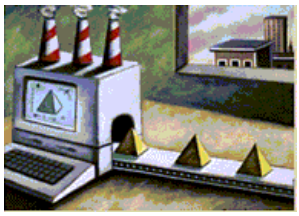
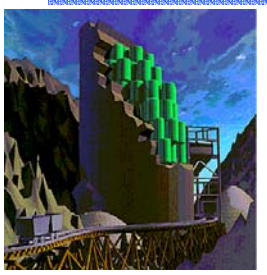
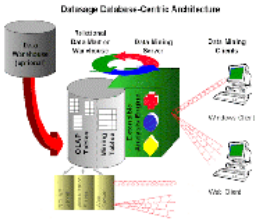


Robert Groth

Métodos e Algoritmos de Data Mining(parte 1)



Usama Fayyad et al



Métodos e Algoritmos de Data Mining

- Soluções distância (K-NN e clustering)
- Naïve-Bayes
- Arvores de decisão
- Regras de associação
- Redes neuronais
- Algoritmos genéticos.
- Combinação de múltiplos métodos de predição.

- **Alguns prós e contras das tecnologias mais comuns; ferramentas mais relevantes e suas características**

Nearest Neighbor e Clustering

Clustering e Nearest Neighbor

Clustering e Técnicas de Predição Nearest Neighbor (vizinho mais próximo) estão entre as técnicas mais velhas de Data Mining.

Intuição de Clustering

- **os registos semelhantes são agrupados ou "clustered" e colocados no mesmo grupo;**
- **permite ter-se uma visão de alto nível do que se passa na base de dados;**
- **também utilizado para significar segmentação.**

Nearest Neighbor

- **técnica preditiva semelhante ao clustering**
- **para saber qual o valor de predição para um dado registo, vamos procurar registos à base de dados com valores de preditores semelhantes e utilizar o valor de predição do mais próximo ou mais comum entre os mais próximos.**

Clustering e Nearest Neighbor

Valor para o negócio:

Medida de Data Mining	Descrição
Automatização	**Técnicas NN são relativamente automáticas, embora requeiram algum pré-procesamento de alguns preditores para valores que possam ser utilizados em medidas de distância. Preditores categóricos não ordenados necessitam de ser definidos em termos da distância entre si. A maioria dos algoritmos são também robustos relativamente a dados sujos ou em falta.**
Careza	**Fácil de utilizar e muito clara: trabalha de forma análoga à forma como as pessoas pensam - detectando exemplos semelhantes. Excelente para explanação clara do porquê da predição ser feita. Um exemplo simples ou conjunto de exemplos podem ser extraídos da base de dados histórica para evidenciar o porquê de algo que deve ou não ser feito. O sistema pode também comunicar quando não está confiante acerca da sua predição. A nível de descrição geral dos dados, não é poderoso, pois que não é criado qualquer modelo.**
ROI	**Boa para efectuar cálculos complexos relativos a ROI, dado que as predições são realizadas a nível local (onde as simulações de negócio podem ser efectuadas por forma a optimizar o ROI). Os registos individuais são transferidos directamente da base de dados sem alterá-la, é possível compreender todas as facetas do comportamento do negócio, não só através da predição, mas de muitos outros factores. Também possui nível semelhante de precisão comprado com outras técnicas.**

Nearest Neighbor (NN)

Tipo de Aplicações:

Tipo de Problema	Descrição
Clusters	**O método subjacente à tecnologia NN é proximidade em alguma característica espacial. É a mesma métrica base utilizada na maioria dos algoritmos de *clustering*, embora, em predição, a característica espacial seja moldada de forma a facilitar uma predição particular.**
Ligações	**Podem ser utilizadas para análise de ligações, desde que os dados sejam pré - formatados por forma a que os valores dos predictores estejam no mesmo registo (ex. na análise de cesto de compras, os artigos de uma dada compra devem estar guardados no mesmo registo - ou seja os registos devem ser de comprimento variável).**
Excepções	**Particularmente boas para detecção de excepções (*outliers*) dado que é criada um espaço dentro do qual é possível determinar quando um registo está fora.**
Regras	**Uma das forças da técnicas NN é que levam em conta todos os preditores em algum grau, o que, sendo bom para predição, leva a que em modelos complexos não possam ser descritos facilmente em regras. O sistema é também geralmente optimizado para a predição de novos registos e não para a extracção de regras de interesse da base de dados.**
Sequências	**Técnicas NN têm sido utilizadas com sucesso para efectuar predições em sequências temporais. Os valores temporais têm de ser codificados em registos.**
Texto	**A maioria dos sistemas de pesquisa textual são baseados em tecnologias NN, e muitos dos avanços em pesquisa em textos são ulteriores refinamentos de algoritmos de predição pesada e cálculo de distância.**

Clustering e NN

Vamos ver clustering e NN em funcionamento: efectuar clusters de amigos?

ID	Nome	Idade	Saldo(€)	Rendimentos	Olhos	Sexo
1	Amanda	62	0	Médio	Castanhos	F
2	António	53	1,800	Médio	Verdes	M
3	Beatriz	47	16,543	Alto	Castanhos	F
4	Baltazar	32	45	Médio	Verdes	M
5	Carla	21	2,300	Alto	Azuis	F
6	Carlos	27	5,400	Alto	Castanhos	M
7	Diva	50	165	Baixo	Azuis	F
8	Daniel	46	0	Alto	Azuis	M
9	Edna	27	500	Baixo	Azuis	F
10	Edgar	68	1,200	Baixo	Azuis	M

Clustering e NN

Não é possível dizer se a 1.ª(pragmática - compatibilidade financeira) ou 2.ª (mais romântica – idade e cor dos olhos) forma de clusters será melhor ou pior. Os clusters não são construídos por razão nenhuma especial, excepto para notar semelhanças entre alguns dos registos e uma visão simplificada da base de dados.

