UNIVERSIDADE
FEDERAL
DE PERNAMBUCO

Centro de Informática
U·F·P·E

Mineração de Dados

Isadora Barros Soares

Disciplina: Banco de Dados

Professores: Ana Carolina Salgado e Fernando Fonseca

Contexto

- Atualmente é gerado um enorme volume de dados a cada instante
- Praticamente todo sistema automatizado gera alguma forma de dado para fins de análise ou diagnóstico.
  - World Wide Web: documentos, logs;
  - Operações financeiras;
  - Tecnologias de sensores, Internet of Things.

Justificativa

- Devido à grande quantidade de dados armazenados a cada dia, sua recuperação por si só não traz todos os benefícios que se deseja.
- A análise e interpretação manual de todo esse volume de dados atual seria ineficiente e praticamente inviável.
- É necessário interpretar esses dados e procurar os padrões que sejam de interesse para cada circunstância.

Justificativa

- Utilizações da mineração de dados
  - Tomada de decisões estratégicas
  - Geração de riquezas
  - Análise de tendências (como consumidores ou o mercado se comportam)
  - Segurança (padrões que ligam a atividades anormais na rede ou no sistema)

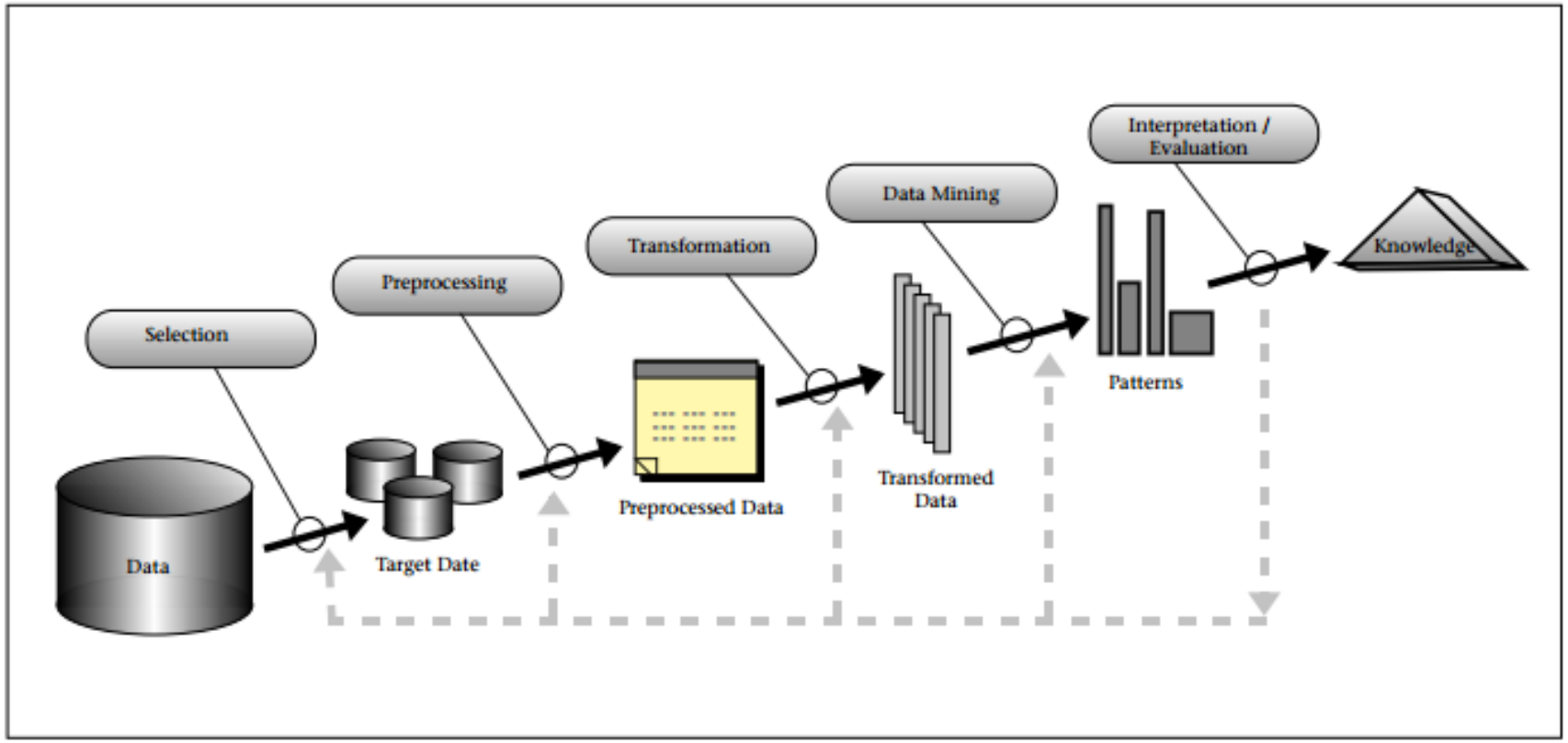
Descoberta de Conhecimento em Bancos de Dados (KDD)

- Processo de descoberta de conhecimento válido e útil a partir dos dados
- Seleção dos dados;
  - Podem ser selecionados dados sobre categorias específicas
- Limpeza dos dados;
  - Remoção de ruídos e redundâncias
  - Ex: eliminação de telefones com prefixos incorretos
- Enriquecimento dos dados;
  - Melhora os dados com fontes de informações adicionais

Descoberta de Conhecimento em Bancos de Dados (KDD)

- Transformação dos dados;
  - Ex: os códigos dos itens podem ser agrupados em termos de categorias dos produtos (áudio, vídeo, acessórios, ...)
- Mineração dos dados;
  - Aplicação de algoritmos específicos para extrair padrões dos dados
  - Proverá informações desconhecidas até então
  - Baseia-se nos critérios de interesse
- Relatórios sobre as informações descobertas.

Descoberta de Conhecimento em Bancos de Dados (KDD)



Descoberta de conhecimento

- Regras de associação
  - Correlacionam a presença de um conjunto de itens com outra faixa de valores para outro conjunto de variáveis.
  - Exemplo: Pessoas que compram uniformes escolares em junho também compram mochilas escolares

Descoberta de conhecimento

- Classificação
  - Criar hierarquia de classes com base em um conjunto existente de eventos ou transações
  - Exemplo: dividir a população em categorias, com base em seu histórico de transações de crédito.

Descoberta de conhecimento

- Padrões sequenciais
  - Associações entre eventos com relacionamentos temporais
  - Exemplo:
    - Se um paciente teve uma parada cardíaca e depois passou a ter alta concentração de ureia no sangue, então é provável que sofra de problemas renais nos próximos 18 meses.

