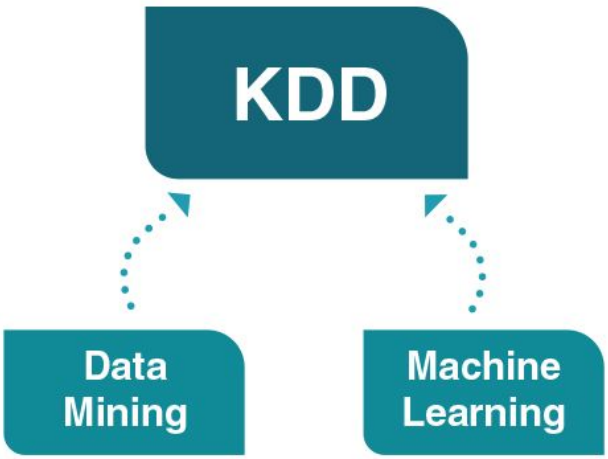
- Estatística;
- Banco de dados;
- Inteligência artificial;
- Aprendizado de máquina.

Sendo o KDD criado a partir dos conceitos destas áreas.

O aprendizado de máquina é a automação do processo de aprendizagem. Assim, usando algoritmos de aprendizado de máquina, são criados padrões de generalizações com base nos dados minerados para análise, possibilitando que eles sejam agrupados e, por fim, sejam criadas regras de associações sobre eles, a fim de inferir conhecimento.

> ⚠
>
> Atenção
>
> O aprendizado de máquina e a mineração de dados não são a mesma coisa, pois cada uma possui suas particularidades e objetivos. Porém, são disciplinas complementares, quando se parte das premissas de que a intenção de ambos é extrair conhecimento de maneira automatizada e otimizada.

A seguir, é possível identificar o relacionamento entre KDD, *data mining* e *machine learning*:



Relacionamento entre KDD, data mining e machine learning.

Caracterização do processo de KDD

O processo de KDD é interativo e iterativo, envolvendo atividades distribuídas por fases ou etapas, com muitas decisões realizadas pelo analista, podendo ser descrito em nove passos:

Primeiro passo

Desenvolver uma compreensão do domínio de aplicação e do conhecimento prévio relevante e identificar o objetivo do processo KDD do ponto de vista do cliente.

Segundo passo

Criar um conjunto de dados-alvo, selecionando um conjunto de dados ou concentrando-se num subconjunto de variáveis ou amostras de dados, sobre os quais a descoberta deve ser realizada.

Terceiro passo

Limpeza e pré-processamento de dados. As operações básicas incluem a remoção do ruído, se apropriado, a obtenção da informação necessária para modelar ou contabilizar o ruído, a decisão de estratégias para lidar com campos de dados em falta e a contabilização da sequência temporal da informação e das alterações conhecidas.

Quarto passo

Redução e projeção de dados. Encontrar características úteis para representar os dados em função do objetivo da tarefa. Com métodos de redução de dimensionalidade ou transformação, o número efetivo de variáveis em consideração pode ser reduzido ou podem surgir combinações de variáveis para os dados.

Quinto passo
Correspondência dos objetivos do processo KDD (primeiro passo) a um determinado método de processamento de dados. Por exemplo, resumo, classificação por classes, regressão, agrupamento etc., que serão descritos em maiores detalhes nos próximos módulos do tema.

Sexto passo
Análise exploratória e seleção de modelos e hipóteses: escolha dos algoritmos de mineração de dados e métodos de seleção a serem utilizados na pesquisa de padrões de dados. Esse processo inclui decidir que modelos e parâmetros podem ser apropriados e a correspondência de um determinado método de mineração de dados com os critérios gerais do processo KDD.

Sétimo passo
Prospecção de dados ao procurar padrões de interesse numa determinada forma representacional ou num conjunto de representações, incluindo regras de classificação ou árvores, regressão e *clustering*.

Oitavo passo
Interpretação de padrões minerados, regressando possivelmente a qualquer dos passos de 1 a 7 para uma maior iteração. Essa etapa pode também envolver a visualização dos padrões e modelos extraídos, ou a visualização dos dados fornecidos pelos modelos extraídos.

Nono passo
Atuar na fronteira do conhecimento descoberto ao utilizar diretamente o conhecimento, classificando o conhecimento em outro sistema para ação futura, ou, simplesmente, documentando e comunicando às partes interessadas. Esse processo inclui também a verificação e a resolução de conflitos potenciais com conhecimentos previamente obtidos (extraídos).

O processo KDD pode envolver interações significantes e conter loops entre quaisquer dos passos. A maioria dos trabalhos sobre o KDD tem-se concentrado no passo 7, a mineração de dados. No entanto, os outros passos são tão importantes (e provavelmente mais) para a aplicação bem-sucedida da prática de KDD.

Técnicas e ferramentas de KDD

As técnicas de KDD podem ser consideradas ferramentas utilizadas para atender aos propósitos do *data mining* (DM).

Não existe uma técnica que resolva todos os problemas de DM. Cada problema exige uma técnica determinada que, por sua vez, tem vantagens e desvantagens na sua aplicação.

Seguem alguns exemplos de técnicas:

1 Descoberta de regras de associação
Estabelece uma correlação estatística entre atributos de dados e conjunto de dados.

2 Árvores de decisão
Hierarquização dos dados baseada em estágios de decisão (nós) e na separação de classes e subconjuntos.

3 Raciocínio baseado em casos
Baseado no método do vizinho mais próximo, combina e compara atributos para estabelecer hierarquia de semelhança.

4 Algoritmos genéticos
Métodos gerais de busca e otimização, inspirados na Teoria da Evolução, na qual, a cada nova geração, soluções melhores têm mais chance de ter "descendentes".

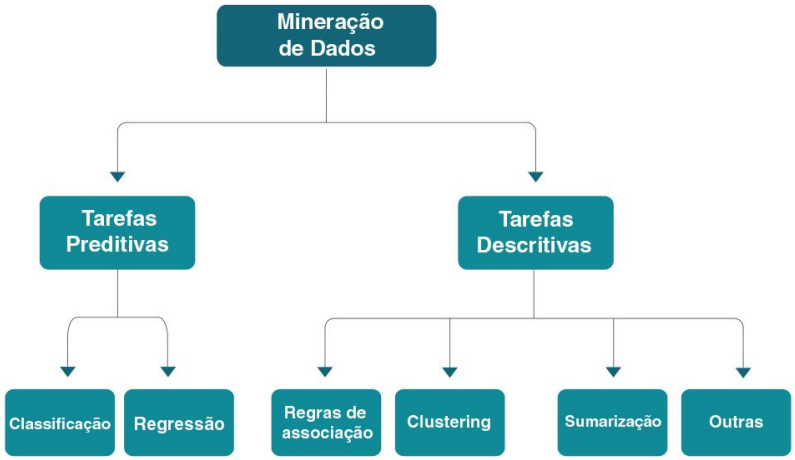
5 Redes neurais artificiais
Modelos inspirados na fisiologia do cérebro, onde o conhecimento é fruto do mapa das conexões neuronais e dos pesos dessas conexões.

As tarefas, também chamadas de funcionalidades, são a maneira como os resultados serão apresentados.

Muitos autores definem uma quantidade diferenciada de tarefas para o *data mining*, uns mais, outros menos, em suas definições, como mostrado a seguir:

- Previsão, identificação, classificação e otimização (ELMASRI; NAVATHE, 2019).
- Descrição e predição (HAN; KAMBER, 2006).
- Classificação, regressão, *clustering*, sumarização, modelo de dependência, escolha e detecção de desvios (FAYYAD, PIATETSKY-SHAPIRO; SMYTH, 1996).
- Classificação, regressão, regras de associação, sumarização, *clustering* e outras (REZENDE, 2005).
- Classificação, regressão, associação, *clustering* e sumarização (DIAS, 2002).