As motivações podem ser diferentes (financeiras x românticas), mas, em geral, as razões são mal definidas, pois que os clusters são fundamentalmente utilizados para exploração e sumarização e não propriamente para predição.

ID	Nome	Idade	Saldo(€)	Rendimentos	Olhos	Sexo
3	Beatriz	47	16,543	Alto	Castanhos	F
5	Carla	21	2,300	Alto	Azuis	F
6	Carlos	27	5,400	Alto	Castanhos	M
8	Daniel	46	0	Alto	Azuis	M
1	Amanda	62	0	Médio	Castanhos	F
2	António	53	1,800	Médio	Verdes	M
4	Baltazar	32	45	Médio	Verdes	M
7	Diva	50	165	Baixo	Azuis	F
9	Edna	27	500	Baixo	Azuis	F
10	Edgar	68	1,200	Baixo	Azuis	M

ID	Nome	Idade	Saldo(€)	Rendimentos	Olhos	Sexo
5	Carla	21	2,300	Alto	Azuis	F
9	Edna	27	500	Baixo	Azuis	F
6	Carlos	27	5,400	Alto	Castanhos	M
4	Baltazar	32	45	Médio	Verdes	M
8	Daniel	46	0	Alto	Azuis	M
7	Diva	50	165	Baixo	Azuis	F
10	Edgar	68	1,200	Baixo	Azuis	M
1	Amanda	62	0	Médio	Castanhos	F
2	António	53	1,800	Médio	Verdes	M
3	Beatriz	47	16,543	Alto	Castanhos	F

Clustering e NN

Qual é a diferença entre clustering e Predição NN?

Clustering: técnica de aprendizagem não supervisionada
não supervisionada no sentido de que, quando são executadas, não há razão nenhuma especial para a criação de modelos – não há nenhuma razão particular para o porquê de certos registos estarem próximos ou porque ficaram no mesmo cluster.

NN: técnica de aprendizagem supervisionada e utilizada geralmente para predição; aqui há uma razão para a a criação de modelos: a predição. Os padrões que são patentes no modelo são sempre os padrões mais importantes da base de dados para executar alguma predição particular.

Clustering

É um compromisso entre homogeneidade e minimização de número de *clusters*.

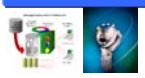
Caso ideal - todos os registos de cada cluster com valores idênticos.

- **Para o garantir, no extremo, teríamos *clusters* de um registo.**

Mas e então "o número razoável de razoável de *clusters*?"

- ***Clusters* de um registos serão inaceitáveis!**

- **Muitos algoritmos deixam os utilizadores especificar o número de *clusters*;**
- **Outros algoritmos encontram o número baseado nos próprios dados e em medidas de proximidade mínima.**



Definição de Proximidade

- **Trabalho num espaço n-dimensional**
- **Há alguma forma de conhecer se um registo está perto ou longe de outro**

Para calcular a proximidade:

- **Distância Manhattan - adicionar as diferenças entre cada preditor dos registos históricos e o registo a ser predito**
- **Distância Euclidiana - calcula a distância como o teorema de Pitágoras: raíz quadrada da soma dos quadrados das distâncias**

Como já foi dito atrás, importa <u>normalizar os valores preditores</u> (o visto fenómeno de escala). Dessa forma não há um preditor que domine claramente todos os outros, tornando o seu impacto na distância determinante.

ID	Name	Prediction	Age	Balance ($)	Income	Eyes	Gender
5	Carla	Yes	21	2300	High	Blue	F
6	Carl	No	27	5400	High	Brown	M

6+3100+0+1+1=3108 ⇨ **Normalizar os valores em cada dimensão por forma a que o valor máximo e mínimo sejam os mesmos (0 e 100).** ⇨ **6+19+0+100+100=225**

NN - Peso das Dimensões

Resta ainda a questão:

⇨ **quando se define <u>próximo</u>, quão importante é o <u>peso</u> de cada dimensão?**

- **A metáfora da subida à montanha é sintomática: é mais difícil subir 1000 m a pique do que fazê-lo em terreno quase plano. A distância vertical deverá ser mais relevante para a medida de distância a utilizar.**

Em text mining, utiliza-se

- **o inverso da frequência com que a palavra é utilizada**
- **a importância da palavra no tópico a ser predito**

Em outros problemas de negócio

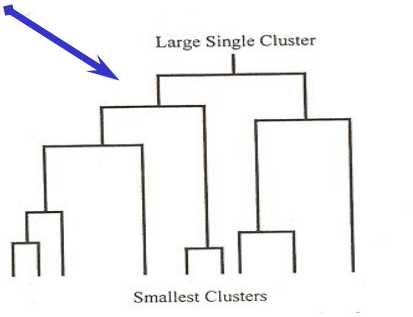
- **correlação entre o preditor e colunas de predição**
- **probabilidade condicional que a predição tenha um certo valor, dado o preditor ter um certo valor (se todas as vezes que um preditor contiver um dado valor vai predizer correctamente o valor predição então deverá ter um peso alto)**

Clustering Hierárquico e não Hierárquico

Dois tipos principais de técnicas de *clustering*:

- **os que criam uma hierarquia**
- **os que não criam essa hierarquia**

A hierarquia de clusters é normalmente visualizada como uma árvore onde os clusters menores se fundem para criar os níveis mais altos de cluster.



Assim o utilizador pode avaliar o número correcto de clusters que sumarizam adequadamente os dados ainda proporcionando informação útil

Tipo de Algoritmos de Clustering Hierárquico

Aglometarivos: Começam com tantos clusters como registos. Os clusters que estão próximos entre si, são fundidos para criar o cluster maior. Esta fusão prossegue até que seja criada uma hierarquia de clusters com um só cluster no topo que contenha todos os registos.

Divisivos: Técnica que seguem a abordagem oposta. Começa com todos os registos num cluster e tenta-se a sua divisão sucessiva em clusters mais pequenos.

