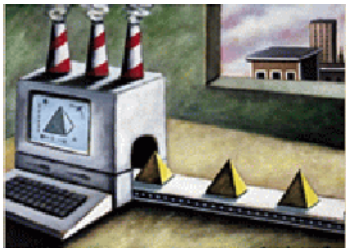
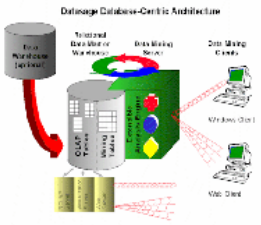
AID (Clementine – 2.ª Parte)

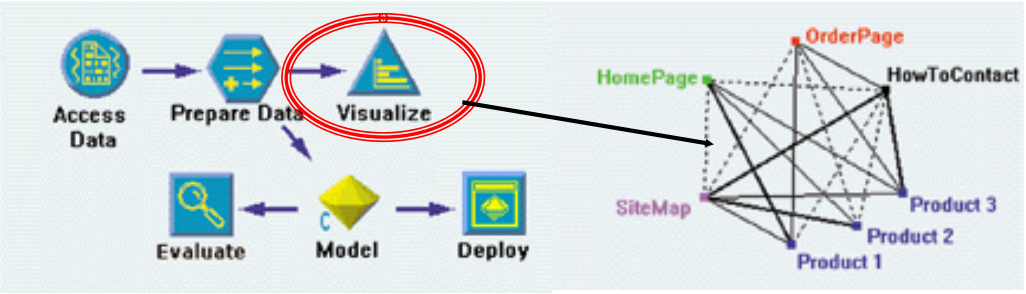




AID (Clementine – 2.ª Parte)

Procurar Relacionamentos nos Dados

- *Introduzir o nó Web e Matrix*
- *Introduzir o nó Plot e Statistics p avaliar a correlação*
- *Usar um histograma para investigar o relacionamento entre campos simbólicos e numéricos*



Introdução

Na maioria dos projectos de DM estaremos, numa fase inicial, interessados em investigar se há relacionamentos entre os campos de dados.

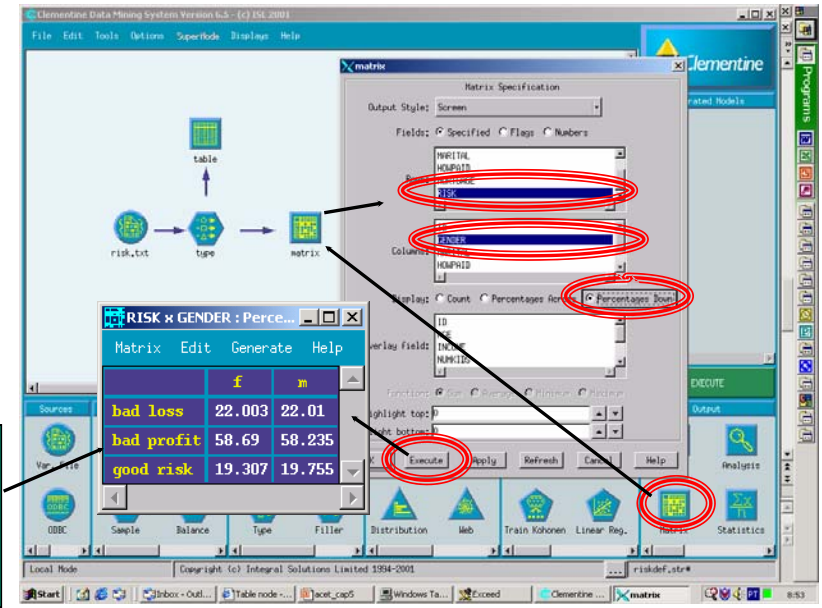
Questões típicas:

- **O risco de crédito está relacionado directamente com o rendimento ou haverá alguma outra medida de rendimento disponível mais apropriada?**
- **O facto de ser homem ou mulher fará diferença em relação ao risco de crédito?**
- **Se um indivíduo tiver um grande número de cartões de crédito, aplicações e cartões de armazéns, isso indicará que poderá ser ter um mau risco de crédito maior ou menor?**

Embora as técnicas de machine learning e modelação do Clementine possam responder a estas questões, importa iniciar com uma fase exploratória e compreender os relacionamentos básicos entre um pequeno número de campos.

Nó Matrix: Relacionar dois Campos Simbólicos (1)

Este nó executa a tabulação cruzada entre dois campos simbólicos, mostrando como os valores de um dos campos se relaciona com o outro.

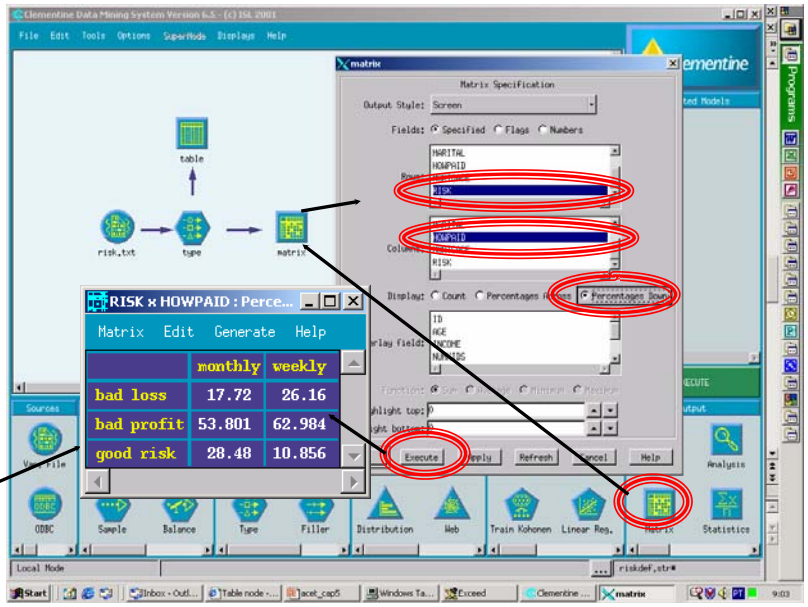


Criada a tabulação cruzada entre o sexo da pessoa e o risco de crédito. Parece não haver diferenças de monta entre os dois grupos.

