Mineração de Dados

# Matlab/WEKA

## Correlação

- load count.dat
- [c,lags] = xcorr(count(:,1),count(:,2),5);
- [d,lags] = xcorr(count(:,1),count(:,3),5);
- [e,lags] = xcorr(count(:,2),count(:,3),5);
- hold on;
- stem(-5:5,c);
- stem(-5:5,d);
- stem(-5:5,e);
- xlim([-10 10])
- Matlab: correlacao, exemplo2_correlacao

# Redução: Compressão de Dados: Análise de Fatores

- 120 alunos foram avaliadados com cinco provas,
  - 2 cobrem a matemática,
  - 2 literatura e,
  - 1 exame global.
- Parece razoável que as cinco graus para um determinado aluno deveriam estar relacionados. Alguns estudantes são bons em ambos os assuntos, algumas são bons em apenas um, etc.
- O objetivo desta análise é verificar se os cinco diferentes exames podem ser substituídos por apenas dois tipos exame de capacidade.

# Redução: Compressão de Dados: Análise de Fatores

- Matlab: exemplo1_AF1
- load examgrades
- [Loadings1,specVar1,T,stats] = factoran(grades,1)

| Loadings1 = | specVar1 = | stats.p = |
|---|---|---|
| 0.6021 | 0.6375 | |
| 0.6686 | 0.5530 | 0.0332 |
| 0.7704 | 0.4065 | |
| 0.7204 | 0.4810 | |
| 0.9153 | 0.1623 | |

Pode-se ver que um fator comum neste modelo coloca grande peso positivo em todos as cinco variáveis, mas mais peso no exame abrangente.

# Redução: Compressão de Dados: Análise de Fatores

- Variância próxima de 0 indicaria que a variável está inteiramente determinada pelos fatores comuns.
- O valor p da estrutura stats rejeita a hipótese nula de um único fator comum, assim pode-se reparar o modelo

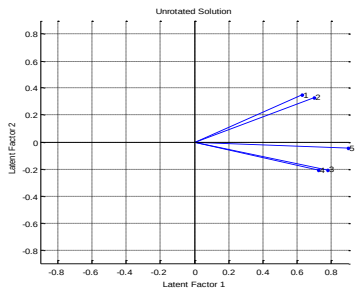
# Redução: Compressão de Dados: Análise de Fatores

- Matlab: exemplo1_AF2
- load examgrades
- [Loadings2,specVar2,T,stats] = factoran(grades,2,'rotate','none');

Loadings2 =

| | |
|---|---|
| 0.6289 | 0.3485 |
| 0.6992 | 0.3287 |
| 0.7785 | -0.2069 |
| 0.7246 | -0.2070 |
| 0.8963 | -0.0473 |

```
biplot(Loadings2, 'varlabels',num2str((1:5)'));
title('Unrotated Solution');
xlabel('Latent Factor 1'); ylabel('Latent Factor 2');
```



## Redução: Compressão de Dados: Análise de Fatores

- Matlab: exemplo1_AF2
- load examgrades
- [Loadings2,specVar2,T,stats] = factoran(grades,2,'rotate','none');

specVar2 =
0.4829
0.4031
0.3512
0.4321
0.1944

stats.p=
0.7061

Esse modelo indica uma variação pouco menor que devido a fatores comuns do que o modelo com um único fator.
Mais uma vez, a variância mínima corre para o quinto exame

## Redução: Compressão de Dados: Análise de Fatores

- Preço de ações de 10 companhias foi analisado sendo que elas foram classificadas como:
  - as 4 primeiras são classificadas como empresas de tecnologia;
  - as próximas 3 como financeiras e;
  - as 3 últimas como varejo;

# Redução: Compressão de Dados: Análise de Fatores

- Parece razoável que os preços das ações para as empresas que estão no mesmo setor deverão variar juntos quando as condições econômicas mudarem.
- Análise de fatores pode fornecer dados quantitativos mostrando que as empresas dentro de cada setor tenham variações semelhantes de preços das ações semana-a-semana .