⇨ **As técnicas aglometarivas são as mais utilizadas e têm mais algoritmos desenvolvidos.**

Técnicas p/ Clustering não Hierárquico

São geralmente mais rápidas do que as de clustering hierárquico, mas necessitam que o utilizador tome decisões acerca do número de clusters desejados ou a "mínima proximidade" requerida para que dois registos fiquem no mesmo cluster.

Podem utilizar:

- **iterações sucessivas, iniciando com clusters arbitrários ou aleatórios e melhorando os clusters por deslocação de alguns registos.**
- **só uma passagem pela BD, adicionando registos a clusters existentes e criando novos quando não exista qualquer cluster que seja um bom candidato para um dado registo.**

Técnicas n/ Hierárquicas x Hierárquicas

- **Os clusters formados pelas primeiras dependem das selecções iniciais dos clusters de arranque que devem ser escolhidos ou quantos clusters utilizar**
- **As primeiras podem revelar-se menos repetíveis do que as segundas**
- **Por vezes criam-se demasiados ou poucos clusters, pois que, nas primeiras, o seu número é pré-determinado pelo utilizador, não somente pelos padrões inerentes da base de dados**

Clustering n/ Hierárquico

Duas técnicas principais:

- **<u>uma só passagem</u> - a base de dados é atravessada uma só vez para serem criados os clusters**
- **<u>método de realocação</u> - envolve o movimento ou "realocação" dos registos de um cluster para outro, por forma a criarem-se melhores clusters; utilizam múltiplas passagens pela base de dados, mas são relativamente rápidos comparados com as técnicas hierárquicas.**

Instituto Superior Politécnico de VISEU
Escola Superior de Tecnologia
Departamento de Informática

Clustering n/ Hierárquico

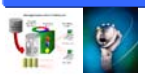
Algoritmo de Passagem Simples

1. **Ler um registo da base de dados e determinar o cluster onde ficará melhor (utilizando alguma medida de proximidade)**
2. **Se o cluster mais próximo estiver algo longe (não há uma boa adaptação) criar um novo cluster com esse registo lá**
3. **Ler o próximo registo**

Prós: **Dado que a leitura de registos da base de dados é, muitas vezes, o aspecto mais caro dos algoritmos de clustering, este algoritmo é bastante rápido.**

Contras:

- **cria muitas vezes grandes clusters muito cedo no processo de clustering**
- **os clusters criados dependem da ordem pela qual os registos se estruturam na base de dados (os registos lidos no início têm um efeito significativo nos clusters formados)**



Clustering n/ Hierárquico

Algoritmo p/ Realocação

1. Pré-seleccionar o número de clusters desejado
2. De forma aleatória seleccionar um registo para se tornar o centro ou "semente" para cada um desses clusters
3. Atravessar a base de dados e assignar cada registo ao cluster mais próximo
4. Recalcular o centro dos clusters
5. Repetir os passos 3 e 4 até que ocorra um mínimo de realocações de registos entre clusters

O que se passa é que:

- os registos inicialmente alocados para clusters podem não ser particularmente bem adaptados
- mas recalculando o centro do cluster no passo 4, vemos que são formados clusters que melhor se adaptem aos dados - os centros dos clusters vão-se movimentando pelo espaço n-dimensional, aproximando-se mais e mais dos centros de alta densidade e afastando-se das excepções

Clustering Hierárquico

- Os clusters são definidos só pelos dados
- O número de clusters pode ser aumentado ou diminuído descendo ou subindo na hierarquia

P/ técnicas aglomerativas:

1. Iniciar com tantos clusters como registos
2. Combinar os dois clusters mais próximos num cluster maior
3. Continuar até só restar um cluster

P/ técnicas divisivas:

1. Iniciar com um cluster que contenha todos os registos da base de dados
2. Determinar a divisão do cluster existente que melhor maximize a similaridade dentro dos clusters e dissimilaridade entre clusters
3. Dividir o cluster e repetir para os dois clusters resultantes
4. Terminar quando algum limiar mínimo de tamanho de cluster ou um número total de clusters tenha sido atingido ou ainda quando só houver um registo no cluster

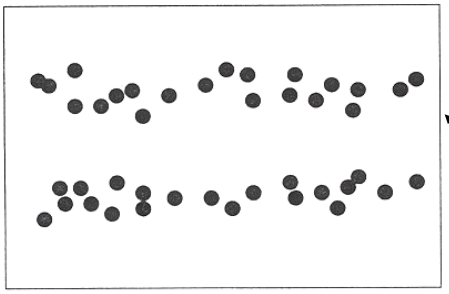
Clustering Hierárquico

- **A técnica divisiva pode ser computacionalmente muito cara**
  - **a base para a divisão é a distância média mínima entre registos dentro do cluster (isto calculado para quaisquer dois possíveis clusters dentro do cluster maior)**
- **Os métodos aglomerativos prevalecem hoje.**
  - **A decisão de fusão pode ser efectuada de várias formas, cada uma delas privilegiando um dado tipo de cluster, desde os compactos esféricos, aos alongados...**

Como fundir Clusters (técnicas aglomerativas)

