AULA 05: Data Mining

1. INTRODUÇÃO

Saudações caros(as) amigos(as),

Hoje vamos à nossa sexta e última aula de **Conhecimentos de Banco de Dados**, encarando o tema Mineração de Dados (Data Mining). Conforme já cheguei a comentar com vocês, trata-se de um assunto de certa forma avançado em Computação, normalmente visto em cursos de pós-graduação. Mas como foi colocado no edital, vamos encará-lo de frente.

Assim como tive dificuldades de delimitar o assunto de Banco de Dados, em Mineração de Dados o problema se repete, e com um agravante. Procurei exaustivamente questões da ESAF sobre Mineração, e

sabem quantas encontrei? Só uma única e mísera questão! Então, fica difícil saber como a Banca vai cobrar esses conceitos. Teremos na aula questões de outras Bancas, principalmente da FCC.

Pelo que pude observar no geral, o assunto deverá ser cobrado em termos de conceitos, quase uma decoreba. É bem provável que a Banca busque observar se vocês sabem do que tratam os conceitos. Então, minha aposta fica na parte conceitual de Mineração de Dados.

Vocês verão que por detrás do assunto Mineração de Dados existem conceitos e cálculos pesados, levando em consideração a teoria da probabilidade, algoritmos genéticos, inteligência artificial e por ai vai. Por óbvio, não vamos mergulhar por esse lado, isso iria requerer um curso bem longo, e pré-requisitos que não vimos por aqui. Bem, vocês vão ver que a parte conceitual não é difícil, e apesar dessa aula estar disponível em data próxima da prova, uma ou duas leituras e entendimento dos principais conceitos devem ser suficientes para responder às questões.

Nos meus tradicionais chutes, acredito que no máximo duas questões saiam desse assunto para a prova. Não desanimem, acredito que nosso curso como um todo vai ser suficiente para vocês terem um bom desempenho nas 10 questões de Conhecimentos de Banco de Dados.

Então, vamos trabalhar.

2. Fundamentos

A área de Mineração de Dados é colocada dentro da grande área de Banco de Dados pela maioria dos autores. Na verdade, Mineração de Dados está dentro de uma área maior, conhecida com KDD (Knowledge Discovery in Databases), ou seja, Descoberta de Conhecimento em Banco de Dados. Atualmente, existe uma tendência de considerar o KDD uma área a parte dentro da Ciência da Computação.

Como já tivemos oportunidade de comentar, a área de Banco de Dados é bem madura, bem estável na Computação. A tecnologia de Bancos de Dados Relacionais surgiu na década de 70, e os SGBDs foram ganhando forma e mercado. Uma das gigantes do setor de TI, a Oracle, tem no SGBD que carrega seu nome o seu carro chefe.

Certo, nós já entendemos qual o processo de criação de um Banco de Dados, desde a modelagem até a implementação no SGBD. Além disso, vimos que com o SQL podemos criar o Banco de Dados, povoá-lo, e extrair informações importantes desses Bancos de Dados. Mas para usar o SQL temos algumas premissas. Em primeiro lugar, devemos saber exatamente qual informação queremos obter do Banco de Dados. Em segundo lugar, temos que conhecer os metadados, ou seja, a estrutura do Banco de Dados.

Acontece que os Bancos de Dados cresceram. Hoje qualquer empresa, mesmo as pequenas, pode ter seu negócio informatizado, com aplicações para manipulação dos dados. Agora imaginem as grandes lojas na Internet, que vendem diariamente milhares de produtos para consumidores em todo Brasil, e às vezes até para o exterior, como é o caso de Saraiva.com. Submarino.com, Americanas.com, Wallmart.com e por ai vai. Essas lojas acumulam muitos dados de suas transações, criando Bancos de Dados gigantescos. Esses Bancos de Dados às vezes são chamado de Very Large Databases (VLDB).

Com o tempo, percebeu-se que era muito difícil encontrar informações importantes para os negócios nesses grandes Bancos de Dados. Afinal, o SQL era limitado a uma espécie de sistemas de perguntas e respostas, onde você deve saber exatamente o que quer saber. Mas dentro desses Bancos de Dados estavam "escondidas" informações que poderiam ser muito úteis para a tomada de decisões nos negócio. Primeiro devemos entender a diferença entre dado e informação.

2.1 Dados versus Informação

Ter um Banco de Dados, por si só, não é garantia de se ter informação. Qual a diferença entre dado e informação?

De forma bem resumida e direta, posso definir dados como fatos que podem ser analisados e que possuem um significado implícito. Por exemplo, se você encontrar uma folha de papel, e ver que nela está escrito o valor 27, o que você pensaria? Bem 27 é um valor numérico que tem algum significado. Mas qual? Pode ser a idade de alguém, pode ser um dia de um mês, pode ser o número de uma casa. Enfim, é apenas um fato, um dado registrado, que possui algum significado, mas que de forma isolada não agrega nenhum valor para quem lê.

Já informação é o resultado do processamento, manipulação e organização de dados, de tal forma que represente uma modificação (quantitativa ou qualitativa) no conhecimento do sistema (pessoa, animal ou máquina) que a recebe. No exemplo acima, se alguém lhe informar que o número escrito no papel é a temperatura máxima em graus Celsius que fará na sua cidade, pronto, nesse momento esse dado, que pouco valor tinha para você, virou uma informação.

A Mineração de Dados surgiu com a motivação de "garimpar" informações relevantes das Bases de Dados. Por exemplo, imaginem um diretor de uma daquelas grandes lojas na Internet que eu citei. Ele precisa tomar decisões para guiar seu negócio. Pelo SQL, poderíamos gerar um relatório com todas as vendas no semestre passado. Mas para que serve um relatório com milhões de linhas, indicando cada venda da loja? Como pode o diretor tomar qualquer decisão com tal relatório? Percebemos que tal relatório não passa de um conjunto de dados para esse diretor, pois pouco ou nenhum valor agregou, e assim não pode ser considerado informação.

Vamos ver um caso curioso que foi resultado da aplicação da Mineração de Dados.

2.2 Caso Prático

É famoso o caso de uma determinada rede de supermercados estadunidense que resolveu aplicar técnicas de Mineração de Dados na sua grande base, que tinha os registros de todas as vendas.

A base era gigantesca, e eles queriam descobrir se existiam relacionamentos entre os produtos vendidos. Após a aplicação da mineração de dados, descobriram que havia um relacionamento entra a venda de cerveja e a venda de fraudas. Sim, isso mesmo, quando a venda de cervejas aumentava, a venda de fraudas acompanhava, e vice-versa.

Passada a fase de Mineração de Dados, que revelou essa informação interessante, eles começaram a analisar, e chegaram à seguinte conclusão: Normalmente as mães ficavam com as crianças pequenas em casa, e quando as fraudas acabavam, os pais saíam para comprar mais no supermercado. Uma vez que já estavam lá, aproveitavam e compravam cerveja também (afinal, ninguém é de ferro). Por isso, esses dois produtos estavam relacionados. O que fez o supermercado para incrementar mais ainda as vendas? Simplesmente transferiu parte de seu estoque de cervejas para perto das fraudas. Com isso, as vendas aumentaram mais ainda.

Por esse exemplo, vemos que descobrir essas informações que não são muito óbvias requer ferramentas mais avançadas do que simples consultas às bases de dados. É ai que entra a Mineração de Dados.

2.3 Conceituação

A Mineração de Dados é o processo de descoberta automática de informações úteis em grandes depósitos de dados. As técnicas de Mineração de Dados são usadas para agir sobre grandes Bancos de Dados com o intuito de descobrir padrões úteis e recentes que poderiam, de outra forma, permanecer ignorados. Elas também oferecem capacidade de previsão do resultado de uma observação futura, com a previsão de se um cliente recém-chegado gastará mais de 100 dólares em uma loja.

O processo de KDD, no qual o Data Mining está inserido, é composto geralmente por 6 fases, que são: seleção de dados, limpeza, enriquecimento, transformação/codificação, mineração, apresentação e interpretação dos resultados. Vocês definitivamente não precisam decorar esses passos para a prova.

A seleção trata de encontrar e selecionar os dados com os quais se quer trabalhar, por exemplo, os dados de vendas de uma loja. O processo de limpeza pode eliminar dados inconsistentes, e fazer correções necessárias. Por exemplo, pode eliminar clientes cadastrados que nunca compraram, ou corrigir dados incompletos, como clientes sem sexo cadastrado etc. A fase de enriquecimento é o acréscimo de informações que podem ser proveniente de outras bases de dados. Por exemplo, pode pegar a base de compras e enriquecer com a base de contas a receber, de forma a verificar clientes que compram e não pagam. A próxima fase é a transformação / codificação, que é deixar os dados de uma maneira que as técnicas de mineração de dados entendem. Por exemplo, no modelo desenvolvido nesse curso, podemos trocar as três datas da tabela empréstimo simplesmente pelo número de dias de atraso. Depois de dar toda essa "ajeitada" nos dados, aplicamos o processo de Data Mining propriamente dito. No final, são gerados relatórios e os resultados são interpretados.