# Redução: Compressão de Dados: Análise de Fatores

- Matlab: exemplo2_AF_1
- load stockreturns
- [Loadings,specificVar,T,stats] = factoran(stocks,3,'rotate','none');
- Loadings = (column corresponds to a common factor)

| | | |
|---|---|---|
| 0.8885 | 0.2367 | -0.2354 |
| 0.7126 | 0.3862 | 0.0034 |
| 0.3351 | 0.2784 | -0.0211 |
| 0.3088 | 0.1113 | -0.1905 |
| 0.6277 | -0.6643 | 0.1478 |
| 0.4726 | -0.6383 | 0.0133 |
| 0.1133 | -0.5416 | 0.0322 |
| 0.6403 | 0.1669 | 0.4960 |
| 0.2363 | 0.5293 | 0.5770 |
| 0.1105 | 0.1680 | 0.5524 |

## Redução: Compressão de Dados: Análise de Fatores

- specificVar = (1 would indicate that there is no common factor component in that variable)
  0.0991
  0.3431
  0.8097
  0.8559
  0.1429
  0.3691
  0.6928
  0.3162
  0.3311
  0.6544

- Variância próxima de 0 indica que a variável está inteiramente determinada pelos fatores comuns.

## Redução: Compressão de Dados: Análise de Fatores

- stats.p

  0.8144

- O valor p da estrutura stats não rejeita a hipótese nula de três fatores comuns, sugerindo que o modelo proporciona uma explicação satisfatória desses dados

## Redução: Compressão de Dados: Análise de Fatores

- Matlab: exemplo2_AF_2
- [Loadings2,specificVar2,T2,stats2] = factoran(stocks, 2,'rotate','none');
- stats2.p

  3.5610e-006

O p-valor para esta segunda análise, rejeita a hipótese de dois fatores, indicando que o modelo mais simples não é suficiente para explicar esses dados

## Redução: Compressão de Dados: Análise de Fatores

- Pode-se usar PCA quando se quer resumir ou aproximar os seus dados com menos dimensões
- Deveria-se utilizar AF quando necessitar de um modelo explicativo para as correlações entre os seus dados.

## Principal Component Analysis

- Considere uma amostra que utiliza nove diferentes índices de qualidade de vida em 329 cidades (USA).
  - clima, habitação, saúde, criminalidade, transporte, educação, artes, lazer e economia.
- Quando maior o indice, melhor.
- Por exemplo, um maior índice de criminalidade significa uma baixa taxa de criminalidade

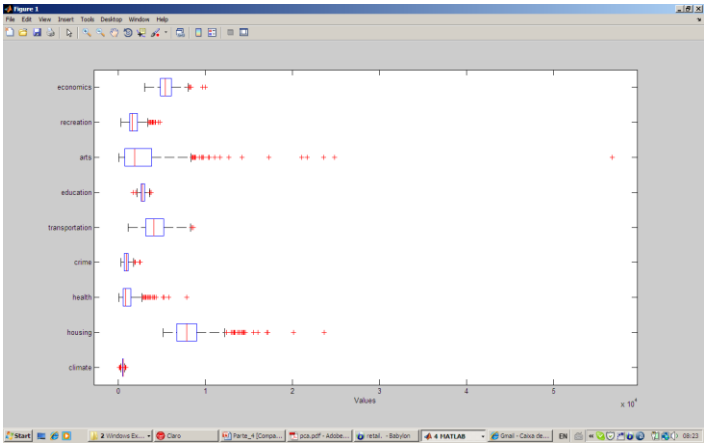
## Principal Component Analysis

- Matlab: exemplo1_PCA1
- load cities
- whos

| Name | Size | Bytes | Class | Attributes |
|---|---|---|---|---|
| categories | 9x14 | 252 | char | |
| names | 329x43 | 28294 | char | |
| ratings | 329x9 | 23688 | double | |

# PCA

- Cont. Matlab: exemplo1_PCA1
- boxplot(ratings,'orientation','horizontal','labels',categories)




# PCA

- Cont. Matlab: exemplo1_PCA1
- stdr = std(ratings);
- sr = ratings./repmat(stdr,329,1);
- [coefs,scores,variances,t2] = princomp(sr);