Porém, a maioria deles concorda que essas tarefas sejam classificadas em dois grandes grupos, como mostra Rezende (2005) na imagem a seguir:



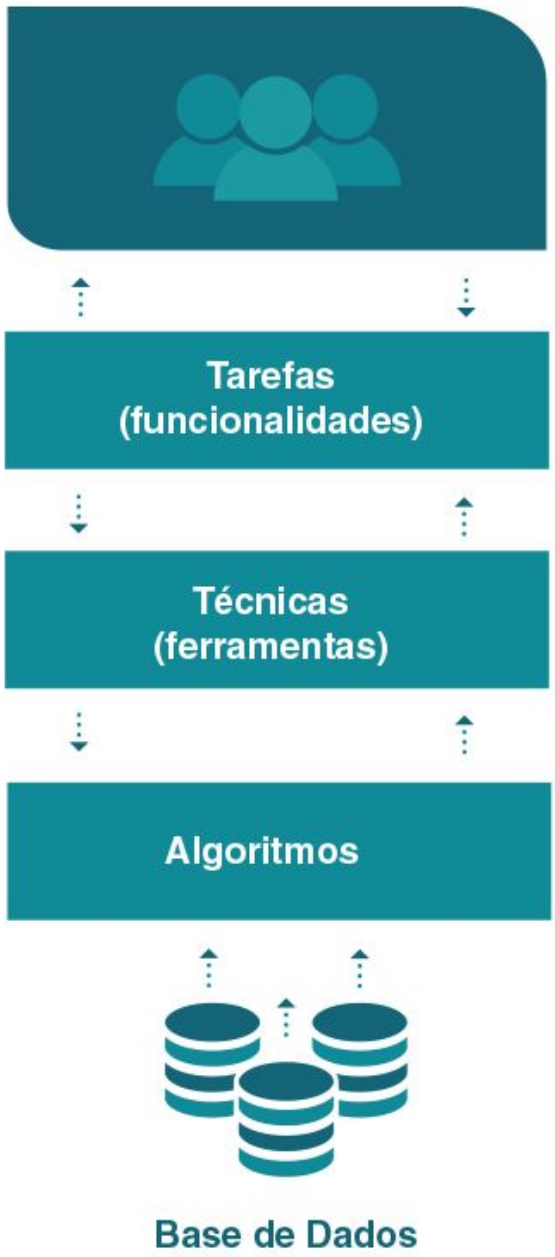
Tarefas de data mining.

As tarefas preditivas envolvem atributos de um conjunto de dados para prever o valor futuro de uma variável meta, visando principalmente a tomada de decisão.

Já as tarefas descritivas procuram padrões interpretáveis pelos humanos, visando o suporte à tomada de decisão.

Técnicas e tarefas são definidas na etapa de extração de padrões. Dependendo da técnica, os algoritmos correspondentes são escolhidos para sua execução.

A imagem a seguir mostra as interações entre técnicas, tarefas e algoritmos:



Interações entre, técnicas, tarefas (funcionalidades) e algoritmos.

Visão geral de descoberta de conhecimento em bases de dados (KDD)

No vídeo a seguir, a especialista Daisy Albuquerque apresenta uma visão geral do KDD.

Conteúdo interativo

Acesse a versão digital para assistir ao vídeo.

Verificando o aprendizado

Questão 1

Em relação aos fundamentos de KDD e Data Mining, é correto afirmar:

A

Data mining é o processo de descobrir conhecimento em banco de dados, que envolve várias etapas. O KDD — *Knowledge Discovery in Database* é uma destas etapas, portanto, a mineração de dados é um conceito que abrange o KDD.

B

A etapa de KDD do *data mining* consiste em aplicar técnicas que auxiliem na busca de relações entre os dados. De forma geral, existem três tipos de técnicas: estatísticas, exploratórias e intuitivas. Todas são devidamente experimentadas e validadas para o processo de mineração.

C

Os dados podem ser não estruturados (bancos de dados, CRM, ERP), estruturados (texto, documentos, arquivos, mídias sociais, *cloud*) ou uma mistura de ambos (e-mails, SOA/web *services*, RSS). As ferramentas de *data discovery* mais completas possuem conectividade para todas essas origens de dados de forma segura e controlada.

D

Estima-se que, atualmente, em média, 80% de todos os dados disponíveis são do tipo estruturado. Existem diversas ferramentas *open source* e comerciais de *data discovery*. Dentre as *open source* está a *InfoSphere Data Explorer*, e entre as comerciais está a Vivisimo, da IBM.

E

As ferramentas de *data mining* permitem ao usuário avaliar tendências e padrões não conhecidos entre os dados. Esses tipos de ferramentas podem utilizar técnicas avançadas de computação, como redes neurais, algoritmos genéticos e lógica nebulosa, dentre outras.

A alternativa E está correta.

As técnicas de KDD podem ser consideradas ferramentas utilizadas para atender aos propósitos do *data mining* (DM). Não existe uma técnica que resolva todos os problemas de DM. Cada propósito exige uma técnica determinada que, por sua vez, tem vantagens e desvantagens na sua aplicação.

Questão 2

As ferramentas e técnicas de mineração de dados (*data mining*) têm por objetivo:

A

Preparar dados para serem utilizados em um *data warehouse* (DW).

B

Permitir a navegação multidimensional em um DW.

C

Projetar, de forma eficiente, o registro de dados transacionais.

D

Buscar a classificação e o agrupamento (clusterização) de dados, bem como identificar padrões.

E

Otimizar o desempenho de um gerenciador de banco de dados.

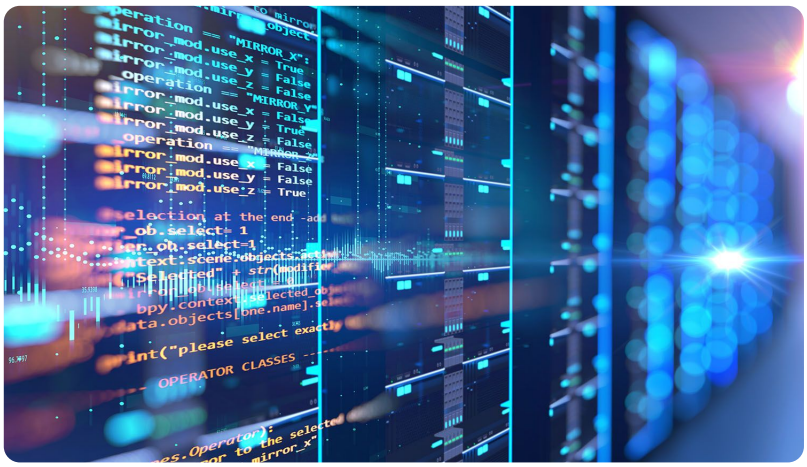
A alternativa D está correta.

As técnicas e ferramentas de *data mining* tem o objetivo de obter conhecimento em base de dados, utilizando dos métodos de classificação e agrupamento para descobrir conhecimento que esteja implícito no agrupamento de dados.

Ligando os pontos

Você sabe o que é o processo de KDD? Sabe a diferença entre dados, informação, conhecimento e sabedoria?

Imagine que você é o mais novo contratado na multinacional Amazon. Lá você trabalhará com sistemas de recomendações de produtos, que nada mais são do que grandes filtros colaborativos de itens de um portfólio, em que usuários emitem suas opiniões. Mediante o cruzamento dessas avaliações, o sistema é capaz de entender o perfil de cada consumidor pelo tipo de compras recorrentes e suas avaliações, bem como entender a similaridade ou dissimilaridade entre perfis.



Você será engenheiro de modelos de aprendizado de máquina, com o treinamento, manutenção e desenvolvimento de modelos SVD (*Singular Value Decomposition*) para recomendações de produtos no carrinho de compras dos usuários.

Nos primeiros momentos, sua maior dificuldade foi em lidar com os dados, pois estamos acostumados com dados tabulares e registros estruturados, mas, no caso de sistemas de recomendações, os dados têm uma característica mais matricial e multidimensional, pois cada linha da matriz representa um usuário; cada coluna, um produto e cada célula, uma nota para determinado produto ou para caso o usuário ainda não tenha comprado determinado produto.

Contudo, ao estudar com seus colegas mais experientes, você foi capaz de lidar com esse formato diferente, produzir seu primeiro SVD com mais de 80% de acurácia e, com ele, contribuir para o bom funcionamento do *e-commerce*.

Após a leitura do caso, é hora de aplicar seus conhecimentos! Vamos ligar esses pontos?

Questão 1

Após esse trabalho, como você determinaria o papel das avaliações dos produtos pelos usuários na trinca dados, informação e conhecimento?

A

Dados.

B

Informações.

C

Conhecimento.

D

Sabedoria.

E

Algoritmos.

A alternativa B está correta.

As avaliações dos produtos são uma interpretação dos símbolos, números emitidos por usuários como classificação de um produto numa escala de 0 a 10, ou de 1 a 5 estrelas etc.

Questão 2

Em que etapa do processo de KDD você extrairia o conhecimento das informações?

A

Mineração de dados.

B

Seleção de dados.

C

Pré-processamento.

D

Transformação de dados.

E

Avaliação.

A alternativa A está correta.

Os algoritmos de mineração de dados recebem, na realidade, informações, pois, quando selecionamos os dados do nosso conjunto, já damos interpretação e afinamos esta ao pré-processarmos e transformá-los. Os modelos de aprendizado, então, buscam a repetição e padrões nesses registros.

Questão 3

Com base no que foi aprendido e discutido no case, qual etapa você identificou como mais relevante para os stakeholders no processo de KDD?