O resultado da Mineração de Dados pode descobrir os seguintes tipos de informação "nova":

- Regras de associação: Por exemplo, se um cliente compra uma máquina fotográfica, pode querer comprar também um cartão de memória.
- Padrões seqüenciais: Por exemplo, determinado cliente pega um empréstimo para comprar um carro. Depois da quarta parcela, começa a atrasar o pagamento. Depois de um ano, deixa de pagar. Isso pode se repetir de forma mais ou menos igual para diversos clientes, e pode definir um padrão. Assim, quando o cliente começa a atrasar muito, a empresa já pode se preparar para ele deixar de pagar a dívida.
- Regras de classificação: Por exemplo, clientes podem ser classificados por freqüência de visitas, por tipo de financiamento utilizado, por quantidade comprada, por afinidades com alguns itens e assim sucessivamente. Algumas estatísticas reveladoras podem ser geradas para cada classe de clientes.

2.4 Tarefas de Mineração de Dados

Vamos agora analisar as principais tarefas relacionadas com Mineração de Dados. Somente uma delas, a classificação, vai ser mais aprofundada, por ser um item do nosso edital. Mas um pequeno entendimento de cada uma dessas tarefas é importante.

- **Predição**: Procura mostrar como certos atributos dos dados vão se comportar no futuro. Por exemplo, podemos analisar transações de compra para prever o que os clientes tendem a comprar se dermos descontos, o volume de vendas que pode

ser alcançado em dado período, e se a exclusão de uma dada linha de produtos pode aumentar os lucros. Parece que os exemplos se aplicam apenas a vendas, mas na verdade outras áreas podem fazer uso dessas técnicas. Por exemplo, a predição pode ser aplicada para, dados certos sintomas que um paciente esteja sentindo, possamos prever problemas que ele possa desenvolver.

- **Identificação**: Padrões de dados podem ser usados para identificar a existência de um item, um evento ou uma atividade. Por exemplo, sistemas de segurança podem identificar a predisposição de alguém para invadir uma casa, ou causar outro tipo de confusão, pela sequência de atitudes tomadas por uma pessoa.
- **Otimização**: Um objetivo de Mineração de Dados pode ser otimizar o uso de recursos limitados (ex tempo, dinheiro, materiais etc), maximizando as variáveis de saída (por exemplo vendas ou lucro) sob determinados conjuntos de restrições. Ou seja, um algoritmo de data mining pode analisar que dado que um empresário tem X reais para investir, tem Y equipamento, funcionários etc, e dado o histórico de vendas, qual seria o melhor caminho (escolha) que deveria ser tomado.
- **Classificação**: é o processo de identificar, entre classes previamente definidas, a qual classe cada elemento pertence. Imaginem que vocês possuem uma série de documentos que baixaram da web com artigos sobre diversas áreas (ex direito constitucional, direito administrativo, Banco de Dados, auditoria etc etc). Classificação em Data Mining é o processo de determinar de forma automática em qual classe cada

elemento se encaixa. Vamos falar mais de classificação em capítulo a parte.

- **Associação**: Determina com certo grau de certeza que determinado item costuma aparecer na presença de outro. É aquele exemplo da loja, que quando alguém compra uma máquina fotográfica, também compra um cartão de memória. Ou quando alguém compra pão, tende a comprar margarina ou manteiga. Essas regras de associação podem ser utilizadas para gerar uma aplicação bastante difundida na Mineração de Dados, que é a recomendação. Ela analisa seu padrão passado de consumo, o padrão de consumidores "semelhantes" a você, e com base nisso recomenda produtos que podem ser do seu interesse. Já presenciaram uma tela parecida com a tela abaixo em uma loja na Internet? Pesquisei um livro, e na mesma tela em que posso comprar esse livro aparecem as recomendações de livros que posso comprar junto. Essa é uma das muitas aplicações de recomendação.

- **Padrões seqüenciais**: Semelhante à associação, mas analisa eventos em determinado período de tempo. Por exemplo, se um paciente fez ponte de safena para artérias bloqueadas e um aneurisma (sai para lá!), e depois desenvolveu uréia alta no sangue no período de um ano, ele está propenso a sofrer problemas renais nos próximos 18 meses.
- **Agrupamento (clustering)**: É semelhante a classificação. Só que cada agrupamento não é predefinido. Assim, posso usar um algoritmo de agrupamento para dividir minha população em agrupamentos, de forma que cada elemento de um agrupamento é mais semelhante com os demais elementos do mesmo agrupamento, do que com elementos de outros agrupamentos. Por exemplo, imagine que um banco pega toda a sua base clientes e quer dividi-lo em agrupamentos. Depois do processo, são encontrado 4 agrupamentos. Em um estão clientes que fazem muitos empréstimos para financiar seus negócios, em outro temos clientes que fazem poupanças para garantir o futuro, em outro temos clientes que fazem aplicações arriscadas, e no último temos clientes que apenas usam a conta para receber salário, sem comprar produtos do Banco. Nos algoritmos de agrupamento, no máximo definimos o número de grupos. Para alguns desses algoritmos, nem isso é definido.

Observem que até agora eu descrevi o que podemos fazer com Mineração de Dados, mas não como fazer. Para o nosso objetivo, basta isso mesmo, saber o que podemos fazer. Das técnicas citadas, só veremos com mais detalhes a classificação. Mas primeiro vamos tratar dos tipos de conjuntos de dados.

3. Dados e Conjunto de Dados

Em primeiro lugar, apresento a vocês uma classificação diferente dos diversos tipos de atributos que podemos ter em um Banco de Dados:

- Atributos Qualitativos: Descrevem qualidades das entidades.
    - Nominais: Os valores de um atributo nominal são apenas nomes diferentes, ou seja, valores nominais fornecem apenas informações suficientes para distinguir um objeto do outro. Ex: CEP, Matricula, Cor dos olhos, sexo.
    - Ordinais: os valores de um atributo ordinal fornecem informações suficientes para criar uma ordem entre os objetos. Ex: Qualidade de uma refeição (ótima, boa, regular, ruim), coloração da água (límpida, turva), notas dos alunos, idade etc
- Atributos quantitativos: Descrevem uma grandeza.
    - Intervalar: As diferenças entre os valores são significativas, existindo uma unidade de medida. Como exemplo temos datas de calendário, temperatura em Celsius e Fahrenheit etc.
    - Proporcional: Tanto as diferenças absolutas quanto as proporções são significativas. Ex: Quantidades monetárias, temperatura em Kelvin.

Para entender melhor a diferença entre atributos intervalares e proporcionais, vamos ver o caso das temperaturas. Quando medida na escala Kelvin, uma temperatura de 2º é, de forma fisicamente significativa, o dobro de uma temperatura de 1º. Já para Celsius e Fahrenheit, isso não é verdade, ou seja, o dobro da temperatura na escala não significa fisicamente o dobro da temperatura, pois são escalas

de certa forma arbitrárias. Bem, e dinheiro? O dobro de um valor é uma proporção exata, não é mesmo.

Bem, outra maneira de classificar atributos é pelo número de valores (domínio) que ele pode receber. Nessa classificação, temos:

- Atributos discretos: Possui um conjunto de valores finitos. Um exemplo são dados numéricos inteiros. Outro exemplo são atributos lógicos, que podem receber valores como verdadeiro ou falso. Por fim, um atributo sexo só comporta os valores masculino ou feminino.
- Atributos contínuos: Normalmente são representados por valores do conjunto dos números reais. Um exemplo é um campo salário, que pode receber uma infinidade de valores. Outro exemplo é o valor total de uma nota fiscal. Não temos como determinar quantos valores distintos um atributo desse tipo pode receber.

3.1 Características de Conjuntos de Dados

Existem três características importantes sobre conjuntos de dados, a seguir descritas:

- **Dimensão**: diz respeito à quantidade de atributos de um conjunto de dados. Por exemplo, imaginem que temos um conjunto de dados com informações sobre clientes de um Banco. Poderíamos ter supostamente 100 atributos nesse conjunto, ou seja, ter uma dimensão muito grande. Na fase de limpeza dos dados, poderíamos reduzir esse número de atributos para 30, pois estes é que realmente interessariam. Então, em todo conjunto de dados a dimensão deve ser considerada, para evitar ter conjuntos de dados com

dimensões que não serão utilizadas no processo de Data Mining.

- **Dispersão**: Ter dado dispersos é ter dados em que a maioria não está preenchida (ou tem NULL ou tem um valor padrão). Com isso, podemos reduzir, em alguns casos, nossa base de análise, dispensando tuplas com determinado campo não preenchido, poupando tempo e recursos do computador. Então, um conjunto de dados é muito disperso quando para um atributo relevante, a maioria dos valores é NULL ou um valor padrão. Por exemplo, em uma base de estudantes de segundo grau (sinal de velhice chamar o ensino médio assim), um campo com o número de filhos teria 0 na maioria dos registros, formando um conjunto disperso. Se eu quiser inferir algo com base no número de filhos, posso pegar um conjunto bem menor, que é daqueles que possuem um filho ou mais.
- **Resolução**: diz respeito à granularidade dos dados. Às vezes pegamos dados agrupados demais, ou agrupados de menos, e isso atrapalha na análise. Se quisermos, por exemplo, descobrir padrões de compras de produtos em determinada região, os dados de cada compra daquela região podem ser necessários. Agora, se queremos encontrar os mesmos padrões para um conjunto de lojas, espalhadas pelo mundo todo, e em um período de 10 anos, talvez seja melhor processar essas compras agrupadas de alguma forma, pois esses dados muito detalhados podem atrapalhar a análise.

