Mineração de Dados

Professor Manoel Mendonça
Núcleo de Pesquisa em Redes de Computadores
Universidade Salvador (UNIFACS)
E-mail: mgmn@unifacs.br
Telefone: (71) 203-2677

Sobre este Curso

- Almeja uma visão abrangente ao invés de se aprofundar em um técnica em particular
- FORMAÇÃO em vez de TREINAMENTO
- Visão geral do processo de mineração

Avaliação

- Apresentação de 1h/aula sobre um dos assuntos a seguir
- Artigo (survey) entre 10 e 25 páginas
- Estudo de caso de Mineração de Dados

Temas para Apresentação em MD

- OLAP e datawarehouses
- Descoberta de associações
- Árvores de classificação
- Redes neurais
- Redes Bayesianas
- Mineração visual e visualização de dados
- Descobertas de aglomerações
- Busca (e mineração) na Web
- Técnicas de limpeza e pré-processamento de dados
- Estudos de casos

Temas para Apresentação em GC

- Sistemas workflow
- Sistemas de suporte ao trabalho cooperativo
- Ontologias e taxonomias
- LDAP e serviços de diretórios distribuídos
- Extração automática de palavras chaves e taxonomias
- Ferramentas para desenvolvimento de portais Web
- Estudos de casos

Estudo de Caso

- Aplicação de Técnica de MD em um conjunto de dados de sua escolha.

- Escolher ferramenta em http://www.kdnuggets.com

- Descrever todo o processo de mineração usado

Trabalhos

- Carolina Passos: mapas de conhecimento
- Gabriela Resende: descoberta de associações
- Grinaldo Oliveira: redes neurais
- Josemeire Moreira: ferramentas para desenvolvimento de Portais Web

Ementa de MD (parte 1)

1. Conceitos Básicos
2. Processo de Mineração de Dados
3. Seleção e Pré-processamento de Dados
4. A Mineração Propriamente Dita
   i. Árvores de Classificação
   ii. Descoberta de Associações
   iii. Aglomeração
   iv. Redes Neurais
   v. Mineração Visual de Dados
5. Assimilação da Informação Minerada
6. Observações Finais

Parte 1: Conceitos Básicos

Motivação

Nos últimos anos tem ocorrido um contínuo barateamento da capacidade de armazenamento dos computadores e uma crescente taxa de integração entre eles. O resultado tem sido um aumento estrondoso no volume de dados disponível para acesso automático em sistemas de informações.

Todavia, a posse de dados não implica em posse de informação útil ...

No momento atual, há uma enorme demanda por ferramentas que transformem dados em informações úteis.

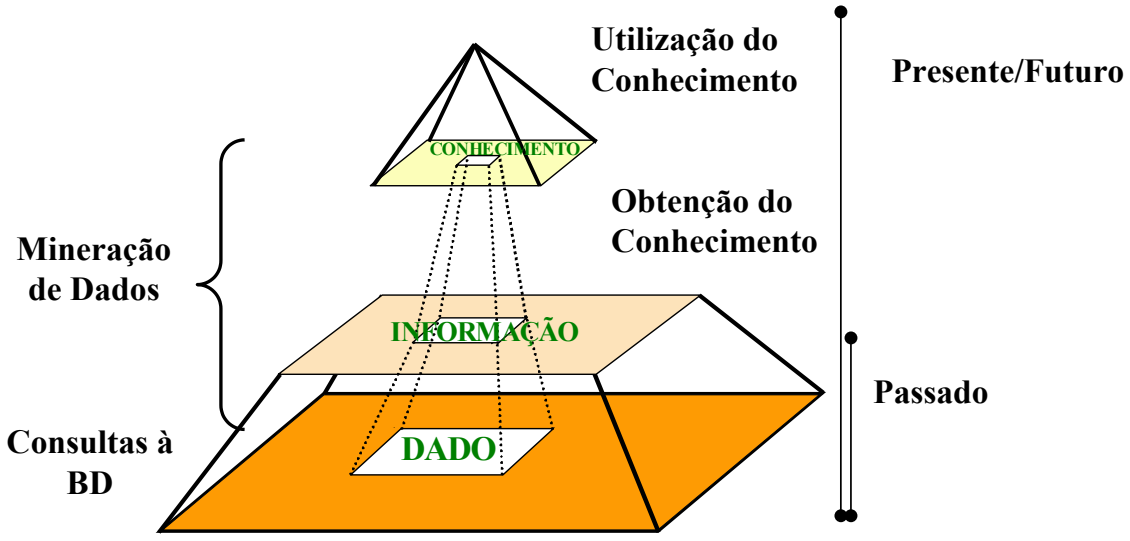
Passado

- Tecnologia limitada
- Armazenamento de pequenos volumes de dados (Mbytes)
- Consultas aos Dados
- Não existiam ferramentas para auxiliar a análise das informações obtidas

Presente/Futuro

- Grandes avanços tecnológicos na área de TI
- Armazenamento de grandes volumes de dados (Tbytes)
- Necessidade de conhecer e entender a BD
- O conhecimento extraído de uma BD deve ser usado para auxiliar as tomadas de decisões.

De Dados a Conhecimento



O Que é Mineração de Dados

Processo de extração de informação nova, não trivial, e útil de repositórios de dados.

Esta definição é abrangente e cobre um largo espectro de métodos, técnicas, e ferramentas. Este curso procura discutir o processo de mineração de dados e alguns dos métodos e técnicas usados para a sua aplicação.

Um Pouco Mais Sobre Terminologia

A definição que demos para mineração de dados equivale ao que é chamado de Descoberta de Conhecimento em Bancos de Dados ou "*Knowledge Discovery in Databases*" (KDD).

Muitos autores usam o termo KDD para se referir a todo o processo de mineração e reservam o termo Mineração de Dados para a fases central deste processo. Aqui não faremos isso.

Principais Atores Envolvidos no Processo de Mineração de Dados

• **Analista de Dados**: é um especialista em análise de dados e no uso de técnicas e mecanismos de mineração de dados.

• **Perito no Domínio**: é um especialista no domínio descrito pelos dados (não necessariamente um perito em análise de dados) que irá definir análises, interpretar a informação extraída pelas ferramentas de mineração de dados, e efetivamente transforma-las em conhecimento.

• **Usuário Final**: é aquele que vai aplicar o conhecimento obtido através da mineração de dados, seja ele expresso num modelo ou em uma descoberta do perito no domínio, no dia a dia de uma instituição.

Princípio da Mineração de Dados

Mineração de dados representa uma mudança de abordagem de análise de dados.

• em vez análise voltada à verificação de hipóteses,

• executa-se uma análise de dados voltada à descoberta.

Análise Voltada à Verificação de Hipóteses

- O Tomador de Decisões levanta uma hipótese sobre a existência de uma informação de interesse,
- coleta dados para tentar corroborar esta informação,
- e testa a hipótese formulada contra os dados coletados.

Análise Voltada a Descoberta (1)

- O Tomador de Decisões possui dados de onde quer extrair mais informações úteis
- ferramentas computacionais são usadas para "filtrar" estes dados procurando encontrar informações úteis que permaneciam escondidas nestes dados.

Este tipo de abordagem tem se tornado cada vez mais útil devido ao crescimento contínuo da complexidade e tamanho dos repositórios de dados disponíveis hoje em dia. Há muita informação útil nestes repositórios que não está sendo utilizada.

Análise Voltada a Descoberta (2)

Abordagens voltadas a descoberta não são novas à análise de dados. A principal razão para o enorme crescimento da mineração de dados são:

• as infra-estruturas de software e hardware estão maduras o suficiente para as técnicas de análise voltada à descoberta;

• a disponibilidade de técnicas de análise tem crescido dramaticamente nos últimos anos;

• o volume de dados disponíveis para análise é enorme.

Descoberta de Conhecimento Através de Análise de Dados

Conhecimento de Domínio é informação empírica, não trivial, e específica a um *domínio de aplicação* que um perito neste domínio acredita ser verdadeira.

Conhecimento de background ou de contexto é *conhecimento de domínio* que o perito já possuía antes de analisar os dados.

Conhecimento novo ou descoberto é o *conhecimento de domínio* ganho pelo perito ao analisar o dados.

Indução x Dedução

Conhecimento pode ser ganho por *indução* ou *dedução*.

• **Dedução** infere informação que é uma conseqüência lógica da informação contida no conjunto de dados. Esta informação é sempre verdadeira se o conteúdo os dados for verdadeiro.

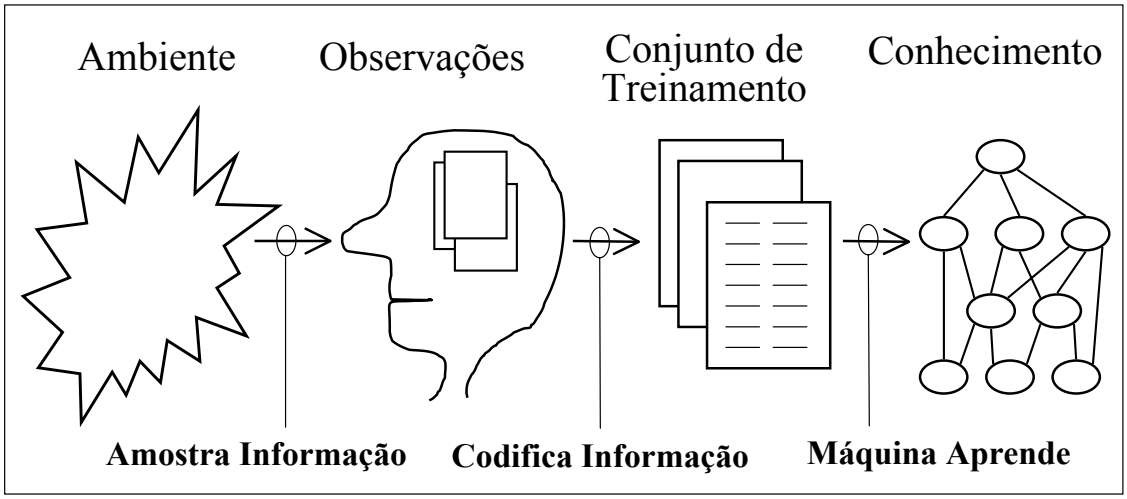
• **Indução** infere informação por generalização da informação contida no conjunto de dados. Esta informação é suposta ser verdadeira e é suportada por padrões que ocorrem no conjunto de dados sob análise.

Aprendizado Indutivo

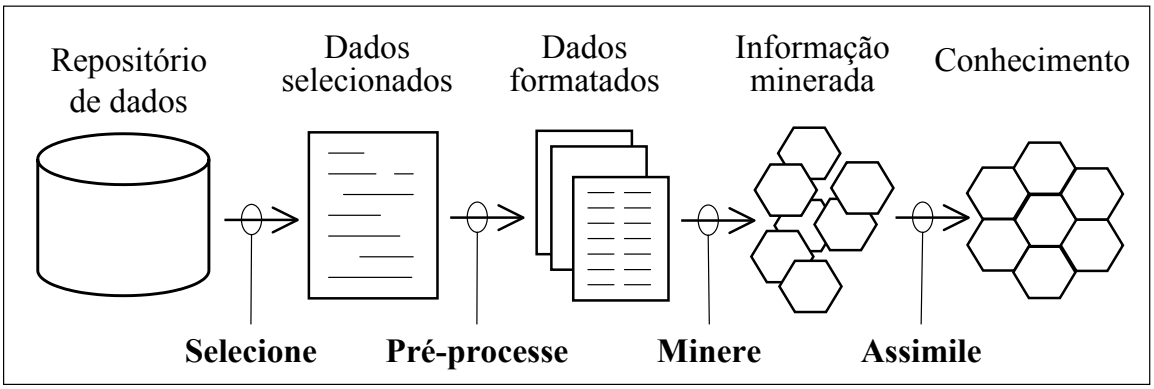
O processo de descoberta *de conhecimento de domínio* é normalmente baseado em **aprendizado indutivo**.

Em computação, a automação dos processos de aprendizado indutivo têm sido originalmente estudados em uma área da inteligência artificial chamada **aprendizado de máquina (*machine learning*)**.

Aprendizado de Máquina



Mineração de Dados



Diferenças entre Mineração de Dados e Aprendizado de Máquina

Os processo de aprendizado de máquina aparentemente similares têm duas diferenças cruciais:

• Os dados crus de repositório de dados para mineração são derivados para finalidades outras do que a mineração de dados

• O produto da mineração não é necessariamente um modelo explícito

As Principais Operações de Mineração

• Estimação e predição

• Classificação

• Aglomeração

• Descoberta de associação

• Visualização (*)

• Consultas iterativas aos dados (*)

Estimação e Predição

• Estimação consiste em examinar atributos (variáveis) de um conjunto de entidades e, baseado nos valores destes atributos, construir modelos que assinalam valores a atributos de uma nova entidade que se quer caracterizar.

• O termo predição é usado quando a estimação é usada para predizer o futuro valor de um atributo.

Classificação

• Classificação consiste em examinar os atributos (variáveis) de uma determinada entidade para, baseada nos valores destes atributos, assinalar esta entidade a uma determinada classe ou categoria.

Aglomeração

• Descoberta de aglomerações ou *clustering* consiste em segmentar uma população heterogênea em subgrupos homogêneos de entidades. Aglomeração difere da classificação no fato de não repartir registros de dados em subgrupos predefinidos. Aglomerações divide uma população na base da auto-similaridade entre registros.

Descoberta de Associações

• Descoberta de associação consiste em identificar quais atributos (variáveis) estão associados - de forma inesperada - com outros em um dado ambiente.

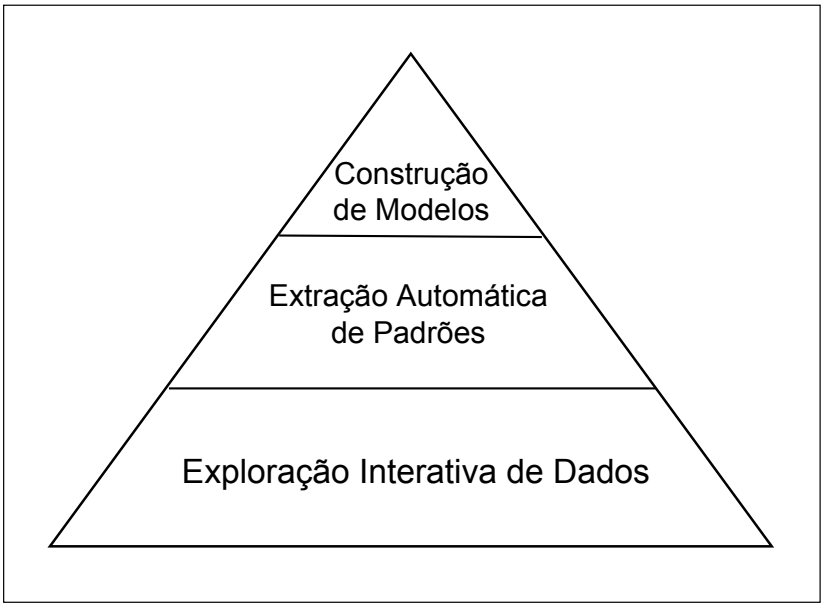
Visualização

• Visualização é a tarefa de descrever conjuntos complexos de dados em telas visuais de fácil interpretação. Visualização é baseada na premissa de que uma boa descrição visual de um conjunto de dados abstratos aumenta drasticamente a capacidade de um perito em entender o seu comportamento dentro de um certo domínio.

Consultas Interativas

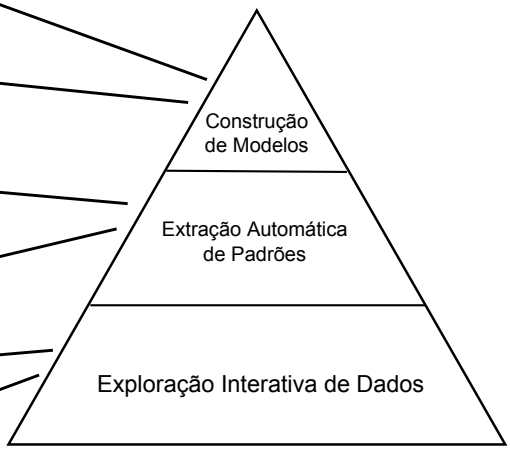
• Consultas interativas aos dados consistem em inspecionar dados através de controles interativos que permitem rapidamente formular consultas sobre os dados. Elas normalmente usam controles gráficos que permitem a formulação de “queries” com simples movimentos do mouse. Este processo normalmente é combinado com visualização para permitir a peritos rapidamente explorar visualmente conjuntos de dados.