# PCA

- Cont. Matlab: exemplo1_PCA1
- c3 = coefs(:,1:3)

```
c3 =
    0.2064    0.2178   -0.6900
    0.3565    0.2506   -0.2082
    0.4602   -0.2995   -0.0073
    0.2813    0.3553    0.1851
    0.3512   -0.1796    0.1464
    0.2753   -0.4834    0.2297
    0.4631   -0.1948   -0.0265
    0.3279    0.3845   -0.0509
    0.1354    0.4713    0.6073
```

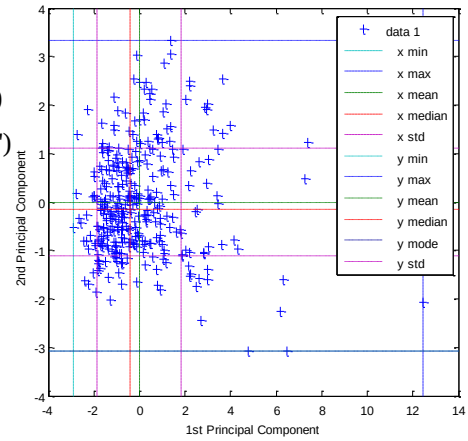
Os maiores coeficientes na primeira coluna (primeiro componente principal) são o terceiro e sétimo elementos correspondentes às variáveis, saúde e artes.

# PCA

- A segunda saída, scores, contém as coordenadas do dados originais no novo sistema de coordenadas definidas pelas principais componentes.
- Esta saída tem o mesmo tamanho que os dados de entrada da matriz.

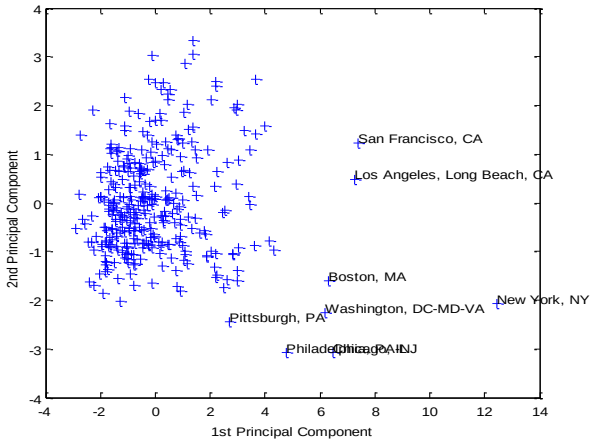
# PCA

- Cont. Matlab: exemplo1_PCA1
- plot(scores(:,1),scores(:,2),'+')
- xlabel('1st Principal Component')
- ylabel('2nd Principal Component')




# PCA

- gnames(name)

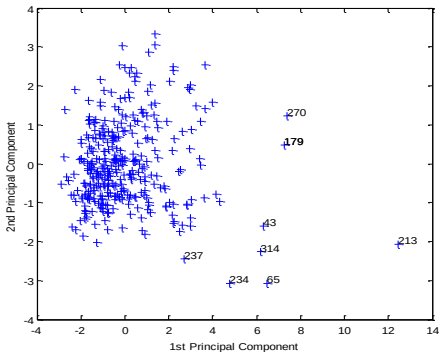


# PCA

- Estas cidades estão definitivamente diferente do resto dos dados, talvez devem ser considerados separadamente.
- Para eliminar estas cidades a partir dos dados, deve-se identificar estes registros
  - Matlab: exemplo1_PCA2
  - plot(scores(:,1),scores(:,2),'+')
  - xlabel('1st Principal Component');
  - ylabel('2nd Principal Component');
  - gname

# PCA

- Matlab: exemplo1_PCA2
- plot(scores(:,1),scores(:,2),'+')
- xlabel('1st Principal Component');
- ylabel('2nd Principal Component');
- gname



# PCA

- Matlab: exemplo1_PCA3
- big= [43 65 179 213 234 270 314];
- names(big,:)
- ans =

  Boston, MA
  Chicago, IL
  Los Angeles, Long Beach, CA
  New York, NY
  Philadelphia, PA-NJ
  San Francisco, CA
  Washington, DC-MD-VA

# PCA

- Matlab: exemplo1_PCA3
- rsubset = ratings;
- nsubset = names;
- nsubset(big,:) = [];
- rsubset(big,:) = [];
- size(rsubset)
- ans =
- 322 9

## PCA

- A terceira saída é um vetor contendo as variâncias correspondentes a cada elemento principal.
- Pode-se facilmente calcular a percentagem da variabilidade total explicada por cada componente principal.