Nó Matrix: Relacionar dois Campos Simbólicos (2)

Outro exemplo:

- Examinar a relação entre o risco de crédito e a periodicidade de recebimento do salário (semanal ou mensalmente)



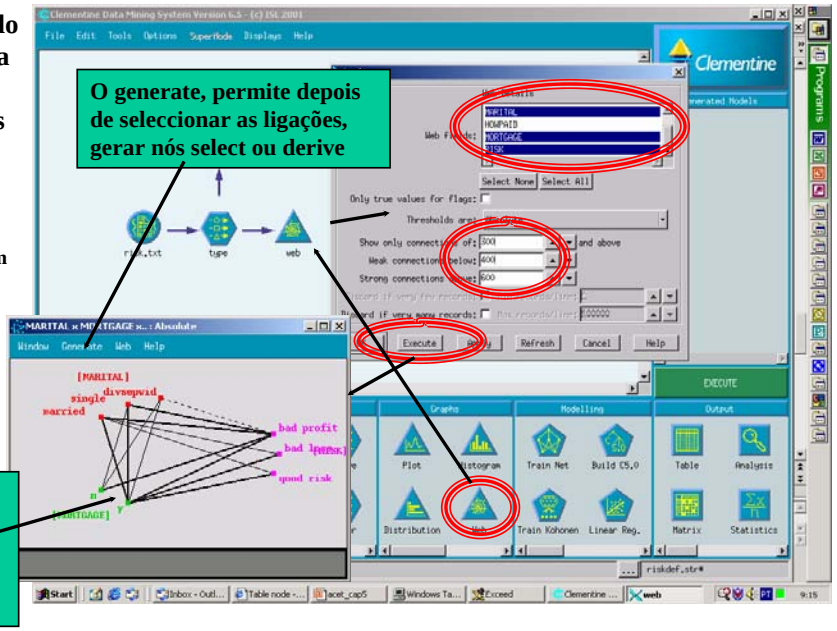
Criada a tabulação cruzada entre a periodicidade de recebimento e salário e o risco de crédito. Parece diferenças: as pessoas pagas mensalmente apresentam um melhor risco de crédito.

Nó Web: Força dos Relacionamentos

Este nó pode ser utilizado para mostrar a força das ligações entre dois ou mais campos simbólicos.

Um gráfico "teia" consiste:

- **em pontos que representam os valores dos campos simbólicos seleccionados**
- **Esses pontos são ligados linhas tracejadas, normais ou grossas, dependendo da força das ligações entre os valores**

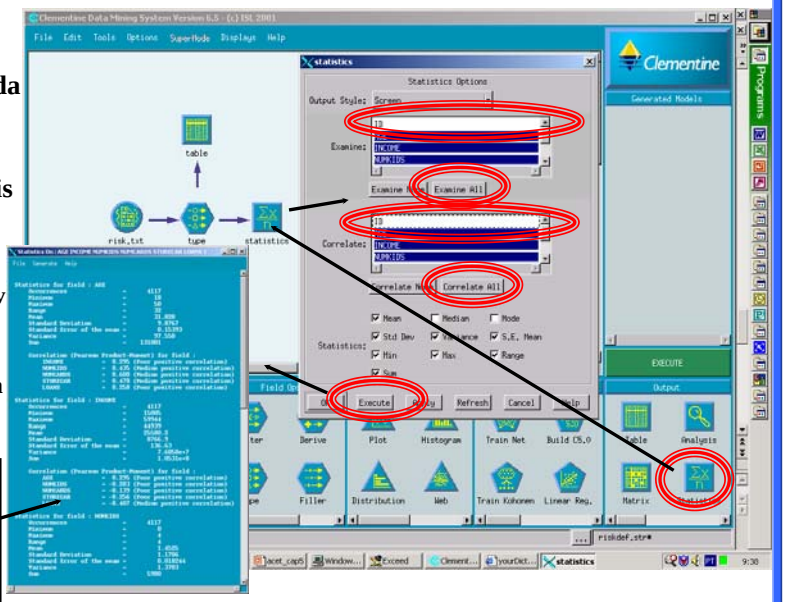


Parece haver uma ligação forte entre hipoteca e todos os riscos de crédito e entre indivíduos casados e risco mau, mas lucrativo.

Correlação entre Campos Numéricos

Na investigação de relacionamentos entre campos numéricos, as correlação linear é utilizada normalmente.

- **Mede a extensão da associação linear entre dois campos numéricos**
- **Varia entre**
  - **1 (relacionamento positiv perfeito)**
  - **e –1 (relacionamento perfeito negativo – se um campo cresce o outro desce à mesma taxa)**

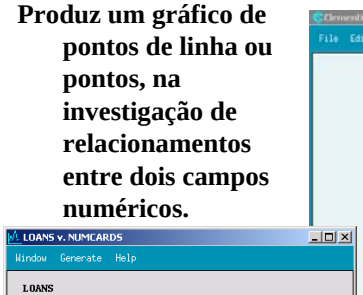


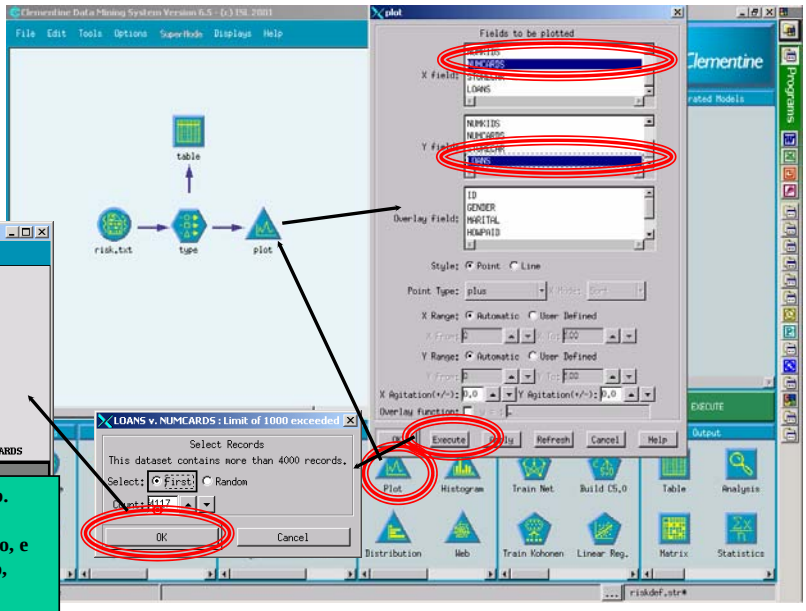
Uma limitação: pode não haver relacionamento linear entre dois campos, mas ele ser de outra tipo

(ex. correlação entre idade e rendimento)

Nó Plot (1)

Produz um gráfico de pontos de linha ou pontos, na investigação de relacionamentos entre dois campos numéricos.