Agora vamos ver os tipos de conjuntos de dados propriamente ditos, que é o assunto escrito literalmente no edital do concurso.

3.2 Tipos de Conjunto de Dados

Existem três tipos principais de conjuntos de dados. Eles são Dados em Registros, Dados Baseados em Grafos e Dados Ordenados. Vamos ver cada um deles na sequência.

Dados em Registros

Esse para nós é o tipo mais fácil. Ora, dados baseados em registros são justamente os conjuntos de dados que vimos até agora. Dados que estão em um Banco de Dados Relacional são o exemplo mais típico de dados baseados em registros.

Uma subespécie de dados em registros são os **dados em transação**. Simplesmente é uma representação onde cada registro envolve um conjunto de itens. Vamos ver um exemplo onde cada linha representa as compras de um cliente em determinada ocasião:

TR. Id	Itens
1	Pão, Refrigerante, Leite
2	Cerveja, Pão
3	Cerveja, Refrigerante, Fralda, Leite

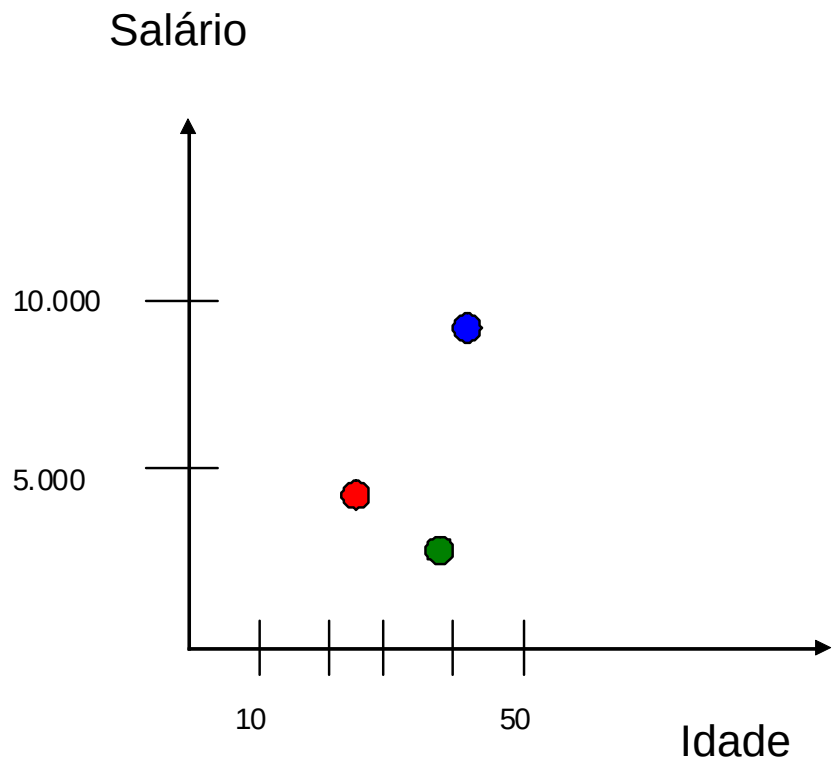
Assim, a transação 1 (as compras do cliente 1) é composta de Pão, Refrigerante e Leite. A transação 2 é composta de Cerveja e Pão e assim sucessivamente.

Outra subespécie de dados em registros são as **matrizes de dados**. São como uma tabela do modelo relacional, mas **todos os atributos são numéricos**. Isso permite que se represente os dados como vetores da álgebra linear. Ih professor, não tinha dito sem matemática!! Pois é, desse conceito não temos como fugir. Vou dar um

exemplo com duas dimensões, que fica mais fácil de observar. Imaginem o seguinte Conjunto de Dados:

Idade	Salário
27	4.500,00
39	2.700,00
42	9.200,00

Nesse conjunto tenho dois atributos, idade e salário. Posso transformar essas informações e colocar em um vetor de duas dimensões da seguinte forma:

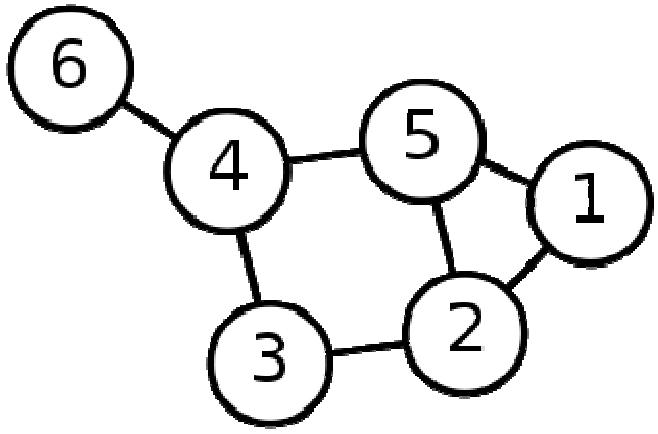


Essa é a representação vetorial da tabela anterior. Ela tem duas dimensões, que é o mesmo número de atributos. Se tivesse três atributos, teriam três dimensões e assim sucessivamente. O ponto vermelho representa a tupla com idade 27 e salário 4.500. O ponto verde

representa a tupla com idade 39 e salário 2.700. Por fim, o ponto azul representa a tupla com idade 42 e salário 9.200. Pelas propriedades da álgebra linear, podemos medir a distância entre dois pontos. O importante é que vocês guardem que na matriz de dados, todos os dados são valores numéricos, podendo assim representar dimensões de um vetor.

Dados baseados em Grafos

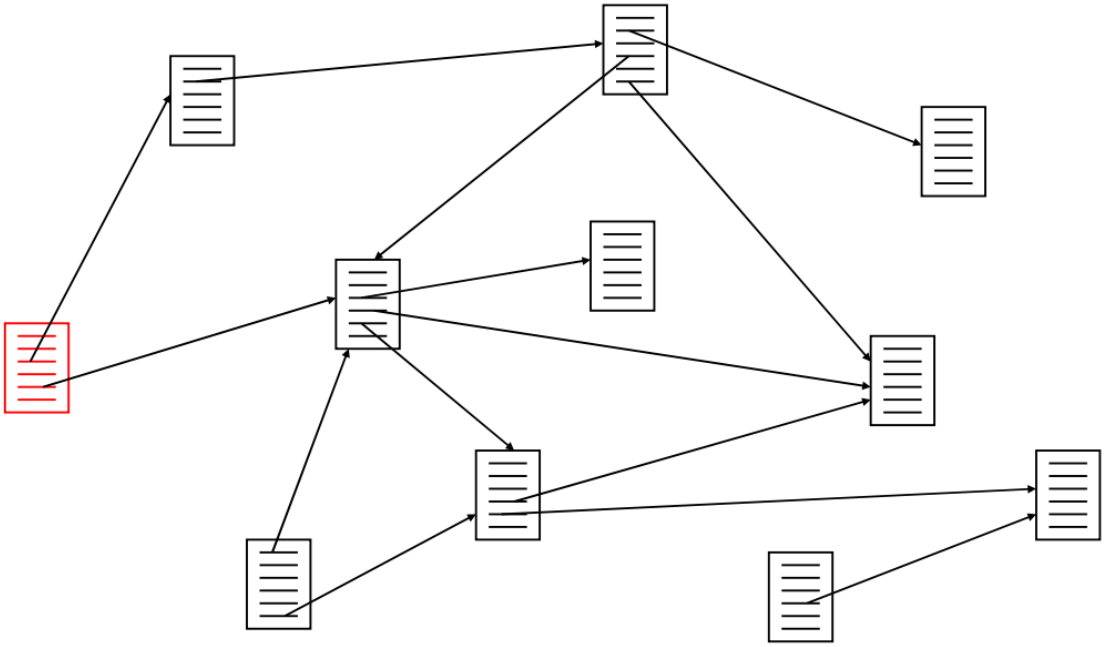
Um grafo é uma estrutura de dados muito usado na Computação e na matemática. Tipicamente, um grafo é representado por um conjunto de pontos (vértices) ligados por retas (grafo não orientado) ou por setas (grafo orientado). Essas retas ou setas são chamadas arestas. Vejam a representação de um grafo não orientado:



Neste exemplo, os vértices, que são os dados propriamente ditos, são representados por círculos com números. As arestas são as linhas que ligam os vértices.