Níveis de Mineração de Dados



As Principais Operações e os Níveis de Mineração

- Estimação e predição
- Classificação
- Aglomeração
- Descoberta de associação
- Visualização
- Consultas iterativas

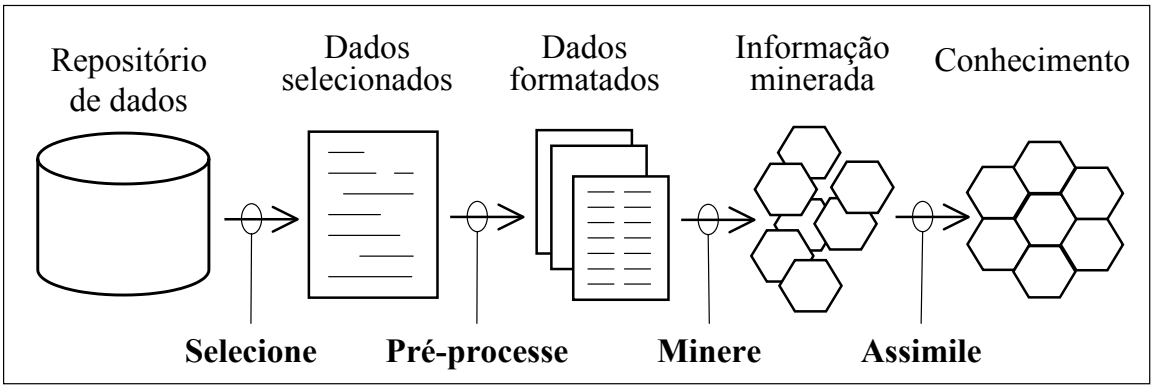


Base Tecnológica Associada

- **Repositório de Dados:** Planilhas, Sistemas de Arquivos, SGBDs, Data Warehouses.
- **Ferramentas OLAP:** análise, síntese e consolidação de dados, e apresentação de resultados.
- **Ferramentas Estatísticas:** construção de modelos, seleção de dados, amostragem de dados, inferências.
- **Ferramentas de Visualização:** seleção de dados, análise de dados, extração visual de padrões, apresentação de resultados.
- **Ferramentas de Aprendizado de Máquina:** construção de modelos, detecção automática de padrões

Parte 2: Processo de Mineração de Dados

O Processo de Mineração de Dados



Os Principais Passos do Processo

- Selecionar dados
- Pré-processar dados para análise
- Minerar os dados
- Assimilar resultados

Seleção de Dados

O primeiro passo do processo de mineração. A seleção consiste em escolher os conjuntos de dados que serão usados pelo algoritmo de mineração.

Neste passo, o analista dos dados terá que identificar aonde os dados desejáveis estão, ganhar acesso a estes dados e, se necessário, transportar os dados do seu local original para o repositório em que ele vai trabalhar.

Pré-processamento de Dados

Via de regra os dados selecionados estão crus, num formato impróprio para análise. Este passo visa "aprontar" os dados para análise.

Este passo pode envolver amostragem, limpeza, formatação, adaptação, e transformação dos dados.

Mineração

Este passo executa a mineração propriamente dita. Nele algum algoritmo é aplicado objetivando extrair informação interessante – potencialmente útil, previamente desconhecida, e não trivial – dos dados. Este passo pode envolver técnicas e ferramentas muito diversas para: construção de modelos, detecção de padrões, e/ou exploração visual de dados.

Assimilação de Resultados

O último passo do processo de mineração de dados é assimilar a informação minerada.

• No caso da construção de modelos, este passo consiste em avaliar a robustez e efetividade dos modelos produzidos. Se aprovados, os modelos devem ser incorporados aos processos operacionais da instituição na qual eles vão ser usados.

• No caso da extração de padrões e exploração visual de dados, este passo consiste em interpretar a informação extraída pelo algoritmo de mineração de dados. Isto é geralmente feito por um *perito* no *domínio da aplicação*.

Parte 2: Seleção e Pré-processamento de Dados

Importância da Seleção e Pré-Processamento de Dados

Fundamental a qualquer empreitada bem sucedida de mineração de dados, estes dois passos são algumas vezes considerados tarefas secundárias no processo de mineração de dados.

Isto é um grave erro !

Não se pode induzir boa informação de dados ruins ou com semântica obscura. Os dados precisam ser bem selecionados, limpos, e seu contexto entendido antes que eles possam ser minerados com sucesso.

Seleção e Obtenção de Dados

A seleção e obtenção de dados pode variar de uma simples consulta a um banco de dados local a uma complexa operação de migração de dados, com conversões de tipos e compatibilização de formatos heterogêneos. Este processo pode envolver ainda negociações para ganhar acesso aos dados e sobre a divulgação de informação proprietárias.

Passos para Seleção e Obtenção dos Dados

Os seguintes passos estão freqüentemente associados a seleção e obtenção de dados:

- Localizar possíveis fontes de dados
- Entender a semântica dos dados armazenados
- Entender como compatibilizar dados de fontes heterogêneas
- Obter autorização para acessar estes dados
- Entender como os dados estão armazenados na fonte de dados
- Entender como eles podem ser acessados
- Extrair os dados

Pré-processamento dos Dados

Dados extraídos diretamente de repositórios geralmente não estão prontos para mineração. Durante e depois do processo de extração, os dados devem ser formatados, limpos, e adaptados para mineração.

Algumas Operações Básicas de Pré-processamento

- Padronização de codificação
- Concatenação e decomposição de campos
- Formato de representação
- Limpeza de caracteres
- Limpeza de dados
- Redução do conjunto de dados

Padronização de Codificação

• Envolve a transformação do fluxo de caracteres de entrada para um padrão aceitável ao algoritmo de mineração de dados. Isto é feito quando a ferramenta de mineração é sensível a certos caracteres especiais, não diferenciam entre letras maiúsculas e minúsculas, ou usam códigos diferentes (ex. UNICODE e ASCII).

Concatenação

• Usada para combinar múltiplos campos de dados em um só o valor para mineração. Um exemplo típico são endereços, normalmente codificados em campos como Rua, Número, Cidade, Estado. Estes campos deve ser concatenados em uma única cadeia de caracteres se o analista deseja tratar endereço como um único atributo.

Decomposição

• Operação oposta à concatenação. Consiste em extrair vários atributos de um único campo de dados. Exemplo típico é um campo de endereço de onde pode se extrair os campos, RUA, ESTADO, CIDADE, e CEP. Note que esta operação pode se tornar bastante sofisticada podendo requerer a interpretação semântica do que está escrito no campo original.

Formatação de Representação

• É usada para padronizar valores que são representados em formatos diferentes por diferentes fontes de dados. Um exemplo típico é data que no estilo brasileiro é representada por DD-MM-YYYY e no estilo americano é representada por MM-DD-YYYY.

Limpeza de Caracteres

• Alguns caracteres são interpretados de forma errônea ou simplesmente não são aceitos por certas ferramentas. Casos típicos é o uso do caractere "$" em valores monetários, ou o uso de pontos e vírgulas em números no padrão inglês e português.

Limpeza de Dados

• Freqüentemente a fonte de dados possui campos com valores faltando, valores que não são de interesse, ou simplesmente valores errados. Estes campos devem ser tratados pelo analista. Pode-se fazer interpolações, entrar códigos especiais (ex. "não se aplica"), colocar código de nulo, ou simplesmente eliminar os registros com estes campos. A ação a ser feita deve considerar o tipo do dado, sua semântica, e seu impacto no processo de mineração.

Redução do Conjunto de Dados

• Alguns conjuntos de dados podem ser grandes demais para certos algoritmos de mineração. Neste caso o analista pode repartir o conjunto de dados em conjuntos menores e mais específicos, ou amostrar o conjunto de dados antes da mineração.

Operações Avançadas de Pré-processamento

Algumas vezes, precisa-se transformar os valores codificados nos dados para que eles possam ser adequadamente usados por certos algoritmos de mineração. Estas operações são geralmente mais complexas que as transformações sintáticas vistas anteriormente pois elas podem afetar a semântica dos dados. Elas devem ser feitas com muito cuidado e, sempre que possível, envolver um perito no domínio de aplicação.

Algumas Operações Avançadas de Pré-processamento

- Redução de escala
- Extensão da escala
- Conversão de unidades
- Normalização de valores
- Adaptação de conjunto de dados

Escalas (1)

- Nominais (Enumeradas e Discretas)

  - **Categórica**: descreve apenas categorias, únicas noções relevante sobre seus valores são a de igualdade ou diferença.
    Exemplos: masculino, feminino
    verde, azul, amarelo

  - **Ordinal**: além de categorias estabelece uma relação de ordem sobre seus valores.
    Exemplos: pequeno, médio, grande
    ensolarado, nublado, chuvoso
    ruim, regular, bom, ótimo

  Note que não existe uma noção de unidade (distância entre dois valores). Ex., distância entre "regular" e "bom" não é necessariamente a mesma que a distância entre "bom" e "ótimo".

Escalas (2)

• Quantitativas (numéricas, contínuas)

• **Intervalo**: introduz a noção de unidades fixas e iguais.
Exemplos: temperatura em graus Celsius
tempo de calendário

• **Razão**: introduz a noção de zero como ausência do atributo quantificado.
Exemplos: temperatura em graus Kelvin
tempo de projeto

• **Outras**: escalas logarítmicas (ótimas para expressar atributos com variações geométricas de valores), ex., volume de som => decibéis.

Redução de Escala

• Alguns algoritmos lidam apenas com escalas nominais (categóricas ou ordinais). Nestes casos se deve mapear campos numéricos de interesse em campos categóricos, para que estes possam ser usados nas análises. Nestes casos tem que se mapear intervalos numéricos para categorias definidas por analistas (perdendo-se a noção de unidade associada à escala numérica original).

Exemplo de Redução de Escala (1)

Suponha que se queira mapear a variável numérica "tamanho da instalação" para a escala <*pequena, média, grande*>. O mapeamento entre as duas escalas pode ser feita de diversas formas. Uma delas é usar a estatística descritiva dos dados numéricos para determinar o que seria "pequena", "média", ou "grande". Por Exemplo:

• O valor é "médio" se ele estiver dentro do intervalo Valor Médio ± 1 Desvio Padrão
• O valor é "pequeno" se ele for menor que o Valor Médio – 1 Desvio Padrão
• O valor é "grande" se ele for maior que o Valor Médio + 1 Desvio Padrão

Exemplo de Redução de Escala (2)

Outra para se mapear a variável numérica "tamanho da instalação" para a escala <*pequena, média, grande*>, seria considerar opinião perita:

• Neste caso, um especialista no domínio seria perguntado o que deveria ser considerado "pequena", "média", e "grande".
• Ele poderia dizer, por exemplo, o que a gerência considera uma instalação, pequena média, e grande.
• Note que esta solução seria mais desejável se estas análises fossem, em última instância, ser usadas por pela gerência geral.

Extensão da Escala

• Alguns algoritmos trabalham apenas com escalas numéricas, nestes casos tem que se transformar campos categóricos de interesse em campos numéricos que possam ser utilizados pelo algoritmo. Como campos categóricos apenas definem a pertinência ou não a uma categoria, a escala numérica usada é normalmente binária. Uma abordagem comum é transformar uma variável categórica em um conjunto de atributos numéricos. Cada novo atributo corresponde a um valor do atributo original. Estes novos atributos são então atribuídos os valores 0 ou 1 a depender do valor do atributo original.

Exemplo Extensão da Escala

• Considere a escala categórica para "tamanho da instalação" discutida anteriormente. Se quisermos tratar os seus valores como números, teríamos que criar três atributos "instalação é pequena?", "instalação é media?", e "instalação é grande?". Cada uma destas variáveis assumiria então o valor 0 ou 1, conforme a categoria do tamanho da instalação. No caso de um registro com *tamanho da instalação = pequena*, teríamos:

instalação é pequena?=1
instalação é media? =0
instalação é grande? =0

Conversão de Unidades

- É comum que diferentes fontes de dados representem a mesma variável em diferentes escalas. O analista deve assegurar que variáveis deste tipo sejam usadas consistentemente durante a análise. Unidades heterogêneas têm que ser convertidas para uma unidade comum, geralmente aquela usada localmente. Esta tarefa aparentemente simples pode se tornar bastante complexa.

Exemplo de Conversão de Unidades

- Considere o caso de uma unidade de mão de obra, pessoa-semana. Precisa-se saber exatamente o que quer se dizer por 1 pessoa-semana de esforço. Uma pessoa-semana no Brasil, onde se trabalha 44 horas por semana, é diferente de uma pessoa semana na Alemanha, onde se trabalha 36 horas por semana. Neste caso precisa-se multiplicar a unidade alemã por 0,818 para se obter um valor compatível com o brasileiro. Este tipo de conversão pode ser muito mais complexo para variáveis cuja semântica não sejam tão claras quanto as do exemplo acima.

Normalização de Valores

• Algumas técnicas de mineração requerem que valores sejam normalizados para um certo intervalo, geralmente de 0 a 1. O valor mínimo é mapeado para 0 e o valor máximo para 1. Todos os valores no intervalo são então mapeados para o intervalo normalizado [0,1]. Isto é normalmente feito para valores numéricos mas pode ser feito também para valores categóricos, como mostrado anteriormente.

Operações de normalização podem ser bastante sofisticadas, com tratamento de pontos fora da curva e mapeamentos não lineares.

Exemplos de Normalização de Valores

• **Normalização Linear**: dada uma variável X, o seu valor normal numa escala [0,1] é: $\overline{X} = \frac{x - X_{\min}}{X_{\max} - X_{\min}}$

• **Normalização Logarítmica**: dada uma variável X, o seu valor normal numa escala [0,1] é:

$$\overline{X} = \frac{LOG(x) - LOG(X_{\min})}{LOG(X_{\max}) - LOG(X_{\min})}$$

Adaptação de Conjunto de Dados

• Alguns pontos fora da curva e/ou conjuntos de dados desbalanceados podem afetar drasticamente os resultados de alguns algoritmos de mineração de dados. Existem operações que podem mitigar este problema. Duas das principais são: (1) a eliminação de pontos fora da curva; (2) a combinação de conjuntos de registros em um "conjunto equivalente" que representa vários registros de uma só vez.

Parte 3: A Mineração Propriamente Dita

Técnicas de Mineração

Neste mini-curso vamos apresentar algumas técnicas de mineração de dados. O objetivo aqui não é ser exaustivo ou discutir qualquer técnica em profundidade, mas dar uma idéia da diversidade de técnicas que estão disponíveis para analistas de dados. As técnicas (brevemente) apresentadas são:

- árvores de classificação
- descoberta de associações
- aglomerações
- redes neurais
- mineração visual de dados

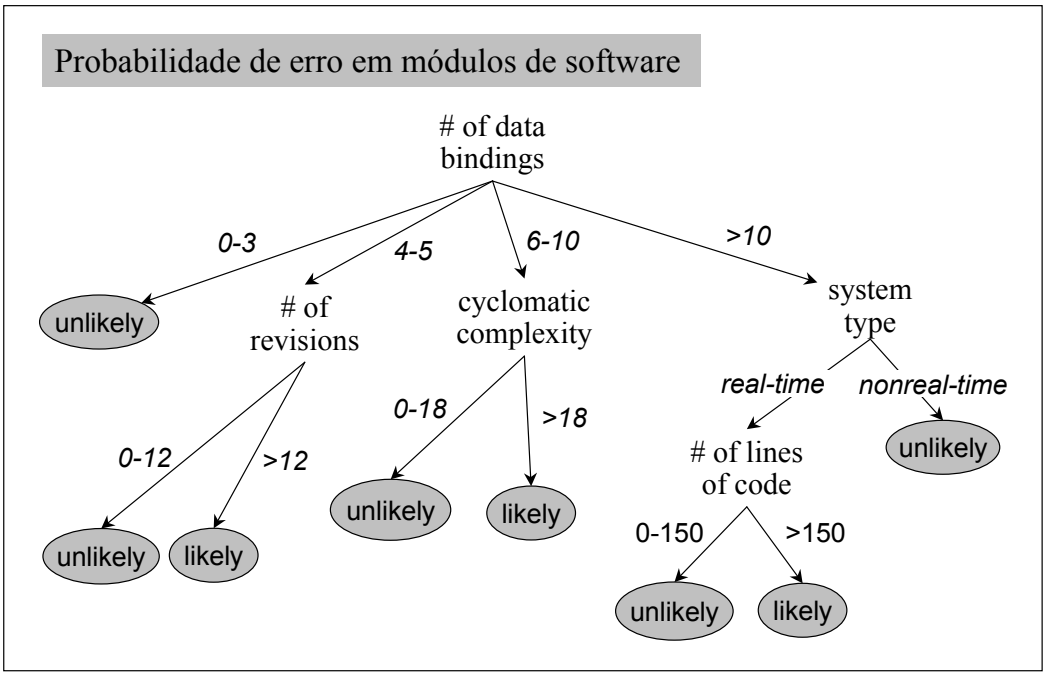
Existem muitas outras técnicas disponíveis ...