Chave de resposta

O processo de KDD é relevante em cada uma de suas etapas, mas, para os executivos e os clientes finais, o valor mesmo está na última etapa após a avaliação, que é a de apresentação. Nela, o conhecimento depois de validado é utilizado ou demonstrado para os interessados, entregando de fato valor para o negócio.

Conceitos de dado, informação, conhecimento e sabedoria

Vivemos em uma era na qual somos expostos a uma quantidade gigantesca de dados e, para compreendê-los, é preciso fazer a hierarquização e interpretação da informação.

A evolução da manipulação dos dados, gerando informações, e, mais recentemente, conhecimento e sabedoria, tem se destacado como fator de competitividade em diferentes tipos de organização.

O gerenciamento desses recursos informacionais subsidia várias atividades, melhorando o planejamento estratégico e o processo de tomada de decisão na organização. Segundo O'Brien (2004):

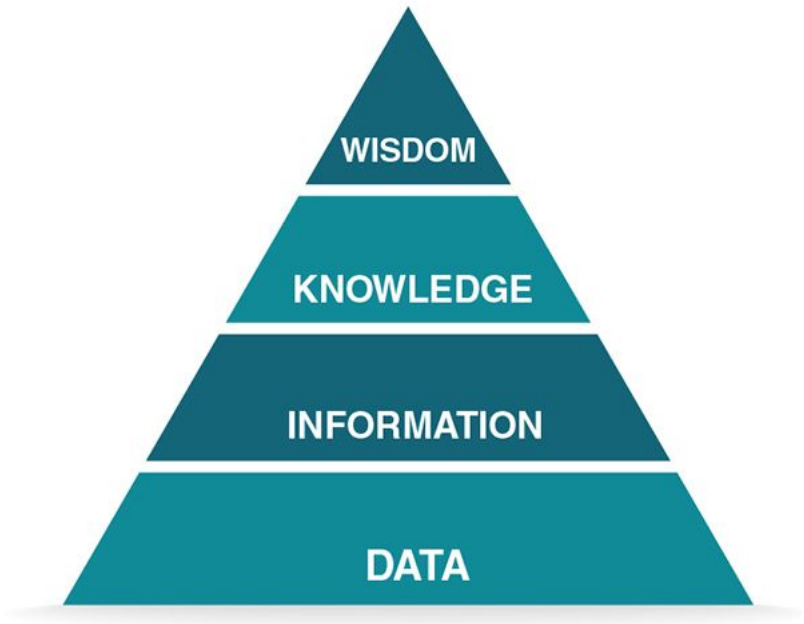
> Dados são fatos ou observações cruas, normalmente sobre fenômenos físicos ou transações de negócios que ainda não foram convertidos em um contexto significativo.
>
> —
>
> (O'BRIEN, 2004)

A informação, componente importante no processo decisório, é formada pelo tratamento dos dados. Sendo assim, ao se tratar o dado, a forma é agregada, manipulada e organizada, seu conteúdo é analisado e avaliado, sendo colocado em um contexto adequado ao usuário.

Porém, um novo componente foi inserido, a contextualização da informação. Quando a informação gerada é introduzida em determinado contexto, gera-se o conhecimento.

Atualmente, diversas técnicas automatizadas permitem contextualizar a informação de forma a proporcionar ao gestor novas maneiras de interpretá-la e validá-la.

Assim, surgiu o termo Pirâmide do Conhecimento, ou Hierarquia DIKW (*Data-Information-Knowledge-Wisdom*), conforme visualizado a seguir:



Pirâmide DIKW (Data-Information-Knowledge-Wisdom).

Um dos primeiros pesquisadores sobre o assunto foi o teórico americano Russel Ackoff, por volta da década de 1980. Segundo ele, um dado sozinho não tem significado total, pois precisa ser interpretado para que possua valor e torne-se, de fato, uma informação.

Os dados, informações, conhecimento e sabedoria se organizam em escala de valores:

Data (dados)

São acessíveis a todos e disponíveis em todos os lugares. Sendo assim, valem menos que a informação.

Information (informação)

É a interpretação dos dados. Ela manipula os dados de relevância e de propósito, atribuindo-lhes um significado.

Knowledge (conhecimento)

É oriundo da contextualização, organização e padronização da informação.

Wisdom (sabedoria)

É o resultado que se alcança através da análise e da formulação de hipóteses perante cenários distintos.

De acordo com Goldschmidt, Bezerra e Passos (2015), os dados podem ser interpretados como itens elementares, captados e armazenados por recursos da Tecnologia da Informação. Sendo assim, podemos definir dados, informações e conhecimento, como:

Dados

São cadeias de símbolos e não possuem semântica.

Informações

Representam os dados processados, com significados e contextos bem definidos.

Conhecimento

Corresponde a um padrão ou conjunto de padrões, cuja formulação pode envolver e relacionar dados e informações.

A pirâmide do conhecimento (pirâmide DIKW) demonstra que a quantidade de dados existentes em um sistema é muito grande. Já o montante de informação é reduzido devido a dados errôneos ou sem expressão. Menor ainda é a quantidade de conhecimento que pode ser extraído dessa informação por meio de técnicas de descoberta de conhecimento.

Exemplo

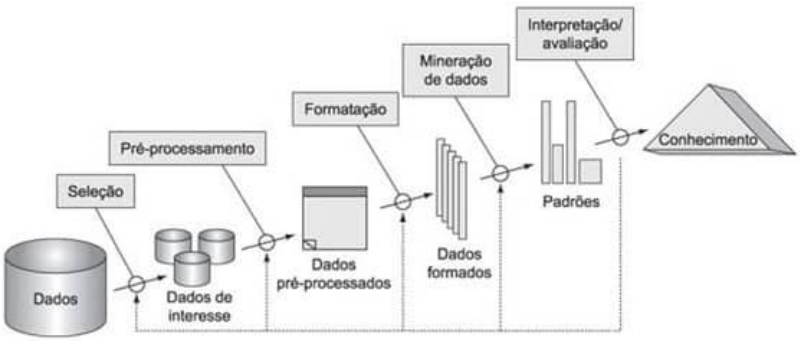
Suponhamos o registro de casos do Coronavírus no país, em que os dados sozinhos apresentam apenas números. Com a interpretação por um determinado período, é possível identificar a parte da população que é mais afetada, em quais regiões, faixa etária, gênero etc. As informações produzem conhecimento para elaboração de estratégias de solução do problema e tomada de decisão, por exemplo, em relação à quarentena e à fabricação de vacinas.

A informação é uma perspectiva mais elaborada de um dado e o conhecimento produz a estratégia e como colocar a informação em prática. A sabedoria é o porquê, a razão e a percepção que se tem do todo, ou seja, os resultados obtidos.

Etapas do processo de KDD

Como vimos no módulo anterior, o processo de KDD inclui vários passos, cada um com seu valor potencial para as dimensões tática e estratégica de uma organização.

Segundo Fayyad, Piatetsky-Shapiro e Smyth, numa visão acadêmica, o processo de KDD é composto por cinco fases, conforme visualizamos a seguir:



Fases do processo de KDD.

Antes de iniciar as fases do processo de KDD, é necessária a definição do problema, ou seja, ter clareza sobre quais são os objetivos da mineração ou as questões sobre as quais se está buscando respostas; nesse caso, as possibilidades são inúmeras e vai depender do tipo de negócio e suas estratégias.

Após essa fase, iniciamos o processo de KDD com cinco fases:

1

Seleção

Diante da enorme massa de dados armazenados em repositórios diversos (banco de dados, relatórios, transações, logs de acessos, redes sociais, sistemas de gestão de relacionamento com o cliente - CRM), cria-se uma segmentação, de acordo com critérios, determinando o conjunto de dados-alvo em que a descoberta deve ser realizada.

2 Pré-processamento e limpeza

Adequação, limpeza e formatação dos dados para ferramenta de mineração, removendo ruídos, redundâncias e realizando manipulações, quando necessárias, reduzindo as variáveis do conjunto de dados, de acordo com os objetivos em questão.

3

Transformação

Os dados são apropriadamente transformados para mineração, por meio da realização de operações de agregação, por exemplo.

4

Data mining

A partir dos repositórios organizados, são feitos ajustes de parâmetros para a execução efetiva de um ou mais algoritmos sofisticados a fim de extrair o padrão de comportamento dos dados, de forma interativa.

5

Interpretação e avaliação

Classificação, consolidação, análise e compreensão aprofundada dos padrões, tendências e seus significados, desconsiderando aquilo que é específico, reconhecendo fatos significativos e valorizando o que é generalizado, de maneira a dar suporte à tomada de decisões.