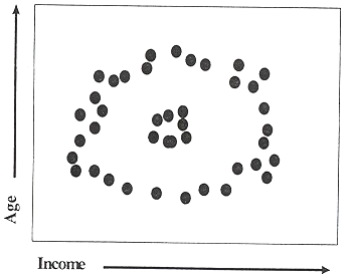
1. **<u>Método Ligação Simples</u> - Fundir os clusters cujos registos mais próximos o estejam o mais possível. Neste caso a fusão pode ser efectuada com base num único par de registos, podendo criar-se assim clusters longos, tipo cobra. Técnica não adequada à extracção dos clássicos clusters esféricos e clusters compactos.**
2. **<u>Método Ligação Completa</u> - Juntar os clusters cujos registos mais distantes estejam tão perto quão possível. O seu nome advém do facto de todos os registos dentro do cluster estarem ligados entre si dentro de uma distância máxima. Esta técnica favorece a criação de clusters compactos e pequenos.**
3. **<u>Método Ligação Média-Grupo</u> - Juntar os clusters onde a distância média entre todos os pares de registos seja tão pequena quanto possível. Como considera todos os registos dentro dos clusters, incluindo o mais próximo e o mais distante, resulta em clusters algo entre os clusters do tipo 1 e 2.**
4. **<u>Método Ward</u> - Fundir os clusters cujo cluster resultante tenha um mínimo de distância total entre todos os registos. Tende a produzir-se uma hierarquia simétrica e é bom na recuperação da estrutura dos clusters, mas é sensível a excepções e tem dificuldade em cobrir clusters alongados.**

Como fundir Clusters



Exemplo de cluster alongado que poderão não ser conseguidos pelos métodos ligação completa ou Ward, mas que o serão pelo método ligação simples. O centro de gravidade do cluster pode estar bastante longe de um dado ponto, mas cada ponto está ligado a todos os outros por uma série de pequenas distâncias.

Exemplo clusters aninhados que não serão conseguidos pelo método de ligação completa ou Ward, mas sê-lo-ão pelo método ligação simples. Neste caso também há a ligação entre cada dois pontos por uma série de pequenas distâncias.



NN como predição

Até aqui, em clustering, a metáfora espacial lidava com dois conceitos:

- **pontos que representavam registos existentes e espaços em branco que representavam registos possíveis mas que não existiam na base de dados.**

Em predição o problema torna-se um pouco mais interessante:

- **os dois conceitos anteriores persistem, mas adiciona-se o conceito de valor de predição.**

Ex. se o valor de predição é simplesmente sim e não, teremos pontos que representam os registos e os espaços em branco, mas os pontos serão neste caso, de dois tipos diferenciados, que simbolizarão registos "sim" e "não".

O problema torna-se agora mais bem definido:

"Criar clusters que sejam tão homogéneos quanto possível de forma a que, quando utilizados em predição, seja minimizado o erro em dados de teste"

K-Nearest Neighbor (K-NN)

Técnica predictiva adequada para classificação.

- **Vizinhança - Registos representados como pontos no espaço de dados que estejam perto uns dos outros devem viver na vizinhança uns dos outros.**

Como prever o comportamento de um conjunto de clientes?

- **Os clientes do mesmo tipo devem manifestar o mesmo tipo de comportamento.**
- **Como:**
  - **em termos de metáfora multidimensional:**
    - **o tipo não será mais do que uma região do espaço de dados**
    - **registos do mesmo tipo estarão perto uns dos outros no espaço de dados**

Poderemos desenvolver um algoritmo de aprendizagem simples, mas poderoso - K-NN

K-Nearest Neighbor (K-NN)

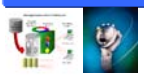
Então:

Para prever o comportamento de um certo indivíduo:

- **iremos olhar para o comportamento de um certo número de indivíduos (ex. 10) que estejam na sua vizinhança no espaço de dados.**
- **calcularemos a média do comportamento desses 10 indivíduos, sendo essa média a previsão para o novo indivíduo.**

Significado de K em K-NN

- **será o número de indivíduos que iremos investigar - se 5, teremos 5-nearest neighbor.**



K-Nearest Neighbor (K-NN)

Não é propriamente uma técnica de aprendizagem, mas mais um método de procura:

- **o data set é utilizado como referência, revelando-se assim uma técnica pura de pesquisa;**
- **não é construído qualquer modelo, ou seja, não cria uma nova teoria que permita melhor compreender a sua estrutura: os dados de treino são o próprio modelo;**
- **quando um novo caso ou instância é apresentado, o algoritmo olha todos os dados para encontrar um subconjunto de dados que sejam mais semelhantes e usa-os para predizer o resultado;**
- **qualquer nova predição obriga à comparação com todos os registos do data set, levando a complexidade quadrática, não desejável para grandes data sets.**

K-Nearest Neighbor (K-NN)

Há dois factores principais no algoritmo k-NN:

- **o número de casos próximos a ser utilizado (k)**
- **uma métrica para medir o que é "próximo"**

Como é feita uma predição, classificando um novo caso:

- **o algoritmo calcula a distância, utilizando a métrica, desde o novo caso a cada caso (linha) dos dados de treino;**
- **o novo caso é predito como tendo o resultado predominante aos k casos mais próximos, nos dados de treino.**

K-Nearest Neighbor (K-NN) - Exemplo (1)

Aplicação do K-NN ao data set relativo à editora de revistas, já utilizado no capítulo anterior.

Revista	Precisão de predição
carros	89% correcto
casas	60% correcto
desporto	74% correcto
música	93% correcto
humor	92% correcto

Análise breve

para revistas de carros, desporto e música, obtem-se uma precisão consideravelmente melhor do que a naïve

para revistas de humor, não se consegue melhor do que a predição naïve, ou seja k-nn não ajuda

é interessante notar que, para revistas de casas tem um menor desempenho que a predição Naïve, podendo indicar que os leitores dessas revistas estão distribuídos aleatoriamente, não havendo padrões reais a ser descobertos ou que o set de dados é demasiado pequeno para revelar o comportamento detalhado específico deste grupo alvo.

K-Nearest Neighbor (K-NN) - Exemplo (2)

Nome	Débito	Rendim.	Casado?	Risco
Joaquim	Alto	Alto	Sim	Bom
Susana	Baixo	Alto	Sim	Bom
João	Baixo	Alto	Não	Pobre
Maria	Alto	Baixo	Sim	Pobre
Frederico	Baixo	Baixo	Sim	Pobre

sem valor para a predição do risco de crédito

colunas independentes, utilizadas para construir o modelo

Obs.: todas as colunas (independentes ou dependente são categóricas)

variável alvo ou coluna dependente: o que se pretende prever

Caso relativo a uma instituição de crédito que procura minimizar o incumprimento dos pagamentos das prestações relativas aos empréstimos concedidos.