Parte 3.1: Árvores de Classificação

Árvores de Classificação ou Decisão

Árvores de classificação ou de decisão são técnicas de indução usadas para descobrir regras de classificação para um atributo a partir da subdivisão sistemática dos dados contidos no repositório sendo analisado.

Um Exemplo



Princípio da Árvore de Classificação

Os algoritmos que constroem árvores de classificação buscam encontrar aqueles atributos e valores que provêem máxima segregação dos registros de dados – com respeito a variável que ser quer classificar – a cada nível da árvore. Na transparência anterior, "# of data bindings" foi selecionada primeiro pois foi a variável que mais claramente dividiu os registros em respeito a "probabilidade de erro" no conjunto de dados analisado.

O Algoritmo ID3 (Quinlan)

1 - Selecione uma variável como raiz da árvore, crie tantos galhos quanto valores existirem para esta variável;

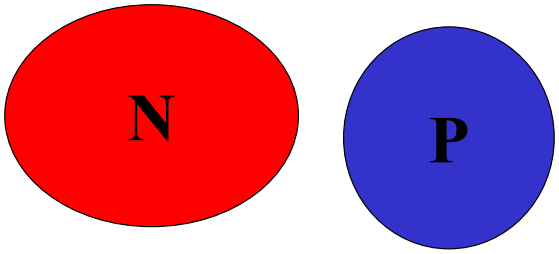
2 - Use a árvore gerada para classificar o conjunto de treinamento. Se todos os exemplos em uma folha tiverem o mesmo valor para a variável sendo classificada, etiquete a folha como este valor. Se todas as folhas forem etiquetadas o algoritmo termina.

3 - Caso contrário, etiquete a folha como um nódulo não terminal com uma variável que ainda não foi usada nos seus nódulos ancestrais, crie todos os galhos possíveis para ele, e retorne ao passo 2.

Seleção da Variável Raiz (1)

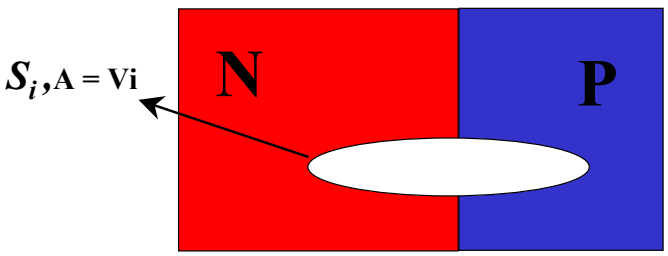
Chave do algoritmo é a seleção da variável de classificação a cada nível. A seleção funciona da seguinte forma:

1. O conjunto de dados é dividido em dois subconjuntos: *P* e *N*. Subconjunto *P* contém todos os elementos positivos. Subconjunto *N* contém todos os negativos. No nosso caso, P contém os módulos que têm e N os que não têm muitos erros.



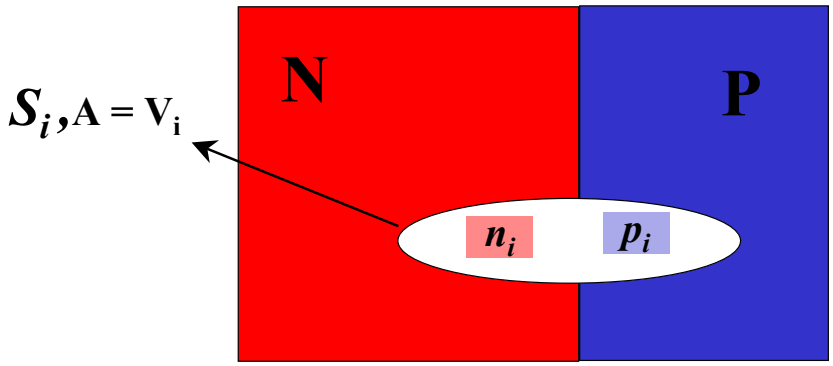
Seleção da Variável Raiz (2)

2. Para cada atributo *A* (ex., número de associações nos dados, número de revisões, etc.) e valor V_i (ex., 0-3, 4-5, 6-10, >10), determine o subconjunto S_i de elementos para os quais *A* está em V_i (ex., número de associações nos dados >10).



Seleção da Variável Raiz (3)

3. Conte o número n_i de elementos que S_i tem em N e o número p_i de elementos que S_i tem em P.



Seleção da Variável Raiz (4)

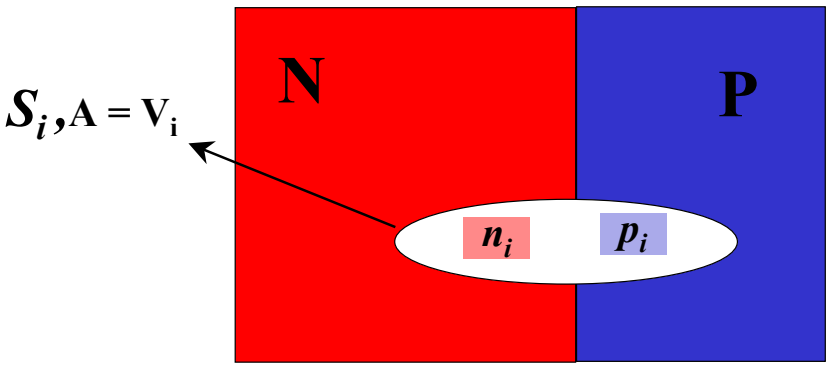
4. Calcule a quantidade de informação necessária para decidir se um elemento arbitrário em S_i pertence a P ou a N usando a seguinte fórmula:

$$I(p_i, n_i) = -\frac{p_i}{p_i + n_i}\log_2\left(\frac{p_i}{p_i + n_i}\right) - \frac{n_i}{p_i + n_i}\log_2\left(\frac{n_i}{p_i + n_i}\right)$$

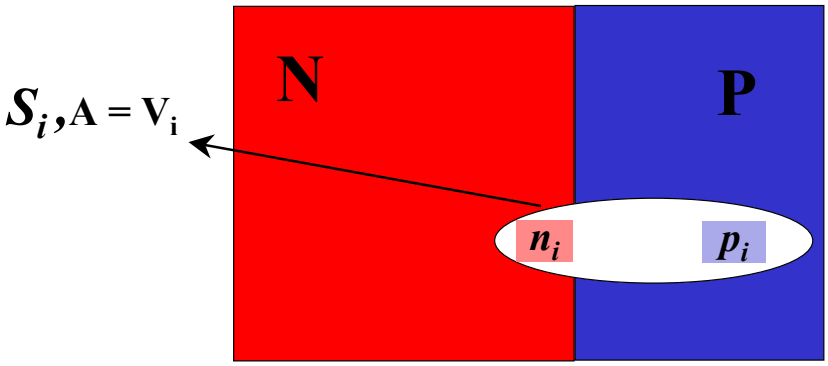
$I(p_i,n_i)$ indica a seletividade do valor V_i do atributo A em relação à variável dependente.

$I(p_i,n_i)$ é mínimo se p_i ou n_i=0, e máximo se p_i=n_i

Seletividade de A=Vi



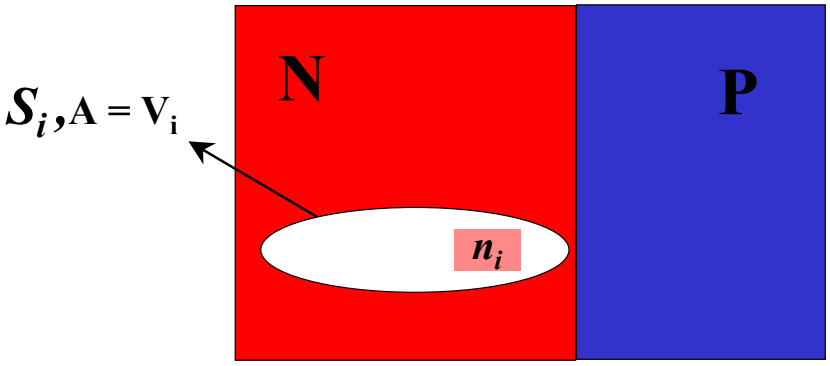
Boa Seletividade de A=Vi



Boa seletividade => $n_i >> p_i$ ou $p_i >> n_i$

$I(p_i,n_i)$ é baixo

Máxima Seletividade de A=Vi



Máxima seletividade => n_i=0 ou p_i=0
I(p_i,n_i) é mínimo

Seleção da Variável Raiz (5)

5. Assumindo que A parte o conjunto de dados *S* nos subconjuntos $\{S_1,S_2,\ldots,S_v\}$ para os seus valores $\{V_1,V_2,\ldots,V_v\}$, e sabendo que a quantidade de informação necessária para se decidir se um elemento de S_i pertence a *P* ou a *N* é $I(p_i,n_i)$, podemos calcular a média ponderada da informação necessária para classificar elementos nas sub-árvores S_i definidas pelos valores $\{V_1,V_2,\ldots,V_v\}$ de A.

$$E(A) = \sum_{i=1}^{v} \frac{p_i + n_i}{p + n} I(p_i, n_i)$$

A informação necessária para se classificar um elemento do conjunto de dados usando um atributo A

Seleção da Variável Raiz (6)

6. Atributo A é selecionado para um nódulo da árvore de classificação se ele precisa da mínima informação $E(A_j)$ entre todos os atributos A_j que podem ser selecionados para este nódulo.

Para todos A_j possível selecione aquele com mínimo $E(A_j)$!

Parte 3.2: Descoberta de Associações

Descoberta de Associações

Descoberta de associações extrai informação útil baseada nas coincidências de valores no conjunto de dados, Descoberta de novo conhecimento ocorre quando estas coincidências são previamente desconhecidas, não triviais, e interpretáveis por um perito no domínio.

Mensagem recebida da amazon.com

We have noticed that many of our customers who have purchased albums by Antonio Carlos Jobim also enjoy music by Caetano Veloso. For this reason, you might like to know that the newest CD by Caetano Veloso, "Noites do Norte," has recently hit the shelves. You can order your copy at a savings of 22% by following the link below:

http://www.amazon.com/exec/obidos/ASIN/B000059LXU/ref=mk_pb_noy

For decades now, Caetano Veloso has been ...

To learn more about "Noites do Norte," please visit the following page at Amazon.com:

http://www.amazon.com/exec/obidos/ASIN/B000059LXU/ref=mk_pb_noy

Happy listening,

Jason Verlinde
Editor, International Music
Amazon.com

Análise de Cesta de Mercado

Técnicas de análise de cesta de mercado permite a descoberta de correlações ou co-ocorrências de eventos transacionais. Elas usam matrizes de correlação cruzada nas quais a probabilidade de um evento ocorrer em conjunção com cada um dos outros eventos é computada na teoria e na prática. Os eventos com maiores (ou menores) correlações "inesperadas" (diferença entre teoria e prática) são selecionados para inspeção.

Matriz de Correlação Cruzada

	Ie	LS	SS	LC	LNR	SNR	LFO	LFI
	38%	53%	15%	55%	67%	18%	45%	48%
Ie ∧		22% (20,1)	12% (5,7)	23% (20,9)	25% (25,5)	8% (6,8)	31% (17,1)	33% (18,2)

Note que eventos transacionais podem ser mapeados para a ocorrência de um (ou um conjunto de) valor(es). Desta forma, o número de associações possíveis cresce rapidamente com o número de variáveis medidas. Necessita-se nesta situação de uma função para automaticamente selecionar as associações mais interessantes para análise dos peritos.

Funções de Interessantismo

Funções de interessantismo são normalmente usadas para quantificar quanto uma associação pode de ser interessante para um perito. Uma associação é interessante se ela é:

- improvável,
- significante,
- nova,
- valiosa,
- e interpretável.

Associações Improváveis

Uma associação é improvável se ela não for conhecimento comum e não for esperada pelo perito. Uma função para quantificar associações improváveis pode usar uma fórmula do tipo:

$| TE(e_i) - TO(e_i) |$, onde TE é a taxa esperada de ocorrência o evento e_i e TO é a taxa observada de ocorrência o evento e_i nos dados.

Um Exemplo (a)

$$Interestingness_1(A,B) = |P(A/B) - P(A)| = \left|\frac{P(A \wedge B)}{P(B)} - P(A)\right| = \left|\frac{P(A \wedge B) - P(A)P(B)}{P(B)}\right|,$$

onde $P(A)$ é a percentagem de ocorrência de A no conjunto de dados e $P(A \wedge B)$ é a percentagem de ocorrência de $A \wedge B$, juntos, no conjunto de dados.

Esta função tende a se aproximar de zero se A e B são estatisticamente independentes, e se tornar grande caso contrário. Todavia, esta função é só uma medida matemática de associação entre valores dos atributos A e B. Uma boa função deve considerar algo do conhecimento do domínio do perito. Pode-se, por exemplo, incluir um filtro para remover associações já esperadas.

Um Exemplo (b)

Um filtro simples pode eliminar todas as associações listadas como já esperadas por peritos, e consequentemente de pouco valor agregado para a análise.

$$Interestingness_2(A,B) = Interestingness_1(A,B) * AssociationFilter(A,B)$$

$$AssociationFilter(A,B) = \begin{cases} 0, \text{ se associação (A, B) está na lista de associações esperadas} \\ 1, \text{ otherwise} \end{cases}$$

Associações Significantes

Funções de interessantismo devem considerar quanto suporte à associação existe no conjunto de dados (significância). Considere, que um evento A e um evento B só ocorram uma vez no conjunto de dados. Se eles ocorrem juntos, sua associação é de 100%. Todavia, esta associação não tem significância alguma pois é baseada em um registro apenas.

Exemplo de Significância

A função do exemplo anterior não considerava quão significante a associação era. A função abaixo acrescenta um termo para levar em conta a significância da associação no cálculo do interessantismo.

$$Interestingness_3(A,B) = Interestingness_2(A,B) * Support(A,B)$$

$$Support(A,B) = FunctionOf(\text{Number of occurrences of } A \wedge B \text{ in the data set})$$

Associações Novas

O fator "novidade" deve também ser fatorado numa FI. Uma forma de se fazer isso é se considerar a freqüência com que as variáveis da associação sendo considerada no momento estão presentes em associações mineradas anteriormente.

Exemplo de Novidade

Considere as principais associações escolhidas por $Interestingness_3(A,B)$, pode-se iterativamente ajustar os seus valores baseado no número de aparições de A e B em outras associações com escores de altos interessantismo:

$$Interestingness_4(A,B) = Interestingness_3(A,B) * Novel(A,B)$$

$$Novel(A,B) = FunctionOf\left(\frac{1}{\text{appearances of A and B on } Interestingness_3(A,B) \text{ top rank}}\right)$$

Valor de Uma Associação

Este fator é subjetivo, altamente dependente do domínio, e da opinião perita. Uma forma de se capturar parte deste valor é permitir que os peritos ordenem os atributos por importância.

Exemplo de Cálculo de Valor

$$Interestingness_5(A,B) = Interestingness_4(A,B) * Value(A,B)$$

$$Value(A,B) = FunctionOf(\text{A and B importance rank according to the domain expert})$$

Facilidade de Interpretação de uma Associação

Como antes, este fator é subjetivo, altamente dependente do domínio, e da opinião perita. Uma forma de se considerar este fator, e assumir que associações "mais simples" são mais fáceis de se interpretar. Associações com menos variáveis e/ou envolvendo variáveis importantes/genéricas devem ser consideradas primeiro.

Parte 3.3: Aglomerações

Aglomerações

Técnicas de aglomeração (clustering) usam o conceito de aglomeração é bastante simples. Deseja-se agrupar entidades similares juntas para se abstrair informações de alto nível sobre cada um destes grupos.

