**I**nstituto **S**uperior de **E**ngenharia de **L**isboa

# Mestrado em Engenharia Informática e Computadores

# Semestre de Verão 2023/2024

**Laboratório 2**

Mineração de Dados em Larga Escala

**Trabalho realizado por:**

Pedro Carvalho nº47113 G01

Nuno Bartolomeu nº47233 G01

Índice

[Part II. Orange 1](#_Toc163463919)

[1) Orange environment and analysis of existing exemples 1](#_Toc163463920)

[2) Feature ranking and selection 2](#_Toc163463921)

[3) Feature reduction with principal component analysis and discretization 4](#_Toc163463922)

[Part III. R Studio 6](#_Toc163463923)

[1) Feature Selection 6](#_Toc163463924)

[2) Feature Reduction 9](#_Toc163463925)

[3) Feature Discretization 10](#_Toc163463926)

# Part II. Orange

## Orange environment and analysis of existing exemples

1. Run the Orange application and select the Examples (Example Workflows) option to see some examples and demos of the use of the software.
2. Run the File and Data Table example and check its key functionalities, using the (default) Iris dataset. Add a Feature Statistics Widget and analyze the four features in the dataset. Identify the list of supported file formats/types for datasets. Show a screen-shot of the statistical analysis.

O formato/tipo do conjunto de dados suportado é Basket, Comma-separated values, Microsoft Excel 97-2004 spreadsheet, Microsoft Excel spreadsheet, Pickled Orange data e Tab-separated values.

Uma imagem com captura de ecrã, texto, diagrama, Saturação de cores

Descrição gerada automaticamente

1. Run the Interactive Visualizations example. Add a Rank widget and report the Information Gain (IG) and FCBF relevance measures for all the features. Explain the purpose of this example. Identify the information provided by the Data Info widget. What are the most relevant features with the IG and FCBF criteria?

O propósito deste exemplo é ajudar na exploração de padrões ocultos, melhorar a intuição durante a análise de dados.

|  |  |  |
| --- | --- | --- |
|  | Information Gain | FCBF |
| petal length | 1.086 | 1.542 |
| petal width | 1.059 | 1.451 |
| sepal length | 0.624 | 0.000 |
| sepal width | 0.361 | 0.255 |

Tabela 1- Information Gain Ratio e FCBF

Segundo a tabela 1 o atributo mais relevante é aquele que apresenta maior número de Information Gain e FCBF, que segundo a tabela 1 é o *petal length*.

## Feature ranking and selection

1. Run the Feature Ranking example (also available on the Web, https://orangedatamining.com/ workflows/Feature-Selection) with the Iris dataset. On the Rank widget, try all the available scoring methods and look for the most relevant feature. What seems to be the most relevant feature?

O que parece ser o atributo mais relevante é o *petal length*.

1. On the Feature Ranking example with the Iris dataset, use the Scatter Plot widget to identify the most relevant feature. Show some screen-shots that of your analysis to find the most relevant feature and justify your answer.

Uma imagem com texto, captura de ecrã, software, diagrama

Descrição gerada automaticamente

Figura 1- Gráfico de dispersão. Eixo x- iris, Eixo y- petal length

Uma imagem com texto, captura de ecrã, software, Ícone de computador

Descrição gerada automaticamente

Figura 2- Gráfico de dispersão. Eixo x- iris, Eixo y- petal width

Uma imagem com texto, captura de ecrã, software

Descrição gerada automaticamente

Figura 3-Gráfico de dispersão. Eixo x- iris, Eixo y- sepal length

Uma imagem com texto, captura de ecrã, software, Ícone de computador

Descrição gerada automaticamente

Figura 4-Gráfico de dispersão. Eixo x- iris, Eixo y- sepal width

Observando as figuras 1, 2, 3 e 4, podemos chegar à conclusão que o atributo mais relevante é o *petal length* pois este é o que apresenta maior precisão e exatidão.

## Feature reduction with principal component analysis and discretization

1. Run the Principal Component Analysis example, https://orangedatamining.com/widget-catalog/ unsupervised/PCA, with the default Brown-Selected dataset. Explain the key actions of this demo and find an adequate number of reduced dimensions.