Trata-se dum problema de simples classificação: predizer se um possível empréstimo será ou não de bom risco.

Para simplificar a data set é reduzido a 5 registos

K-Nearest Neighbor (K-NN) - Exemplo (2)

Nome	Débito	Rendim.	Casado?	Risco
Joaquim	1	1	1	1
Susana	0	1	1	1
João	0	1	0	0
Maria	1	0	1	0
Frederico	0	0	1	0

Codificação:

os valores categóricos são codificados como 0 e 1

Exemplo de cálculo da métrica

distância entre o Joaquim e Susana:

1 x 0 1 x 1 1 x 1

1 + 0 + 0 = 1

Métrica a utilizar:

vamos somar as diferenças

será a soma dos pontos calculados como 1 se o valor da coluna nos dois registos for diferente e 0, se igual.

K-Nearest Neighbor (K-NN) - Exemplo (2)

Nome	Joaquim	Susana	João	Maria	Frederico
Joaquim	0	1	2	1	2
Susana	1	0	1	2	1
João	2	1	0	3	2
Maria	1	2	3	0	1
Frederico	2	1	2	1	0

Quadro obtido, aplicando a métrica definida atrás e calculando a distância entre os registos

Para K=3

os vizinhos mais próximos do Joaquim, como o set de treino e teste é o mesmo, será ele próprio, a Susana e a Maria

Para estes vizinhos, o valor do risco é Bom, Bom e Pobre

Risco previsto: Bom

K-Nearest Neighbor (K-NN) - Exemplo (2)

Nome	Joaquim	Susana	João	Maria	Frederico
Joaquim	0	1	2	1	2
Susana	1	0	1	2	1
João	2	1	0	3	2
Maria	1	2	3	0	1
Frederico	2	1	2	1	0

Quadro obtido, aplicando a métrica definida atrás e calculando a distância entre os registos

Para K=3

Relativamente à Susana, os seus vizinhos mais próximos, será ela mesma e Joaquim, João e Frederico à mesma distância. Vamos considerá-los os três, mas com valor 2/3

Para estes vizinhos, o valor do risco é Bom (Susana), 2/3 Bom (Joaquim), 2/3 Pobre (João) e 2/3 Pobre (Frederico)=1/3 Bom = Bom

Risco previsto: Bom

K-Nearest Neighbor (K-NN) - Exemplo (2)

Nome	Débito	Rendim.	Casado?	Risco	Predicção
Joaquim	Alto	Alto	Sim	Bom	Bom
Susana	Baixo	Alto	Sim	Bom	Bom
João	Baixo	Alto	Não	Pobre	Bom
Maria	Alto	Baixo	Sim	Pobre	Pobre
Frederico	Baixo	Baixo	Sim	Pobre	Pobre

Predições para o algoritmo 3-NN

Precisão: 80% (falhou para o João)

João Susana Joaquim e Frederico

Pobre Bom 1/2 Bom 1/2 Bom = Bom

K-Nearest Neighbor (K-NN)

Qual a influência de K? Qual será um bom valor para K?

Se k=1

- **será procurado o caso mais próximo**
- **ao encontrar teremos o valor da predição**
- **para o set de treino teremos uma precisão de 100%**

mas

- **será enormemente susceptível ao ruído e não reflectirá qualquer tipo de padrão nos dados**

Muitos dos algoritmos comerciais, começam com k=10

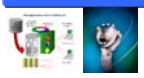
O melhor valor para k:

- **requer alguma experiência**
- **comparar resultados para diversos valores de k, relativamente a set de teste**

Avaliação do k-NN

É possível <u>não só efectuar predições</u>, mas também obter informação acerca da <u>confiança na predição</u>.

- **a <u>distância ao vizinho mais próximo</u> proporciona um nível de confiança: se a vizinhança for muito próxima ou uma perfeita sobreposição, deverá haver uma maior confiança na predição. O contrário acontecerá na situação inversa.**
- **<u>o grau de homogeneidade</u> entre as predições dos K vizinhos próximos pode também dar a noção de confiança. Se todos os k registos mais próximos fizerem a mesma predição, então a confiança será muito maior do que se o valor da predição for quase 50-50.**



NN como predição - fraquezas

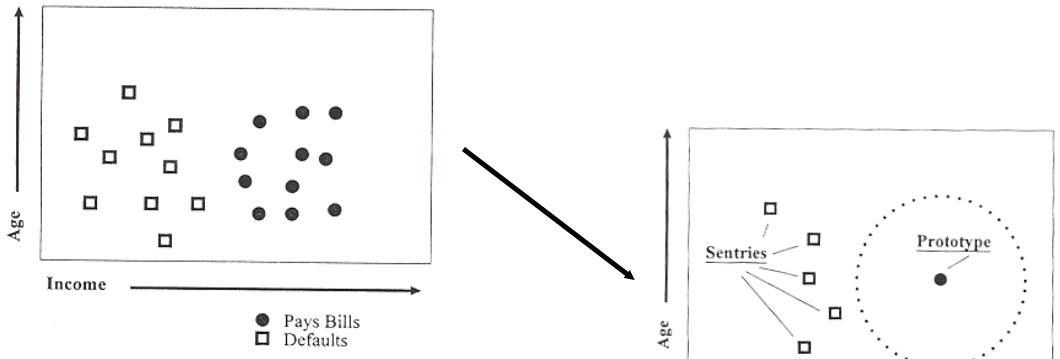
1. **Não é criado um modelo (descrição mais compacta dos dados), toda a base de dados será o próprio modelo. É assim muito grande, comparado com outros modelos preditivos gerados por outras técnicas (redes neuronais ou árvores de decisão).**
2. **Também não há uma maneira formal de prevenir o sobreajustamento, pois que a totalidade da base de dados é utilizada como parte do modelo preditivo.**