Tal como o gráfico Web, este é tb. Interactivo, permitindo indicar através de um rectângulo a região, e depois, através do menu contexto, gerara um nó select ou derive ou apagar a região.

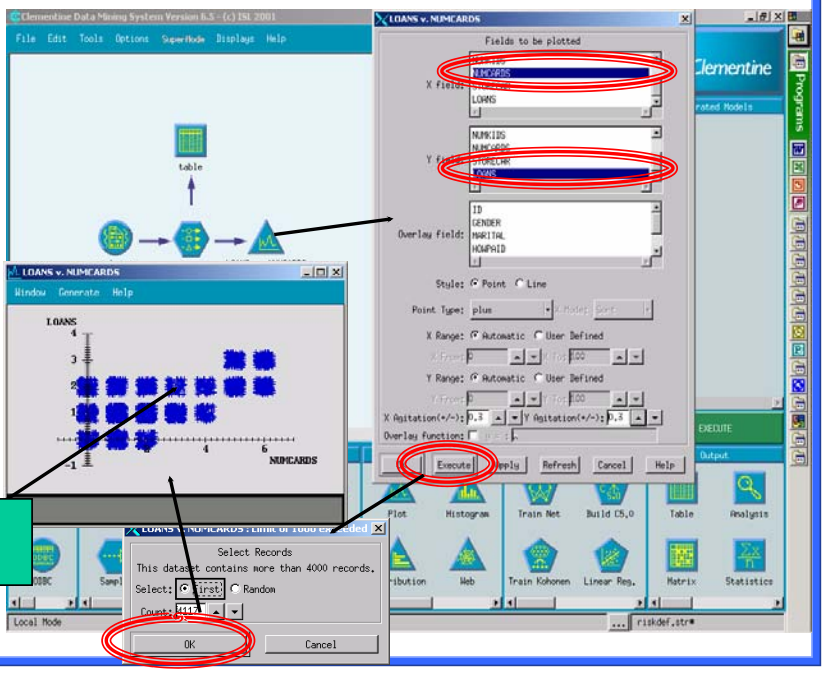
Nó Plot (2)

O gráfico anterior, produziu apenas um pequeno número de pontos, apesar dos 4000 registos

- **Os dois campos seleccionados têm um número de valores limitado**

É útil utilizar a agitação

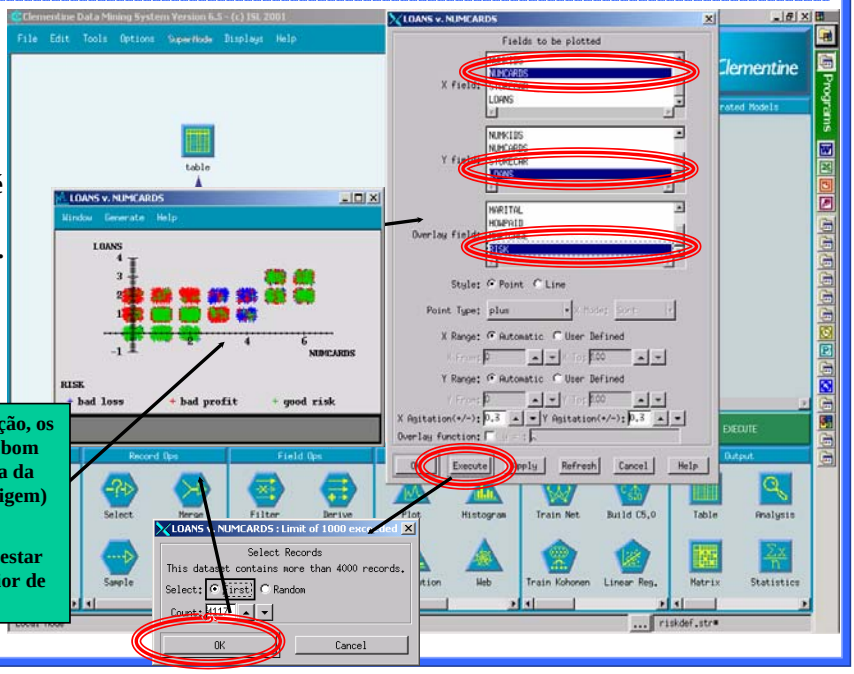
- **avaliar se alguns clusters serão maiores do que outros**



O gráfico sugere um ligeiro maior número de registos nos extremos do que no meio de ambos os intervalos.

Nó Plot (3)

Pode agora ser útil avaliar se o relacionamento entre os empréstimos e o número de cartões é diferente para os três grupos de risco.