Distâncias

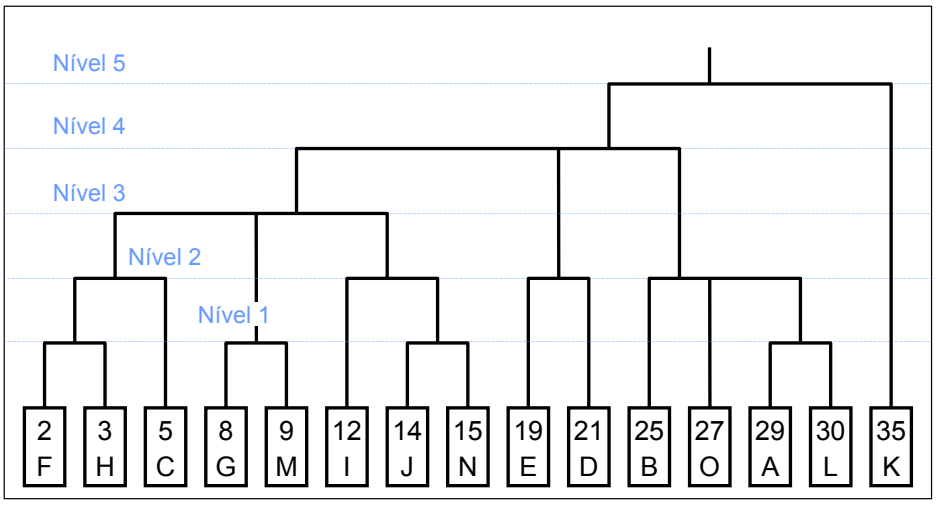
O conceito mais importante em algoritmos de aglomeração é a da distância entre dois registros. As diferenças entre os valores das variáveis que caracterizam cada registro podem ser usadas para se calcular a distância absoluta entre eles.

Exemplo de Cálculo de Distância e Aglomeração em uma Dimensão (1)

No sua instância mais simples, a distância entre estes registros é calculada a partir da diferença de valores de uma só variável. Considere o uso da variável "número de modificações" para agrupar os módulos de software listados abaixo:

Modulos	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
Modificações	29	25	5	21	19	2	8	3	12	14	35	30	9	15	27
Número Ciclomático	122	132	21	85	87	23	19	24	34	84	134	110	124	89	129

Exemplo de Cálculo de Distância e Aglomeração em uma Dimensão (2)



Exemplo de Cálculo de Distância e Aglomeração em uma Dimensão (3)

Na figura anterior um algoritmo de aglomeração hierárquico chamado de “ligação simples” é usado para agrupar os registros. O algoritmo aglomera registros em grupos repetidamente até que só reste um grupo com todos um registros juntos.

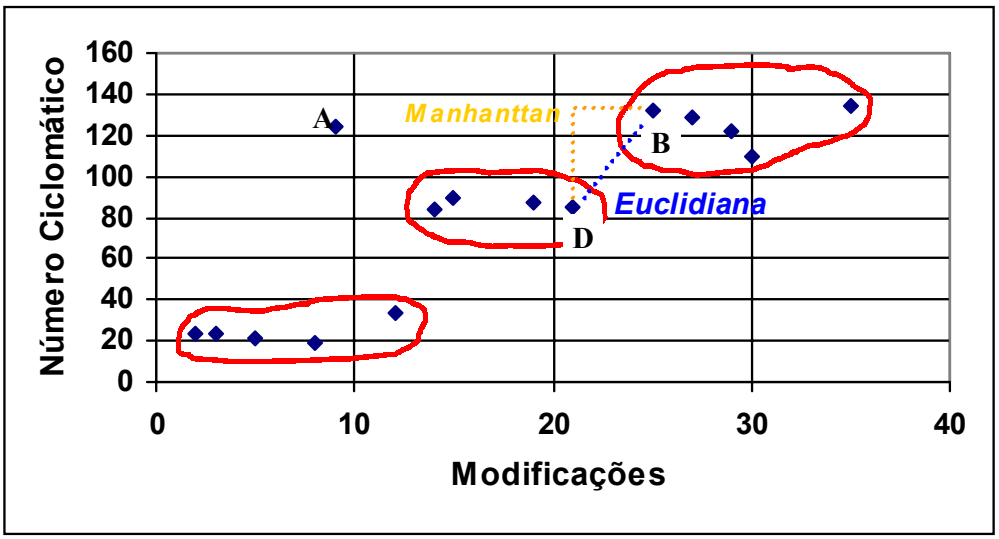
A medida de distância usada para aglomerar entre grupos é a menor distância entre pares de registros de grupo diferentes no nível inferior.

Calculando Distâncias Usando Mais de Uma Variável (1)

Modulos	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
Modificações	29	25	5	21	19	2	8	3	12	14	35	30	9	15	27
Número Ciclomático	122	132	21	85	87	23	19	24	34	84	134	110	124	89	129

Qual a distância entre B e D ?

Calculando Distâncias Usando Mais de Uma Variável (2)



Calculando Distâncias Usando Mais de Uma Variável (3)

$$\text{Euclidiana}(B,D) = \sqrt{(a_1(B)-a_1(D))^2 + \ldots + (a_n(B)-a_n(D))^2}$$

$$\text{Manhattan}(B,D) = |a_1(B) - a_1(D)| + \ldots + |a_n(B) - a_n(D)|$$

Outros Cálculos de Distâncias

$$\text{Chebychev}(B, D) = \textit{Máximo}\left|a_i(B) - a_i(D)\right|$$

$$\text{Potência}(B, D) = \sqrt[r]{\sum_i \left|a_i(B) - a_i(D)\right|^{p_i}}$$

$$\text{Discordância}(B, D) = \left(\text{número de } a_i(B) \neq a_i(D)\right)/i$$

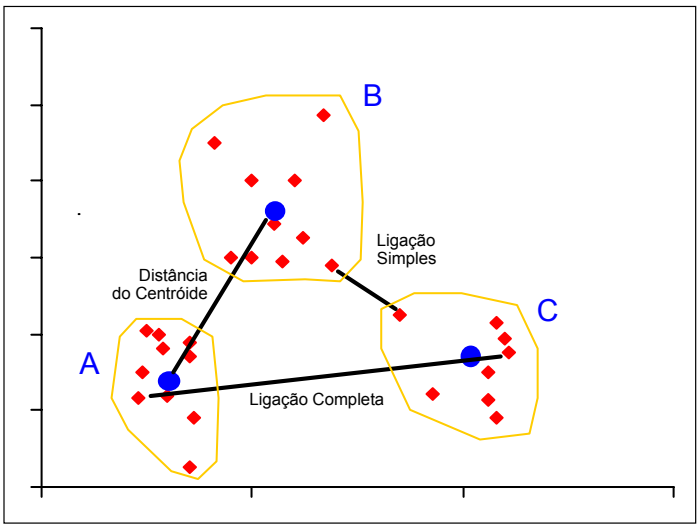
Alguns Critérios de Aglomeração Hierárquica

Ligação simples: a distância entre aglomerados é calculada como a distância entre os seus dois registros mais próximos. Este foi o critério usado no nosso exemplo anterior.

Ligação Completa: a distância entre aglomerados é calculada como a distância entre os seus dois registros mais distantes.

Distância do Centróide: a distância entre aglomerados é calculada como a distância entre os seus centróides.

Visualização de Critérios de Aglomeração Hierárquica



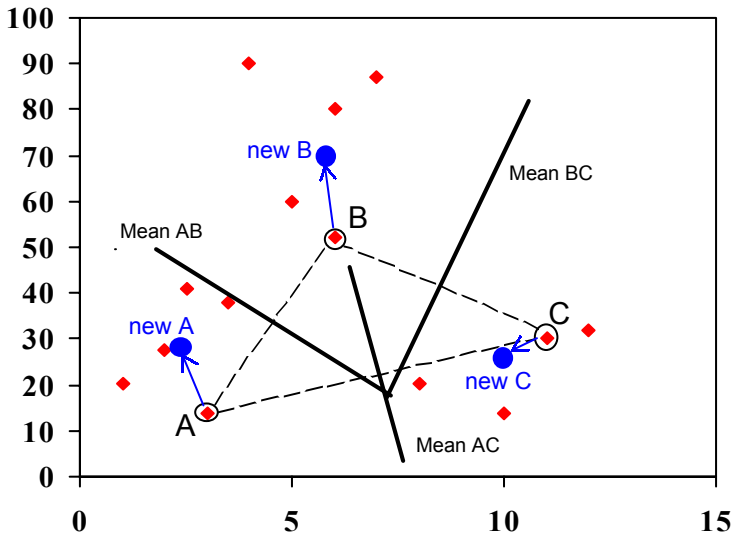
Algoritmos Não Hierárquicos

Os algoritmos hierárquicos são geralmente muito caros computacionalmente, especialmente em termos de espaço ocupado. Existem técnicas de aglomeração não hierárquicas que são mais baratos computacionalmente, estas técnicas são normalmente baseadas em métodos de re-alocação. Nestas técnicas, **um número pré-estabelecido de aglomerados é definido** e os registros são alocados (e re-alocados) a estes aglomerados, até que uma alocação estável seja alcançada. O algoritmo básico de re-alocação, conhecido por K-médias, é mostrado a seguir.

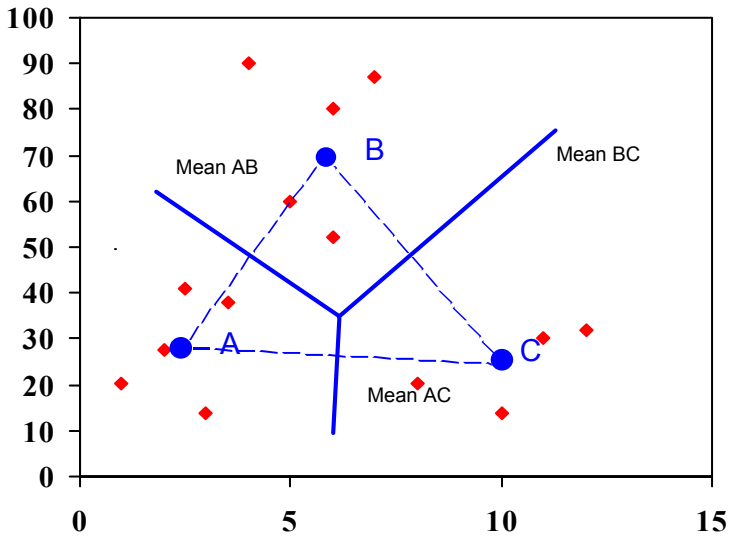
Algoritmo K-Médias (1)

1 - Selecione o número K de aglomerados que se deseja produzir.
2 - Escolha um registro como o centróide de cada um dos k aglomerados.
3 - Varra o conjunto de dados e assinale cada registro ao centróide (aglomerado) mais próximo.
4 - Recalcule o centróide como a média dos registros contidos nos aglomerados resultantes do passo 3.
5 - Repita os passos 3 e 4 até que o número de re-alocações de registros entre aglomerados seja mínimo.

Algoritmo K-Médias (2)



Algoritmo K-Médias (3)

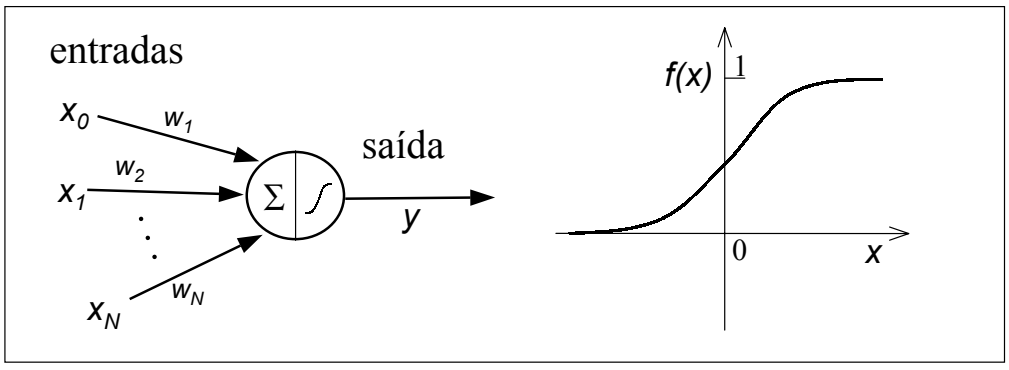


Parte 3.4: Redes Neurais

Redes Neurais

Redes neurais artificiais têm sido uma das técnicas mais difundidas para construção de modelos de classificação e estimação. Elas são redes fortemente conectadas de elementos computacionais simples, chamados nódulos ou neurônios artificiais.

Neurônio Artificial



Funcionamento do Neurônio (1)

O neurônio tem N entradas x_1, $x_{2, \ldots,}$ x_N e uma saída y, todos tendo valores contínuos em um determinado domínio, geralmente [0,1]. Cada entrada tem também um peso (w_1, $w_{2, \ldots,}$ w_N) que determina o quanto cada entrada contribui para a saída y.

O neurônio determina sua saída calculando a somatória ponderada de suas entradas e passando o resultado por uma função não linear de filtragem $f(x)$.

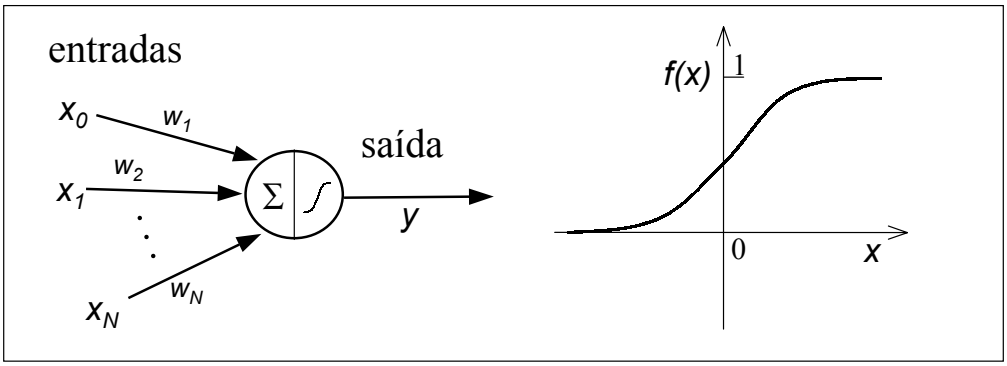
Funcionamento do Neurônio (2)

Uma função de filtragem muito usada é a sigmoide. Usando esta função, a saída do neurônio seria calculada da seguinte forma:

$$y = f\left(\sum_{i=0}^{N} w_i x_i\right), \text{ onde a função sigmoide é } f(x) = \frac{1}{1+e^{-x}}$$

$$\therefore \qquad y = \frac{1}{1+e^{-\sum_{i=0}^{N} w_i x_i}}$$

Cálculo da Saída

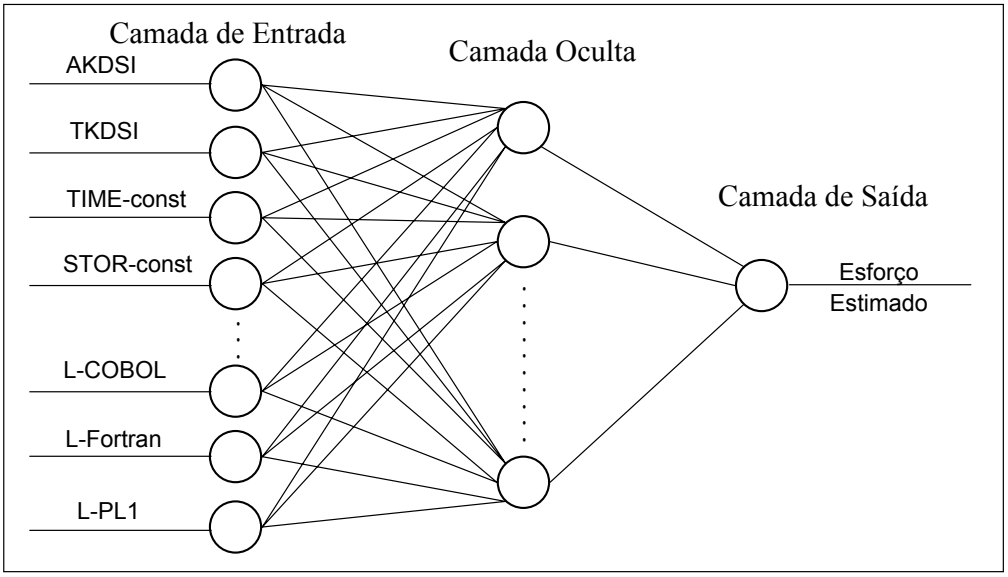


$$y = \frac{1}{1 + e^{-\sum_{i=0}^{N} w_i x_i}}$$

Construção da Rede

Redes neurais são construídas se conectando a saída de um neurônio a entrada de um ou mais neurônios. A conexões de entradas são assinaladas ao uma camada de nódulos, chamada de camada de entrada, e as saídas finais são assinaladas a outra camada de nódulos, chamada de camada de saída.