Soluções:

- **Protótipos são formados pela fusão de registos próximos num registo médio (protótipo) que representa ambos os registos. O processo é repetido enquanto a precisão global não ficar abaixo da conseguida com a totalidade da base de dados.**
- **Sentinelas são formadas removendo os registos não necessários a fazer uma predição precisa. Eles representam as fronteiras do espaço n-dimensional dentro do qual o valor de predição é homogéneo. O seu nome advém do facto de funcionarem como sentinelas a guardar as fronteiras geográficas.**

NN como predição – Generalização

Protótipos e Sentinelas



Protótipos e Sentinelas reduzem dramaticamente o número de registos que são necessários para manter a precisão predictiva e muitas vezes melhoram a robustez e generalidade do modelo

NN como predição - forças

- possibilidade de utilização da metáfora espacial para agrupar dados
- o algoritmo pode basear-se directamente nos dados históricos
- permite a compreensão não só do que é o modelo preditivo e de como trabalha, mas também de como é derivado
- permite visualizar os k registos que contribuíram para uma predição particular
- dado que o modelo está fortemente ligado à base de dados, à medida que são acrescentados novos registos ou colunas, são automaticamente incorporadas na base de dados
- também se a infraestrutura da base de dados for melhorada, as vantagens do ganho de desempenho serão imediatamente passadas ao algoritmo
- muito adequado a tratamento de bases de dados com grande número de colunas predictor de natureza binária ou categórica

K-Nearest Neighbor (K-NN) - Resumo

- O K-NN não faz uma passagem de aprendizagem pelo data set
- Todo o processamento é adiado até que as predições sejam feitas:
  - cada novo caso obriga a uma nova passagem pelos dados de treino para calcular a distância entre a instância alvo e cada instância do set de dados
  - o algoritmo vai registando as K instâncias mais próximas, à medida que progride na avaliação do data set
  - o resultado é obtido quando a passagem for completada

K-Nearest Neighbor (K-NN) - Resumo

Quanto à métrica:

1. **Para dados categóricos, tais como as variáveis de dois valores do último exemplo, a métrica 0-1 que utilizámos é normalmente suficiente;**
2. **Para dados categóricos ordenados (grupos de idades, p.ex.), deveremos empregar uma métrica que reconheça a diferença entre grupos /entre 21-30 e 51-60 > 21-30 e 41-50;**
3. **Quanto a dados contínuos, será a diferença suficiente?**

- **aqui há que precaver relativamente à escala correcta, sob pena do valor de uma determinada coluna se sobrepor a todas as outras.**
- **já vimos atrás, a utilização de uma medida euclidiana de distância.**

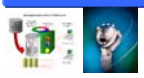
Apreciação Geral do K-NN

Tem algumas desvantagens, quando comparado com outros métodos:

- **Falta de descrição dos resultados, ainda que proporcione um conjunto de casos onde se baseia a predição;**
- **Dependente de uma medida de distância arbitrária (tornando a técnica subjectiva, restringida a aplicações que tenham medidas naturais de distância);**
- **Desempenho (tem de se tratar todo data set a cada novo caso), só aliviada pela utilização de sentinelas ou protótipos;**

Qualidades:

- **Algoritmo bastante simples;**
- **Só necessita de uma passagem pelos dados;**
- **Pode ser utilizado para comparar predições geradas através de outros métodos;**
- **Em casos onde haja medida natural de distância, produz bons resultados.**



Naïve-Bayes

Naïve-Bayes

Deve o nome a Thomas Bayes (1702-1761) que, numa sua obra póstuma, demonstrou o teorema de Bayes, utilizado na técnica de Naïve-Bayes para calcular as probabilidades que são utilizadas para fazer as predições.

Técnica de classificação que é simultaneamente predictiva e descritiva. Analisa o relacionamento entre cada variável independente e a variável dependente para derivar uma probabilidade condicional para cada relacionamento.

- **Quando um novo caso é analisado,**
  - **é feita a previsão combinando o efeito das variáveis independentes na variável dependente.**

Naïve-Bayes (avaliação genérica)

Em teoria, uma predição Naïve-Bayes só será correcta se todas as variáveis independentes forem estatisticamente independentes umas das outras, o que é frequentemente falso.

- **Ex. os dados acerca de pessoas conterão normalmente múltiplos atributos (peso, grau de educação, rendimento, etc.) que estão correlacionados com a idade. Neste caso a utilização deste algoritmo irá sobrevalorizar o efeito da idade.**

Apesar dessas limitações, a prática mostra que Naïve-Bayes produz bons resultados e a sua simplicidade e velocidade tornam-no numa ferramenta ideal para modelação e investigação de relacionamentos simples.

Naïve-Bayes (avaliação detalhada)

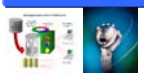
Vantagens:

- **só requer uma passagem pelos dados de treino para gerar um modelo de classificação (é a técnica de data mining mais eficiente)**

Desvantagens:

- **não trata dados contínuos:**
  - **assim, qualquer variável dependente ou independente deve ser transformado em intervalos ("binned")**
    - **ex. no caso da idade, teremos de ter intervalos como <20, 21-30, 31-40, etc.**

A transformação em intervalos é tecnicamente simples, mas a selecção dos intervalos podem ter um impacto dramático na qualidade do modelo produzido.