Tudo bem, mas para que serve essa representação? Bem, às vezes os objetos representados tem relacionamentos que ficam melhor representados por grafos. Um exemplo típico é um conjunto de páginas web. Nas páginas web, nos encontramos links, que são pontos na página que você clica para chegar à outra página. Cada página tem um número indeterminado de links, e cada link direciona para outra página que

também tem um número indefinido de links. As páginas web, no formato de grafo, podem ser representadas assim:



As páginas são as arestas, e cada link de uma página para outra representa um vértice. Essa representação por grafos tem diversas aplicações na Computação. O que interessa para o nosso estudo é entender o que é um grafo, e saber que esse é um dos tipos de representação de conjuntos de dados.

Dados Ordenados

Os dados ordenados são nosso terceiro tipo de dado. Nessa representação, a ordem dos dados tem importância, pois eles perdem sentido se não estiverem na ordem correta. Vamos ver cada subtipo deles.

Os **dados de transações seqüenciais, ou dados seqüenciais, ou dados temporais**, podem ser pensados como uma extensão dos dados de registro, mas com cada registro possuindo um tempo associado a ele. Como exemplo temos um conjunto de transações de compra, onde também se armazene o momento (tempo) em que os eventos ocorreram. Essa informação de tempo permite encontrar um padrão do tipo "o pico

de vendas de gravatas ocorre antes do dia dos pais". Vejam o exemplo a seguir:

Ocasião	Cliente	Itens Comprados
t1	1	A,B
t2	3	A,C
t3	1	C,D
t4	2	A.D
t5	2	E
t6	1	A,E
t7	3	B

O próximo subtipo são os **dados de sequência**. Eles são bastante semelhantes aos dados seqüenciais, mas não possuem marcação de tempo. Exemplo de dados de sequência são conjuntos de palavras ou de letra em um documento. Claro, se trocarmos a sequência das palavras, o documento não fará mais sentido. Outro exemplo são sequências de DNA.

Por fim, temos os **dados de séries temporais**. São um tipo especial de dados seqüenciais no qual cada registro é uma série de tempos, ou seja, é uma série de medições feitas no decorrer do tempo. Por exemplo, podemos ter um conjunto de medições de temperatura em determina região, nem determinado período. Ao se trabalhar com esse tipo de dado, deve-se considerar que duas medições efetuadas de forma próxima no tempo normalmente não têm valores muito diferentes.

Amigos, isto é o mais importante no assunto Tipos de Conjuntos de Dados. Não encontrei nenhuma questãozinha sobre esse assunto, então

não tenho idéia se e como vão cobrar. Mas duvido que a ESAF vá se aprofundar nisso. Bem, o jeito é tentar ler e entender o assunto, mesmo que por alto, e esperar as questões da prova.

4. Classificação

Classificação é a tarefa de organizar objetos em uma entre diversas categorias. É uma das aplicações mais utilizadas em Mineração de Dados, pois os algoritmos de classificação podem ser usados em diversos contextos.

Vamos a um exemplo. Todos vocês tem caixas de e-mail. Entre as pastas disponíveis, temos mensagens recebidas, mensagens enviadas, e uma pasta para Spam. Mas como o programa de correio eletrônico sabe o que é e o que não é Spam? Simplesmente ele utiliza a técnica de classificação de mensagem, e dependendo do resultado obtido, classifica ou não a mensagem como Spam.

Outro exemplo. Um Banco tem diversos clientes, com os mais variados perfis. Uma forma do Banco decidir se vai ou não aprovar um empréstimo é classificando o cliente em uma entre as diversas categorias (alto risco, risco médio, baixo risco). São usados uma série de critérios para classificar esses clientes, afinal como vocês devem saber, a melhor maneira de conseguir um empréstimo é provando ao Banco que você não precisa dele.

Existe uma abordagem geral para a resolução de um problema de classificação. Essas etapas podem ser assim resumidas:

O primeiro passo para resolver um problema de classificação é o Treino. Nessa fase, separamos um conjunto dos nossos dados, e submetemos esse conjunto ao classificador, de forma que ele "aprenda" como classificar.

A idéia é eu ter um conjunto de dados que represente a maior parte das situações que ele pode encontrar. No exemplo do email eu submeteria uma série de mensagens que não é Spam, e informaria isso ao classificador. Também submeteria uma série de mensagens que são Spam, de forma que ele "entenda" o que caracteriza uma mensagem como Spam.

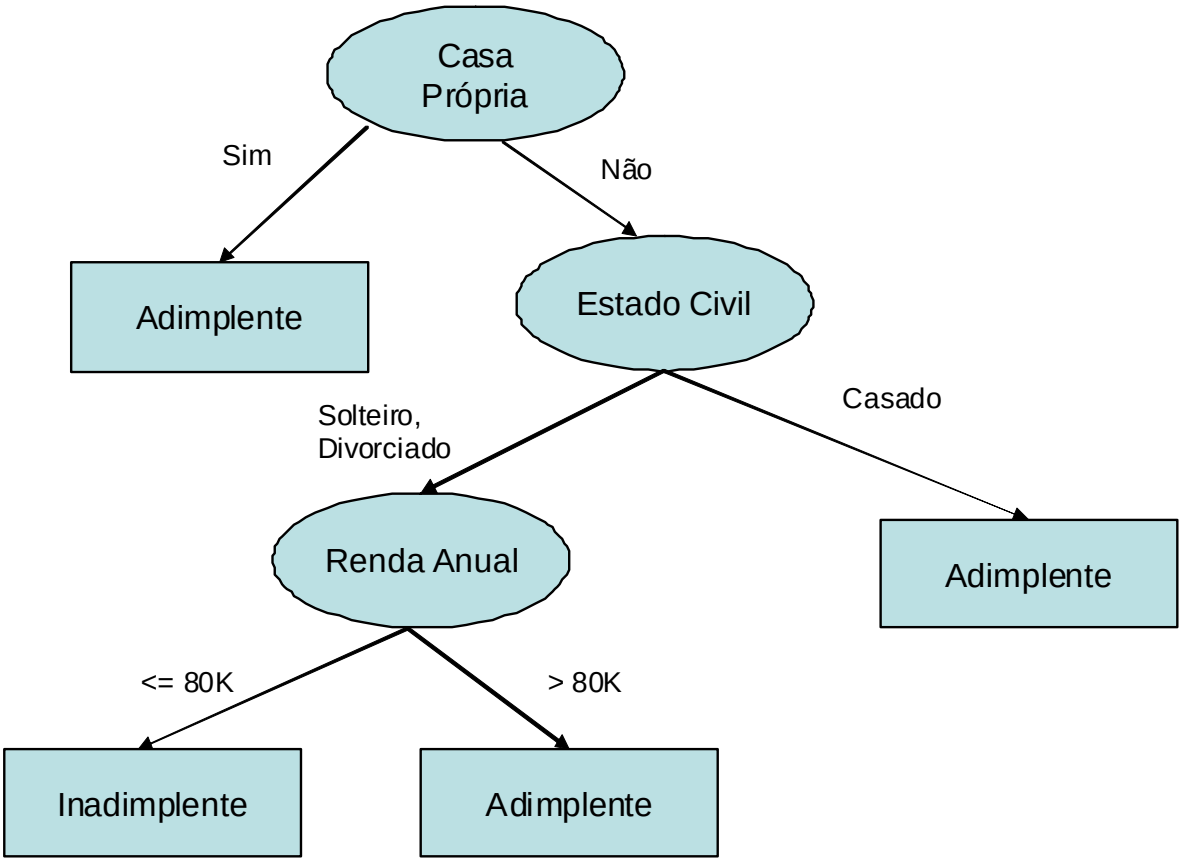
Como é que o professor na escola sabe se o aluno aprendeu? Fazendo um teste ou prova, não é? É a mesma coisa para o classificador. Depois da fase de treino vem a fase de teste. Eu simplesmente separo outro subconjunto de dados, e vejo como ele se sai. Ou seja, eu separaria uma série de mensagens Spam e não Spam, e veria se o Classificador acerta. Quando uma mensagem que não é Spam é classificada de forma errada como Spam, temos um falso positivo. Se uma mensagem que é Spam não for classificada como tal, temos um falso negativo. Com base nesse resultado, sei se meu classificador vai funcionar corretamente ou não. Se não estiver legal, treino de novo.

Por fim, como meu classificador está pronto, aplico o modelo no dia-a-dia, ou seja, coloco meu classificador para funcionar em uma aplicação real.

Visto isso, vamos passar ao estudo dos principais algoritmos de classificação.

4.1 Árvore de decisão

Uma árvore de decisão é uma técnica de classificação composta de nodos terminais e nodos não terminais. Os nodos terminais, que vou representar como quadrados, representam as classes. Os nodos não terminais, que vou representar por elipses, são as decisões a serem tomadas para se chegar nas classes. Vamos ver na prática.



Nessa árvore de decisão, que poderia ser usada por um Banco para decidir sobre conceder ou não um empréstimo, ou mesmo usar uma taxa maior ou menor, em cada elipse é tomada uma decisão. Toda vez que um retângulo é encontrado, um objeto é classificado. A idéia é bem simples mesmo, o classificador percorre a árvore de encontra a classe na qual o objeto se encaixa.