Descoberta de conhecimento

- Padrões em séries temporais
  - Sequências de dados obtidos em intervalos regulares
  - Detecção de semelhanças entre esses dados
  - Exemplo:
    - Dois produtos possuem o mesmo padrão de venda no verão, mas um diferente no inverno.

Descoberta de conhecimento

- Agrupamentos
  - Segmentação de eventos em conjuntos de elementos semelhantes.
  - Exemplo:
    - Um conjunto de pacientes podem ser divididos em grupos, com base nas semelhanças dos efeitos colaterais apresentados.

Regras de associação

- Regra de associação
  - $X \Rightarrow Y$, onde $X = \{x_1, x_2, ..., x_n\}$, e $Y = \{y_1, y_2, ..., y_m\}$ são conjuntos de itens, com x_i e y_j sendo itens distintos para todo i e todo j.
- Modelo cesta de mercado
  - A cesta de mercado corresponde ao conjunto de itens que um cliente compra em uma visita ao supermercado.
  - Se um cliente compra X, é provável que também compre Y.

Regras de associação

- Em geral, tem a forma LHS (left-hand side) => RHS (right-hand side)
  - LHS e RHS são conjuntos de itens
- O conjunto LHS U RHS é chamado de itemset
- A regra deve satisfazer alguma medida de interesse
  - Duas medidas comuns são: Suporte e Confiança

Regras de associação

- Suporte
  - O suporte de uma regra LHS => RHS está relacionado ao itemset.
  - Qual a frequência do itemset no banco de dados
  - Porcentagem de transações que contêm todos os itens do itemset LHS U RHS
- Confiança
  - Está relacionada à implicação presente na regra
  - Calculada como: suporte(LHS U RHS)/suporte(LHS)
  - Probabilidade de RHS ser comprado, dado que LHS está sendo comprado

Regras de associação

ID da transação	Horário	Itens comprados
101	6:35	Leite, pão, cookies, suco
792	7:38	Leite, suco
1130	8:05	Leite, ovos
1735	8:40	Pão, cookies, café

- Regra leite => suco
  - Suporte de {leite, suco} = 50%
  - Confiança da regra = 66,7%
- Regra pão => suco
  - Suporte de {pão, suco} = 25%
  - Confiança da regra = 50%

Regras de associação

- O objetivo é gerar o máximo possível de regras que:
  - Excedam limiares mínimos de suporte e confiança especificados pelo usuário
- Gerar todos os itemsets com um suporte superior ao limiar
  - São chamados de itemsets grandes ou frequentes
- Para cada itemset frequente, as regras devem ter um mínimo de confiança:
  - Para um itemset frequente X e $Y \subset X$, seja $Z = X - Y$
  - Se suporte(X)/suporte(Z) > confiança minima
    - então a regra Z => Y (ou seja, X – Y => Y) é valida

Regras de associação

- Gerar as regras utilizando itemsets frequentes é relativamente simples
- O problema é descobrir todos os itemsets frequentes, com seus valores de suporte, caso a cardinalidade do conjunto de itens seja muito alta.
- Para reduzir o espaço de busca combinatorial, os algoritmos para encontrar regras de associação utilizam os seguintes algoritmos:
  - Fechamento para baixo
  - Antimonotonicidade

Regras de associação

- Fechamento para baixo (downwards closure)
  - Um subconjunto de um itemset grande também deve ser grande
  - Ou seja, cada subconjunto excede o suporte mínimo necessário
- Antimonotonicidade
  - Um superconjunto de um itemset pequeno também é pequeno
  - Ajuda a reduzir o espaço de soluções possíveis.

Regras de associação

- Algoritmo Apriori
  - Utiliza as propriedades de fechamento para baixo e antimonotonicidade.
  - Calcula os valores de suporte dos itemsets candidatos com diferentes quantidades de itens para cada iteração.
  - O suporte deve ser superior ao suporte mínimo estabelecido.

ID da transação	Horário	Itens comprados
101	6:35	Leite, pão, cookies, suco
792	7:38	Leite, suco
1130	8:05	Leite, ovos
1735	8:40	Pão, cookies, café

- Suporte mínimo = 0.5
- 1-itemsets
  - C1 = {{leite}, {pão}, {suco}, {cookies}, {ovos}, {café}}
  - Valores de suporte: 0.75, 0.5, 0.5, 0.5, 0.25, e 0.25
  - L1 = {{leite}, {pão}, {suco}, {cookies}}
- 2-itemsets
  - C2 = {{leite, pão}, {leite, suco}, {pão, suco}, {leite, cookies}, {pão, cookies}, {suco, cookies}}
  - Valores de suporte: 0.25, 0.5, 0.25, 0.25, 0.5, e 0.25
  - L2 = {{leite, suco}, {pão, cookies}}

Regras de associação

- Algoritmo de amostragem
  - Selecionar uma pequena amostra das transações, que caiba na memória, e determinar os itemsets frequentes daquela amostra.
  - Se os itemsets frequentes da amostra, é calculado o suporte exato para todo o banco de dados.
  - O superconjunto de itemsets frequentes pode ser encontrado usando o algoritmo Apriori com um suporte mínimo reduzido.

Regras de associação

- Algoritmo de amostragem
  - Borda negativa: Para decidir se items frequentes foram perdidos.
  - A borda negativa de um conjunto de itemsets frequentes contém os itemsets mais próximos que também poderiam ser frequentes.
    - X não está contido no conjunto de subsets frequentes
    - Todos os subconjuntos de X estão contidos no conjunto de subsets frequentes
    - X está na borda negativa