Como funciona o algoritmo de Naïve-Bayes

1. Calcular a probabilidade "*a priori*" da variável dependente
   ex. se bom risco ocorre 2 em 5 casos, então a probabilidade a priori do risco ser bom será de 0.4 ou 40%, ou seja, se não soubermos mais nada acerca da aplicação de empréstimo, poderemos dizer que há uma probabilidade de 0.4 de que seja bom.
2. Calcular quão frequentemente cada variável independente ocorre em combinação com cada valor da variável dependente (saída). Desta forma é calculada a probabilidade condicional, que se irá combinar com a probabilidade "*a priori*" para fazer previsões.
3. Aplicar o teorema de Bayes, calculando assim as previsões para a variável dependente (utilizar a probabilidade condicional para modificar a probabilidade "*a priori*").

Exemplo com Naïve-Bayes

Voltando ao set relativo ao caso de empréstimo da instituição de crédito.

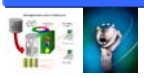
Nome	Débito	Rendim.	Casado?	Risco
Joaquim	Alto	Alto	Sim	Bom
Susana	Baixo	Alto	Sim	Bom
João	Baixo	Alto	Não	Pobre
Maria	Alto	Baixo	Sim	Pobre
Frederico	Baixo	Baixo	Sim	Pobre

0 - converter todas as variáveis para categóricas ou intervalos (já são)

1 - Calcular a probabilidade a priori

Risco Bom = 2 / 5 = 0.4

Risco Pobre = 3 / 5 = 0.6



Exemplo com Naïve-Bayes

2.a) calcular o resultado do risco (bom ou fraco) para cada valor das variáveis independentes

Nome	Débito	Rendim.	Casado?	Risco
Joaquim	Alto	Alto	Sim	Bom
Susana	Baixo	Alto	Sim	Bom
João	Baixo	Alto	Não	Pobre
Maria	Alto	Baixo	Sim	Pobre
Frederico	Baixo	Baixo	Sim	Pobre

ex. para a linha 3, relativo a rendimento alto, temos como resultado 2 riscos bons e 1 mau

Variável independente	Valor	Contagem Risco Bom	Contagem Risco Pobre
Débito	Alto	1	1
Débito	Baixo	1	2
Rendimento	Alto	2	1
Rendimento	Baixo	0	2
Casado	Sim	2	2
Casado	Não	0	1
Total p/ Risco		2	3

há três riscos pobres e dois riscos bons

Exemplo com Naïve-Bayes

2.b) calcular a probabilidade condicional utilizando essas contagens, dividindo-as pelo total p/ risco.

ex. para a linha 3, a probabilidade condicional é p/ bom risco de 2/2 = 1.0 e de 1/3 = 0.33 para risco pobre

Variável independente	Valor	Contagem Risco Bom	Contagem Risco Pobre	Probabilidade Bom Risco	Probabilidade Risco Pobre
Débito	Alto	1	1	0.5	0.33
Débito	Baixo	1	2	0.5	0.67
Rendimento	Alto	2	1	1.0	0.33
Rendimento	Baixo	0	2	0	0.67
Casado	Sim	2	2	1.0	0.67
Casado	Não	0	1	0	0.33
Total p/ Risco		2	3		

Exemplo com Naïve-Bayes

Neste ponto, temos as probabilidades das variáveis independentes e não da dependente, como pretenderíamos. Ou seja, aparentemente temos o inverso do que almejamos.

É aqui que entra a beleza do teorema de Bayes, permitindo combinar os efeitos das variáveis independentes na variável dependente.

Se bem se lembram, o teorema de Bayes estabelece que:

Probabilidade de A sobre B ocorrerem =

Probabilidade de A e B / Probalidade de B

o valor mais alto será o valor de predição

ou visto de outro modo,

Probabilidade de cada valor das variáveis =
Probabilidade condicional das variáveis independentes x Probabilidade à priori da variável dependente

Exemplo com Naïve-Bayes

3. Aplicando o teorema de Bayes à nossa tabela, teremos:

Nome	Débito	Rendim.	Casado?	Risco	Cálculo p/ Risco Bom	Cálculo p/ Risco Pobre
Joaquim	Alto	Alto	Sim	Bom	0.2	0.044
Susana	Baixo	Alto	Sim	Bom		
João	Baixo	Alto	Não	Pobre		
Maria	Alto	Baixo	Sim	Pobre		
Frederico	Baixo	Baixo	Sim	Pobre		

Tb. para o Joaquim,

P/ risco pobre:

0.33*0.33*0.67=0.072963

A probabilidade a priori é 0.6

Então 0.072963*0.6=0.0437778 = 0.044

P/ o Joaquim: débito alto, rendimento alto e casado=sim

P/ risco bom

na tabela anterior, as probabilidades condicionais associadas a esses valores é de 0.5, 1 e 1 (linha 1, 3 e 5) = 0.5

A probabilidade a priori = 0.4, então 0.4 * 0.5 = 0.2

Como a probabilidade de risco bom> probabilidade de risco mau

prediz-se risco bom

Exemplo com Naïve-Bayes

Para os restantes valores, ficamos com a tabela:

Nome	Débito	Rendim.	Casado?	Risco	Cálculo p/ Risco Bom	Cálculo p/ Risco Pobre	Risco Previsto
Joaquim	Alto	Alto	Sim	Bom	0.2	0.044	Bom
Susana	Baixo	Alto	Sim	Bom	0.077	0.089	Bom
João	Baixo	Alto	Não	Pobre	0	0.044	Pobre
Maria	Alto	Baixo	Sim	Pobre	0	0.096	Pobre
Frederico	Baixo	Baixo	Sim	Pobre	0	0.137	Pobre

Obs.:

A previsão relativa ao caso de treino é de 100% correcta.

Embora seja um bom sinal, deve ser validada com um conjunto de dados de teste.

Exemplo com Naïve-Bayes

Os valores calculados podem ser facilmente convertidos para probabilidades, dividindo cada valor pela soma total (de risco bom+pobre)

Ex. a probabilidade do Joaquim ter um bom risco é de 82%=0.2/0.244 (0.2+0.044 - soma de risco bom e risco pobre)

Note-se que não é necessário termos todos os valores das variáveis independentes para fazer uma predição. Poderemos não ter mesmo nenhum, empregando-se a predição a priori.