Embora haja alguma sobreposição, os indivíduos com risco de crédito bom parecem estar agrupados à volta da parte inferior (mais perto da origem) de ambos os campos;

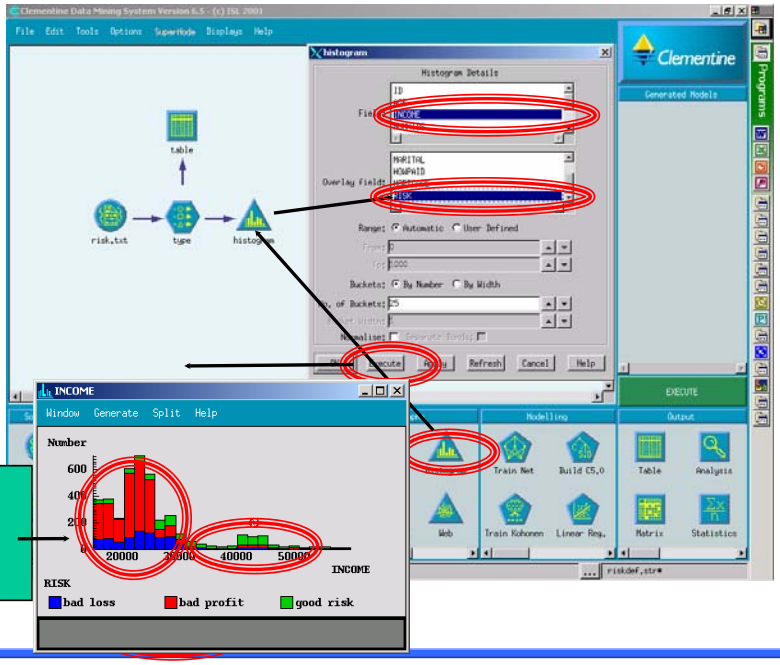
os indivíduos a evitar, parecem estar concentrados nas franjas superior de ambos os intervalos.

Nó Histograma

Visualização dos relacionamentos entre Campos Numéricos e Simbólicos.

Permite:

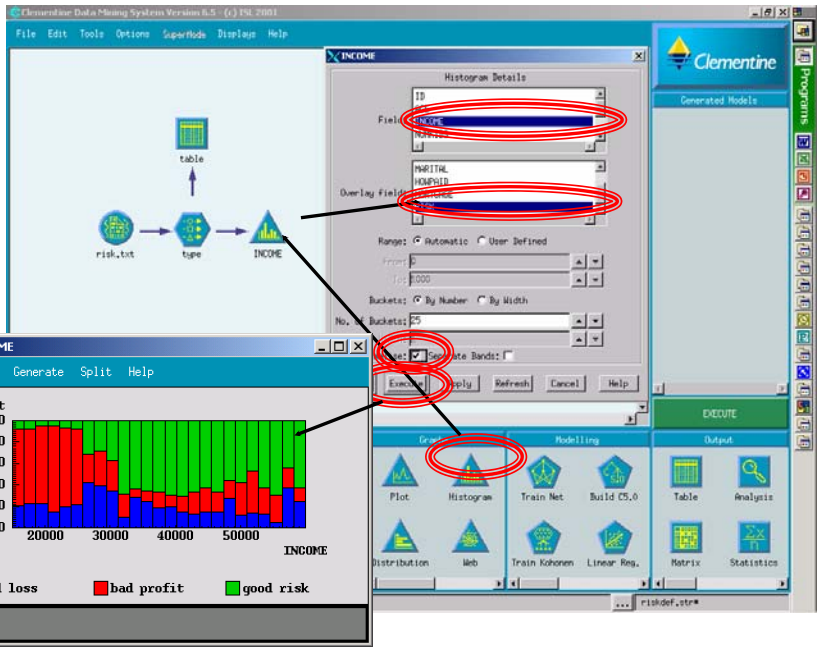
- **examinar a distribuição de um dado campo**
- **em opção, incluir um campo simbólico no Overlay e assim, produzir um gráfico sobreposto com cores que representam os valores diferentes do campo simbólico.**



Parece que o risco bom está mais associado a rendimentos altos, enquanto que os restantes tipos de risco, estão associados a baixos rendimentos

Nó Histograma (normalizado)

A normalização tem o efeito de mostrar as proporções de cada valor do campo simbólico dentro de cada intervalo do campo numérico.

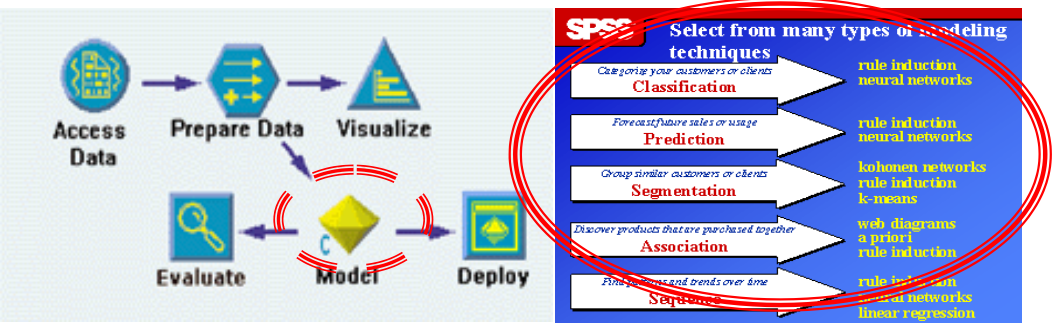


O gráfico normalizado permite uma ideia clara das proporções relativas das três categorias de risco de crédito. Contudo a informação acerca da concentração relativa dos registos pelos vários rendimentos, perde-se.

AID (Clementine – 2.ª Parte)

Técnicas de Modelação em Clementine

- *Dar uma breve introdução acerca das técnicas de modelação disponíveis no Clementine*
- *Qual técnica utilizar? Quando e como?*



Introdução

Clementine inclui várias técnicas de machine learning e modelação. Embora sejam possíveis diversas taxonomias, poderemos classificar essas técnicas em três grandes abordagens:

- **Preditiva**
  - **tb. chamada de aprendizagem supervisionada**
  - **as entradas são utilizadas para predizer o valor de uma saída (disponíveis: redes neuronais, indução de regras, regressão e análise logística)**
- **Clustering**
  - **tb. chamada de aprendizagem não supervisionada, não existe o conceito de campo de saída;**
  - **o propósito é a segmentação dos gados em grupos de indivíduos com padrões semelhantes de campos de entrada (disponíveis: Kohonen, k-means e two-steps)**
- **Associações**
  - **Técnicas gerais de modelação preditiva; os campos podem actuar com entradas e saídas simultaneamente;**
  - **As regras de associação tentam associar uma dada conclusão a um conjunto de condições (disponíveis: APRIORI e GRI)**

Redes Neuronais

Tentam resolver problemas mimetizando a forma como o cérebro opera.