Exemplo de Uma Rede



Passos na Construção de uma Rede

(1) identificar as entradas e saídas da rede;

(2) processar os valores de entrada para que eles caiam em um intervalo numérico bem definido, normalmente entre 0 e 1;

(3) escolher uma topologia apropriada para a rede definindo cuidadosamente seus níveis intermediários;

(4) treinar a rede com um conjunto de dados representativo;

(5) testar a rede em um conjunto independente do conjunto de treinamento e retreinar a rede se necessário;

(6) aplicar o modelo gerado e avaliar os seus resultados na prática.

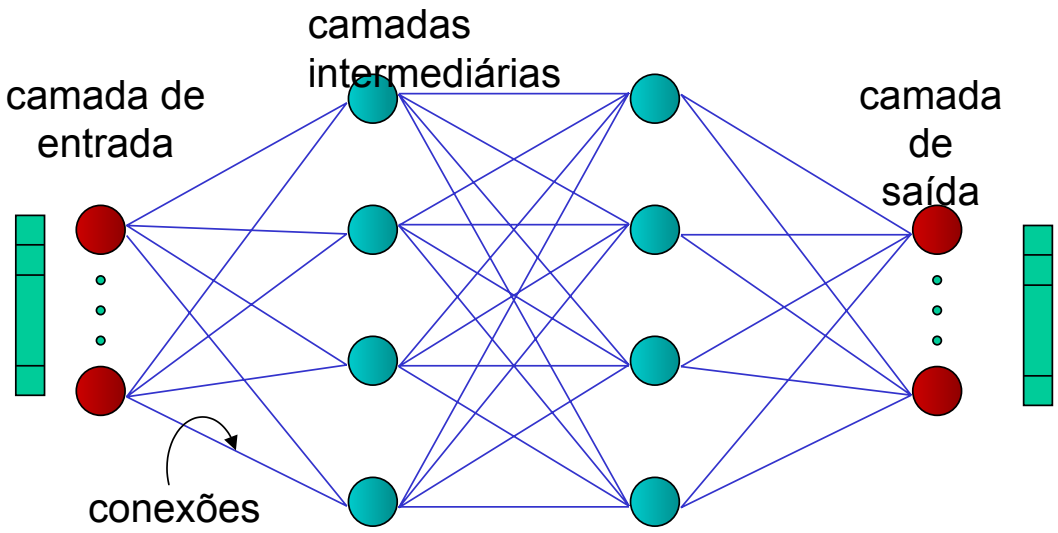
Treinamento de uma Rede

O principal desafio do processo descrito anteriormente é o treinamento da rede (passo 4). Isto é feito, ajustando-se os pesos das conexões até que a rede produza os padrões apropriados de saídas (estimativas de esforço no nosso exemplo) para os correspondentes padrões de entradas (atributos do software e fatores de custo no nosso exemplo). A idéia é usar um conjunto de treinamento para ajustar os pesos das conexões da rede até chegar ao comportamento preditivo (ou classificatório) correto.

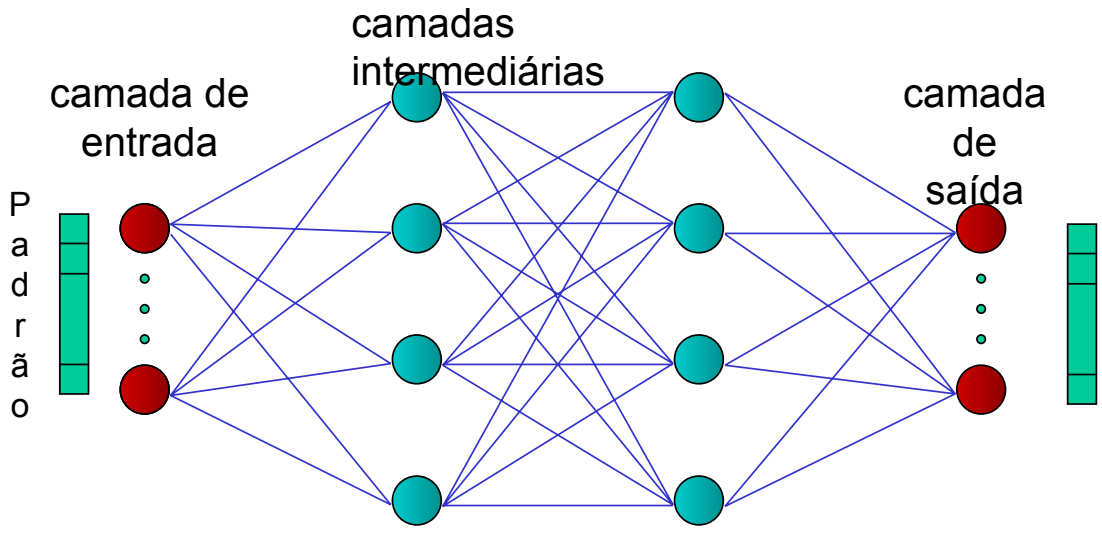
Algoritmo de Propagação Para Trás

A abordagem mais comum para se treinar redes neurais é a propagação para trás (backpropagation). O algoritmo começa com um conjunto aleatório de pesos, usa um exemplo do conjunto de treinamento para estimar sua saída, e compara esta estimativa com o valor real. Esta comparação é usada para calcular o erro de estimação (ou classificação). Este erro é então propagado para trás pela rede, com os pesos das conexões sendo recalculados para minimizá-lo.