4.2 Classificador Baseado em regras

Um classificador baseado em regras é uma técnica para classificar registros usando um conjunto de regras do tipo "se .... então". Vamos ao exemplo.

R1: Se possui casa própria → Adimplente

R2: Se não possui casa própria E é casado → Adimplente

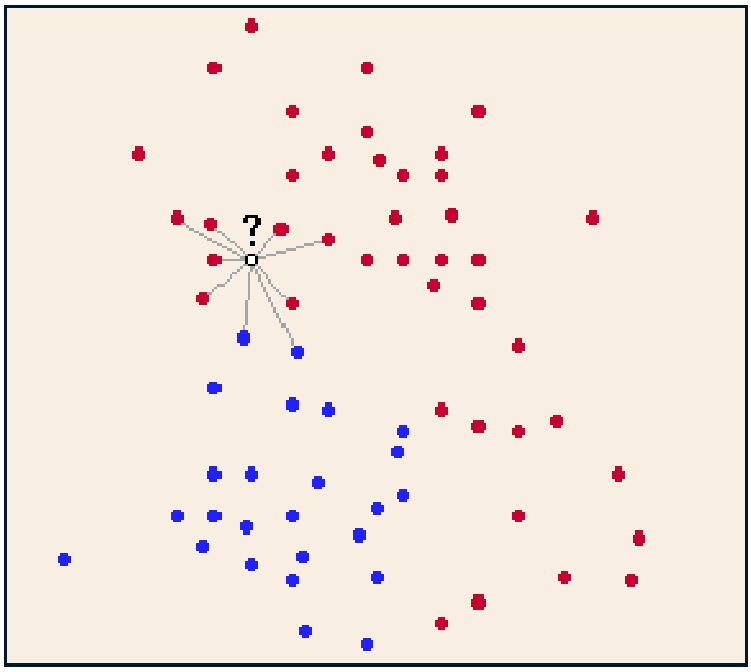
R3: Se não possui casa própria E é solteiro OU é divorciado E tem renda menor que 80 K → Inadimplente

R4: Se não possui casa própria E é solteiro OU é divorciado E tem renda maior que 80 K → Adimplente

Pronto, representamos as mesmas regras da árvore de decisão em um classificador baseado em regras. A árvore de decisão é mais fácil de visualizar, mas o resultado nesse caso seria o mesmo.

4.3 Classificador do Vizinho Mais Próximo

Lembram daquela representação de um Conjunto de Dados em um espaço vetorial. Os classificadores de vizinhos mais próximos usam aquela representação, determinando que a classificação de um objeto depende da distância entre este e seus vizinhos. Esses classificadores são conhecido como KNN, ou seja, classificam de acordo com os K nearest-neighor (os k vizinhos mais próximos). Vamos ver isso graficamente que fica mais fácil.



Eu tenho um ponto em um espaço vetorial, que na figura está com uma interrogação em cima. Tenho eu classificá-lo como da classe azul ou da classe vermelha. O que esse KNN faz é ver a distância dos k pontos mais próximos. Na figura essa distância está definida como uma linha. Então, em que classe está o ponto? Como temos mais vizinhos vermelhos próximos do que vizinhos azuis, então o classificador vai defini-lo como vermelho. Na prática, é assim que funciona.

Esse classificador segue meio que aquela máxima. Diga-me com quem andas e te direi quem és.

4.4 Classificador Bayesiano

Esse classificador utiliza uma abordagem probabilística, baseado no Teorema de Bayes. Não veremos detalhes aqui. O que esse classificador faz na prática? Bem, um item que precisa ser classificado tem uma série de atributos, correto? Então, esse classificador calcula a probabilidade do item pertencer a cada classe, dados aquele atributos que ele tem. Dessa

forma, a classe que alcançar a maior probabilidade é aquela na qual o item vai ser classificado.

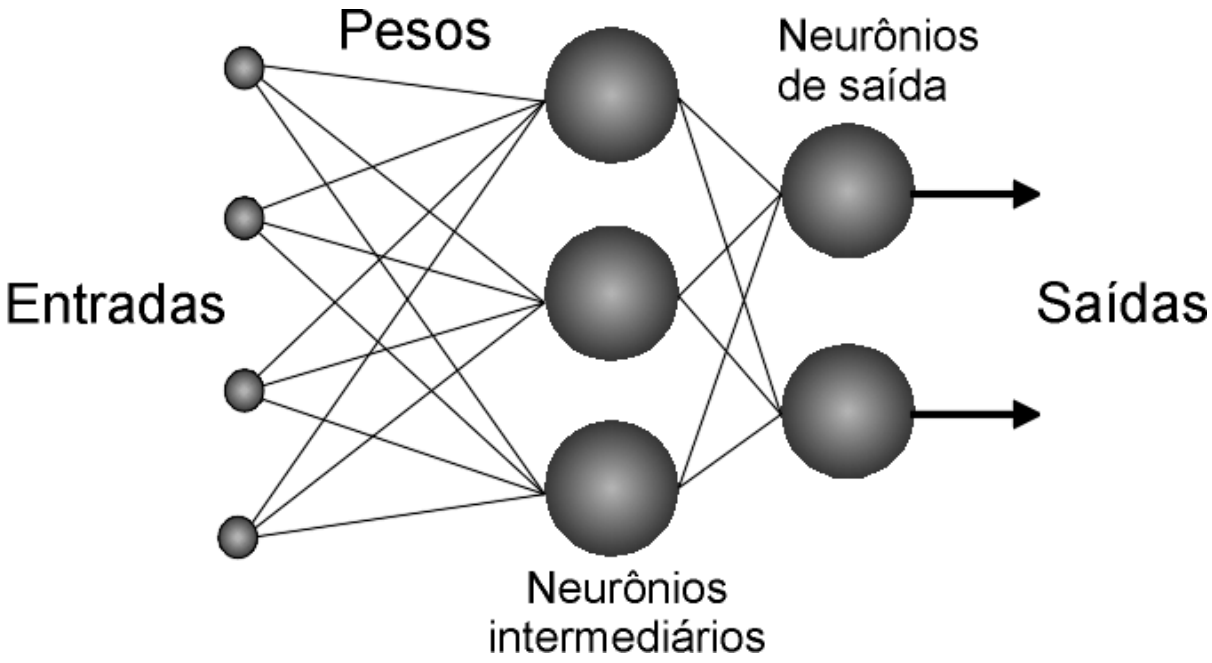
Por exemplo, imagine que eu tenho uma série de sintomas, e dados esses sintomas eu quero classificar a pessoa em alguma doença. Então, imaginem que os sintomas são secreção, cansaço, tosse, febre e dor muscular. Bem, com base no treinamento anterior (aquelas fases valem para todos os classificadores), meu classificador bayesiano vai classificar em resfriado, câncer, problema renal ou cardiopatia. Ele pega essa lista de sintomas, e com base no que aprendeu, calcula a probabilidade desses sintomas casarem com as doenças indicadas. No final, meu classificador escolhe a doença com maior probabilidade de gerar aqueles sintomas.

Vejam que estou sendo bem conciso na descrição dos classificadores. Acho que não teríamos aqui tempo para ver detalhes de implementação, cálculos probabilísticos e etc. Mas espero que vendo essas idéias centrais possamos resolver as questões.

4.5 Classificador de Redes Neurais Artificiais

As redes neurais artificiais são uma técnica de inteligência artificial que tentam imitar o funcionamento do cérebro humano. De forma simplificada, essa técnica liga uma série de computadores como se fossem neurônios, e a saída de um neurônio é ligada à entrada de outro, de forma a simular a passagem de pulsos elétricos no cérebro. Assim, uma das aplicações dessas redes neurais é a classificação.

Mais uma vez, não vou entrar em detalhes. Nem a ESAF seria louca de cobrar detalhes disso. De qualquer forma, vocês devem saber que isso existe e que pode ser usado como classificador. Segue uma representação gráfica de redes neurais.



4.6 Support Vector Machine (SVM)

Esse é o cara, o tal. Se uma questão pedir o classificador mais arrojado e tiver esse, podem marcá-lo, pois ele é o estado da arte em classificação.

E como funciona? Ah, nem esquentem com isso. Nem eu sei direito, confesso. Utiliza um conceito de hiperplano de margem máxima, que por mais que tentasse nunca consegui entender. Mas professor, não tem vergonha de colocar isso numa aula não? Nenhuma meus caros, quando precisei utilizar SVM peguei programas prontos e usei. Implementar isso nem pensar.

Bem, o que vocês devem guardar do SVM então. Devem guardar a informação que é um classificador, e que é o top utilizado por ai. Isso mais do que basta.

5. Exercícios

Vamos às questões que encontrei.