Regras de associação

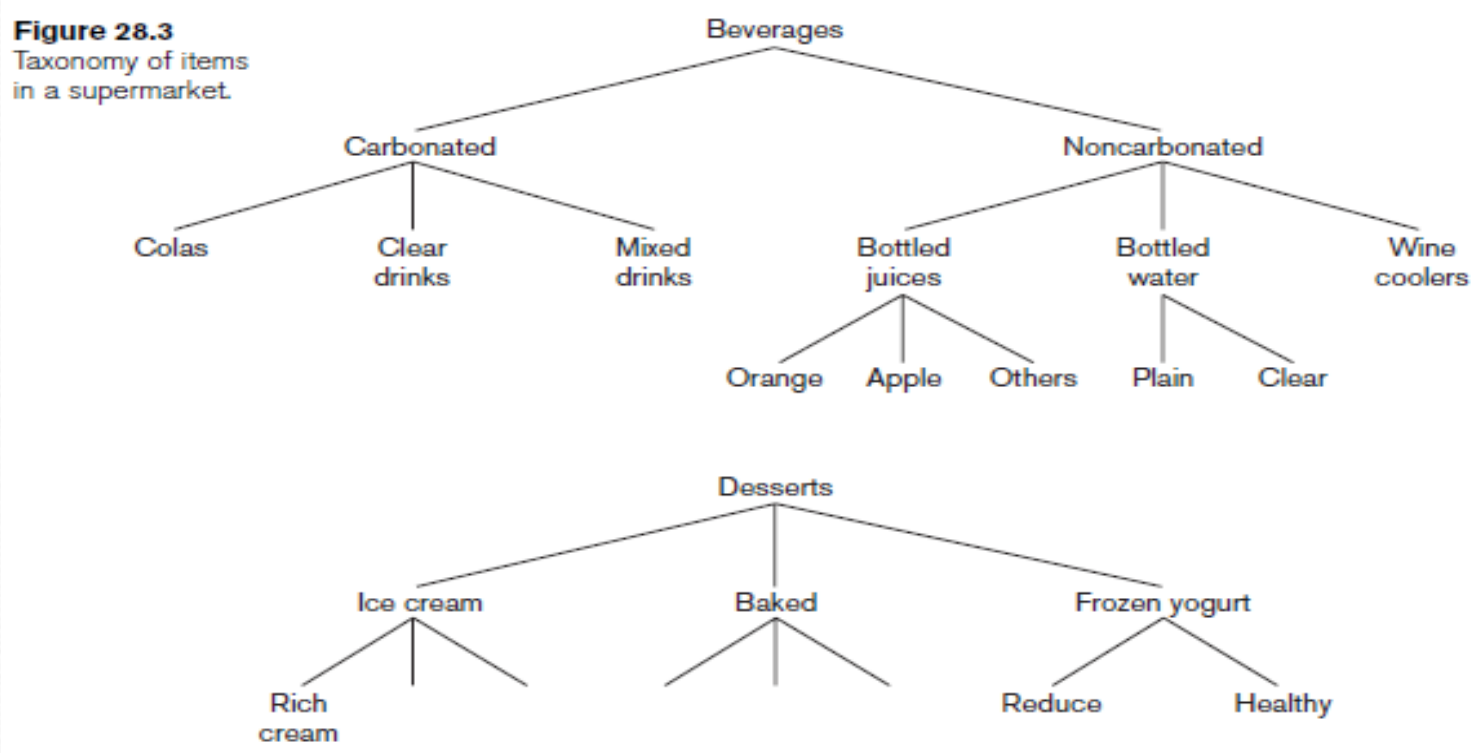
- Algoritmo de amostragem
  - O suporte da borda negativa é calculado quando o restante do banco é escaneado
  - Se um itemset que está na borda negativa pertencer ao conjunto de itemsets frequentes:
    - É possível que um superconjunto de X também seja frequente
    - É necessário analisar o banco de dados outra vez para garantir que todos os itemsets frequentes foram encontrados.

Regras de associação

- Algoritmo de amostragem
  - Considere o conjunto de itens I = {A, B, C, D, E}
  - Sejam os itemsets frequentes de tamanho 1 a 3:
    - S = {{A}, {B}, {C}, {D}, {AB}, {AC}, {BC}, {AD}, {CD}, {ABC}}
  - A borda negative será {{E}, {BD}, {ACD}}.
    - {E} é o único 1-itemset não contido em S
    - {BD} é o único 2-itemset que não está em S, cujo todos subconjuntos 1-itemset estão
    - {ACD} é o único 3-itemset que não está em S, cujo todos subconjuntos 2-itemset estão

Regras de associação

- Regras de associação entre hierarquias
  - Tipicamente é possível dividir itens entre hierarquias disjuntas com base na natureza do domínio.
  - Nesse caso, as associações de interesse são entre hierarquias, e não dentro da mesma hierarquia.



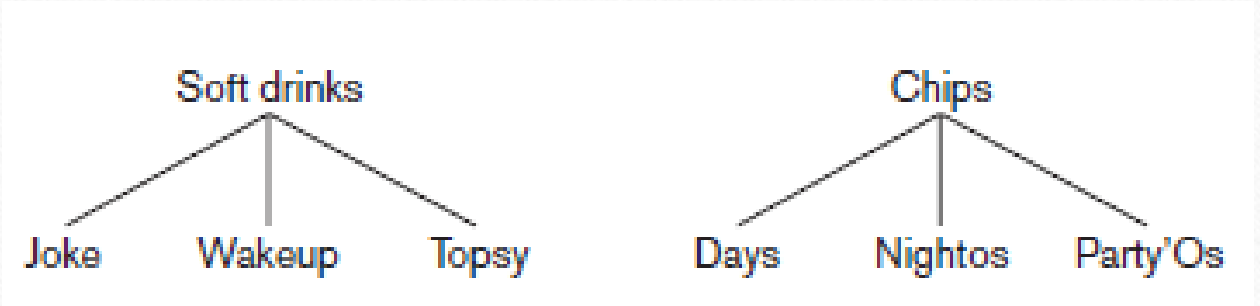
- Associações do tipo Frozen Yogurt Healthy => Água Plain podem produzir confiança e suporte suficientes para serem regras de associação de interesse válidas.

Regras de associação

- Associações multidimensionais
  - Os algoritmos mostrados até o momento envolviam apenas uma dimensão: Item comprado
    - Item_comprado(leite) => item_comprado(suco)
  - Pode ser interessante obter regras de associação que envolvam mais de uma dimensão
    - horário(6:30-8:00) => item_comprado(leite)
  - As dimensões representam atributos, que podem ser categóricos ou quantitativos.
    - Para aplicar o algoritmo apriori, pode-se transformar os atributos quantitativos em categóricos, atrelando nomes a intervalos não sobrepostos.

Regras de associação

- Associações negativas
  - É mais complicado do que descobrir uma associação positiva.
  - A maioria das combinações entre itens provavelmente não aparece no banco de dados
    - Se considerarmos que todas essas combinações significam uma associação negativa, poderemos ter milhões de associações com RHS sem interesse algum.
  - O problema é encontrar as associações negativas interessantes.
  - Uma abordagem é utilizar hierarquias, com base no conhecimento sobre os itens.