Para melhor entendimento de todo o processo, Goldschmidt, Bezerra e Passos (2015), sugerem que as fases do processo KDD sejam agrupadas em três grandes etapas denominadas etapas operacionais do processo KDD. A imagem a seguir ilustra essas etapas:

Etapas operacionais do processo de KDD.

Pré-processamento

A etapa de pré-processamento compreende todas as funções relacionadas com a captação, organização e tratamento dos dados. Esta etapa tem como objetivo a preparação dos dados para os algoritmos de mineração de dados que serão aplicados na etapa seguinte.

As principais funções de pré-processamento são:

Seleção de dados para pré-processamento

A seleção de dados é constituída por um agrupamento organizado de uma massa de dados, alvo da prospecção. Nessa fase, é feita a escolha dos dados que realmente serão utilizados no processo. A seleção dos dados pode ter dois enfoques distintos: a seleção de atributos ou a seleção de registros que devem ser submetidos ao processo de extração de conhecimento. Nesse estágio, é interessante a participação de um especialista no conjunto de dados, uma vez que essa seleção é de fundamental impacto nos resultados das análises.

Limpeza dos dados

A limpeza dos dados tem como finalidade eliminar essas adversidades de forma que elas não influenciem nos resultados. Algumas das causas desses problemas são erros humanos (erros de digitação, por exemplo), indisponibilidade da informação no momento do levantamento dos dados ou quando alguma mudança no sistema operacional ainda não tenha sido refletida no ambiente da mineração de dados.

Integração dos dados

A integração dos dados analisa profundamente os dados coletados para que se possa integrá-los de forma consistente e apta à obtenção de resultados satisfatórios. Redundâncias, dependências entre as variáveis e valores incompatíveis (categorias diferentes para os mesmos valores ou regras diferentes para os mesmos dados, por exemplo) são algumas técnicas observadas e corrigidas na integração dos dados.

Transformação dos dados

A transformação dos dados é a etapa mais importante do pré-processamento, pois é por meio dela que se constitui um padrão para os dados, facilitando o uso das técnicas computacionais de análise. Na etapa de Transformação, os dados pré-processados passam por tratamento para um armazenamento adequado, visando facilitar o uso das técnicas de *data mining*, pois existem diversos tipos de algoritmos e cada um necessita de uma entrada específica, além das conversões de dados, criação de novas variáveis e categorização de variáveis contínuas.

Redução dos dados

A redução dos dados é essencial para que análises sejam feitas de forma a se obter resultados satisfatórios, ainda que feita a seleção dos dados no início, pois quando as bases de dados são muito grandes, a mineração torna-se inviável.

Mineração de dados

Esta etapa, basicamente, utiliza os repositórios organizados para a execução de algoritmos de complexos para extrair padrões de comportamento e associações entre os dados de forma interativa.

Sendo assim, esta etapa consiste no uso de técnicas que permitem automatizar a busca em grandes volumes de dados por padrões e tendências que não são detectáveis por análises mais simples, auxiliando gestores na tomada de decisões estratégicas.

As tarefas mais comuns aplicadas na prática de mineração de dados incluem:

- Sumarização;
- Classificação;
- Regressão e estimação;
- Predição;
- Agrupamento ou clusterização;
- Associações;
- Detecção de desvios;
- Descoberta de sequências.

Atenção

Cada técnica de mineração de dados ou cada implementação específica dos algoritmos que são utilizados para conduzir as operações dessas técnicas, adequa-se melhor a alguns problemas que a outros, o que dificulta a existência de um método inteiramente melhor.

Um conceito muito importante e muito utilizado em mineração de dados é a **noção de medidas de interesse**. As medidas são essenciais ao processo de KDD e podem ser usadas após a etapa de MD a fim de ordenar ou filtrar os padrões descobertos de acordo com o grau de interesse, ou podem ser usadas para guiar ou restringir o espaço de busca da MD.

Pós-processamento

A etapa de pós-processamento é referente a **interpretação e avaliação dos dados minerados**.

Segundo Goldschmidt, Bezerra e Passos (2015), a etapa do pós-processamento abrange o tratamento do conhecimento adquirido na etapa de mineração de dados. Entre as principais funções desta etapa, destacam-se a elaboração e organização do conhecimento obtido, podendo incluir a simplificação de gráficos, diagramas e relatórios, além da conversão da forma de representação em conhecimento obtido.

Há muitas propostas na literatura para “minerar” o conhecimento descoberto, ou seja, pós-processar o conhecimento descoberto pela etapa de *data mining*.

Em geral, as propostas se enquadram em duas categorias básicas:

Métodos subjetivos

Nestes métodos, é preciso que o usuário estabeleça previamente o conhecimento ou crenças, a partir do qual o sistema irá minerar o conjunto original de padrões descoberto pelo algoritmo de *data mining*, buscando por aqueles padrões que sejam surpreendentes ao usuário.

Métodos objetivos

O método objetivo não necessita que um conhecimento prévio seja estabelecido. Pode-se dizer que o método objetivo é dirigido por dados (*data-driven*) e o subjetivo é dirigido pelo usuário (*user-driven*).

Papel dos usuários no processo de KDD

Para entendermos adequadamente o papel dos usuários no processo de KDD, vamos adaptar a pirâmide de conhecimento (pirâmide DIKW):

Pirâmide de conhecimento para a tomada de decisão.

Para cada camada da pirâmide, encontram-se diferentes usuários com diferentes funções.

Administrador da base de dados

Este profissional, por exemplo, no nível operacional, trabalha primariamente com dados sobre operações diárias e de rotina encontrados em arquivos e bases de dados, na base da pirâmide informacional. Nessa camada, ocorre basicamente a criação dos dados.

Analistas de negócios e executivos

São responsáveis pela tomada de decisão ao indicar qual direção seguir, formulam estratégias e táticas, além de supervisionar a sua execução. Para esse tipo de usuário do processo de KDD, é necessário o acesso a informações de maior qualidade. Esses usuários preocupam-se com tendências, padrões, ameaças, pontos fortes e fracos, oportunidades, informação de mercado, entre outros.

Tanto informações internas como externas são levadas em consideração por esse tipo de usuário. Nesse caso, há demanda de dados analisados com alto valor agregado, e estes se localizam no topo da pirâmide, na camada de **tomada de decisão**.

Visão geral do processo de KDD

Neste vídeo, a especialista Daisy Albuquerque apresenta uma visão das etapas do processo de KDD.

Conteúdo interativo

Acesse a versão digital para assistir ao vídeo.

Verificando o aprendizado

Questão 1

Dado, informação e conhecimento são elementos fundamentais para a comunicação e a tomada de decisão nas organizações, mas seus significados não são tão evidentes. Eles formam um sistema hierárquico de difícil delimitação. O que é um dado para um indivíduo pode ser informação e/ou conhecimento para outro. Com isso, podemos afirmar que o grande desafio dos tomadores de decisão:

A

Minimizar e transformar as interferências individuais em dados e dados em informação, nesse processo de transformação do conhecimento.

B

Transformar conhecimento em informação e informação em dados, minimizando as interferências individuais nesse processo de transformação.

C

Transformar dados em informação e informação em conhecimento, minimizando as interferências individuais nesse processo de transformação.

D

Transformar informação em dados e conhecimento em informação, minimizando as interferências individuais nesse processo de transformação.

E

Interferir o mínimo na transformação da informação e dos dados nesse processo de conhecimento.

A alternativa C está correta.

O grande desafio é criar meios para gerar sabedoria de forma tempestiva, eficaz e eficiente, ao transformar dado em informação e informação em conhecimento, para chegar no topo, à sabedoria.

Questão 2

Na visão acadêmica, o processo de KDD é composto por cinco etapas ou fases ordenadas da seguinte forma:

A

Mineração de Dados – Seleção – Processamento – Transformação e Interpretação — Avaliação.

B

Seleção – Mineração de Dados — Processamento – Transformação e Interpretação — Avaliação.

C

Seleção — Processamento – Transformação — Interpretação e Avaliação – Mineração de Dados.

D

Mineração de Dados – Processamento – Seleção - Interpretação e Avaliação — Transformação.

E

Seleção – Processamento – Transformação – Mineração de Dados – Avaliação e Interpretação.