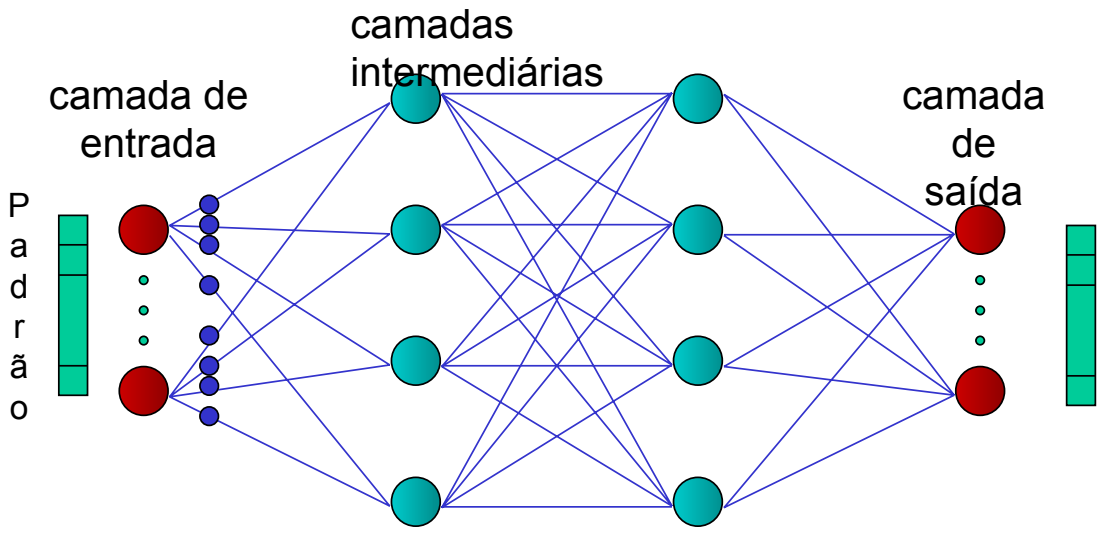
Um Exemplo Animado



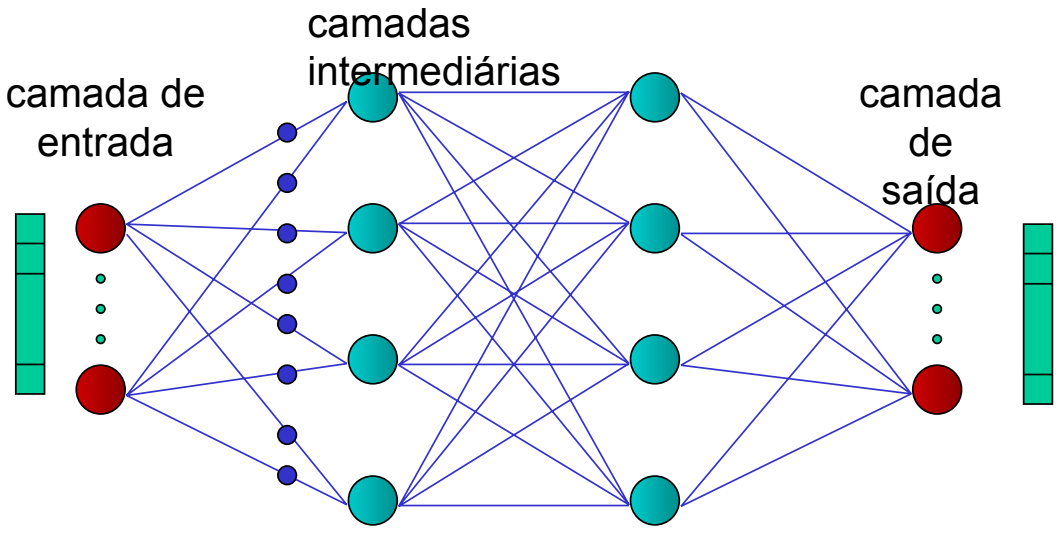
Um Exemplo Animado



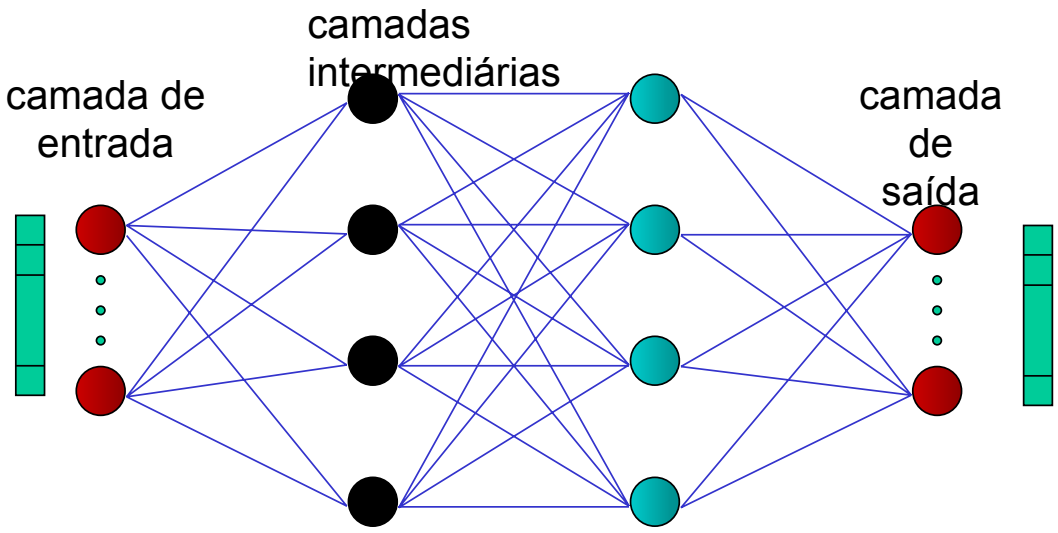
Um Exemplo Animado



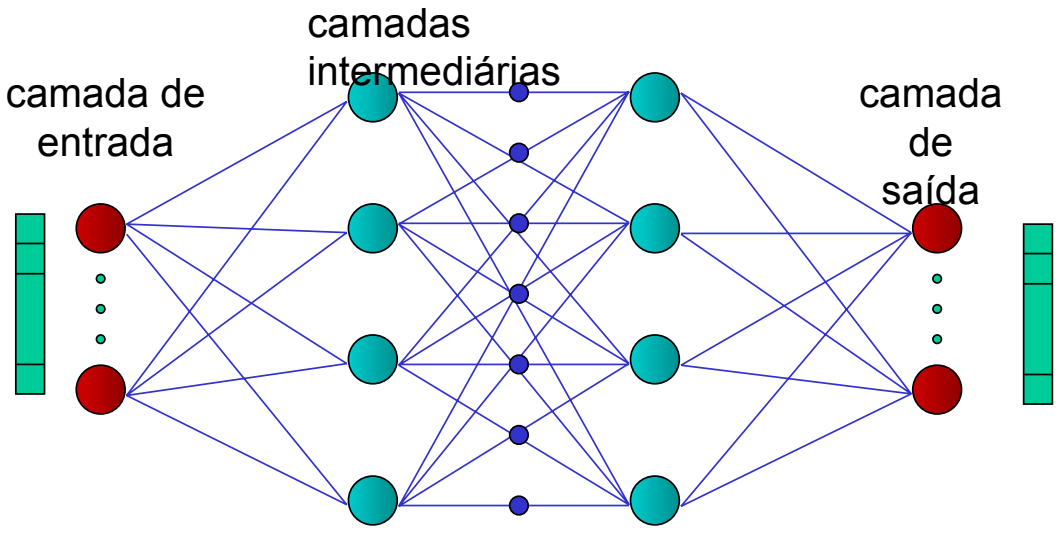
Um Exemplo Animado



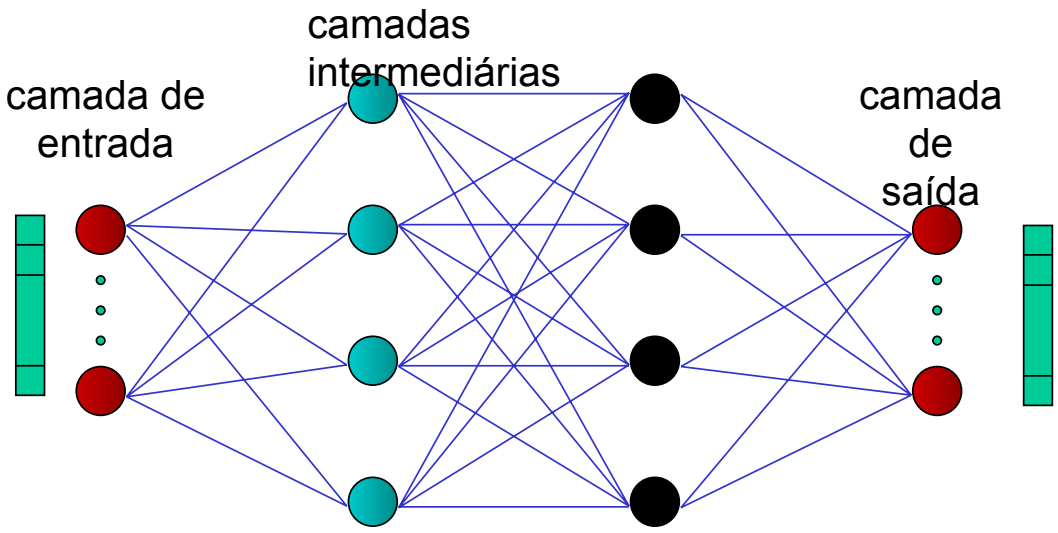
Um Exemplo Animado



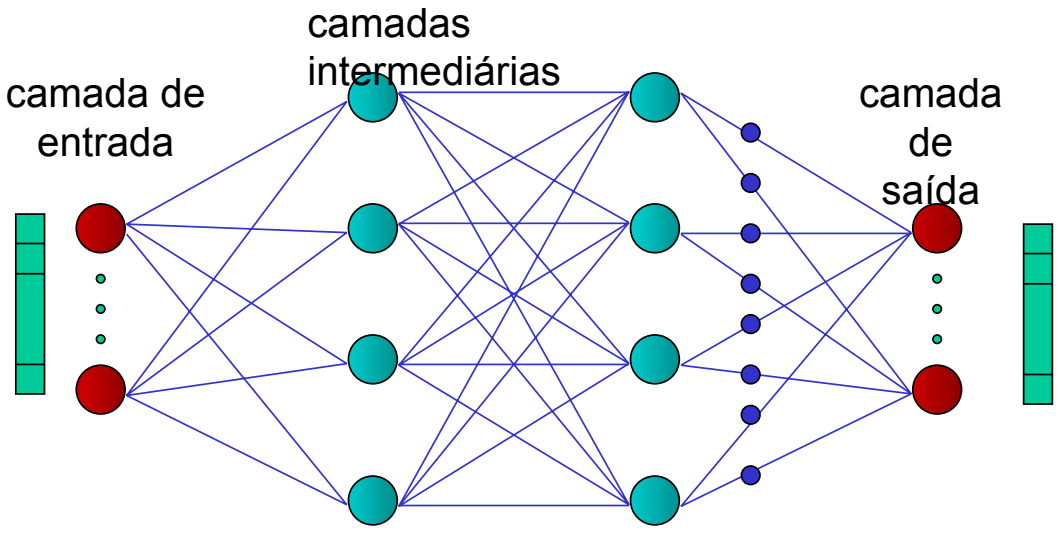
Um Exemplo Animado



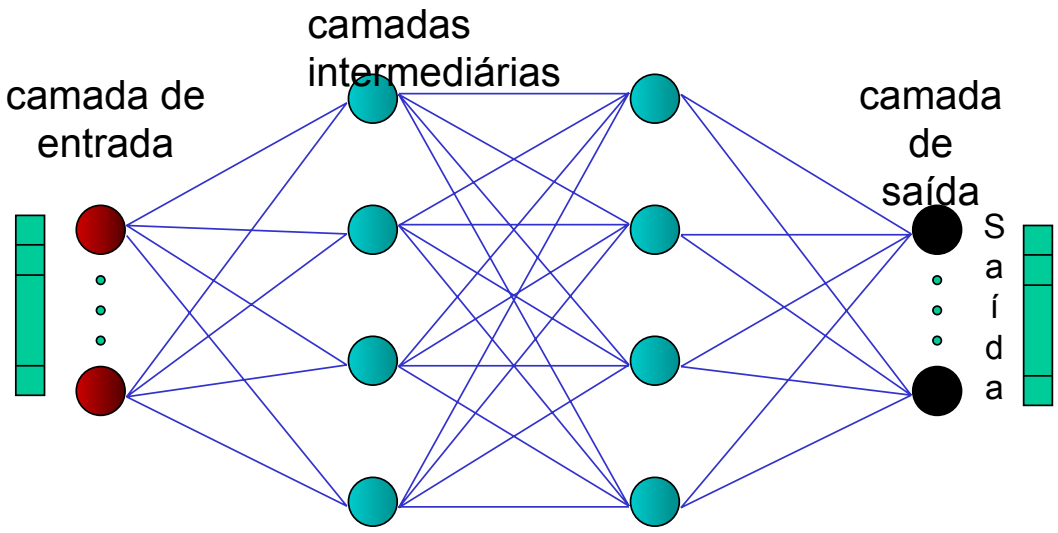
Um Exemplo Animado



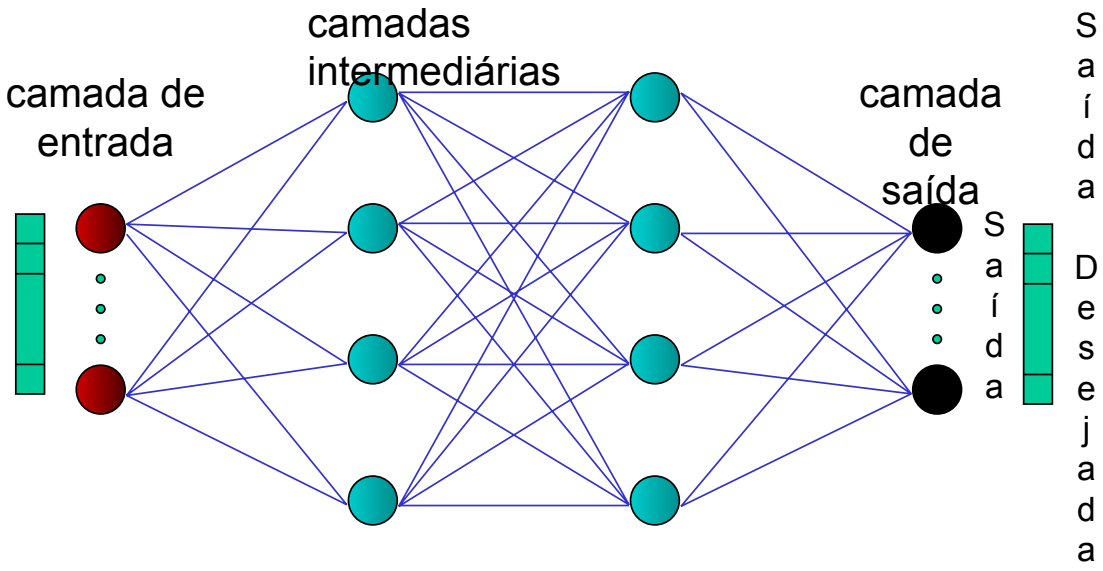
Um Exemplo Animado



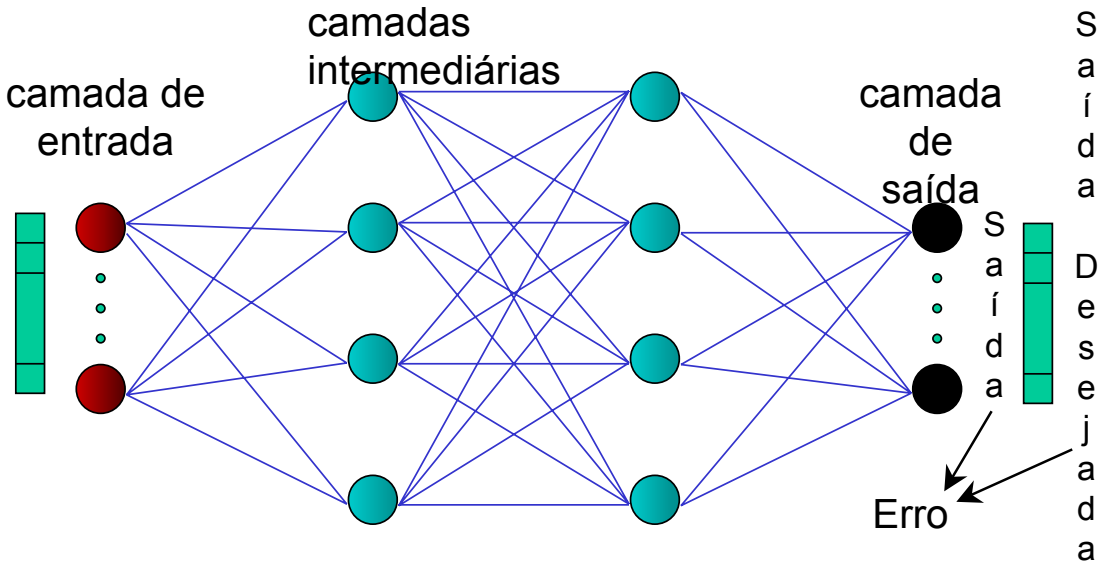
Um Exemplo Animado



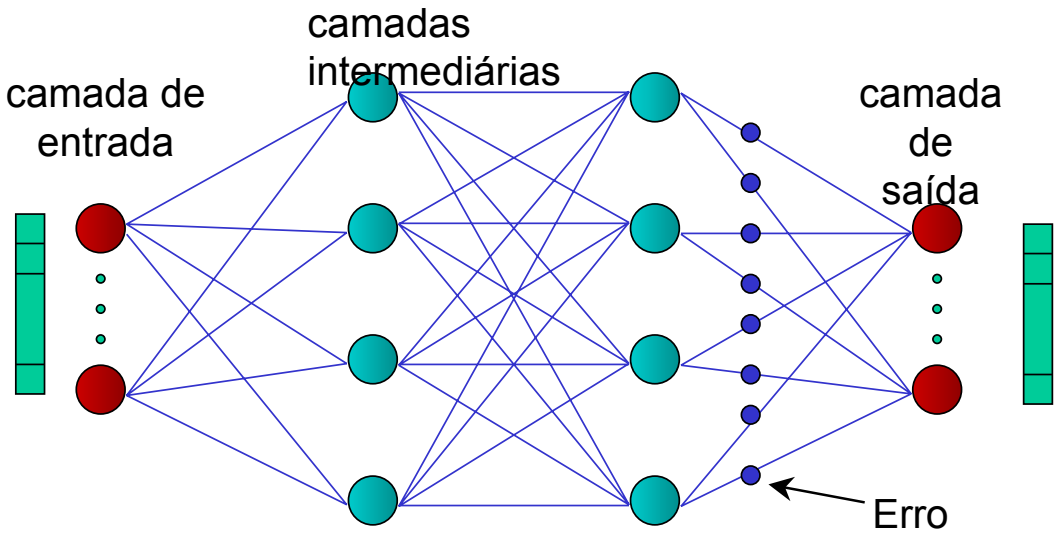
Um Exemplo Animado



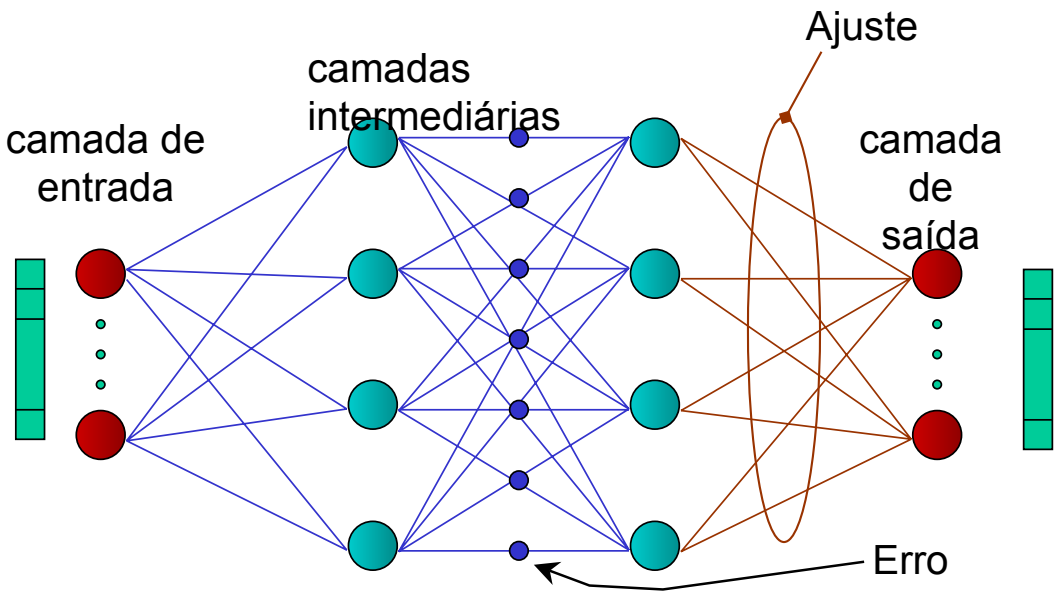
Um Exemplo Animado



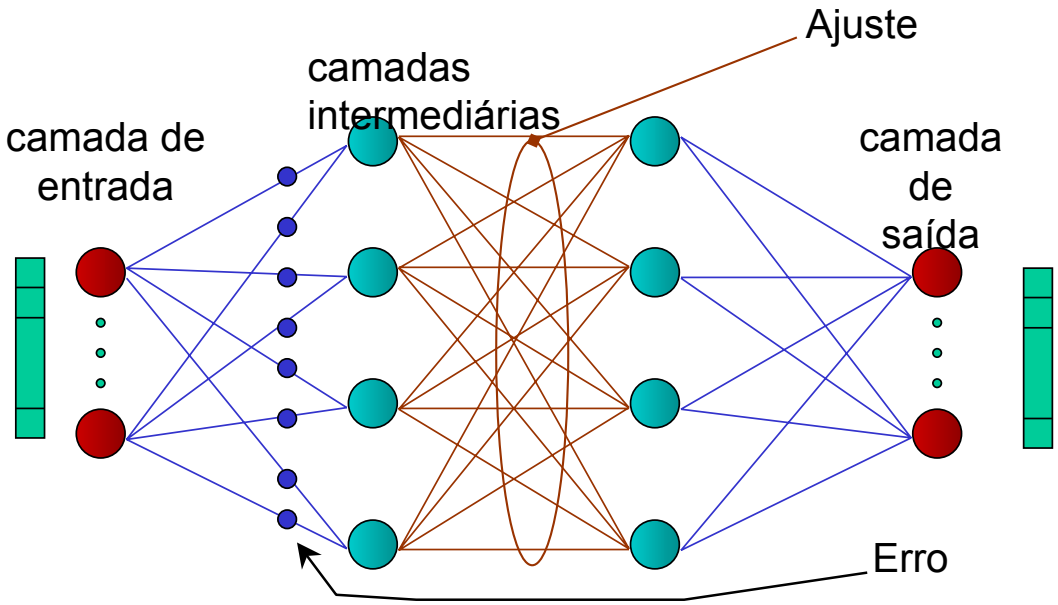
Um Exemplo Animado



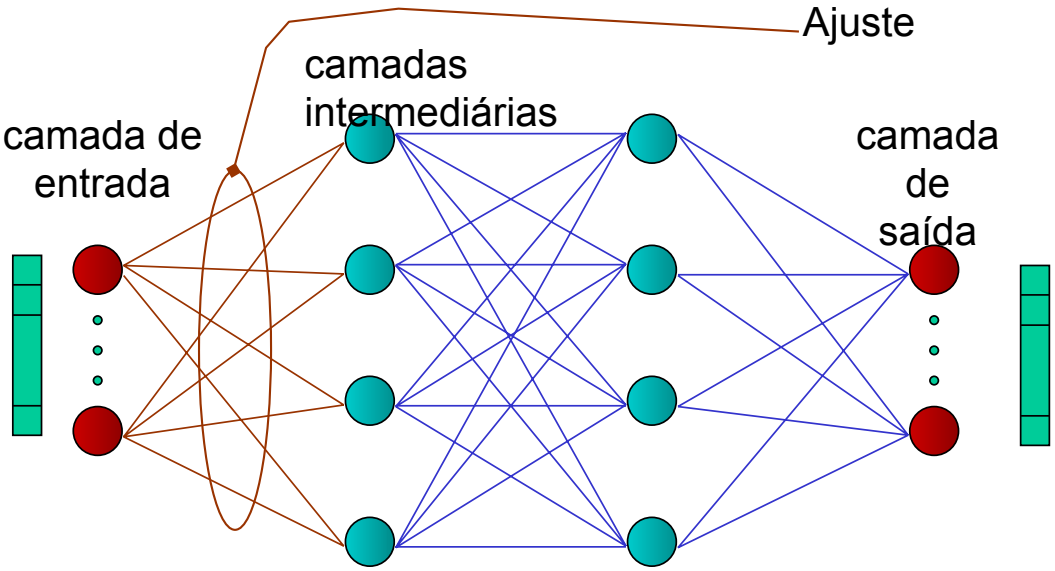
Um Exemplo Animado



Um Exemplo Animado

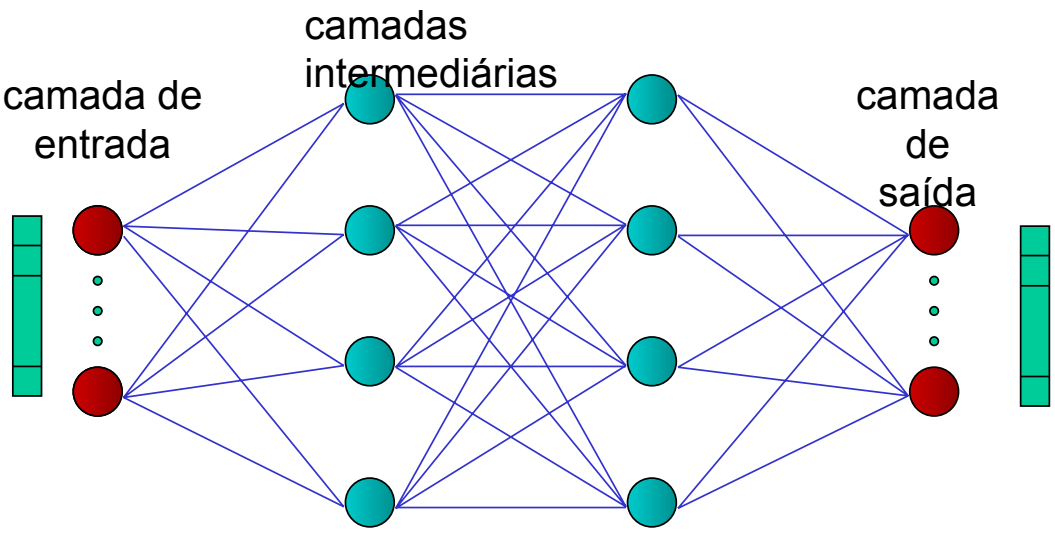


Um Exemplo Animado

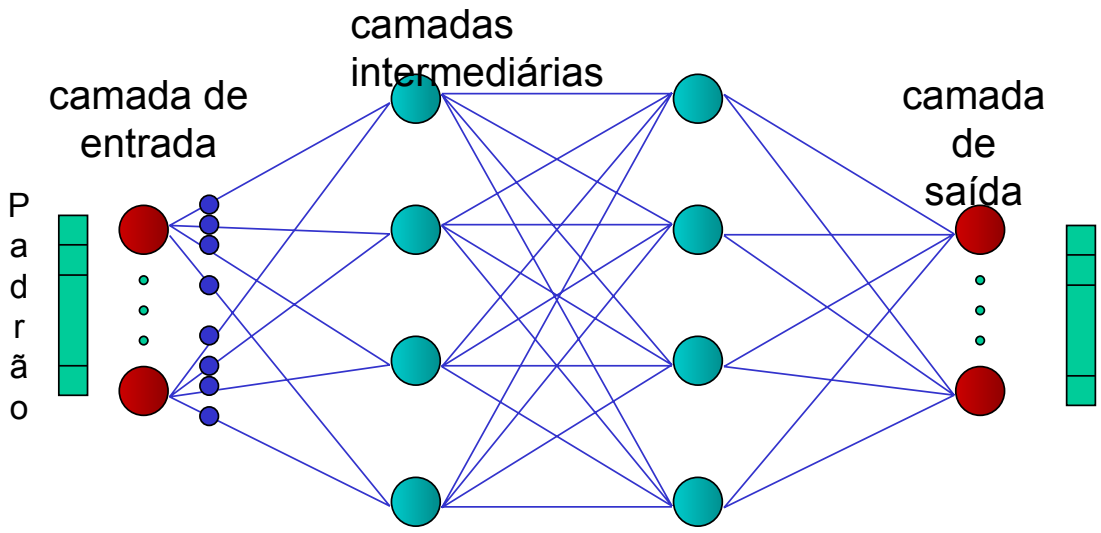


Um Exemplo Animado

Rede recalculada

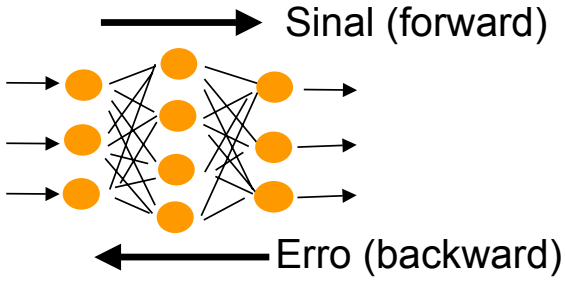


Um Exemplo Animado



Algoritmo de Propagação Para Trás (Sumário)