- Sabe-se que há uma forte associação entre Soft drinks e Chips
  - Se for grande o suporte para a compra de chips Daisy e bebidas Topsy, mas não para as bebidas Wakeup e Topsy, isso pode ser interessante.
  - Como são bebidas semelhantes, essa associação negativa pode trazer algum valor para análise posterior.
- O escopo é limitado pelo conhecimento sobre as hierarquias dos itens
  - O crescimento de associações negativas é exponencial

Classificação

- Processo de aprendizagem de um modelo que descreve diferentes classes de dados
- Os dados são particionados com base em uma amostra de treinamento pré-classificada
- As classes são definidas previamente
- Exemplo: classificar os clientes do banco conforme o histórico de pagamento
- É uma forma de aprendizagem supervisionada

Classificação

- Primeiramente utiliza-se um conjunto de dados já classificados para o treinamento.
- Cada registro deve conter um atributo indicando a classe à qual pertence
- O modelo é produzido geralmente na forma de uma árvore de decisão ou de um conjunto de regras.

Classificação

- Árvore de decisão
  - É uma representação gráfica simples da descrição de cada classe
  - Cada nó interno possui duas ou mais ramificações, com base nos possíveis valores dos atributos
  - Cada nó folha possui uma classe associada a si

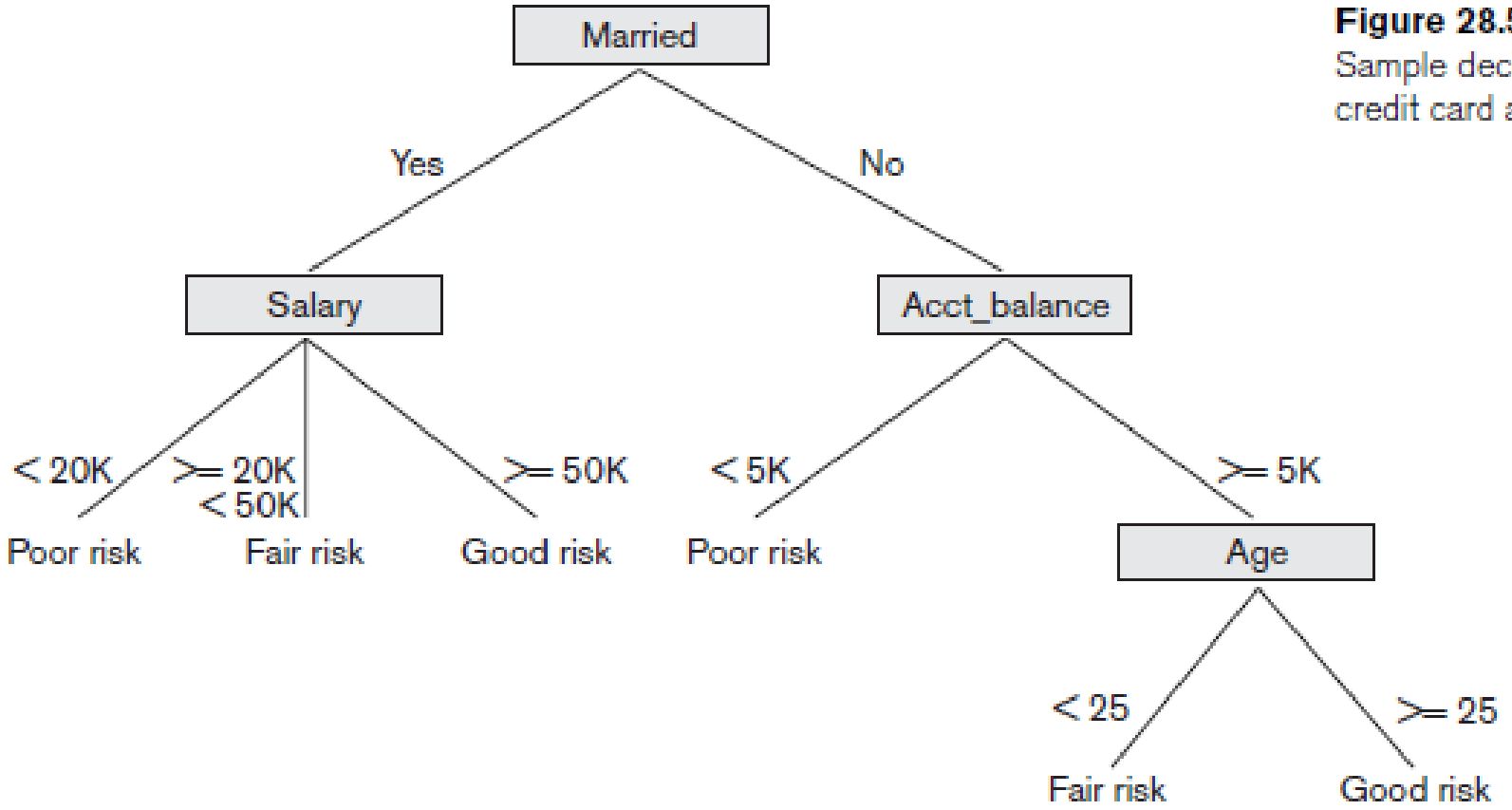


Figure 28.5
Sample decision tree for credit card applications.

Classificação

- Redes Neurais
  - Guiada por uma amostra de testes usada para a inferência e aprendizagem inicial.
  - Classificadas em duas categorias: supervisionadas e não supervisionadas
  - São auto adaptativas, ou seja, aprendem com as informações sobre um problema específico.
  - Têm um bom desempenho em tarefas de classificação
  - Desvantagem: Suas saídas são altamente quantitativas e difíceis de compreender

Agrupamento

- Também conhecido como clusterização
- Útil para particionar dados quando não há uma amostra de treinamento
- Forma de aprendizagem não-supervisionada
- Objetivo: dividir os registros em grupos, de forma que a semelhança entre os membros de um cluster é maior que a semelhança entre membros de diferentes clusters.
- Os grupos geralmente são disjuntos

Agrupamento

- Quando os dados são numéricos, geralmente é usada uma função de similaridade baseada em distância.
  - Quanto menor for a distância entre dois dados, maior a similaridade entre eles.
- Diferentes escolhas de atributos, medidas de proximidade, critérios de agrupamento e algoritmos de clusterização levam a resultados diferentes.
- Aplicações
  - Negócios: determinar grupos de consumidores com padrões de consumo similares
  - Medicina: determinar grupos de pacientes com reações aos medicamentos prescritos similares.