Derivam o seu nome do paralelismo da sua arquitectura e a arquitectura do cérebro humano:

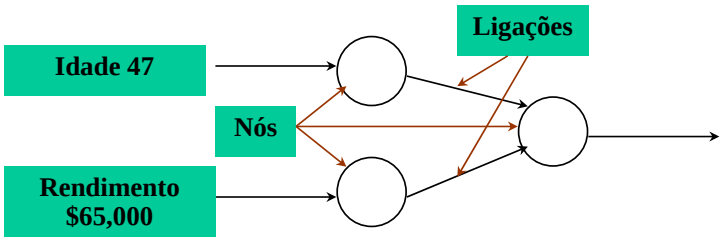
- **O cérebro humano consiste num grande número de neurónios (cerca de 10^{11}), ligados através de um largo número de sinapses (1 neurónio pode dispor de milhares delas).**
- **A rede de neurónios e as suas sinapses constitui um excelente modelo que pode ser implementado artificialmente:**
  - **através de hardware especial**
  - **utilizando programas de software**

Redes Neuronais

- **Uma rede neuronal típica consiste em vários neurónios arranjados em camadas de forma a criar uma rede;**
- **Cada neurónios pode ser pensado como um elemento processador ao qual é dada uma pequena parcela de uma tarefa;**
- **As ligações entre os neurónios proporcionam à rede a capacidade de apreender os padrões e relacionamentos nos dados.;**
- **Uma rede neuronal é uma técnica de caixa preta, o que quer dizer que é difícil perceber o raciocínio que está por detrás de uma dada predição.**

Aspecto da Rede Neuronal

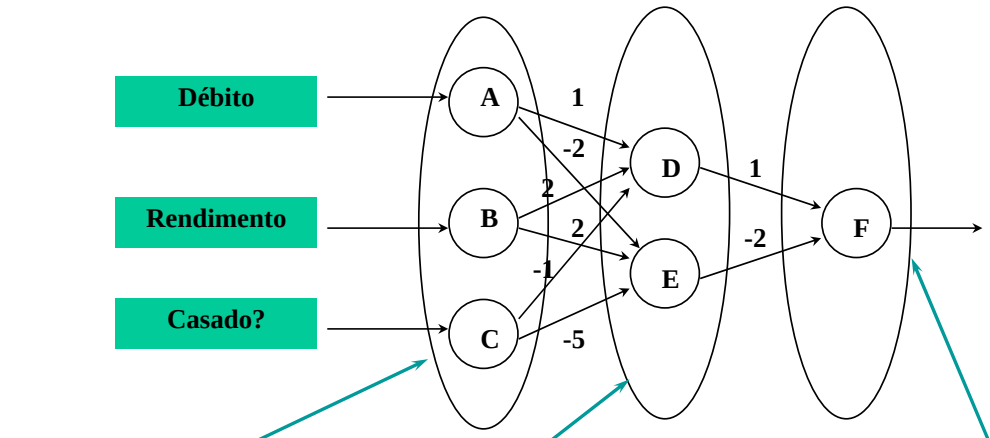
- Constituída por uma série de nós (que correspondem de perto aos neurónios do cérebro humano):
  - nós de entrada - recebem os sinais de entrada,
  - nós de saída - fornecem os sinais de saída;
- Um potencialmente número ilimitado de camadas intermédias que contém nós intermédios;
- Ligações - que correspondem às ligações entre neurónios (dendritos e sinapses) no cérebro humano.



Redes Neuronais

- Quando se utiliza uma rede neuronal em modelação preditiva, a camada de entrada contém todos os campos de entrada (utilizados para predizer o resultado)
- A camada de saída contém o campo de saída: o alvo de predição
- Os campos de entrada e saída podem ser numéricos ou simbólicos (estes últimos são transformados em numéricos - através de codificação de conjunto binária - pois que a rede só pode tratar campos numéricos)
- A camada escondida (intermédia) contém um número de nós que combinam as saídas da camada anterior - pode haver uma ou mais camadas intermédia - embora sejam mantidas num mínimo)
- Todos os nós de uma camada estão normalmente ligados a todos os nós da próxima camada (rede completamente conectada)

Outra Rede Neuronal



Nós de Entrada

(correspondem às variáveis independentes)

Nós Escondidos (hidden)

Podem ser várias camadas (normalmente aumenta com o crescimento do número de nós de entrada e complexidade do problema)

Nó de Saída

Pode ser mais do que um (aqui diferem da maioria das técnicas de Data Mining)

Topologia ou Arquitectura de Rede Neuronal

- **O número de nós de entrada, saída e escondidos é muitas vezes designado como topologia da rede ou ainda, arquitectura da rede.**
- **O arranjo dos nós em camadas é comum, mas não essencial.**
- **Pesos:**
  - **cada seta dos dois modelos apresentados mostra números (excepto nas setas de entrada e saída)**
  - **podem ser positivos ou negativos**
  - **normalmente são restringidos a valores pequenos**
  - **são tipicamente números reais com decimais**

Redes Neuronais (Treino)

Treino da rede neuronal - processo de aprendizagem dos relacionamentos entre os dados e os resultados

- **depois do treino, a rede pode ser utilizada para tomar decisões ou predições sobre dados novos, baseada na sua experiência.**

Como aprende uma criança:

- **são apresentados padrões de letras e ela faz uma tentativa:**
  - **se acertar é recompensado - da próxima vez que encontra a mesma combinação de letras, provavelmente lembrar-se-à da resposta correcta**
  - **se não acertar - é-lhe dita a resposta correcta e tenta ajustar a sua resposta baseada neste feedback.**

Redes Neuronais

A aprendizagem da rede neuronal processa-se da mesma forma que a criança mostrada anteriormente:

- **numa MLP (Multi-Layer Perceptron) cada neurónio da camada escondida recebe uma entrada baseada numa combinação pesada das saídas dos neurónios da camada anterior;**
- **os neurónios da camada final, produzem uma saída;**
- **este valor é comparado com a saída correcta;**
  - **a diferença (erro) é transmitido para trás, daí o nome de retropropagação,**
  - **a actualização do peso de cada ligação é efectuada de acordo com a dimensão do erro e a sua responsabilidade.**

Algoritmo Backpropagation

Quando parar o treino?