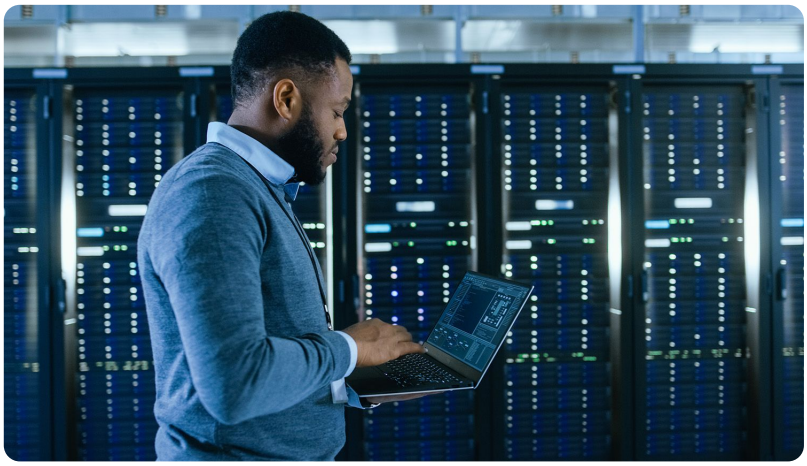
A alternativa E está correta.

KDD é o processo de transformação de dados em conhecimento. Este processo possui cinco etapas ou fases: Seleção, Pré-processamento, Transformação, Mineração de Dados, ou Data Mining, e Interpretação e Avaliação.

Ligando os pontos

Você sabe o que é mineração de dados? Já ouviu falar de processamento de linguagem natural?

Suponha que você trabalha em um escritório de advocacia que está se modernizando. Sua função é engenheiro de modelos de aprendizado de máquina e tem como projeto criar um sistema gerador de contratos baseado no histórico de outros contratos.



Documentos de texto, como contratos e notícias, são representantes muito característicos de dados semiestruturados, uma vez que textos não têm uma estrutura fixa e definida como um registro num banco de dados, mas, ainda assim, seguem regras mínimas da linguagem. Basicamente o que você terá que fazer é pré-processar esses documentos extraindo tokens, n-gramas, sentenças, termos-chave, e sentimentos para, assim, poder caracterizar e dar estrutura a esse tipo de dado. Sendo assim, você poderá treinar um modelo de aprendizado semissupervisionado de rede neural profunda e gerará novos contratos baseados nos contratos passados e no seu conjunto de características e termos-chave. Para isso, você precisou buscar, no registro geral de cópias de contratos do escritório, um colega advogado experiente para lhe prestar consultaria de especialista e ajudar a ajustar o conjunto de treinamento.

Com isso, você conseguiu um modelo bem acurado e otimizou o processo da criação de novos contratos, encontrando templates automáticos para cada tipo de situação, que pudessem ser customizados pelos advogados do escritório com seus clientes.

Após a leitura do caso, é hora de aplicar seus conhecimentos! Vamos ligar esses pontos?

Questão 1

Durante o processo de treinamento, você precisou separar os dados em dois subconjuntos mutuamente excludentes, que são chamados respectivamente de

A

peso e viés.

B

treinamento e teste.

C

acurácia e R2.

D

tramite e transação.

E

dados e informações.

A alternativa B está correta.

Os conjuntos de dados sempre são divididos em treino e teste para que os modelos possam aprender a mapear a função de associação com o conjunto de treinamento e possam ser avaliados sobre sua performance com o conjunto de testes. Em alguns casos mais raros, podemos ainda dividir o conjunto de testes em teste e validação, numa proporção 80:10:10, ou 70:20:10 nesse caso especial para simular um teste com dados do mundo real.

Questão 2

De acordo com a sua análise, a extração de características determinantes e estruturais de um documento textual é componente essencial de que tipo de técnica de mineração de dados?

A

NLP

B

KDD

C

CRISP-DM

D

SVM

E

KNN

A alternativa A está correta.

A mineração de dados textuais ou simplesmente mineração de textos é também conhecida por Processamento da Linguagem Natural ou PLN. A partir dessa metodologia, os cientistas de dados podem extrair informações-chave dos documentos de texto, como tokens, sentenças, sentimentos, n-gramas etc.

Questão 3

Qual o papel do advogado consultor que auxiliou você nesse trabalho? Por que você precisou ter um participante especialista no assunto trabalhando com você?

Chave de resposta

De modo geral, o cientista de dados ou o engenheiro de modelos de aprendizado de máquina não necessariamente tem conhecimento prévio na área do problema a ser resolvido com o modelo de aprendizado de máquina. Sendo assim, o especialista é essencial para curar os resultados do modelo, principalmente neste caso em que o resultado gerado é um novo texto criado pela máquina. Podemos ver situação semelhante em sistemas especialistas de diagnósticos médicos e prescrição.

Conceitos básicos de mineração de dados

Nos últimos anos, os avanços computacionais, tanto no poder como na velocidade de processamento, levaram à substituição das práticas manuais, tediosas e lentas para a análises de dados rápidas, fáceis e automatizadas.

> Quanto mais complexas são as bases de dados coletadas, mais potencial há para delas extrair insights relevantes.

Varejistas, bancos, fabricantes, operadoras de telecomunicações, seguradoras, etc., estão usando a mineração de dados (MD) para descobrir relações entre seus dados — desde preços, promoções e demografias até a economia, o risco, a concorrência e as mídias sociais estão afetando seus modelos de negócio, receitas, operações e relacionamentos com os clientes.

A mineração de dados é um termo usado para generalizar todas as técnicas e métodos computacionais usados para analisar e extrair informação de bases de dados, ou seja, é uma extração de informação com o

objetivo de descobrir padrões válidos e potencialmente úteis, o que seria impossível por meio de uma observação mais superficial.

A tecnologia de MD está em constante evolução para acompanhar tanto o potencial ilimitado do Big Data como a computação de baixo custo. Essa tecnologia é composta por técnicas de análise exploratória de dados, desenvolvidas no decorrer dos anos, desde que o termo foi cunhado em 1977 por John Tukey, em seu livro *Exploratory Data Analisys*.

Diversas definições de MD podem ser encontradas na literatura. Entre as diversas definições, destacamos os autores a seguir:

SHOLOM; NITIM, 1999

"Mineração de dados é a busca de informações valiosas em grandes bases de dados. É um esforço de cooperação entre homens e computadores. Os homens projetam bancos de dados, descrevem problemas e definem seus objetivos. Os computadores verificam dados e procuram padrões que se casem com as metas estabelecidas pelos homens".

MICHAEL; GORDON, 1997

"Mineração de dados é a exploração e análise de dados, por meios automáticos ou semiautomáticos, em grandes quantidades de dados, com o objetivo de descobrir regras ou padrões interessantes".

JESUS, 1999

"Mineração de dados, em poucas palavras, é a análise de dados indutiva".

BHAVANI, 1999

"Mineração de dados é o processo de proposição de várias consultas e extração de informações úteis, padrões e tendências, frequentemente desconhecidos, a partir de grande quantidade de dados armazenada em bancos de dados".

JIAWEI; MICHELINE, 2001

"Mineração de dados, de forma simples, é o processo de extração ou mineração de conhecimento em grandes quantidades de dados"

Na verdade, o termo mineração de dados refere-se a uma das fases de um processo maior de descoberta de conhecimento em base de dados (KDD), o qual possui uma metodologia própria para preparação e exploração dos dados, interpretação de seus resultados e assimilação dos conhecimentos minerados.

> A mineração de dados pode ser considerada a etapa de aplicação de técnicas ou ferramentas para apresentar e analisar os dados.

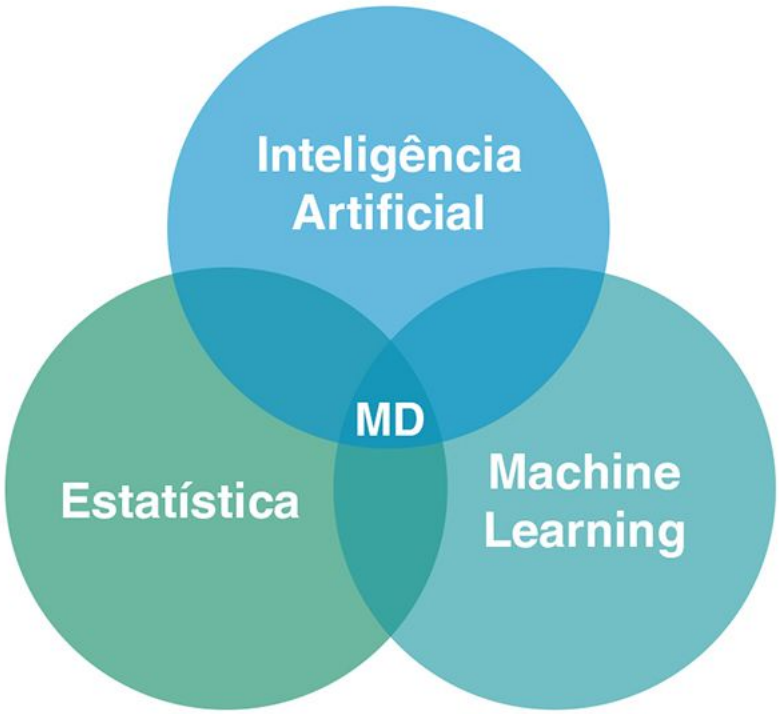
Conceitualmente, mineração de dados é uma área de pesquisa multidisciplinar, que consiste essencialmente em extrair informação de gigantescas bases de dados (Big Data), analisar e descobrir padrões ocultos, procurando encontrar relações entre dados não explícitas que possam ser usadas em modelos matemáticos

com capacidade preditiva e explanatória, incluindo tecnologia de bancos de dados, inteligência artificial, aprendizado de máquina, redes neurais, estatística (análise exploratória de dados), reconhecimento de padrões, sistemas baseados em conhecimento, **lógica *fuzzy***, recuperação da informação, computação de alto desempenho e visualização de dados.