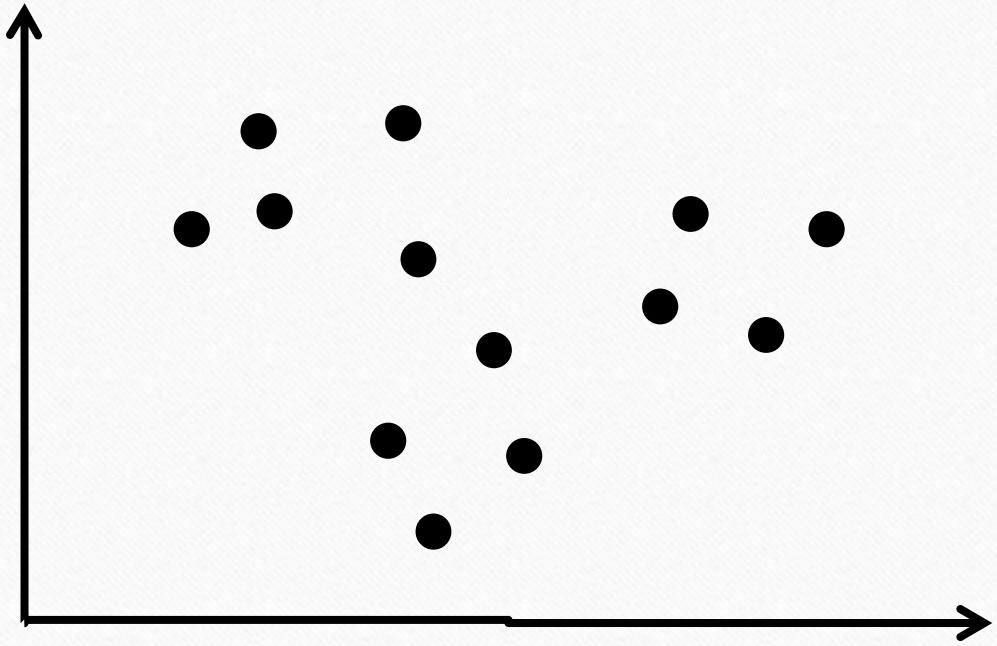
Agrupamento

- K-means
  - É um algoritmo clássico de agrupamento
  - Inicialmente escolhe de forma randômica k registros para representar os centroides (means), m_1, m..., m_k, dos clusters C_1, ..., C_k.
  - Todos os registros são posicionados em determinado cluster, com base na distância entre o registro e o centroide do cluster.
    - Se a distância entre m_i e o registro r_j for a menor entre todos os centroides, então o registro r_j fará parte do cluster C_i.

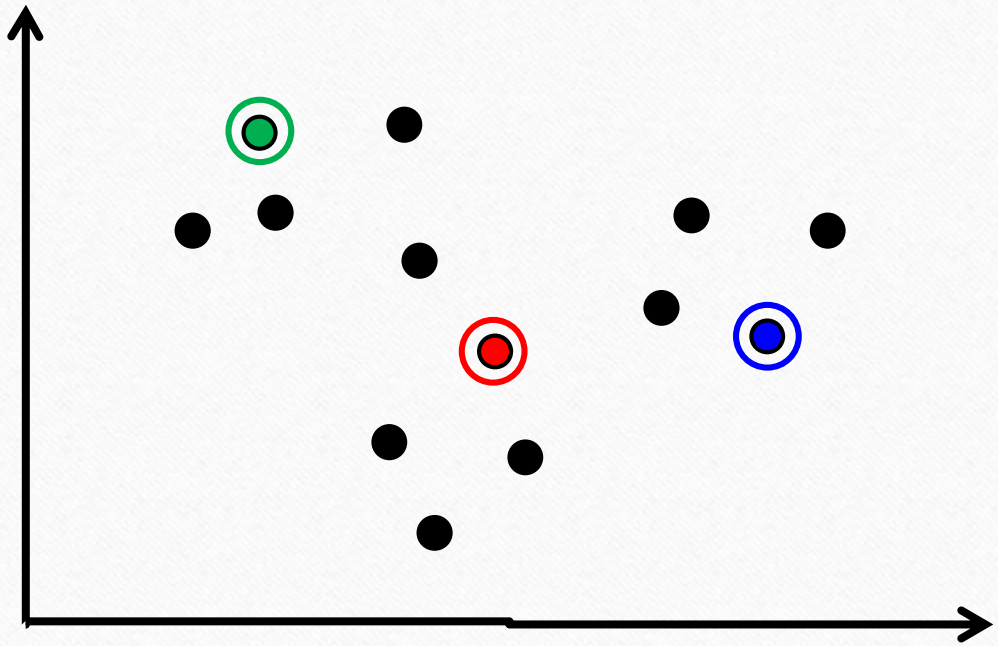
Agrupamento

- K-means
  - Uma vez que todos os registros forem colocados em um cluster inicialmente, o centroide para cada cluster é recalculado.
  - Então o processo se repete, avaliando cada registro novamente e posicionando no cluster com centroide mais próximo.
  - Diversas iterações podem ser necessárias até o algoritmo convergir
  - Seu principal problema é a dependência de uma boa inicialização