## PCA

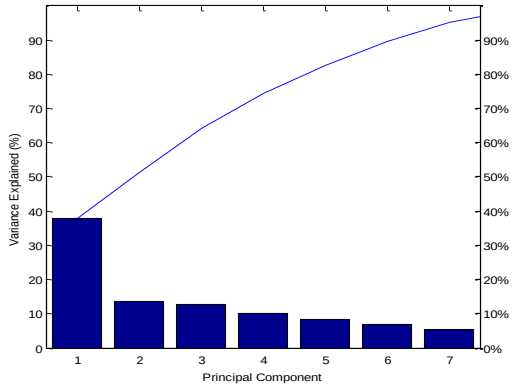
- Matlab: exemplo1_PCA4

- variances =
  3.4083
  1.2140
  1.1415
  0.9209
  0.7533
  0.6306
  0.4930
  0.3180
  0.1204

percent_explained = 100*variances/sum(variances)
percent_explained =
  37.8699
  13.4886
  12.6831
  10.2324
  8.3698
  7.0062
  5.4783
  3.5338
  1.3378

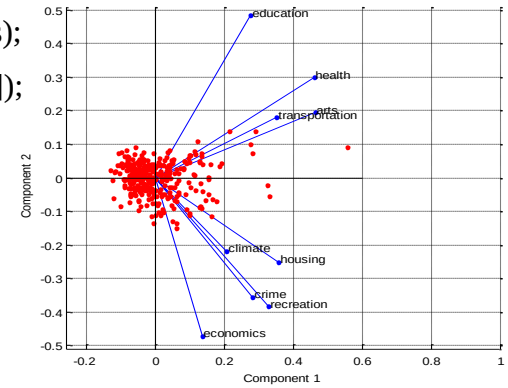
# PCA

- Matlab: exemplo1_PCA4
- pareto(percent_explained)
- xlabel('Principal Component')
- ylabel('Variance Explained (%)‘)




# PCA

- Matlab:
- biplot(coefs(:,1:2), 'scores',scores(:,1:2),...
- 'varlabels',categories);
- axis([-.26 1 -.51 .51]);



## PCA

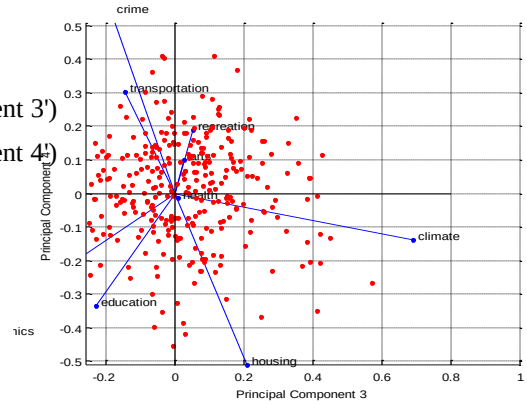
- As nove variáveis são representadas neste plano por um vetor, e o sentido e o comprimento do vetor indicam como cada variável contribui para os dois principais componentes.
- Pode-se constatar que o primeiro componente principal, representados nesta figura no eixo horizontal, tem coeficientes positivos para todos os nove variáveis.
- Também pode ser visto que o segundo principal componente, representada pelo eixo vertical, tem coeficientes positivos para a educação, saúde, das artes, transporte, e negativa para os 5 restantes.

## PCA

- Isso indica que este componente distingue entre as cidades que têm alto valor para o primeiro conjunto de variáveis e baixa para o segundo, e as cidades que têm o contrário

# PCA

- Matlab: exemplo1_PCA6
- biplot(coefs(:,3:4), 'scores',scores(:,3:4),...
- 'varlabels',categories);
- axis([-.26 1 -.51 .51]);
- xlabel('Principal Component 3')
- ylabel('Principal Component 4')




# LSE

- load V.txt
- calcula_LSE

## Software Weka

- Software para data mining/machine learning escrito em Java (distribuído sob GNU Public License)
- Utilizado em pesquisa e educação
- Principais características:
  - Extenso conjunto de rotinas para pré-processamento, esquemas de aprendizagem, além de métodos de avaliação
  - GUIs (inclusive para visualização dos dados)
  - Ambiente para comparação de algortimos de aprendizagem.