A base da MD é composta por três disciplinas científicas entrelaçadas, conforme visualizado a seguir:

Lógica fuzzy

A lógica fuzzy permite expressar uma sentença lógica, uma verdade, por meio de valores reais que estejam compreendidos no intervalo entre 0 e 1, por sua vez na lógica booleana os valores lógicos podem assumir explicitamente os valores 0 ou 1 para determinada condição, em ambos os casos, tanto na lógica fuzzy quanto na booleana, 0 representa o valor falso, e o 1 representa o valor verdadeiro.



Disciplinas científicas da mineração de dados.

Sendo assim, podemos dizer que a MD descende fundamentalmente de três linhagens:

Estatística

Estudo numérico das relações entre dados. Esta é a disciplina mais antiga; sem a estatística não seria possível termos a mineração de dados, pois ela é a base da maioria das tecnologias a partir das quais a MD é construída.

Inteligência artificial

É a inteligência inserida em máquina que se assemelha com a inteligência humana. Essa disciplina é construída a partir dos fundamentos da heurística, ou seja, em oposição à estatística, e sua tarefa consiste em imitar a maneira como o homem pensa na resolução dos problemas estatísticos.

Machine learning

É o uso de algoritmos que aprendem com os dados para realizar previsões. Pode ser considerada como o casamento entre a estatística e a inteligência artificial ao combinar heurística e análise estatística. ML é uma disciplina científica que se preocupa com o design e desenvolvimento de algoritmos que permitem aos computadores aprenderem com as bases de dados. Um dos principais objetivos da ML é automatizar o aprendizado para reconhecer padrões e tomar decisões inteligentes baseadas nos dados.

Algoritmos baseados em modelos preditivos e modelos descritivos

A mineração de dados possui como premissa básica a argumentação ativa, ou seja, em vez do usuário definir o problema, selecionar os dados e as ferramentas para analisá-los, as ferramentas de MD extraem automaticamente dos dados anomalias e possíveis associações para identificar problemas que não tinham sido visualizados pelo usuário.

> As ferramentas de mineração de dados analisam os dados, descobrem problemas ou oportunidades de melhorias nas associações entre os dados, para diagnosticar o comportamento dos negócios, requerendo a mínima intervenção do usuário.

As ferramentas de mineração são baseadas em algoritmos, responsáveis pela inserção da inteligência artificial no processo.

Os objetivos principais da mineração de dados estão inseridos no **modelo descritivo** e no **modelo preditivo**. No entanto, a escolha do modelo entre previsão e descrição varia de acordo com o sistema de mineração de dados utilizado. Os objetivos são conseguidos por meio de vários algoritmos, incorporados em vários métodos de mineração de dados.

Modelo preditivo

Representa a área de investigação de dados que busca inferir resultados a partir dos padrões encontrados na análise descritiva, ou seja, prognosticar o comportamento de um novo dado. Podemos subdividi-lo em:

1

Classificação

Prediz associações dos dados entre as classes. Por exemplo, por esse modelo, é possível predizer a probabilidade de uma pessoa ir ou não jogar tênis, responder ou não a uma solicitação, analisar se tem um baixo ou alto risco de crédito etc.

2

Predição

A regressão prediz um número — por exemplo, qual o valor do aluguel de um imóvel, qual a receita gerada pelo cliente no próximo ano ou qual a vida útil de um equipamento (número de meses antes de falhar).

3

Estimação

Prediz algum valor baseado num padrão já conhecido — por exemplo, conhecendo o padrão de despesas e a idade de uma pessoa, é possível estimar seu salário e seu número de filhos.

Os três algoritmos de modelagem preditiva mais utilizados são: Árvore de decisão, regressão e redes neurais artificiais. Vamos, agora, conhecer acada um deles.

Árvores de decisão

Os algoritmos de árvores de decisão, ou *decision trees*, são algoritmos de *machine learning* largamente utilizados, com uma estrutura de simples compreensão e que costumam apresentar bons resultados em suas previsões. Eles também são a base do funcionamento de outros algoritmos, como o Random Forest.

Em uma árvore de decisão, cada ramo representa uma escolha entre um número de alternativas e cada folha, uma classificação ou decisão. Esse modelo analisa os dados e tenta encontrar a variável que divide os dados em classes diferentes.

Imagine a utilização de uma árvore de decisão para predizer se iremos ou não à praia. Ao analisarmos a imagem, a seguir, a árvore de decisão prediz que iremos para à praia em dia de SOL e em dia sem VENTO.

Árvore de decisão Vou à praia.

Dessa maneira, podemos extrair três associações:

1. SE campo Sol com valor SIM e campo Vento com valor NÃO, ENTÃO Vou para praia.
2. SE campo Sol com valor SIM e campo Vento com valor SIM, ENTÃO Não vou para praia.
3. SE campo Sol com valor NÃO, ENTÃO Não vou para praia.

Regressão

A análise de regressão é um método estatístico que permite examinar a relação entre duas ou mais variáveis. Sendo assim, este algoritmo trabalha com dados contínuos que podem seguir uma distribuição normal, encontrando padrões essenciais em grandes bases de dados. Por exemplo, usa-se esse algoritmo quando se deseja determinar quanto cada fator específico, como o preço, influência o movimento de um ativo.

Curiosidade

Os algoritmos de regressão linear e logística são alguns dos algoritmos mais populares na disciplina de estatística.

Com a análise de regressão, é possível predizer um número, ou seja, na regressão linear a resposta será uma variável contínua. Podemos inclusive classificar em relação à quantidade de variáveis independentes utilizadas para realizar a predição, sendo estas:

Regressão linear

Uma variável independente é usada para prever o resultado.

Regressão linear múltipla

Usa duas ou mais variáveis independentes para prever o resultado.

Em resumo, o algoritmo de **regressão linear** simplesmente acha a reta que melhor se encaixa entre os pontos. Assim, é possível prever um valor de y dado um valor de x.

Por exemplo, vamos supor dados aleatórios de uma tabela sobre duas variáveis: **x** e **y**. Ao criar o par (x,y) em um gráfico, teremos a representação com os pontos vermelhos.

Para prever o valor de y caso x fosse 1, basta olhar na linha qual valor de y quando x assume o valor 1. Na imagem a seguir, y seria aproximadamente 2.5 (ponto amarelo).

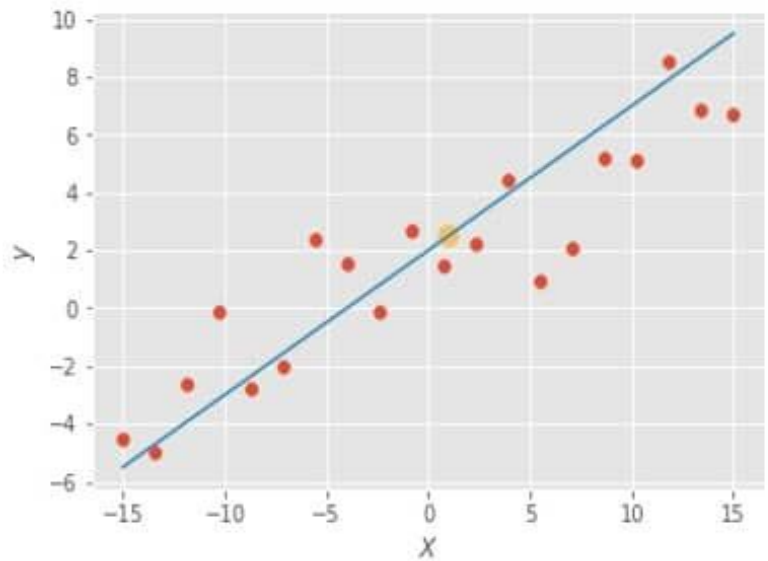


Gráfico gerado da tabela aleatória.

Na **regressão logística**, valores desconhecidos de uma variável discreta são previstos com base no valor conhecido de outras variáveis.

Ou seja, a regressão logística mede a relação entre a variável dependente categórica e uma ou mais variáveis independentes, estimando as probabilidades usando uma função logística.