- K-means

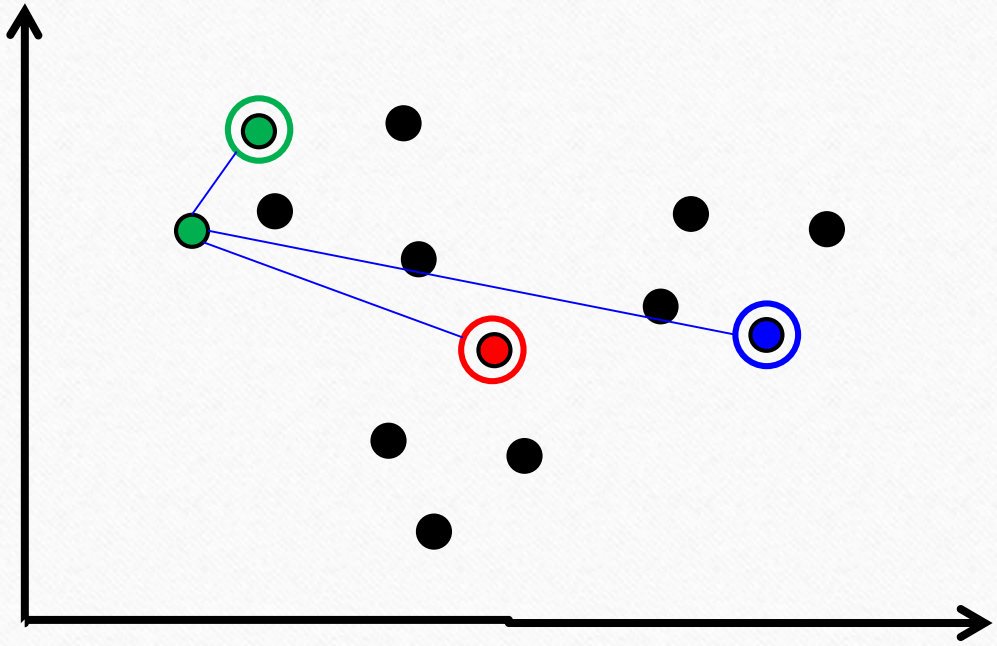


- K-means

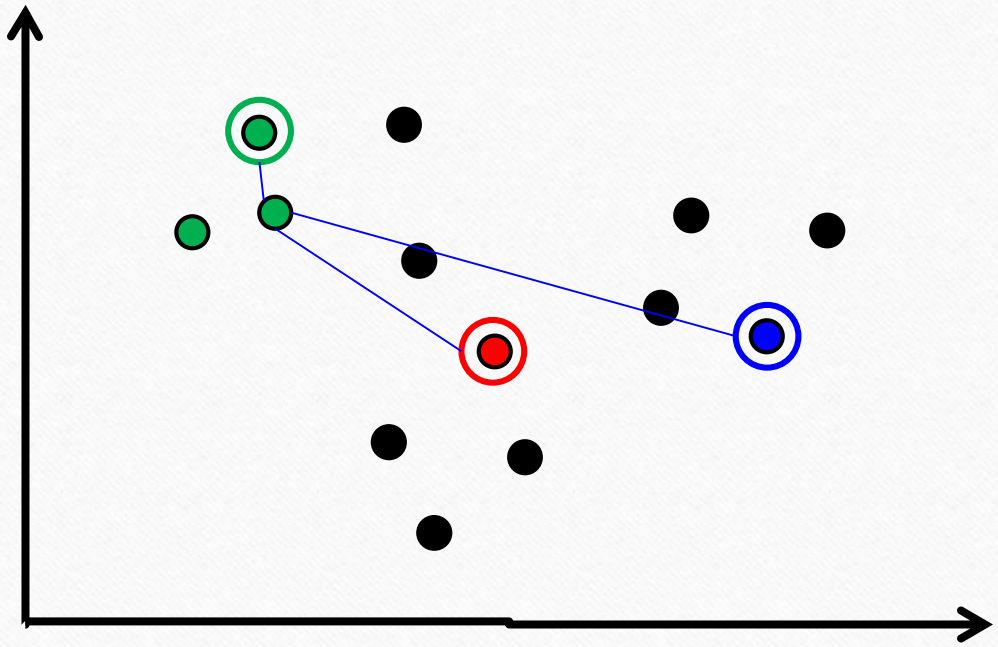


- Selecionam-se os k centroides iniciais (k = 3)