## Weka trabalha com flat files

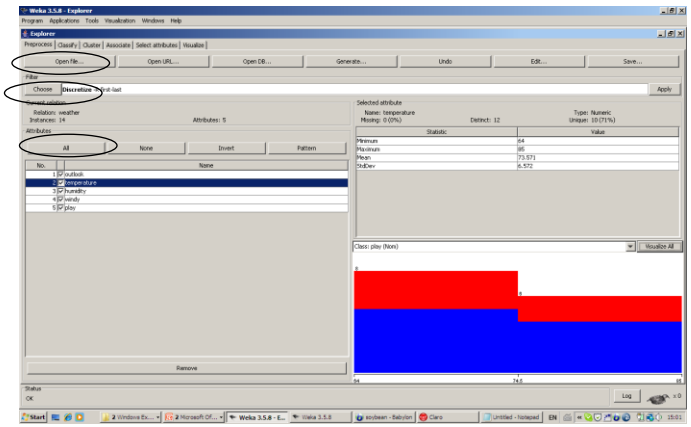
```
@relation heart-disease-simplified

@attribute age numeric
@attribute sex { female, male}
@attribute chest_pain_type { typ_angina, asympt, non_anginal,
    atyp_angina}
@attribute cholesterol numeric
@attribute exercise_induced_angina { no, yes}
@attribute class { present, not_present}

@data
63,male,typ_angina,233,no,not_present
67,male,asympt,286,yes,present
67,male,asympt,229,yes,present
38,female,non_anginal,234,no,not_present
```

Flat file in ARFF format

# Explorer: Pre-processing

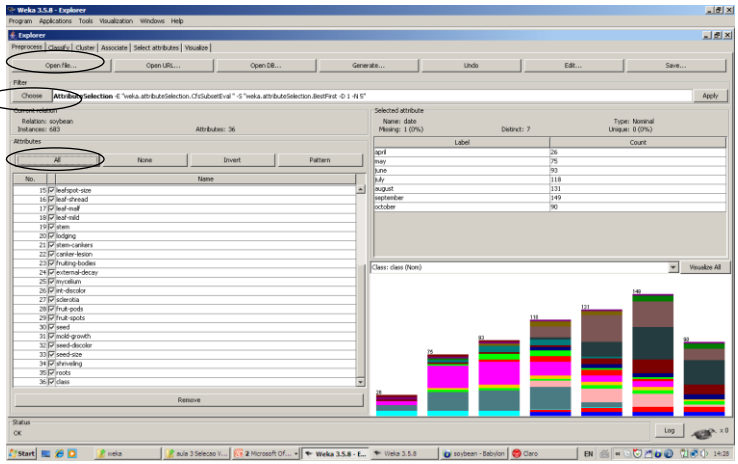
- Importação dos dados em vários formatos: ARFF, CSV, C4.5, binary
- Dados também podem ser lidos de uma URL ou de um banco de dados (utilizando o pacote JDBC)
- Rotinas de pré-processamento no Weka são chamados de filtros
- Weka tem filtros para:
  - Discretização, normalização, amostragem, seleção de atributos, transformação e combinação de atributos, entre outros.

# Weka



- weather.arff

# Weka



- OPEN soybean.arff

# Mineração de Dados

## Alguns Softwares

- **Weka: software de domínio público (Java), desenvolvido** pela Universidade de Waikato, contem uma série de algoritmos de DM. (www.cs.waikato.ac.nz/ml/weka)
- **Intelligent Miner: desenvolvido pela IBM, é uma** ferramenta de DM interligado diretamente com o banco de dados DB2 da IBM.
- **Oracle Data Miner: desenvolvido pela Oracle que** permite interligação direta com o banco de dados Oracle Enterprise 9i.
- **Enterprise Miner: desenvolvido para DM** tradicionalmente utilizado na área de negócios, marketing e inteligência competitiva
- **Statistica Data Miner: acrescenta as facilidades de** mineração de dados ao tradicional pacote utilizado em aplicações de estatística.