1. (ESAF/AFC-TI/ STN 2005) O Data Warehouse é um conjunto de dados orientado por assuntos, não volátil, variável com o tempo e integrado, criado para dar suporte à decisão. Considerando essa tecnologia e suas derivadas ou variantes é correto afirmar que

a) a premissa do Data Mining é uma argumentação ativa, isto é, em vez de o usuário definir o problema, selecionar os dados e as ferramentas para analisar tais dados, as ferramentas do Data Mining pesquisam automaticamente os mesmos à procura de, por exemplo, possíveis relacionamentos, identificando assim problemas não identificados pelo usuário.

b) um Data Mining é considerado Top-down quando uma empresa, por desconhecer a tecnologia do Data Warehouse, prefere primeiro criar um banco de dados para somente uma área. Com isso, os custos são bem inferiores de um projeto de Data Warehouse completo. A partir da visualização dos primeiros resultados, parte para outra área e assim sucessivamente até resultar num Data Warehouse.

c) um Data Mining é considerado Bottom-up quando a empresa cria um Data Warehouse e depois parte para sua segmentação, isto é, dividindo o Data Warehouse em áreas menores, gerando assim pequenos bancos orientados por assuntos departamentalizados.

d) o propósito de uma análise de dados com a tecnologia Data Mart é descobrir, previamente, características dos dados, sejam relacionamentos, dependências ou tendências desconhecidas.

e) as ferramentas de Data Mart analisam os dados, descobrem problemas ou oportunidades escondidas nos relacionamentos dos dados, e então diagnosticam o comportamento dos negócios, requerendo a mínima intervenção do usuário.

Comentários:

Bem, essa é a famosa única questão que a ESAF fez relacionada a Data Mining, pelo menos que eu encontrei.Ela mistura conceitos de Data Mining com Data Warehouse. Mas professor, por que não foi falado nada de Data Warehouse? Ora, simplesmente porque o programa não previu. Em poucas palavras, Data Warehouse é uma forma de se visualizar grandes bancos de Dados históricos, de forma a encontrar informações relevantes.

Opa, espera ai, mas o Data Mining não procura informações relevantes. Sim, procura, **mas de forma automática**. Na tecnologia de Data Warehouse (armazém de dados) as informações são buscadas no olhômetro de quem procura. Os Data Warehouses simplesmente agregam essas informações e as colocam de uma maneira mais fácil de visualizar do que com comandos SQL.

Bem, vemos que na letra (a) temos elementos importantes para a definição de Data Mining, ou seja, afirma que as ferramentas do Data Mining pesquisam automaticamente os dados à procura de, por exemplo, possíveis relacionamentos, identificando assim problemas não identificados pelo usuário.

Dessa forma, esse é o gabarito da nossa questão, e os demais itens não vou entrar no mérito porque Data Warehouse não está no nosso programa.

Gabarito: Letra a

2. (FCC/Analista-DBA/INFRAERO 2011) Funcionalidade cujo objetivo é encontrar conjuntos de dados que não obedecem ao comportamento ou modelo dos dados. Uma vez encontrados, podem ser tratados ou descartados para utilização em mining. Trata-se de

a) descrição.

b) agrupamento.

c) visualização.

d) análise de outliers.

e) análise de associações.

Comentários:

Não comentamos sobre isso na aula também, mas como não poderia perder nenhuma questão, resolvi colocar. Ora, falamos que existe uma fase de limpeza dos dados antes do Data Mining propriamente dito. Nesta fase, podem ser identificados pontos fora da curva, ou outliers. O que é isso? Ora, são transações ou registros atípicos. Imaginem que eu tenha uma revenda de carros e quero fazer uma série de análise do comportamento padrão dos meus clientes. A maioria vai lá e troca o carro a cada dois ou três anos, comprando um ou dois carros no máximo. De repente chega o Eike Batista na minha loja, compra logo 20 carros de luxo de uma vez. O que isto significa? Um ponto fora da curva, ou outlier. Se eu não eliminar esse ponto antes da análise, minha resposta pode ser distorcida, porque vou começar a acreditar que meus clientes trocam de carro a cada 6 meses em média. Então, é importante tratar outliers na fase de pré-processamento do Data Mining.

Gabarito: Letra d.

3. (FCC/Analista-DBA/INFRAERO 2011) No âmbito da descoberta do conhecimento (KDD), a visão geral das etapas que constituem o processo KDD (Fayyad) e que são executadas de

forma interativa e iterativa apresenta a seguinte sequência de etapas:

a) seleção, pré-processamento, transformação, data mining e interpretação/avaliação.

b) seleção, transformação, pré-processamento, interpretação/ avaliação e data mining.

c) data warehousing, star modeling, ETL, OLAP e data mining.

d) ETL, data warehousing, pré-processamento, transformação e star modeling.

e) OLAP, ETL, star modeling, data mining e interpretação/ avaliação.

Comentários:

Essa questão pode ser respondida com base na aula. Pelo que vimos, na letra (a) temos a sequência de passos que formam o KDD.

Vou aproveitar para pelo menos apresentar alguns termos que aparecem na questão, apesar de estarem relacionados ao Data Warehouse (DW).

OLAP ou On-line Analytical Processing é a capacidade para manipular e analisar um grande volume de dados sob múltiplas perspectivas. É a forma de processamento que os Data Warehouses oferecem para que os dados sejam visualizados. O OLAP contrasta com o OLTP (Online Transaction Processing ou Processamento de Transações em Tempo Real), que é a tecnologia para tratar informações que os SGBDs relacionais utilizam. Então, OLAP é uma tecnologia para organizar e consultar grandes Bancos de Dados por meio dos Data Warehouses, enquanto OLPT é na prática nossa velha DML, usada para inserir, alterar, excluir e consultar dados em um SGBD tradicional.

Outro conceito que vou mencionar é o ETL, do inglês Extract Transform Load (Extração Transformação Carga), que são ferramentas de software cuja função é a extração de dados de diversos sistemas, transformação desses dados conforme regras de negócios e por fim a carga dos dados em um Data Warehouse. Ou seja, são ferramentas que fazem aquele pré-processamento dos dados que falamos no processo de KDD, mas normalmente não se usa esse termo ETL em Data Mining, só no mundo dos Data Warehouses.

Repito mais uma vez, a ESAF não colocou DW no programa. Se cobrar cabe recurso, porque DW e Data Mining são coisas diferentes.

Gabarito: Letra a

4. (CESPE/Pesquisador/INMETRO 2010) As técnicas de datamining incluem

a) polarização, classificação, estimativa, granularidade e análise de afinidade.

b) análise de agrupamentos, análise de tipos, estimativa, previsão e ponto médio.

c) polarização, classificação, estimativa, previsão e análise de afinidade.

d) análise de agrupamentos, análise de tipos, estimativa, previsão e ponto médio.

e) análise de agrupamentos, classificação, estimativa, previsão e análise de afinidade.

Comentários:

Temos na letra (e) as técnicas análise de agrupamento, classificação e previsão, que vimos. A estimativa é o que chamamos de predição. E análise de afinidade é o que chamamos de associação. Então, essa é a nossa resposta.

Gabarito: Letra e

5. (CESPE/Pesquisador/INMETRO 2010) Assinale a opção correta, com relação às classes de datamining do ponto de vista de processo orientado.

a) A modelagem de prognóstico é o processo de análise em um banco de dados para encontrar padrões escondidos sem uma ideia ou hipótese predeterminada sobre o que são esses padrões.

b) Na análise prévia, os padrões descobertos no banco de dados são usados para prognosticar o futuro.

c) O descobrimento é o processo de análise em um banco de dados para encontrar padrões escondidos sem uma idéia ou hipótese predeterminada sobre o que são esses padrões.

d) Na classe strategic mining, grande conjunto de dados corporativos é examinado com o objetivo de se obter o seu conhecimento global.

e) Na classe episodic mining, são buscados dados de um fato específico.

Comentários:

Meus amigos e minhas amigas, o CESPE viaja legal. Mas vamos direto ao ponto. A letra (c) traz uma boa definição do processo de descobrimento, como processo de análise em um banco de dados para encontrar padrões escondidos sem uma idéia ou hipótese predeterminada sobre o que são esses padrões. Foi o que falamos de KDD e Data Mining, então é nossa resposta.

Gabarito: Letra c

6. (FGV/DBA/DETRAN-RN 2010) Sobre Data Mining, pode-se afirmar que:

a) Refere-se à implementação de banco de dados paralelos.

b) Consiste em armazenar o banco de dados em diversos computadores.

c) Relaciona-se à capacidade de processar grande volume de tarefas em um mesmo intervalo de tempo.

d) Permite-se distinguir várias entidades de um conjunto.

e) Refere-se à busca de informações relevantes a partir de um grande volume de dados.

Comentários:

Questão bem conceitual de Data Mining, e podemos perceber que só a letra (e) se encaixa no conceito. As demais não chegam nem perto.

Gabarito: Letra e

7. (CESPE/Analista Judiciário-TI/TRT21 2010) Acerca de sistemas de suporte a decisão e data warehousing, julgue os itens a seguir:

[I] O data mining é um processo automático de descoberta de padrões, de conhecimento em bases de dados, que utiliza, entre outros, árvores de decisão e métodos bayesianos como técnicas para classificação de dados.

Comentários:

Olha ai essa questão. Além de conceituar Data Mining, citou dois métodos de classificação que vimos. Questão correta.

Vejam que as questões como um todo são bem conceituais. Então, apesar do assunto um pouco pesado, a cobrança é mais de conceitos.