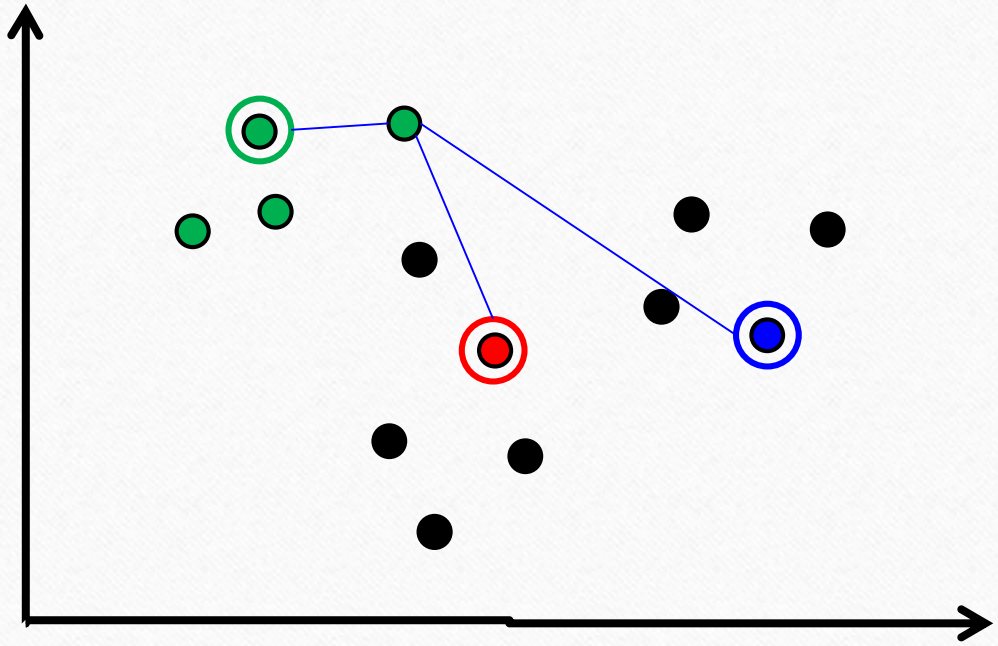
- K-means



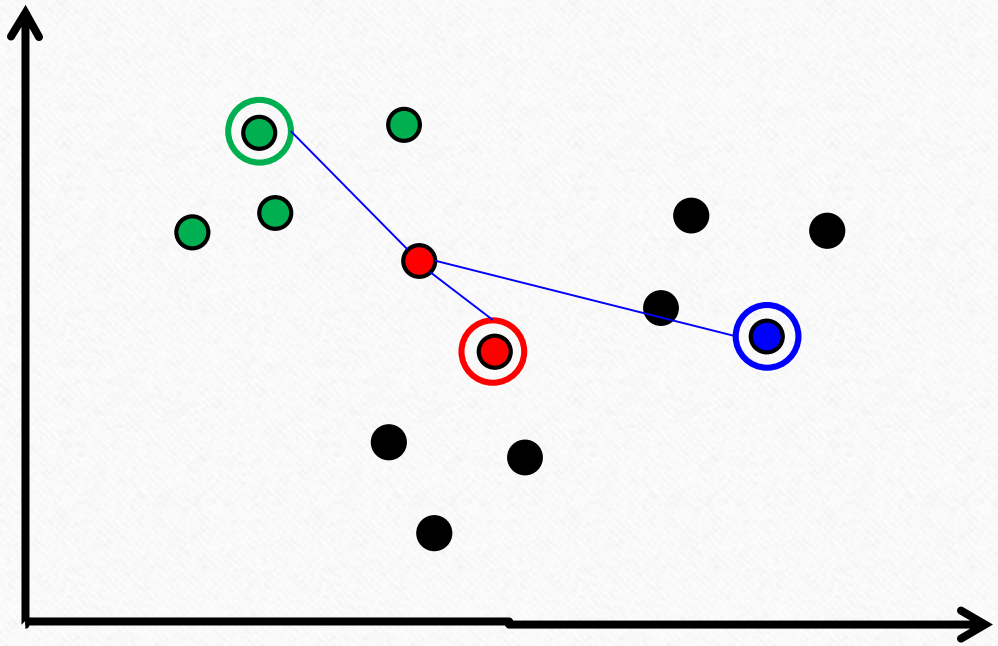
- K-means



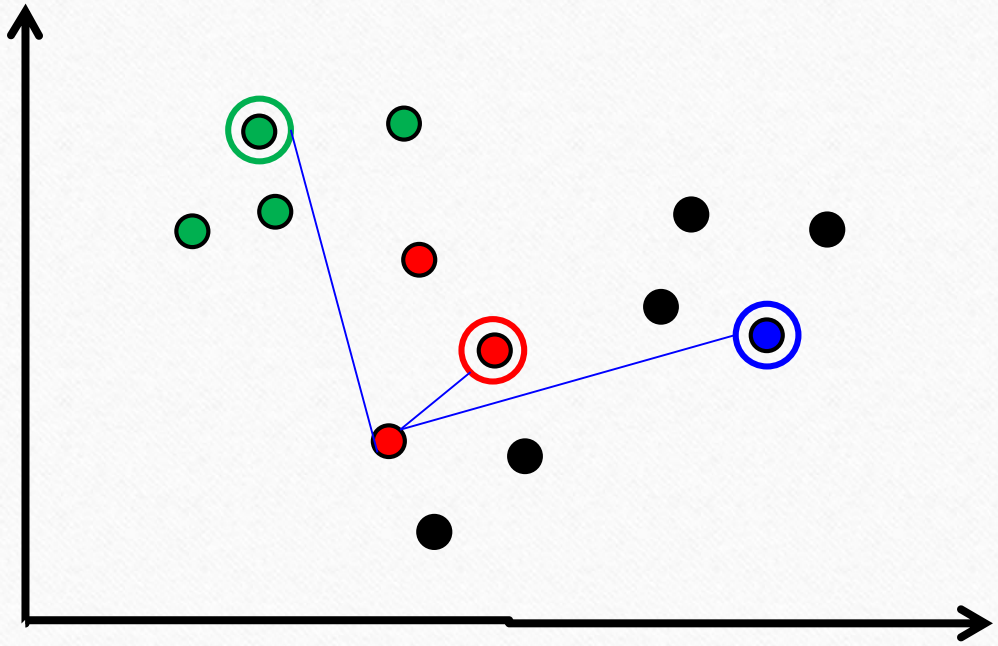
- K-means



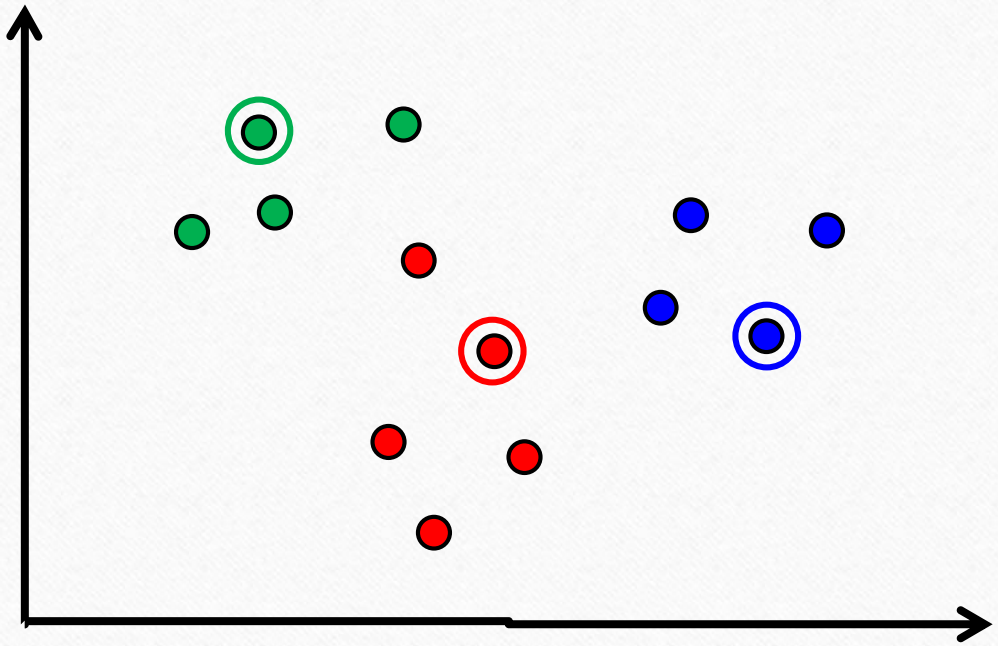
- K-means



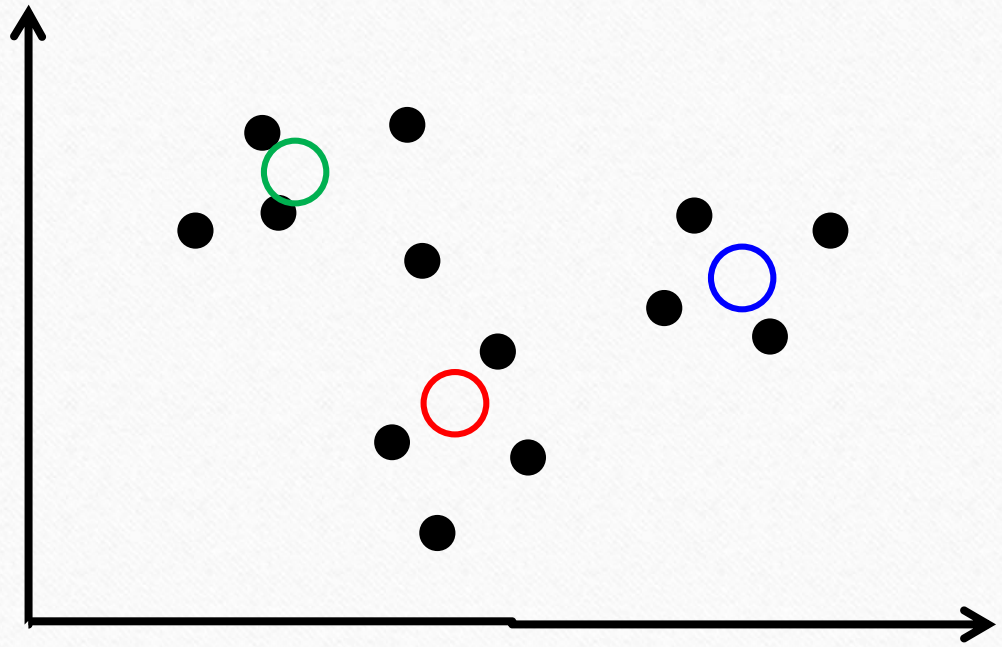
- K-means



- K-means



- K-means



- Repetem-se os passos anteriores até os centroides convergirem

Descoberta de padrões sequenciais

- Baseada no conceito de uma sequência de itemsets
- As transações são ordenadas por um valor temporal
  - Exemplo: {leite, pão, suco}, {pão, ovos}, {cookies, leite, café} pode ser uma sequência de itemsets baseada em três compras do mesmo consumidor.
- A sequência S1, S2, S3, ..., é um indicador do fato de que um cliente que compra o itemset S1 provavelmente comprará o itemset S2, depois o S3, e assim por diante.
- Essa predição é baseada na frequência (suporte) dessa sequência no passado.

Descoberta de padrões em séries temporais

- Séries temporais são sequências de eventos.
- Por exemplo, o preço de fechamento de uma ação é um evento que ocorre todo dia útil para cada ação.
- A sequência desses valores por ação constitui uma série temporal.
- Aplicações:
  - Encontrar o período durante o qual uma ação fechou em alta por n dias;
  - Encontrar o trimestre durante o qual a ação teve maior porcentagem de lucro ou perda

Aplicações

- Marketing
  - Análise de comportamento do consumidor com base em padrões de compra;
  - Definição de estratégias de marketing incluindo publicidade, localização das lojas e mensagens direcionadas
- Finanças
  - Análise da credibilidade de clientes;
  - Análise de desempenho de investimentos financeiros, como ações e títulos;
  - Avaliação de opções financeiras;
  - Detecção de fraudes.

Aplicações

- Indústria
  - Otimização de recursos como máquinas, mão-de-obra e materiais;
  - Design de processos de produção e de produtos conforme os requisitos dos consumidores.