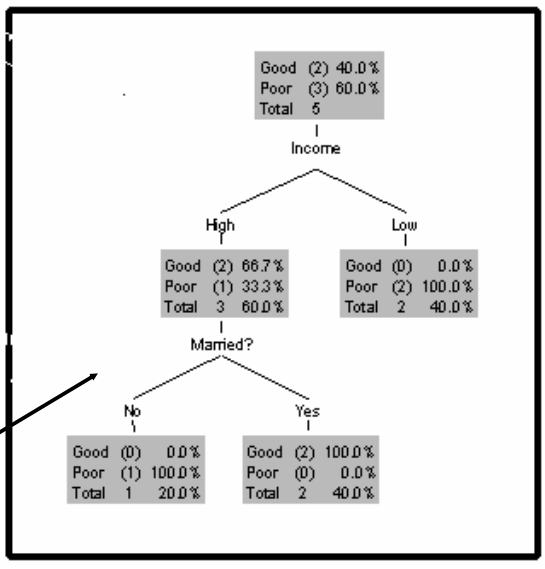
Podem ser utilizadas diversas regras, sendo as mais comuns:

- **parar após um número especificado de épocas;**
- **parar quando o erro atingir um determinado limiar;**
- **parar quando o erro não conhecer nenhum melhoramento depois de um certo número de épocas;**
- **parar quando a medida de erro num conjunto de controlo se desvia mais do que um determinado valor relativamente ao medido no set de treino; esta regra é particularmente orientada à prevenção do sobretreinamento;**
- **possibilitar feedback visual, por forma a que o utilizador possa intervir na paragem do processo.**

Árvores de Decisão

Trata-se de um modelo que é simultaneamente predictivo e descritivo.

- **o seu nome deriva do facto do modelo resultante ser apresentado na forma de uma estrutura em árvore;**
- **cada ramo da árvore traduz uma questão de classificação;**
- **as folhas da árvore são partições dos dados com a sua classificação.**



Aspectos descritivos:
•o débito parece não ter qualquer papel na determinação do risco
•as pessoas com baixo rendimento são sempre de risco pobre
•o rendimento é o factor mais significativo na determinação do risco

Indução de Regras ou Árvores de Decisão

Não sofre do problema que enferma a rede neuronal ser técnica tipo "caixa preta"

O Clementine disponibiliza dois algoritmos:

- **Build C5.0 e C&R Tree (classification and regression tree).**

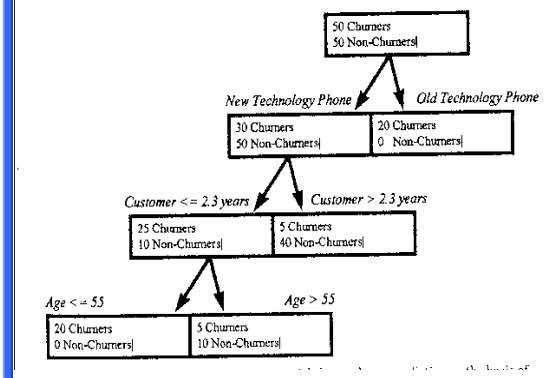
Ambos criam uma árvore de regras que tentam descrever segmentos diferentes dentro dos dados em relação a um resultado ou campo de saída.

A estrutura da árvore mostra perfeitamente o raciocínio e permite assim ser utilizado para compreender o processo de tomada de decisão que leva a um dado resultado.

Uma outra vantagem é que o processo elimina quaisquer campos que não sejam importantes na tomada de decisões.

- **proporciona informação útil**
- **permite a redução do número de campos a utilizar numa rede neuronal, p. ex.**

Árvores de Decisão



Outro exemplo de árvore de decisão, relativa a exemplo de classificação de clientes que não renovam os contratos relativos a telemóveis (ditos "churn")

Observações:
•divide os dados em cada nó, sem perder qualquer dado (o nº de registos total num dado nó, é igual à soma dos registos contidos nas suas duas folhas)
•o n.º de clientes que não renovaram e renovaram o contrato é conservado, à medida que se sobre ou desce na árvore
•é fácil compreender como o modelo está a ser criado (em contraste com os modelos das redes neuronais ou estatística standard)
•é fácil utilizar este modelo ao pretender-se atingir os clientes em vias de não renovar o contrato, com ofertas de marketing
•pode criar algumas intuições acerca da base de dados de clientes:
Ex. "clientes há vários anos e que tenham telemóveis actuais, são muito leais"

Indução de Regras ou Árvores de Decisão

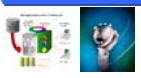
O algoritmo Build C5.0 (da responsabilidade de Ross Quinland) permite visualizar as regras em dois formatos:

- **a apresentação em árvore de decisão (embora não em formato gráfico)**
  - **de interesse se o utilizador quer visualizar como o preditor divide os dados em subconjuntos**
- **apresentação tipo regras (colecções de IF... Then) organizadas por resultado**
  - **de interesse se precisarmos de saber como grupos particulares de valores de entrada estão relacionados com um valor da saída.**

Modelos Estatísticos de Predição

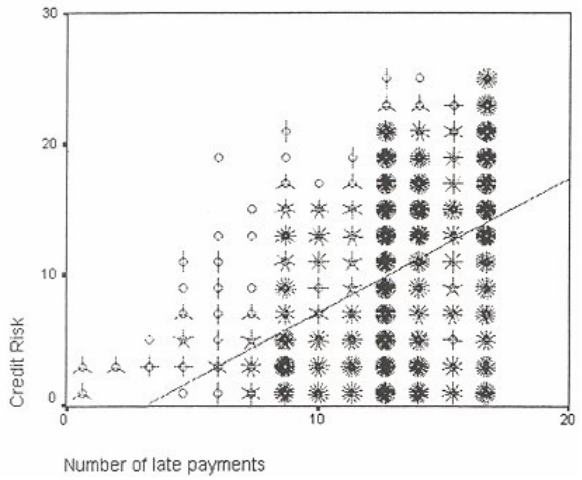
Os procedimentos de modelação estatística do Clementine usam suposições mais fortes sobre os dados do que as técnicas de machine learning