Gabarito: Certo

8. (CESGRANRIO/Analista de Sistema Funcional /ELETROBRÁS 2010) Em uma reunião sobre prospecção de novos

pontos de venda, um analista de TI afirmou que técnicas OLAP de análise de dados são orientadas a oferecer informações, assinalando detalhes intrínsecos e facilitando a agregação de valores, ao passo que técnicas de data mining tem como objetivo

a) captar, organizar e armazenar dados colecionados a partir de bases transacionais, mantidas por sistemas OLTP.

b) facilitar a construção de ambientes de dados multidimensionais, através de tabelas fato e dimensionais.

c) melhorar a recuperação de dados organizados de forma não normalizada em uma base relacional conhecida como data warehouse.

d) extrair do data warehouse indicadores de controle (BSC) para apoio à tomada de decisão por parte da diretoria da empresa.

e) identificar padrões e recorrência de dados, oferecendo conhecimento sobre o comportamento dos dados analisados.

Comentários:

Mais uma questão bem conceitual e que fica fácil nós encontrarmos a definição de Data Mining. Vejam que as outras não chegam nem perto do conceito. Minha maior curiosidade nessa questão é descobrir que raio de cargo é esse, analista de sistema funcional.

Gabarito: Letra e

9. (FCC/Analista Judiciário-Informática/TRF4 2010) Sobre data mining, é correto afirmar:

a) Não requer interação com analistas humanos, pois os algoritmos utilizados conseguem determinar de forma completa e eficiente o valor dos padrões encontrados.

b) Na mineração de dados, encontrar padrões requer que os dados brutos sejam sistematicamente "simplificados", de forma a

desconsiderar aquilo que é genérico e privilegiar aquilo que é específico.

c) É um grande banco de dados voltado para dar suporte necessário nas decisões de usuários finais, geralmente gerentes e analistas de negócios.

d) O processo de descobrimento realizado pelo data mining só pode ser utilizado a partir de um data warehouse, onde os dados já estão sem erros, sem duplicidade, são consistentes e habilitam descobertas abrangentes e precisas.

e) É o processo de descoberta de novas correlações, padrões e tendências entre as informações de uma empresa, por meio da análise de grandes quantidades de dados armazenados em bancos de dados usando técnicas de reconhecimento de padrões, estatísticas e matemáticas.

Comentários:

Mais uma questão que basta conhecer o conceito de Data Mining para acertar. O resto são aquelas afirmativas para pegar os músicos, aquelas alternativas que "soam bem".

Gabarito: Letra e

10. (CESPE/Analista TI-BD/EMBASA 2010) Com relação aos sistemas de apoio à decisão, julgue os itens seguintes:

[I] A regra de classificação pode ser aplicada de várias formas, sendo uma delas a regressão. A análise de regressão é comumente aplicada quando se tem apenas um único domínio de pesquisa. No entanto, o resultado dessa análise não equivale a uma operação de data mining para a previsão da variável destino.

Comentários:

Nem falamos de regressão, mas fica aqui o registro de regressão é em princípio uma técnica de previsão, e não de classificação. A característica principal da regressão é que trabalho com dados contínuos (valores do conjunto dos reais). Assim, a afirmativa esta incorreta.

Gabarito: Errado

11. (CESPE/Analista TI-Desenvolvimento/EMBASA 2010) Com relação aos sistemas de apoio à decisão, julgue os itens seguintes:

[I] Data mining é o processo de extração de conhecimento de grandes bases de dados, sendo estas convencionais ou não, e que faz uso de técnicas de inteligência artificial. ERRADO

Comentários:

O CESPE viajou nessa questão, colocando ela como errada. Não há nada de errado nela. Data Mining extrai conhecimentos de grandes bases de dados? Sim. As bases podem ser convencionais ou não? Claro que sim, apesar de eu não saber o que o examinador entende por bases convencionais. Uma base não convencional poderia ser, por exemplo, uma base com dados de localização de um satélite no espaço, ou dados de pontos de desmatamento na Amazônia. Por fim, Data Mining utiliza técnicas de IA? Sim. Então, não tem nada de errado com o item. Sabe o que acontece, às vezes o examinador nem conhece bem o assunto, pega um texto em um livro, muda uma palavra e acha que está errado. Vai ver que o CESPE pegou um texto que não tinha o "ou não", acrescentou isso e deu a questão como errada. Enfim, deixa para lá, o importante é vocês verificarem que a questão está correta.

Gabarito: Para o CESPE ERRADO, mas está CERTO

12. (FCC/Produção e BD/TCE-SP 2010) NÃO é um objetivo da mineração de dados (mining), na visão dos diversos autores,

a) garantir a não redundância nos bancos transacionais.

b) conhecer o comportamento de certos atributos no futuro.

c) possibilitar a análise de determinados padrões de eventos.

d) categorizar perfis individuais ou coletivos de interesse comercial.

e) apoiar a otimização do uso de recursos limitados e/ou maximizar variáveis de resultado para a empresa.

Comentários:

Pela aula, podemos ver que de (b) até (e) temos objetivos da mineração de dados. Garantir não redundância não tem nada a ver com Data Mining, é assunto do Modelo Relacional.

Gabarito: Letra a

13. (FCC/Produção e BD/TCE-SP 2010) Considere uma dada população de eventos ou novos itens que podem ser particionados (segmentados) em conjuntos de elementos similares, tal como, por exemplo, uma população de dados sobre uma doença que pode ser dividida em grupos baseados na similaridade dos efeitos colaterias produzidos.

Como um dos modos de descrever o conhecimento descoberto durante a data mining este é chamado de

a) associação.

b) otimização.

c) classificação.

d) clustering.

e) temporização.

Comentários:

Falamos da técnica que agrupa os itens de acordo com a sua similaridade. Trata-se de agrupamento, ou clustering.

Gabarito: letra d

6. Palavras finais

Bem meus amigos e minhas amigas. Percorremos aqui um longo caminho nessa disciplina. Acredito que encerro esse curso muito orgulhoso do trabalho que desenvolvi, e a parceria com todos vocês foi essencial. Escrever um curso do nada, depois de lançado o edital, e com essas características, que procurou abrigar todos aqueles que nenhum conhecimento têm de TI, foi realmente um desafio muito grande, que foi muito prazeroso em muitos momentos, mas me levou a uma pequena internação no meio do caminho, por estafa total. Bem, vocês pagaram pelo material e nada tem a ver com isso, mas como foi meu primeiro trabalho em cursos on-line, peço desculpas pelas diversas falhas que cometi nesse caminho.

Recebi muitos emails bastante carinhosos e motivadores, e agradeço a todos. Outros foram mais ríspidos, mas conheço bem a tensão dos concursandos, sei que os nervos ficam a flor da pele, isso é natural nesse processo. Daqui a pouco passa o concurso e as tensões diminuem. Lamento não ter conseguido entregar as aulas antes.

Desejo a todos vocês muito sucesso nesse e em eventuais outros concursos. Vou tentar responder às dúvidas o mais rápido que conseguir, e torcer por todos no dia da prova. Alguns me perguntaram sobre carta na manga para a prova. Conto rapidamente um caso prático, o meu. Em 2009 teve o concurso para AFRFB, e vinha me preparando a mais de um ano. Estava muito afiado. Mas também estava muito cansado, chateado e decepcionado com o TRT. Então minha ansiedade para sair era muito grande. Com isso coloquei uma pressão imensa nos meus ombros. Resultado: nos últimos dois meses, dormia de três a quatro horas por dia, tinha insônia e ansiedade. Estudei até o último segundo para a prova.

Mesmo bem preparado, e concorrendo para um monte de vagas, estava exausto na prova, e errei as piores besteiras do mundo. Fiquei a um ponto de ir para a discursiva.

Agora em 2011 encarei o TCU, um sonho antigo. Concorri para uma única vaga no Amazonas. Me considerava um azarão na disputa. Mas não fiz nenhum esforço sobre-humano. Aproveitei anos de preparação, atualizei o que tinha que atualizar de conhecimento, estudei as novidades, fiz exercícios, enfim, uma preparação padrão, mas sem esperar nada. Nos dias anteriores fui ao cinema com a esposa e filha, fiz passeio nas cachoeiras e por ai vai. No dia da prova de manhã estava fazendo uma das coisas que mais me relaxam, cozinhando e ouvindo meus rocks. Fiz as duas provas em tardes de sábado e domingo totalmente relaxado, mas muito focado no que eu sabia (afinal, no CESPE chutar é perigoso). O resultado meus amigos foi esse, me sinto o cara mais realizado profissionalmente, feliz demais por alcançar o TCU, de certa forma orgulhoso por ter passado para uma única vaga, e em paz comigo.

Então, meu único conselho, se me permitem dizer, é estudar até onde der, confiar no que vocês estudaram, e descansar a cabeça pelo menos na véspera.

No mais, muito sucesso a todos, que Deus lhes abençoe, obrigado pela preferência. Mais uma vez desculpem as falhas, e que a Força esteja com vocês. Para aqueles que passarem, parabéns, mandem o convite para o churrascão da posse, e lembrem, o Brasil conta com vocês e confia demais nos órgãos de controle. Para aqueles que não conseguirem, é questão de tempo, logo chegarão lá.