Sendo assim, o algoritmo de Machine Learning analisa diferentes aspectos ou variáveis de um objeto para depois determinar uma classe na qual ele se encaixa melhor.

Há três modelos de regressão logística:

Regressão logística binominal

Os dados são classificados em dois grupos ou categorias. Por exemplo, a mensagem é *spam* ou não.

Regressão logística ordinal

Os dados são classificados em três ou mais classes em uma ordem já determinada. Por exemplo, o desempenho do atleta é ruim, justo ou excelente.

Regressão logística multinomial

Os dados são classificados em três ou mais categorias sem ordem entre si. Por exemplo, o animal é um gato, um leão ou um tigre.

Redes neurais artificiais (RNA)

São modelos mais complexos e sofisticados, capazes de modelar relações extremamente complexas.

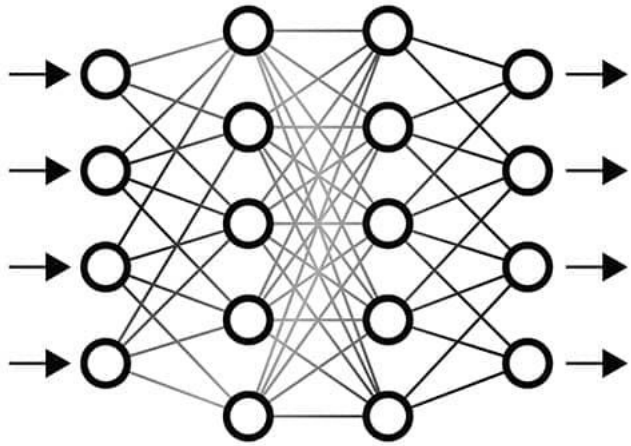
Os algoritmos de RNA apresentam um modelo matemático inspirado na estrutura neural de organismos inteligentes e adquirem conhecimento por meio da experiência.

Uma grande RNA pode ter centenas ou milhares de unidades de processamento; já o cérebro de um mamífero pode ter muitos bilhões de neurônios.

> Os algoritmos de RNA também são populares porque são poderosos e flexíveis. Seu poder vem de sua capacidade de lidar com relações não lineares em dados, o que é cada vez mais comum à medida que coletamos mais dados.

Uma RNA é composta por várias unidades de processamento, cujo funcionamento é bastante simples. Essas unidades, geralmente, são conectadas por canais de comunicação que estão associados a determinado peso. As unidades fazem operações apenas sobre seus dados locais, que são entradas recebidas pelas suas conexões. O comportamento inteligente de uma Rede Neural Artificial vem das interações entre as unidades de processamento da rede.

Arquiteturas neurais são tipicamente organizadas em camadas, com unidades que podem estar conectadas às unidades da camada posterior, conforme é visualizado a seguir:



Exemplo de arquitetura de uma rede neural.

Modelo descritivo

Este modelo representa a área de investigação de dados que busca tanto descrever fatos relevantes, não triviais e desconhecidos dos usuários, como analisar a base de dados, principalmente pelo seu aspecto de qualidade, para validar todo o processo da mineração.

O modelo consiste em estudar tudo que tem a ver com o passado. Sendo assim, é usado para descrever os eventos que ocorreram de acordo com os parâmetros e referências usados na tomada de decisão. Podemos subdividi-lo em:

Agrupamento
Os dados são agrupados de acordo com a sua similaridade.

Sumarização
O objetivo é encontrar uma descrição simples e compacta para os dados.

Associação
Deve-se encontrar padrões para associar os atributos dos dados.

Alguns algoritmos de modelagem descritiva mais utilizados são:

Expectation-Maximization (EM)

A técnica da Expectância Máxima é usada para estimar funções de máxima semelhança. Esta técnica retorna bons resultados sobre bases com dados incompletos, pois pode-se utilizar os valores aprendidos com os dados preenchidos para estimar os valores ausentes (SILVA; JÚNIOR; SILVA, 2016).

K-means

O algoritmo k-Médias possui uma implementação relativamente simples, o que ajuda a tornar seu uso mais difundido. Esta técnica utiliza centroides, que representam a instância média de um grupo. Para realizar o agrupamento, o método utiliza alguma medida de similaridade entre as instâncias. Normalmente, essa medida é a distância euclidiana (SILVA; JÚNIOR; SILVA,2016).

Hierárquico (H)

O algoritmo hierárquico, que pode ser aglomerativo ou divisivo, trabalha criando uma sequência de partições, como uma árvore (FACELI *et al.*,2015). No divisivo, as instâncias são colocadas em um único grupo e são divididas até que o número de grupos seja o desejado. Já no aglomerativo, as instâncias iniciam sendo, cada uma, um grupo e são agrupadas iterativamente até que se tenha o número de grupos desejado (SILVA; JÚNIOR; SILVA,2016).

Técnicas de mineração de dados não estruturados

Como já vimos, a mineração de dados é o processo de extrair padrões de uma grande quantidade de dados e transformá-los em conhecimento.

Os dados podem ser representados de uma forma **estruturada** ou **não estruturada**.

A forma mais comum de apresentação são os dados estruturados em tabelas, em que as colunas definem os tipos de dados e as linhas, os valores associados a cada uma de suas entradas.

A tabela a seguir mostra uma estrutura para o registro dos dados de pacientes e seus respectivos diagnósticos:

Idade	Sexo	Peso	Altura	Estado Civil	Profissão	Jornada	Intervalo	Horas	Dieta	Atividade física
12	F	40	1,58	Solteiro	Estudante	0	Sim	2	?	Sim
15	F	50	1,57	Solteiro	Estudante	0	Sim	2	Não	Não
32	F	71	1,65	Casado	Auxiliar de Limpeza	8	Sim	1	?	?
14	M	57	1,72	Solteiro	Estudante	0	Sim	2	?	?
31	M	61	1,71	Casado	Técnico Eletrônico	8	Sim	1	Não	Sim
36	M	90	1,83	Solteiro	Chaveiro	11	Não	0	Não	Sim
20	M	78	1,69	Solteiro	Estudante	0	Sim	2	?	Sim
50	M	88	1,70	Casado	Motorista	12	Sim	2	Não	Sim
43	F	58	1,70	Divorciado	Representante Comercial	44	Sim	2	?	?

Tabela: Cadastro de pacientes e seus respectivos diagnósticos.
Daisy Albuquerque.

Nesse caso, as técnicas de mineração poderiam ser aplicadas à base de dados para encontrar associações entre suas colunas e descobrir as causas do diagnóstico recebido pelo paciente.

Uma outra forma de apresentação dos dados são os textos.

A extração de conhecimento em texto, também conhecida como **mineração de texto (MT)**, é um dos assuntos mais abordados atualmente, devido ao volume desse tipo de informação circulando na Internet, ambientes empresariais e pesquisa científica.

Esses dados podem ser registrados pela interação entre os usuários, como a troca de mensagens por e-mail ou registro das conversas em um aplicativo, ou no registro de informação para o uso de sistemas, como o registro de um problema encontrado ou o detalhamento de uma configuração. Segundo Feldman e Sanger (2002):

> “O texto é definido como uma unidade de dados textuais discretos dentro de uma coleção que normalmente, mas não necessariamente, correlaciona-se com algum documento real, como, por exemplo um e-mail, um artigo científico, uma notícia.
>
> —
>
> (FELDMAN; SANGER, 2002)

Podemos observar que cada tipo de texto possui um padrão diferente de escrita. Alguns textos usam termos que não pertencem ao vocabulário formal da língua e outros apresentam um vocabulário mais específico para seus assuntos. Todos os textos, porém, passam uma mensagem e devem ser entendidos durante a interação com seus receptores.

> A mineração de textos possui o propósito de extrair padrões não triviais e relevantes para que o processamento seja computacionalmente viável, com menor perda possível de valor.

Para a mineração de dados não estruturados, são utilizadas as seguintes técnicas:

Processamento de linguagem natural (PLN)

É um ramo da inteligência artificial que auxilia as máquinas a entenderem, interpretarem e manipularem a linguagem humana.

Por exemplo, quando você fala para a Alexa (assistente virtual desenvolvida pela Amazon) que gosta de uma determinada música, o assistente virtual irá entender e executar a ação de salvar sua classificação, retornando uma mensagem de voz que simula a de um ser humano.

Recuperação de informações

É uma técnica que utiliza métodos e medidas estatísticas ou semânticas para processar textos e encontrar quais documentos possuem respostas para a questão.

Extração de informação

É a técnica que busca partes relevantes de um texto em um documento e extrai informações específicas.