- Saúde
  - Descoberta de padrões em radiografias;
  - Análise de efeitos colaterais de medicamentos e da eficácia de certos tratamentos;
  - Relacionamento entre o estado do paciente e as qualificações do médico.

NETFLIX

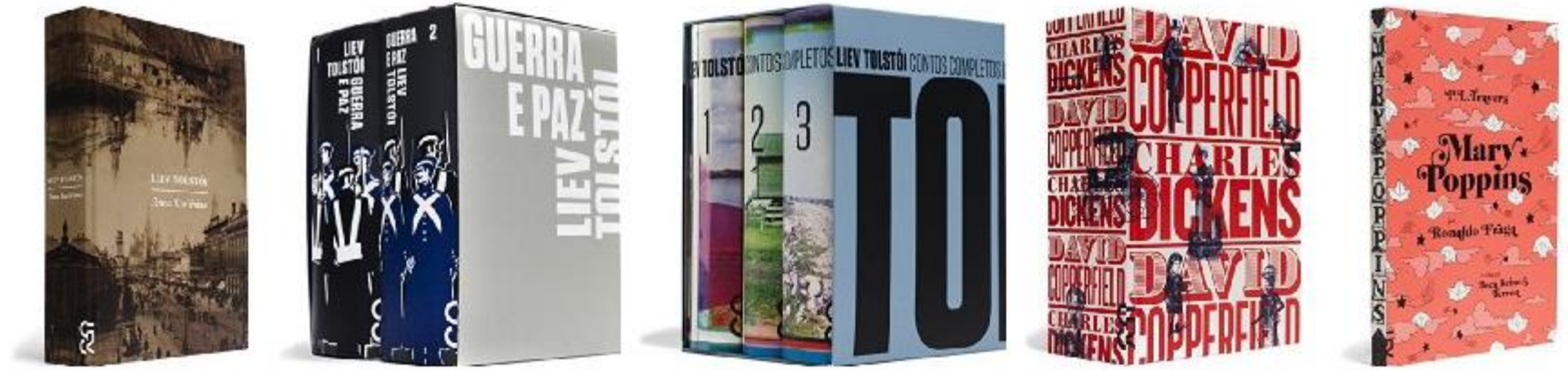
Porque você assistiu a Marvel - Demolidor

Assistir novamente ›

Lançamentos

amazon.com.br

Sugerido com base em seu histórico de navegação Veja mais



Mais Vendidos em Livros Veja mais

MIROCULUS

Desafios

- Grandes volumes de dados, que exigem mecanismos rápidos e eficazes
- Fontes de dados e tipos de dados heterogêneos, que precisam ser integrados
  - Comunicação com diferentes tipos de dispositivos e sistemas
- Dados incompletos e com ruídos
- Questões éticas

Tecnologias

- R
  - Linguagem de programação e ambiente de software projetado para computação e visualização estatística
- Weka
  - Coleção de algoritmos de aprendizagem de máquina escritos em Java, para aplicação em mineração de dados
- Orange
  - Suíte baseada em componentes para mineração de dados e aprendizagem de máquina escrita em Python

Referências

AGGARWAL, Charu C. **Data mining: The textbook**. Springer, 2015.

FAYYAD, Usama; PIATETSKY-SHAPIRO, Gregory; SMYTH, Padhraic. From data mining to knowledge discovery in databases. **AI magazine**, v. 17, n. 3, p. 37, 1996.

HAND, David J.; MANNILA, Heikki; SMYTH, Padhraic. **Principles of data mining**. MIT press, 2001.

NAVEGA, Sergio. Princípios essenciais do data mining. **Anais do Infoimagem**, 2002.

REZENDE, Solange O. et al. Mineração de dados. **Sistemas inteligentes: fundamentos e aplicações**, v. 1, p. 307-335, 2003.

TSAI, Chun-Wei et al. Data mining for internet of things: a survey.**Communications Surveys & Tutorials, IEEE**, v. 16, n. 1, p. 77-97, 2014.

WITTEN, Ian H.; FRANK, Eibe. **Data Mining: Practical machine learning tools and techniques**. Morgan Kaufmann, 2005.

Dúvidas?