Para todos, que logo fiquem com muito dinheiro no bolso e que estejam como eu quero estar em 10 anos:

7. Questões apresentadas nesta aula

1. (ESAF/AFC-TI/ STN 2005) O Data Warehouse é um conjunto de dados orientado por assuntos, não volátil, variável com o tempo e integrado, criado para dar suporte à decisão. Considerando essa tecnologia e suas derivadas ou variantes é correto afirmar que

a) a premissa do Data Mining é uma argumentação ativa, isto é, em vez de o usuário definir o problema, selecionar os dados e as ferramentas para analisar tais dados, as ferramentas do Data Mining pesquisam automaticamente os mesmos à procura de, por exemplo, possíveis relacionamentos, identificando assim problemas não identificados pelo usuário.

b) um Data Mining é considerado Top-down quando uma empresa, por desconhecer a tecnologia do Data Warehouse, prefere primeiro criar um banco de dados para somente uma área. Com isso, os custos são bem inferiores de um projeto de Data Warehouse completo. A partir da visualização dos primeiros resultados, parte para outra área e assim sucessivamente até resultar num Data Warehouse.

c) um Data Mining é considerado Bottom-up quando a empresa cria um Data Warehouse e depois parte para sua segmentação, isto é, dividindo o Data Warehouse em áreas menores, gerando assim pequenos bancos orientados por assuntos departamentalizados.

d) o propósito de uma análise de dados com a tecnologia Data Mart é descobrir, previamente, características dos dados, sejam relacionamentos, dependências ou tendências desconhecidas.

e) as ferramentas de Data Mart analisam os dados, descobrem problemas ou oportunidades escondidas nos relacionamentos dos dados, e então diagnosticam o comportamento dos negócios, requerendo a mínima intervenção do usuário.

2. (FCC/Analista-DBA/INFRAERO 2011) Funcionalidade cujo objetivo é encontrar conjuntos de dados que não obedecem ao comportamento ou modelo dos dados. Uma vez encontrados,

podem ser tratados ou descartados para utilização em mining. Trata-se de

a) descrição.

b) agrupamento.

c) visualização.

d) análise de outliers.

e) análise de associações.

3. (FCC/Analista-DBA/INFRAERO 2011) No âmbito da descoberta do conhecimento (KDD), a visão geral das etapas que constituem o processo KDD (Fayyad) e que são executadas de forma interativa e iterativa apresenta a seguinte sequência de etapas:

a) seleção, pré-processamento, transformação, data mining e interpretação/avaliação.

b) seleção, transformação, pré-processamento, interpretação/ avaliação e data mining.

c) data warehousing, star modeling, ETL, OLAP e data mining.

d) ETL, data warehousing, pré-processamento, transformação e star modeling.

e) OLAP, ETL, star modeling, data mining e interpretação/ avaliação.

4. (CESPE/Pesquisador/INMETRO 2010) As técnicas de data mining incluem

a) polarização, classificação, estimativa, granularidade e análise de afinidade.

b) análise de agrupamentos, análise de tipos, estimativa, previsão e ponto médio.

c) polarização, classificação, estimativa, previsão e análise de afinidade.

d) análise de agrupamentos, análise de tipos, estimativa, previsão e ponto médio.

e) análise de agrupamentos, classificação, estimativa, previsão e análise de afinidade.

5. (CESPE/Pesquisador/INMETRO 2010) Assinale a opção correta, com relação às classes de datamining do ponto de vista de processo orientado.

a) A modelagem de prognóstico é o processo de análise em um banco de dados para encontrar padrões escondidos sem uma ideia ou hipótese predeterminada sobre o que são esses padrões.

b) Na análise prévia, os padrões descobertos no banco de dados são usados para prognosticar o futuro.

c) O descobrimento é o processo de análise em um banco de dados para encontrar padrões escondidos sem uma idéia ou hipótese predeterminada sobre o que são esses padrões.

d) Na classe strategic mining, grande conjunto de dados corporativos é examinado com o objetivo de se obter o seu conhecimento global.

e) Na classe episodic mining, são buscados dados de um fato específico.

6. (FGV/DBA/DETRAN-RN 2010) Sobre Data Mining, pode-se afirmar que:

a) Refere-se à implementação de banco de dados paralelos.

b) Consiste em armazenar o banco de dados em diversos computadores.

c) Relaciona-se à capacidade de processar grande volume de tarefas em um mesmo intervalo de tempo.

d) Permite-se distinguir várias entidades de um conjunto.

e) Refere-se à busca de informações relevantes a partir de um grande volume de dados.

7. (CESPE/Analista Judiciário-TI/TRT21 2010) Acerca de sistemas de suporte a decisão e data warehousing, julgue os itens a seguir:

[I] O data mining é um processo automático de descoberta de padrões, de conhecimento em bases de dados, que utiliza, entre outros, árvores de decisão e métodos bayesianos como técnicas para classificação de dados.

8. (CESGRANRIO/Analista de Sistema Funcional /ELETROBRÁS 2010) Em uma reunião sobre prospecção de novos pontos de venda, um analista de TI afirmou que técnicas OLAP de análise de dados são orientadas a oferecer informações, assinalando detalhes intrínsecos e facilitando a agregação de valores, ao passo que técnicas de data mining tem como objetivo

a) captar, organizar e armazenar dados colecionados a partir de bases transacionais, mantidas por sistemas OLTP.

b) facilitar a construção de ambientes de dados multidimensionais, através de tabelas fato e dimensionais.

c) melhorar a recuperação de dados organizados de forma não normalizada em uma base relacional conhecida como data warehouse.

d) extrair do data warehouse indicadores de controle (BSC) para apoio à tomada de decisão por parte da diretoria da empresa.

e) identificar padrões e recorrência de dados, oferecendo conhecimento sobre o comportamento dos dados analisados.

9. (FCC/Analista Judiciário-Informática/TRF4 2010) Sobre data mining, é correto afirmar:

a) Não requer interação com analistas humanos, pois os algoritmos utilizados conseguem determinar de forma completa e eficiente o valor dos padrões encontrados.

b) Na mineração de dados, encontrar padrões requer que os dados brutos sejam sistematicamente "simplificados", de forma a desconsiderar aquilo que é genérico e privilegiar aquilo que é específico.

c) É um grande banco de dados voltado para dar suporte necessário nas decisões de usuários finais, geralmente gerentes e analistas de negócios.

d) O processo de descobrimento realizado pelo data mining só pode ser utilizado a partir de um data warehouse, onde os dados já estão sem erros, sem duplicidade, são consistentes e habilitam descobertas abrangentes e precisas.

e) É o processo de descoberta de novas correlações, padrões e tendências entre as informações de uma empresa, por meio da análise de grandes quantidades de dados armazenados em bancos de dados usando técnicas de reconhecimento de padrões, estatísticas e matemáticas.

Comentários:

Mais uma questão que basta conhecer o conceito de Data Mining para acertar. O resto são aquelas afirmativas para pegar os músicos, aquelas alternativas que "soam bem".

Gabarito: Letra e

10. (CESPE/Analista TI-BD/EMBASA 2010) Com relação aos sistemas de apoio à decisão, julgue os itens seguintes:

[I] A regra de classificação pode ser aplicada de várias formas, sendo uma delas a regressão. A análise de regressão é comumente aplicada quando se tem apenas um único domínio de pesquisa. No entanto, o resultado dessa análise não equivale a uma operação de data mining para a previsão da variável destino.

11. (CESPE/Analista TI-Desenvolvimento/EMBASA 2010) Com relação aos sistemas de apoio à decisão, julgue os itens seguintes:

[I] Data mining é o processo de extração de conhecimento de grandes bases de dados, sendo estas convencionais ou não, e que faz uso de técnicas de inteligência artificial. ERRADO

12. (FCC/Produção e BD/TCE-SP 2010) NÃO é um objetivo da mineração de dados (mining), na visão dos diversos autores,

a) garantir a não redundância nos bancos transacionais.

b) conhecer o comportamento de certos atributos no futuro.

c) possibilitar a análise de determinados padrões de eventos.

d) categorizar perfis individuais ou coletivos de interesse comercial.

e) apoiar a otimização do uso de recursos limitados e/ou maximizar variáveis de resultado para a empresa.

13. (FCC/Produção e BD/TCE-SP 2010) Considere uma dada população de eventos ou novos itens que podem ser particionados (segmentados) em conjuntos de elementos similares, tal como, por exemplo, uma população de dados sobre uma doença que

pode ser dividida em grupos baseados na similaridade dos efeitos colaterias produzidos.

Como um dos modos de descrever o conhecimento descoberto durante a data mining este é chamado de

a) associação.

b) otimização.

c) classificação.

d) clustering.

e) temporização.

8. Gabaritos

1	a	2	d	3	a	4	e	5	c
6	e	7	CERTO	8	e	9	e	10	ERRADO
11	X	12	a	13	d				