Nesse contexto, podemos entender que a mineração de dados não estruturados consiste em um conjunto de técnicas e processos que descobrem conhecimento inovador em objetos como textos, criando relações entre os atributos (palavras).

Devido à sobrecarga de informação disponível, estima-se que a mineração de texto apresenta um apelo comercial maior do que a mineração de dados, o que é facilmente visualizado quando estudos mostram que a maior parte das informações de uma instituição, cerca de 80%, é armazenada em documentos de texto.

Estudo de caso de mineração de dados

No vídeo a seguir, a especialista apresenta um exemplo simples de classificação com algoritmo clássico de árvore de decisão.

Conteúdo interativo

Acesse a versão digital para assistir ao vídeo.

Verificando o aprendizado

Questão 1

A aplicação de técnicas de mineração de dados (data mining) pode ser de grande valia para uma auditoria. No caso das pesagens de caminhões em estradas, por exemplo, uma ação típica de mineração de dados, passível de ser tomada com o auxílio de instrumentos preditivos, é:

A

Quantificar as ocorrências de possíveis pesagens fraudulentas ocorridas durante todo o trimestre que antecede a data da análise, em alguns postos selecionados, mediante parâmetros comparativos preestabelecidos.

B

Analisar o percentual de ocorrências das menores permanências de caminhões nos postos, no último ano, em relação ao movimento total.

C

Relacionar os postos onde ocorreram, nos últimos seis meses, as menores permanências das empresas suspeitas e informar o escalão superior para a tomada de decisão.

D

Realizar uma abordagem surpresa em determinado posto, com probabilidade significativa de constatar ocorrência fraudulenta.

E

Reportar ao escalão superior as características gerais das pesagens e permanências de todos os caminhões, nos cinco maiores postos do Estado, no mês que antecede a data de análise.

A alternativa D está correta.

O modelo preditivo prediz uma ação de acordo com a análise dos dados passados. No caso, a opção retrata a visita surpresa ao posto, baseada em interpretações sobre os dados extraídos da base de dados.

Questão 2

Considere que uma empresa de saneamento busca realizar a gestão de recursos hídricos subterrâneos com base em parâmetros conhecidos que determinam a poluição das águas subterrâneas. Um desses parâmetros, para exemplificar, seria o nitrato, um indicador de poluição difusa de água subterrânea. Criando-se regras para realizar o aprendizado supervisionado do sistema de *data mining* e utilizando-se uma certa técnica, chegar-se-á a um resultado que considera os diversos parâmetros para se descobrir se um certo aquífero tem água potável ou não, comparando-se com uma definição conhecida. Nesse cenário, a técnica aplicada é denominada:

A

Associação.

B

Classificação.

C

Clustering.

D

Regressão

E

Prediction.

A alternativa B está correta.

O modelo preditivo pode ser subdividido em classificação, estimação e predição. A classificação prediz associações dos dados entre as classes. Nesse modelo, é possível predizer a probabilidade de a água ser potável ou não baseados nas definições já conhecidas.

Considerações finais

Ao longo do tempo, as empresas se desenvolveram e se tornaram mais diversificadas e, consequentemente, aconteceu um crescimento da exigência da relevância da informação.

As operações e atividades desenvolvidas pelas empresas nos últimos anos, tanto no setor privado quanto no público, são registradas computacionalmente e, dessa maneira, vem sendo gerada uma ampla base de dados.

Um dos grandes desafios é a geração de informações úteis e ágeis para a empresa, partindo da ampla base de dados que foi criada e alimentada pelas atividades e operações da empresa.

A necessidade de manipular a base de dados, gerenciando, recuperando e analisando os dados, se torna relevante para a prática de gestões estratégicas de conhecimento. Dessa forma, surge a necessidade da utilização ferramentas de auxílio que permitam otimizar essa busca por informações mais concretas e que sejam de real valia para a análise das atividades da organização.

Nesse contexto, este tema apresentou como o processo de KDD e a mineração de dados, em conjunto com o aprendizado de máquina, se enquadram ofertando suas técnicas e ferramentas no auxílio da geração e descoberta de conhecimento em bases de dados estruturadas e não estruturadas.

Podcast

Ouça o podcast com a especialista Daisy Albuquerque. Ela traz um resumo do tema mineração de dados, abordando a sua aplicação pelo profissional de Ciência de Dados.

Conteúdo interativo

Acesse a versão digital para ouvir o áudio.

Explore +

Para aprimorar seus conhecimentos sobre o conteúdo abordado:

- Visite o canal **Programação Dinâmica**, produzido pelos cientistas de dados Kizzy Terra e Hallison Paz, e que traz o tema de mineração de dados nos seus vídeos de uma forma descontraída, além de outros temas afins como a inteligência artificial e *machine learning*, promovendo discussões sobre novas tecnologias e seus impactos na sociedade.
- Veja a pesquisa do IDC (International Data Corporation) sobre a quantidade de dados produzida e replicada no mundo.
- Saiba mais sobre a arquitetura da primeira rede neural profunda conhecida, treinada, em 1965, por Alexey Grigorevich Ivakhnenko.

- Conheça o software WEKA (Waikato Environment for Knowledge Analysis), desenvolvido para resolver problemas de mineração de dados e aprendizado de máquina. Disponível no site da Universidade de Waikato, Nova Zelândia.

Referências

ADRIAANS, P.; ZANTINGE, D. **Data Mining**. Boston: Addison-Wesley, 1996.

CARVALHO, D. R.; DALLAGASSA, M. R. **Mineração de dados**: aplicações, ferramentas, tipos de aprendizado e outros subtemas. *In*: AtoZ: novas práticas em informação e conhecimento, v. 3, n. 2, p. 82–86, 31 dez. 2014. Consultado na internet em: 3 fev. 2021.

BHAVANI T. **Data Mining**. Boca Raton: CRC Press, 1999.

DIAS, M. M. **Parâmetros na escolha de técnicas e ferramentas de Mineração de Dados**. *In*: Acta Scientiarum, v. 24, n. 6, p. 1715, Maringá, 2002. Consultado na internet em: 3 fev. 2021.

ELMASRI, R.; NAVATHE, S. B. **Sistemas de banco de dados**. 7. ed. São Paulo: Pearson Addison Wesley, 2019.

FAYYAD, U. M.; PIATETSKY-SHAPIRO, G.; SMYTH, P. **From data mining to knowledge discovery in databases**. *In*: AI Magazine 17(3), p. 37-54, 1996. Consultado em na internet em: 3 fev. 2021.

FAYYAD, U.; PIATETSKY S. G.; SMYTH, P. **The KDD process for extracting useful knowledge from volumes of data**. *In*: Communications of the ACM, v. 39, n. 11, p. 27-35, 1996. Consultado na internet em: 3 fev. 2021.

FELDMAN, R.; SANGER, J. (2002), **The text mining handbook**: advanced approaches in analyzing unstructured data. 1. ed. Cambridge: Cambridge University Press, 2006.

GOLDSCHMIDT, R.; BEZERRA, E.; PASSOS, E. **Data mining**: conceitos, técnicas, algoritmos, orientações e aplicações**.** Rio de Janeiro: Elsevier, 2015.

HAN, J.; KAMBER, M. **Data mining**: concepts and techniques. 2. ed. San Francisco: Morgan Kaufmann, 2006.

INMON, W. H. **Como construir um data warehouse**. 2. ed. Rio de Janeiro: Campus, 1997.

JESUS, Mena. **Data mining your website**. Digital Press, 1999.

JIAWEI, H.; MICHELINE K., **Data mining – concepts and techniques**. Burlingnton: Morgan Kaufmann Publishers, 2001.

MICHAEL J. A. B.; GORDON, L. **Data mining techiques for marketing, sales, and customer support**. Nova Jersey: John Wiley & Sons, 1997.

O'BRIEN, J. A. **Sistemas de Informação e as decisões gerenciais na era da Internet**. 2. ed. São Paulo: Saraiva, 2004.

REZENDE, S. O. **Sistemas inteligentes** – fundamentos e aplicações. Barueri: Manole, 2005.

SILVA, A. R. da; JÚNIOR, J.; SILVA, I. **Rede de sensores para controle inteligente de ambientes**. *In*: 8 o SBCUP-Simpósio Brasileiro de Computação ao Ubıqua e Pervasiva — XXXVI CSBC —Congresso da Sociedade Brasileira

de Computação. [S.l.: s.n.], 2016. Consultado na internet em: 3 fev. 2021.

SHOLOM M. W. NITIM, I. **Predict data mining**, Burlingnton: Morgan Kaufmann Publishers, 1999.

WITTEN, I. H.; FRANK, E.; HALL, M. A.; PAL, C. J. **Data mining**: Practical machine learning tools and techniques. 3. ed. San Francisco: Morgan Kaufmann, 2011.