- Rede é treinada com pares entrada-saída
- Cada entrada de treinamento está associada a uma saída desejada
- Treinamento em duas fases, cada uma percorrendo a rede em um sentido
  - Fase forward
  - Fase backward



Pontos Importantes do Algoritmo

Na propagação para trás, cada nódulo vê suas entradas como conselheiros, quanto mais uma entrada contribui com um mau conselho mais degradado é seu peso. O cálculo dos novos pesos incluem uma função de filtragem e uma taxa de aprendizado que dita a intensidade das modificações. Depois que os pesos ajustados são propagados para trás, dos nódulos de saída para os nódulos de entrada, um novo exemplo é mostrado a rede. Depois que suficiente exemplos sejam mostrados, os pesos da rede passarão a variar mais suavemente. É então que o treinamento pára e a rede é dita ter "aprendido o conceito".

Taxa de Aprendizado

Crítico ao processo de treinamento é a taxa de aprendizado, ou quanto os pesos são modificados a cada ciclo. Taxas muito agressivas podem conduzir a uma instabilidade e impedir a rede de convergir para um conjunto de pesos. Uma taxa muito baixa de aprendizado irá, por outro lado, alongar o tempo de treinamento, aumentando o número de casos de treinamento necessários. Usualmente, começa-se com uma taxa agressiva de aprendizado e se reduz esta taxa à medida que a rede é treinada.

Máximo Locais

Redes neurais podem convergir para máximos locais. Um máximo local é uma solução ótima local, em que a rede produz bons resultados com pesos que não conseguem ser melhorados pelo algoritmo de treinamento. Podem existir, todavia, pesos, muitas vezes bem diferentes, que convergiriam para um modelo muito melhor. Treinar a rede várias vezes com pesos iniciais diferentes pode mitigar este problema. Neste caso pode-se selecionar o melhor entre os modelos encontrados.

Overfitting

Overfitting (vestir os dados) acontece quando o modelo funciona muito bem no conjunto de treinamento, mas muito mal nos conjuntos de teste. Este é um problema que pode ocorrer em toda modelagem preditiva, mas é especialmente problemático em redes neurais. Isto ocorre por que redes neurais grandes podem facilmente "vestir" um conjunto de dados pequenos de treinamento. A rede simplesmente cria um modelo que descreve os dados ao invés de criar um modelo induzindo casos genéricos.

Evitando Overfitting

(1) deve-se usar conjuntos de treinamento grandes quando comparados ao número de nódulos de entrada;

(2) deve-se usar um número limitado de camadas intermediárias na rede e poucos nódulos ocultos (alguns sugerem 2/3 do número de nódulos de entrada);

(3) deve-se sempre usar bons casos de teste para validar a rede obtida após o treinamento.

Parte 3.5: Mineração Visual de Dados

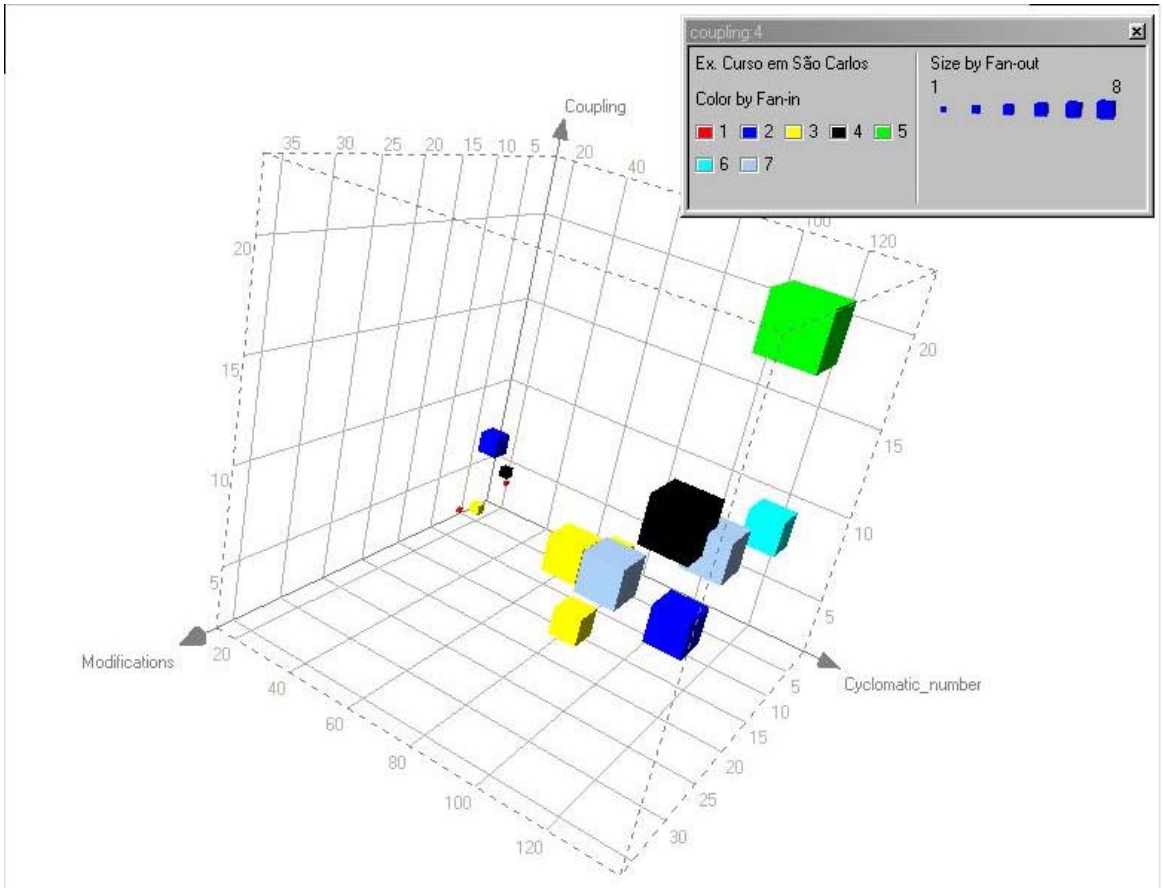
Visualização (1)

Visualização pode ser pensada como a ciência de mapear volumes de dados multidimensionais para a tela bidimensional de um computador. Visualização é uma técnica importante para mineração de dados pois seres humanos são excelentes para processar informação visual mas péssimos para processar informação tabular.

Dados em Formato Tabular

Modules	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
Fan-out	7	8	4	6	5	1	1	2	2	5	5	6	6	4	6
Fan-in	4	5	2	3	3	1	1	4	3	3	2	7	6	3	7
Coupling	14	22	7	8	4	4	3	5	5	6	12	11	10	7	13
# of Modif.	29	25	5	21	19	2	8	3	12	14	35	30	9	15	27
Cyclo. #	122	132	21	85	87	23	19	24	34	84	134	110	124	89	129

Dados em Formato Visual



Visualização (2)

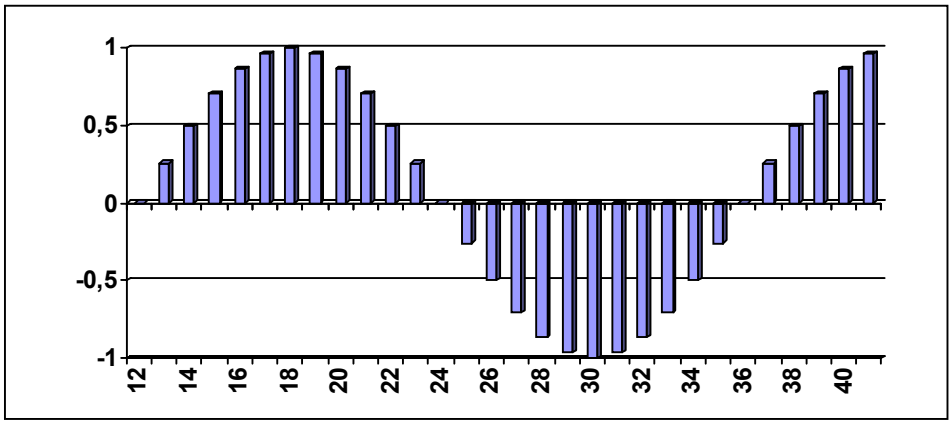
• Seres humanos são capazes de extrair as características fundamentais de uma cena visual complexa em questões de milisegundos.

• Boas técnicas de visualização jogam com essa nossa habilidade natural mostrando conjuntos complexos de dados em um formato visual que pode ser rapidamente processado pelo cérebro humano.

Um Padrão em Formato Tabular

Y	0	0.259	0.5	0.707	0.866	0.966	1	0.966	0.866	0.707
X	12	13	14	15	16	17	18	19	20	21
Y	0.5	0.259	0	-0.259	-0.5	-0.707	-0.866	-0.966	-1	-0.966
X	22	23	24	25	26	27	28	29	30	31
Y	-0.866	-0.707	-0.5	-0.259	0	.0259	0.5	0.707	0.866	0.966
X	32	33	34	35	36	37	38	39	40	41

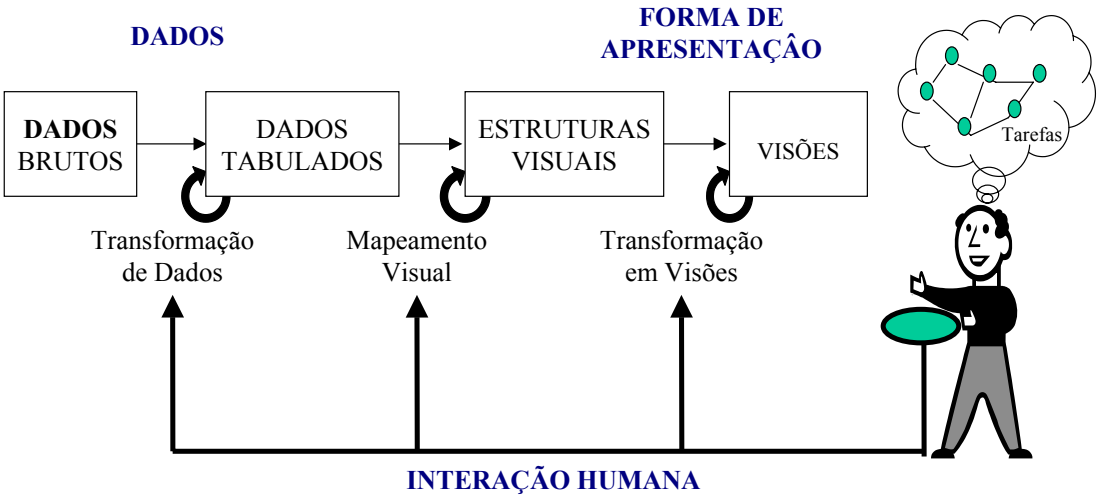
O Mesmo Padrão em Formato Visual



Exploração Iterativa e Mineração Visual de Dados

• Muitas das ferramentas modernas de visualização, combinam a capacidade de construir complexas cenas visuais com controles de seleção iterativa dos dados mostrados. Estas funcionalidades combinadas permitem que um perito possa ele próprio iterativamente explorar os dados. Este tipo de exploração interativa de dado é chamada de **mineração visual de dados**.

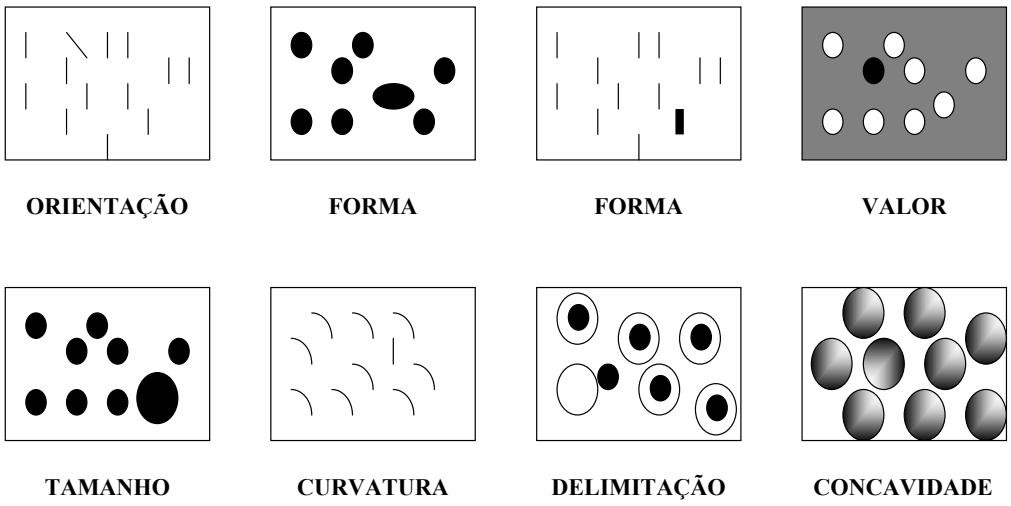
Transformando Dados em Formas Visuais



Forma de Apresentação Visual

- Para se mostrar uma informação é chave se planejar como a informação será exibida na tela visual.
- Existem vários tipos de atributos visuais que se pode utilizar para isso:
  - Forma (ex.: largura, tamanho, curvatura, orientação)
  - Cor (ex.: tonalidade, intensidade),
  - Movimento (ex.: piscar, direção do movimento)
  - Posição Espacial (ex.: côncavo/convexo, 2D, 3D)

Alguns Atributos Visuais



Características Básicas de Uma Ferramenta de Mineração Visual de Dados

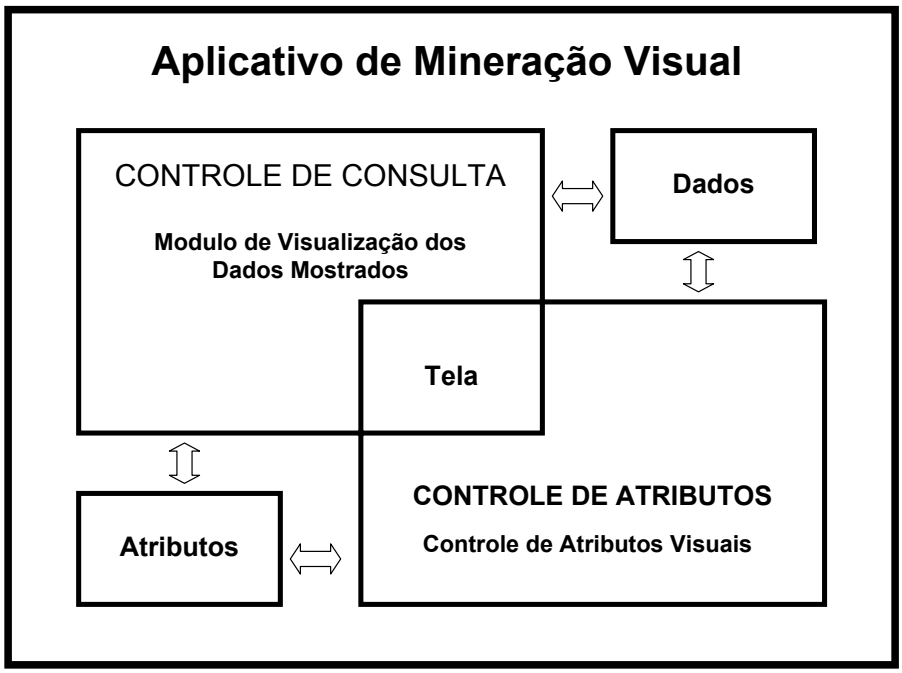
Controle de Atributos Visuais - permite o controle interativo dos formatos de apresentação e do atributos visuais dos gráficos mostrados.