- **os modelos podem ser expressos sob a forma de equações simples**
- **permitem fácil interpretação**
- **mas não são capazes (na sua forma standard) de capturar relações complexas ou não lineares como as redes neuronais.**



Regressão Linear

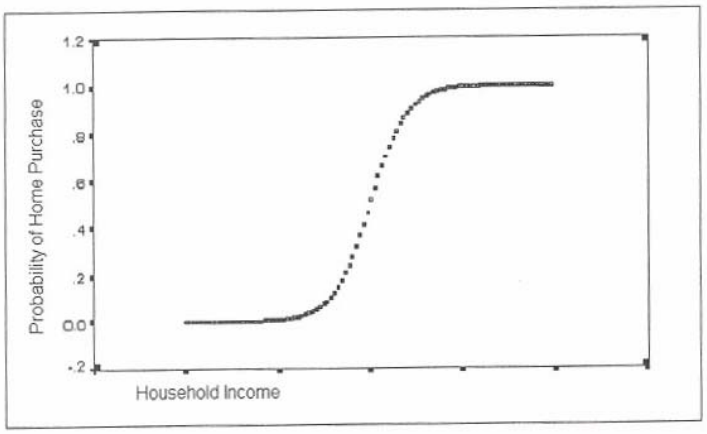
- **Assume que os dados podem ser modelados através de uma relacionamento linear**
- **Na sua forma clássica é capaz de predizer um resultado numérico através de um conjunto de preditores também numéricos**
- **Para utilizar campos não numéricos é necessária codificação**
- **É uma técnica bastante rápida: uma só passagem nos dados**
- **O modelo é uma única equação linear - a interpretação é fácil**



Regressão Logística

Tenta calcular coeficientes de regressão por forma a colocar os casos numa linha, mas não uma linha recta.

Tenta predizer uma função contínua que representa a probabilidade associada a ser um a dada categoria de resultado.



Clustering

Ajudam a descobrir grupos de registos de dados com valores de padrões semelhantes

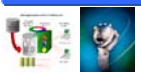
Técnica utilizada em marketing (segmentação de clientes) e outras aplicações de negócio (registos que caem em clusters de um só registo, podem ser erros ou casos de fraude)

- **Algumas vezes é executada antes da modelação preditiva**
  - **cada segmento é depois modelados individualmente (partindo da hipótese que cada cluster é único)**
  - **ou o grupo de clusters pode ser tratada como uma entrada adicional para o modelo**
- **Disponíveis três métodos:**
  - **redes Kohonen**
  - **k-means**
  - **two-steps**

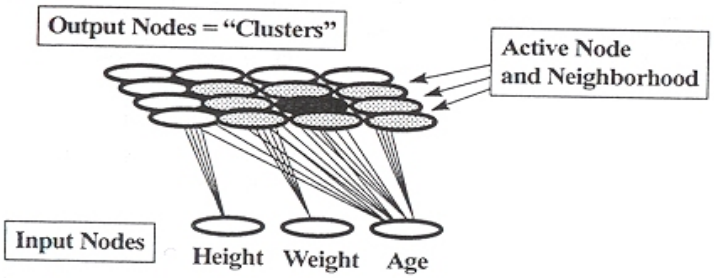
Clustering: redes Kohonen

Rede neuronal que possibilita aprendizagem não supervisionada:

- **não há uma saída ou resultado a predizer**
- **utilizada para gerar clusters ou segmentar dados, baseada nos padrões dos campos de entrada**
- **parte da suposição básica de que os clusters são formados a partir de padrões que partilham características semelhantes e agruparão conjuntos de padrões semelhantes**
- **estas redes são normalmente grelhas uni ou multidimensionais de arrays de neurónios artificiais.**
- **cada neurónio é ligado a cada uma das entradas**
- **os pesos de cada neurónio representam um perfil para o cluster relativamente aos campos utilizados na análise**
- **não há camada de saída**
- **mas os neurónios que constituem o chamado mapa Kohonen podem ser pensados como uma saída**



Clustering: redes Kohonen



Arranjo dos nós de saída na rede Kohonen numa grelha bi-dimensional que simulam o reforço positivo de neurónios do cérebro.

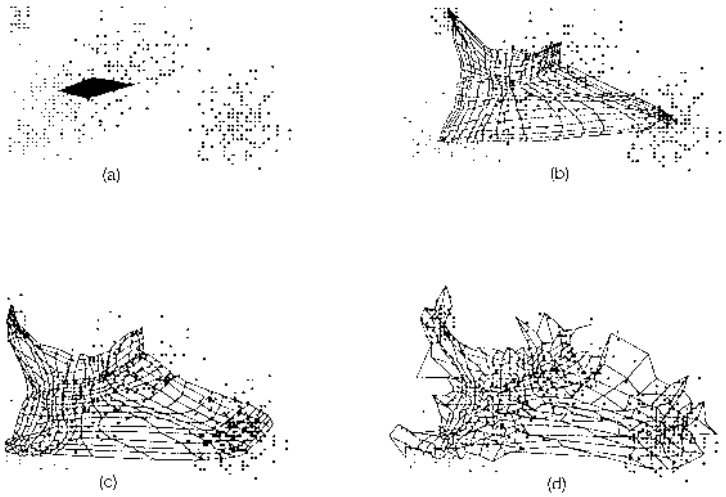
Clustering: redes Kohonen

Teino - apresentado um registo à grelha

- **os seus padrões de entrada são comparados com os dos neurónios da grelha**
- **o neurónio com o padrão mais semelhante, ganha a entrada**
- **os pesos do neurónio vencedor vão ser alterados por forma a que se torne mais semelhante ao padrão de entrada**
- **os pesos dos neurónio que rodeia o vencedor são também ajustados ligeiramente**

- **o efeito combinado deste mecanismo de aprendizagem é a movimentação do neurónio mais semelhante e os nós das redondezas, num menor grau, para a posição do registo do espaço do dado de entrada.**
- **depois dos dados passarem pela rede várias vezes, será gerado um mapa que contém clusters dos registos correspondentes aos diversos tipos de padrões existentes nos dados.**

Redes Neuronais: Kohonen



Visualização do processo de treino da rede Kohonen. A rede comporta-se como uma superfície elástica que é estendida pelo espaço de amostragem.