- **Se só soubermos a probabilidade condicional relacionada com o rendimento, poderemos utilizar essa para modificar a probabilidade a priori e fazer a predição.**

- **Vantagem: possibilidade de, com este algoritmo, fazer a predição a partir de informação parcial.**

Limitações do Naïve-Bayes

A aplicação do teorema de Bayes é valida, assumindo independência estatística entre as variáveis independentes, o que usualmente não acontece.

Desta suposição surge o adjectivo Naïve ao nome do algoritmo, presumivelmente porque o assumir dessa independência, muitas vezes, não é correcto.

Respondendo a esta limitação, poderemos capturar interacções entre as variáveis não independentes por forma a levar e conta a sua influência na predição efectuada.

Apesar disso, obtêm-se bons resultados práticos, mesmo quando a suposição de independência é falsa.

Extensão à Técnica Naïve-Bayes

A captura de interacções entre as variáveis não independentes responde à limitação atrás focada.

- **Captura-se a interacção entre pares, 3 a 3, 4 a 4 ..., de colunas não independentes.**
  - **as combinações de agrupamentos de colunas cria uma explosão combinatória: casos onde tivermos muitas colunas e com vários valores por coluna, teremos centenas de milhões de combinações.**

De qualquer forma,

- **nem todas as possíveis combinações ocorrerão;**
- **além disso, continuamos a ter apenas uma passagem pelos dados.**

Aqui, poderemos ter interacção humana, especificando que combinações de colunas considerar, diminuindo assim o número de pares eu serão contados.

Apreciação do Naïve-Bayes

A técnica de classificação Naïve-Bayes, na sua forma pura, revela-se uma excelente ferramenta exploratória:

- **Rápida, sendo a técnica mais eficiente, relativamente ao treino;**
- **Permite fazer predições, a partir de informação parcial;**
- **A informação descritiva que proporciona, pode constituir uma ajuda importante na compreensão (alguns utilizadores preferem a informação proporcionada com esta técnica, relativamente às árvores de decisão);**
- **A maior desvantagem é a necessidade de independência entre colunas, ainda que, na prática, isto não constitua, normalmente, um problema, utilizando a aludida extensão.**

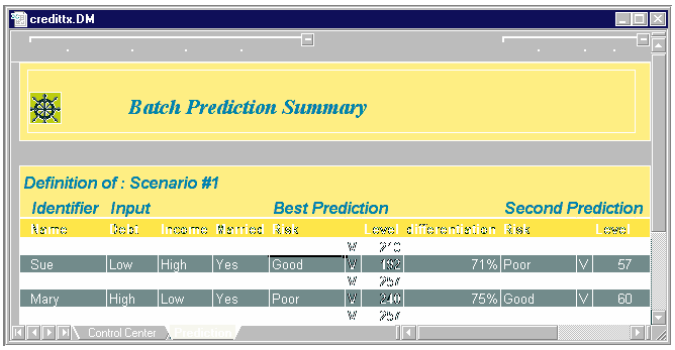
Depois de treinado, o modelo deve ser testado com um data set independente, para assegurar que os relacionamentos encontrados são geralmente aplicáveis antes de utilizar o modelo para fazer predição.

Comparação K-NN x Naïve-Bayes

Como predictor,

Naïve-Bayes:

- **usa valores de frequências para calcular probabilidades condicionais**
- **pode ser utilizado para predizer múltiplos casos**
- **pode incluir valores (chamados níveis) para cada classe possível, para avaliar da certeza da predição (exemplo, mostra-se na figura seguinte, relativamente ao produto Data Mine Builder da RedBrick)**



Comparação K-NN x Naïve-Bayes

Naïve-Bayes (continuação)

- **tem um aspecto descritivo que falta no K-NN; como calcula frequências, estas podem ser utilizados para saber quais os relacionamentos forte ou fracamente suportados.**

Neste gráfico vê-se que Risco pobre predomina para baixos rendimentos

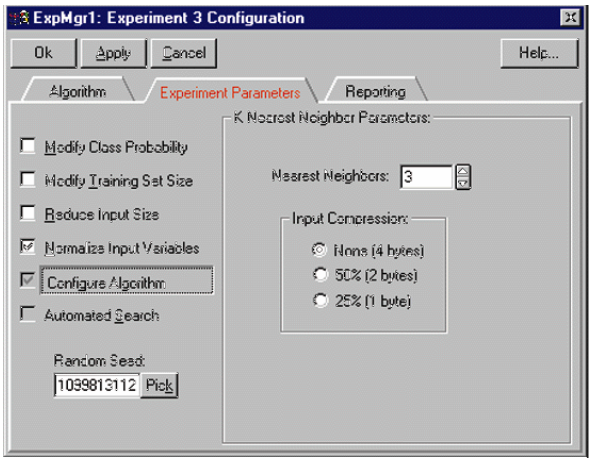
A torta de rendimentos altos (2ª torta da linha dos rendimentos) é 1/4 amarela.

Além da classificação, há uma visualização da força dessa previsão, através da mostra das probabilidades condicionais

Comparação K-NN x Naïve-Bayes

K-NN

- **utiliza os próprios dados para identificar casos semelhantes**
- **utilizável para pequeno número de casos, pois que obriga a nova passagem para efectuar nova predição**
- **dependente de uma medida de distância arbitrária (tornando a técnica subjectiva, restringida a aplicações que tenham medidas naturais de distância);**
- **o K pode influenciar bastante o resultado, requer vários testes, com set de teste para avaliação**



Exemplo com K-NN, onde é deixado o utilizador seleccionar o K (parâmetro para o número de casos mais próximos que serão utilizados para predizer o resultado relativo a um caso novo)