Controles de Consultas - permite a consulta interativa ao conjunto de dados disponível, habilitando as pessoas a olharem os dados de uma perspectiva geral ou rapidamente mergulhar nos detalhes de um subconjunto de dados.

Obtenção de Detalhes sob Demanda - permite a obtenção de detalhes sobre um subconjunto de dados ou um registro específico

Demonstração de Ferramentas para Exploração Visual de Dados ...

Arquitetura Típica Deste Tipo de Sistema



Parte 4: Assimilação da Informação Minerada

Padrões e Modelos

- Padrão:

  ABABABABAB ...

- Modelo:

  Se “A” então “B” o seguirá. Se “B” então “A” o seguirá

Existem técnicas que produzem modelos a partir dos dados, e existem técnicas que apenas revelam padrões nos dados.

Modelos

- Precisam ser cuidadosamente validados.
- Estão “prontos” para serem usados no dia a dia da organização.
- Precisam ser incorporados operacionalmente na organização.
- Úteis quando se precisa tomar decisões (fazer classificações, estimações, ou predições) rápidas.

Padrões

- Não são informação útil por se só, têm que ser inspecionados por peritos que vão tentar extrair conhecimento dos padrões minerados.
- Permitem a peritos trazer seu conhecimento de domínio ao processo de mineração, ex. A, AB, ABC, ABCD, ABCDE, ...
- Permite aos peritos aprender sobre os dados.
- Úteis em domínios pouco conhecidos ou que mudam muito com o passar do tempo.

Interpretação de Padrões

- Deve-se trabalhar metodicamente de "cima para baixo"
- Deve-se focar em padrões envolvendo variáveis globais ou cruciais antes de olhar para as variáveis secundárias ou mais específicas.
- Deve-se seguir uma linha de raciocínio por vez quando estiver "mergulhando" nos dados.
- Deve-se registrar explicitamente padrões interessantes e descobertas assim que elas ocorrerem.
- Sempre que possível, deve-se testar formalmente os padrões descobertos para garantir sua validade estatística.

Validação de Modelos

Validação de modelos deve ser usada junto com técnicas que produzem modelos a partir dos dados minerados. Pode-se usar:

Validação cruzada: um conjunto de dados de teste é separado e usado para se assegurar que o modelo produzido terá bons resultados no mundo real.

v-dobras: o próprio conjunto de dados de treinamento é usado como conjunto de testes.

Validação Cruzada

• Divide-se o conjunto de dados em dois pedaços. O primeiro, chamado de conjunto de treinamento, é usado para se produzir o modelo. O segundo, chamado de conjunto de teste, é usado para se testar o modelo.

• O modelo produzido a partir o conjunto de treinamento é usado para estimar o valor de cada registro no conjunto de teste.

• A precisão média dos resultados é considerada então ser a futura precisão do modelo no mundo real.

Validação com *v*-dobras (v-fold)

1 - Divida o conjunto de dados em *v* subconjuntos de tamanho similar, $v \geq 10$

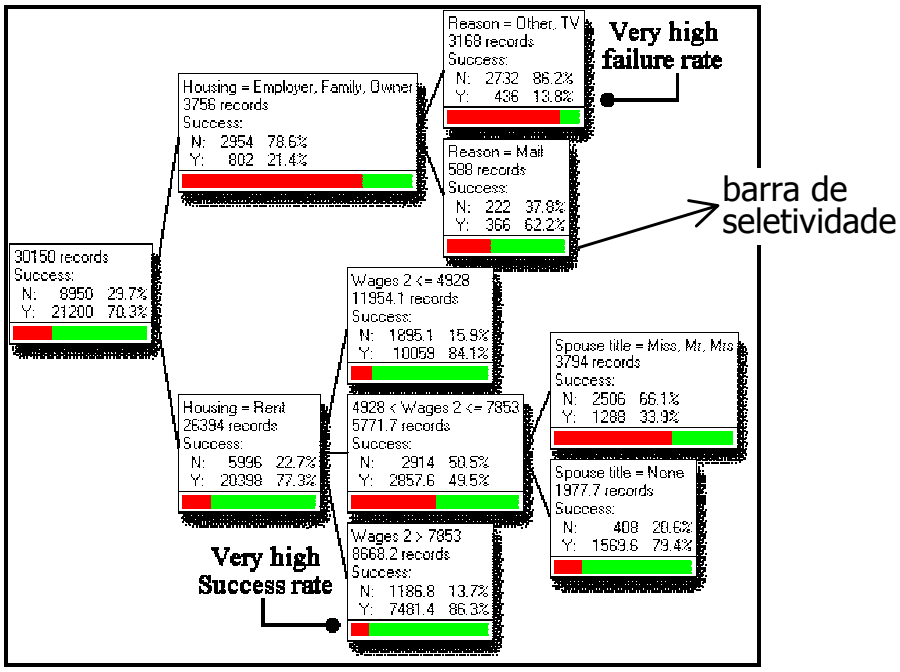
2 - Escolha um dos subconjuntos como o conjunto de teste e construa um modelo usando os outros conjuntos.

3 - Use o modelo para estimar os valores do conjunto de teste, e estime sua precisão. Enquanto existir subconjuntos não usados como subconjunto de teste, volte ao Passo 2.

4 - Depois que todos os conjuntos forem usados como conjuntos de teste, calcule a precisão média dos modelos produzidos.

5 - Este valor será considerado ser uma estimativa da precisão média do modelo a ser construído se usando todos os conjuntos de teste.

Visualizando Modelos



Parte 5: Observações Finais

Sites Interessantes

- http://www.kdnuggets.com
- http://www.andypryke.com/university/TheDataMine.html
- http://www.dw-institute.org/
- http://www.acm.org/sigkdd/
- http://www.computer.org
- http://www.almaden.ibm.com/cs/quest
- http://www.sas.com/technologies/data_mining/index.html
- http://www.oracle.com/ip/analyze/warehouse/datamining/
- http://www.sgi.com/solutions/sciences/chembio/resources/mineset/general/index.html

Ferramentas

- Várias ferramentas comerciais
  - Relativamente caras.
  - Maioria não apresenta suporte para todas as fases da mineração de dados, e se foca algumas abordagens de mineração.
- Centros de pesquisas e empresas desenvolvem ferramentas de domínio público.
- http://www.kdnuggets.com/software/ tem uma lista bastante completa de ferramentas disponíveis.

Conclusões (1)

• A qualidade dos dados é fundamental. Não existe descoberta de conhecimento em dados ruins. Apesar de não serem considerados parte da mineração, esforços de coleta de dados e medição são muitíssimo importantes para ela.

• Deve-se entender a semântica dos dados antes de se começar qualquer esforço de mineração. Não se pode minerar o que não se entende.

• Seleção e pré-processamento de dados são tarefas trabalhosas, mas cruciais à mineração bem sucedida de dados.

Conclusões (2)

• Todas as técnicas têm pontos fortes e pontos fracos. Deve-se escolher a técnica mais adequada para o problema e não o contrário.

• Deve-se estar muito atento a significância dos padrões e modelos minerados. Eles devem ser suportados por um volume significativo de registros de dados.

• Modelos podem sobre vestir (overfitting) os dados. Isto é perigoso. Eles podem explicar um conjunto de dados sem serem uma generalização deles (ex., o caso de reconhecimento de tanques de guerra).

Conclusões (3)

• Modelos interpretáveis têm a vantagem de poderem ser revisados por peritos (e serem combináveis ao seu conhecimento de domínio).

• Técnicas de detecção de padrões e de mineração visual de dados são cruciais em domínios voláteis ou que ainda não são bem entendidos. Isto é, são ótimos para se aprender sobre o comportamento dos dados.

• A Estatística está viva e muito bem! Não deve-se reinventar a roda. Existem muitas técnicas estatísticas que podem (ou devem) ser combinadas a esforços de mineração de dados.

Conclusões (4)

• Grandes fabricantes de sistema de gestão de banco de dados estão incorporando facilidades de mineração dentro de seus sistemas.

• Mesmo assim ainda há muito a ser feito. Estas facilidades não são de forma alguma exaustivas.

• Boa parte da boa mineração não pode ser automatizada, por exemplo: a escolha das técnicas, a consideração dos fatores específicos ao domínio da aplicação, a gestão do processo como um todo, etc.

Conclusões (5)

- O processo de mineração tem muitas partes: seleção, pré-processamento, mineração, e assimilação. O seu sucesso não depende só destas partes, mas também de como ele é executado como um todo. De como estas partes são colocadas juntas, e de como se interage entre elas.

- Atualmente há um acumulo crescente de dados que não são adequadamente explorados nos repositórios de dados. Há uma enorme demanda por técnicas que transformem dados em informações úteis. Técnicas que possam ser usadas por leigos. Existe muita coisa a ser feita, e tremendas oportunidades, no campo da mineração de dados.

Apêndice 1:
Aprendizado de Máquina

Características de Sistemas de Aprendizado de Máquina

Tipo de Aprendizado	Paradigmas de Aprendizado	Linguagens de Descrição	Integração de Novos exemplos
- Supervisionado - Não-supervisionado	- Simbólico - Estatístico - Instance-Based - Conexionista - Genético	- Exemplos ou Instâncias - Hipóteses ou Conceitos Aprendidos -Teoria de Domínio ou Conhecimento de Fundo	- Incremental - Não Incremental

Alguns Algoritmos de Aprendizado de Máquina

Indutor	Tipo de Aprendizado	Paradigma de Aprendizado	Linguagem de Descrição	Integração de Novos Exemplos
Aha-IB	Supervisionado	Instance-Based	Exemplos	Incremental
C4.5	Supervisionado	Proposicional	Árvore de decisão	Não-Incremental
CART	Supervisionado	Estatístico	Árvore de decisão	Não-Incremental
CN2	Supervisionado	Proposicional	Regras de Produção	Não-Incremental
Decision Tables	Supervisionado	Instance-Based	Exemplos	Incremental
ID3	Supervisionado	Proposicional	Árvore de decisão	Não-Incremental
Native Bayes	Supervisionado	Estatístico		
Nearest-neighbor	Não-Supervisionado	Estatístico		Incremental
OC1	Supervisionado	Proposicional	Árvore de decisão	Não-Incremental
Perceptron	Supervisionado	Conexionista		Não-Incremental
T2	Supervisionado	Proposicional	Árvore de decisão	Não-Incremental
Foil	Supervisionado	Relacional	Clausulas de Horn	Não-Incremental
Golem	Supervisionado	Relacional	Clausulas de Horn	Não-Incremental

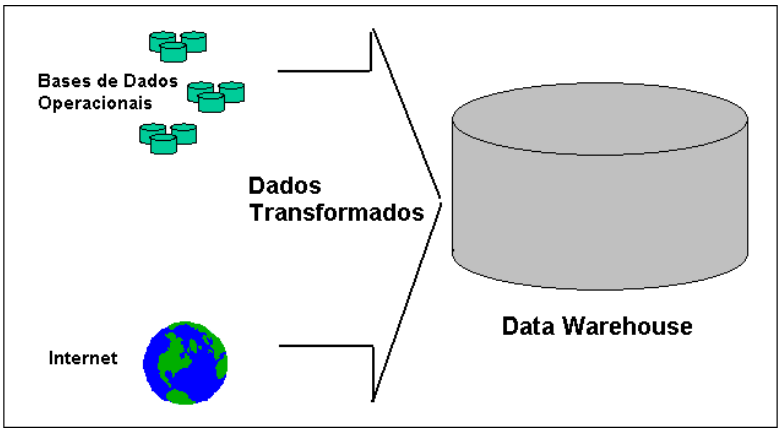
Apêndice 2:
Data Warehousing

Definição

"Data Warehousing é um processo, não um produto, para montar e gerenciar dados de várias fontes com o propósito de ganhar uma visão detalhada e singular de parte ou do todo de um negócio" Gardner

Definição de Data Warehouse

Data Warehouse (DW) é um sistema de bancos de dados integrados, derivado de diversos outros bancos (ou fontes) de dados, voltado ao suporte à tomada de decisão.



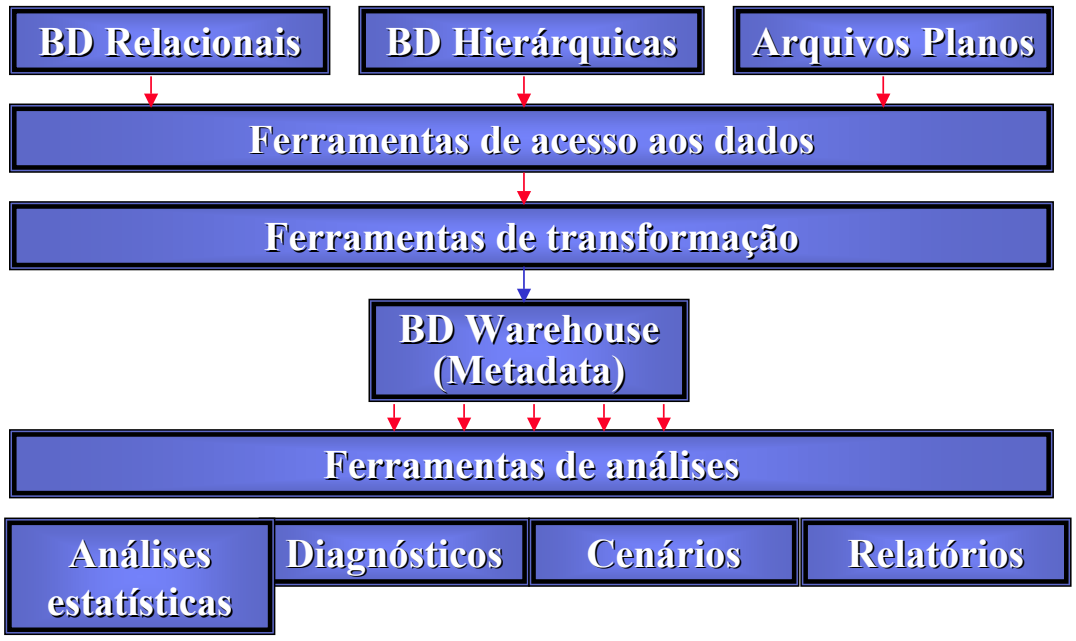
Data Warehouse e Ferramentas de Apoio à Tomada de Decisão

- Data Warehouse
  - Dados coletados de diferentes fontes
  - Base de Dados Operacionais x Data Warehouse
- Ferramentas de Apoio à Tomada de Decisão
  - OLAP
  - Estatística
  - KDD/Data Mining

DW são indicadas para:

- Empresas que precisam processar e armazenar uma grande quantidade de dados.
- Empresas que precisam organizar de forma integrada a sua estrutura de informação.
- Empresas que precisam de informação no momento para a tomada de decisões

Arquitetura de uma DW



Soluções Apoiadas por uma DW

- Sistemas de grande porte de Suporte à Tomada de Decisões
- Base de Dados Multidimensionais complexas
- Sistemas de Banco de Dados com alta taxa de crescimento
- Sistemas Multidepartamentais integrados

Apêndice 3:
On-Line Analytical Processing (OLAP)

Definição de OLAP

"On-line Analytical Processing (OLAP) é uma tecnologia que focaliza a análise, síntese e consolidação de grandes volumes de dados por meio de métodos multidimensionais" Codd

Ferramenta geralmente utilizada para a análise de Data Warehouse

DW + OLAP ⇨facilitam a análise/ obtenção da informação

Definição de OLAP

- **SQL normal**
  - O usuário/analista tem que ter suporte e compreender SGBD. É uma lista simples onde os dados podem ser analisados.
- **Ferramenta OLAP**
  - É voltada para gerentes. Resultado é um relatório mais sofisticados mas sem a necessidade de conhecimentos prévios de SGBD.
  - Apresenta dados relacionais de forma a facilitar a compreensão dos dado.