A 1ª. figura mostra o estado inicial dos vectores;

As figuras seguintes mostram estados subsequentes que permitem visualizar o mapa que se organiza gradualmente em direcção a uma melhor cobertura do espaço de amostragem.

Clustering: K-Means

Método relativamente rápido de exploração de clusters nos dados

Algoritmo:

- **O utilizador estabelece o número de clusters (k)**
- **O procedimento selecciona k registos de dados como clusters iniciais**
- **Cada registo é depois assignado ao mais próximo dos k clusters**
- **Os centros dos clusters (média dos campos utilizados no clustering) são actualizados por forma a acomodar os novos membros**
- **Passagens de dados adicionais são feitas, se necessário**
  - **À medida que o centro de um dado cluster se desloca, um registo de dados pode necessitar de se mover para um cluster mais próximo**

Como o número de clusters é especificado pelo utilizador:

- **Este método é executado normalmente várias vezes, avaliando-se para cada caso os resultados, fazendo-se variar o K**

Clustering: Two Steps

Solução híbrida entre clustering hierárquico e não hierárquico.

O utilizador especifica um intervalo (Mínimo - Min e Máximo - Max) para o número de clusters.

- **1.º Passo: todos os registos são classificados como pertencentes a um dos Max clusters (clustering não hierárquico)**
- **2.º Passo: Um método de cluster aglomerativo (hierárquico) é utilizado para combinar sucessivamente os clusters gerados no passo anterior**

Resultado: o número final de clusters estará compreendido entre Max e Min, de acordo com um critério estatístico.

Vantagem deste método:

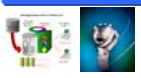
- **selecção automática do número de clusters (dentro do intervalo especificado)**
- **não obriga a enormes recursos de máquina (como é o caso do clustering não hierárquico)**

Regras de Associação

Trata-se da forma de Data Mining que de mais perto se assemelha ao processo que a maioria das pessoas lhe associa "Minar uma grande base de dados à procura da pepita de ouro"

Aqui a pepita de ouro significa:

"Uma regra que diga algo sobre a base de dados que não se saiba e que provavelmente não sejamos capazes de articular explicitamente, além de ser algo de interesse"



Regras de Associação

Os gestores de marketing gostam de regras como:

"90% das mulheres possuidoras de carros de desporto vermelhos e cães pequenos, usam Chanel Nº5"

- **Este tipo de regras descritivas mostra perfis claros dos clientes que poderão constituir alvos das suas acções de marketing.**

Estas regras são chamadas de Regras de Associação, sendo um dos outputs possíveis de várias técnicas de Data Mining.

- **as regras de associação são sempre definidas em termos de atributos binários**
- **é possível encontrar associações em grandes bases de dados**

mas

- **além de associações de interesse**
- **encontrar-se-ão muitas associações de pouco valor, já que o número de associações será quase infinito.**

Regras de Associação

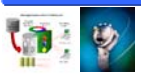
Procuram coisas (eventos, compras, atributos) que, tipicamente, ocorram em conjunto)

Encontram automaticamente padrões nos dados que podem manualmente ser encontrados utilizando técnicas visuais (nó web p. ex.), mas a uma muito maior velocidade e podendo explorar um muito maior número de padrões complexos.

As regras encontradas associam uma cada categoria de resultado (chamada conclusão) com um conjunto de condições.

Os campos de resultado podem variar de regra para regra:

- **o utilizador não pode, na maioria das vezes, focar num determinado campo de saída.**



O que é uma Regra?

"Se são comprados pickles, então também é comprado Ketchup"

Antecedente: pode ser uma ou mais condições, que devem ser todas verdadeiras por forma ao consequente seja verdadeira, dada a precisão

Consequente: Geralmente só uma condição e não uma combinação múltipla.

Para uma regra ser útil, além da própria regra, há que fornecer dois suplementos:

1. **Precisão - Quantas vezes está correcta a regra? É a probabilidade de que se o antecedente for verdadeiro, então o consequente será verdade. Precisão alta, significa estar em presença de uma regra de alta confiança, daí ser também chamada de confiança.**
2. **Cobertura - Quantas vezes se aplica a regra? Relacionado com a % da base de dados que é coberta pela regra. Alta cobertura significa que a regra se aplica muitas vezes e que também é pouco provável que se trate de algo espúrio introduzido pela técnica de amostragem ou ideosincrasias.**

Regras de Associação

Vantagem sobre as árvores de decisão:

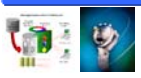
- **são procuradas associações podem existir entre quaisquer dois ou mais campos.**

Desvantagem:

- **como tentam encontrar padrões num espaço de pesquisa muito grande, são de execução bastante lenta**

Algoritmos disponíveis no Clementine:

- **APRIORI**
- **GRI**



Que Técnica? Quando?

Se estiver disponível um campo de dados que claramente se pretende predizer:

- **utilizar qualquer das técnicas de aprendizagem supervisionada**
- **um dos métodos de modelação estatística**

Se é pretendido encontrar grupos de indivíduos que se comportam de forma semelhante:

- **utilizar qualquer dos métodos de clustering será adequado**

Regras de associação - não permitem directamente a predição, mas:

- **são muito úteis para a compreensão dos vários padrões existentes nos dados**

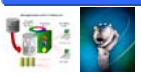
Que Técnica? Quando?

Mas... dentro de cada tipo de utilização, que técnica particular utilizar?

- **dependem dos dados sobre os quais se está a trabalhar**
- **dependem dos campos a predizer e como estão relacionados com as várias entradas**

há no entanto algumas sugestões que podem guiar a selecção que serão indicadas à medida que se forem estudando as diversas técnicas

- **mas estas sugestões não passam disso, não são regras: podem não se adequar**



Que Técnica? Quando?

Mas, dado que o Clementine permite uma grande simplicidade na construção de modelos:

- **Podemos criar redes neuronais, árvores de decisão e modelos de regressão podem ser criados com grande facilidade e velocidade e**
- **Comparar os resultados obtidos**

O Data Mining é um processo iterativo:

- **São criados modelos que depois podem ser combinados ou eliminados, antes de o analista de negócio fique contente com os resultados**