As ações chave desta demonstração são:

* Cálculo do PCA

O algoritmo PCA é aplicado ao conjunto de dados pré-processado para identificar os componentes principais. Esses componentes são autovetores que representam as direções de máxima variância nos dados.

* Redução de Dimensionalidade

Após calcular os componentes principais, a demonstração provavelmente realizará a redução de dimensionalidade selecionando um subconjunto desses componentes. O número de componentes selecionados determina a dimensionalidade reduzida do conjunto de dados.

* Visualização

Para entender o impacto da redução de dimensionalidade, a demonstração pode fornecer visualizações, como gráficos de dispersão ou gráficos de variância explicada. Essas visualizações ajudam na interpretação dos dados no espaço de dimensões reduzidas.

O número adequado para dimensões reduzidas é o menor número possível com a variância coberta o mais alto possível, logo, para uma variância de 90% tem-se uma dimensão reduzida de 25.

1. Modify the example to discretize the data with the EFB technique, in the reduced dimensionality space. Save the discretized data into a file. Show the Orange widget that performs these actions as well as the resulting file.

Uma imagem com logótipo, Tipo de letra, Gráficos, símbolo

Descrição gerada automaticamente

Figura 5- Widget usado para a discretização usando a técnica EFB

Uma imagem com texto, captura de ecrã, número, Tipo de letra

Descrição gerada automaticamente

Figura 6- Configuração do widget

Uma imagem com padrão, captura de ecrã, tecido, arte

Descrição gerada automaticamente

Figura 7- Ficheiro resultante

# Part III. R Studio

## Feature Selection

For both datasets:

1. Compute the (unsupervised) relevance of each feature, using variance and mean-median, as the relevance measures. Plot the sorted relevance values in decreasing order. Comment on the resulting plot. Compare on the smallest and the largest relevance value.

Após escrita do código pode-se observar os seguintes gráficos:

Uma imagem com texto, captura de ecrã, diagrama, Tipo de letra

Descrição gerada automaticamente

Figura 8- Relevância dos atributos de acordo com variância e média-mediana

Observando o primeiro gráfico da figura 8, apenas existe um atributo que se destaca, *insulin*, sendo esse o atributo mais relevante, enquanto os restantes têm valores muito baixos. No segundo atributo encontramos, dois atributos que se destacam enquanto os outros têm valores muito baixos e outros não têm valores.

Para descobrir o atributo mais relevante usando a variância e média-mediana, é necessário que o tipo do atributo seja numérico para poder fazer os cálculos, se ele for do tipo categórico não se consegue obter os valores necessários para saber a sua relevância, é o que acontece com o gráfico dos dados de Lisbon,não foi possível calcular as suas medidas(variância média-mediana) de certas classes para determinar a sua relevância.

1. For the relevance values found in (a), compute an adequate number of features, m, by the cumulative sorted relevance criterion, with three different thresholds. State the value of the considered thresholds as well as the corresponding values of m.

Os valores considerados para os *thresholds* foi 0.5, 0.75 e 0.9.

Os valores m correspondentes foram:

* Pima dataset

m0.5= 1

m0.75= 1

m0.9= 2

* Lisbon\_ 2023-01-01\_2023-01-31 dataset

m0.5= 1

m0.75= 2

m0.9= 2

1. Repeat (a) and (b) using the Fisher ratio as the relevance measure instead of the variance/mean-median relevance. Comment on the results.

Uma imagem com texto, captura de ecrã, diagrama, Tipo de letra

Descrição gerada automaticamente

Figura 9- Fisher Ratio for both datasets

Os valores considerados para os *thresholds* foi 0.5, 0.75 e 0.9.

Os valores m correspondentes foram:

* Pima dataset

m0.5= 2

m0.75= 3

m0.9= 5

* Lisbon\_ 2023-01-01\_2023-01-31 dataset

m0.5= NA

m0.75= NA

m0.9= NA

Observando os gráficos da figura 9 podemos concluir que a glucose apresenta uma grande relação com se o sujeito tem ou não têm diabetes. No segundo gráfico, é possível observar que não existe relação entre os atributos e classe severerisk, após observar o ficheiro .csv correspondente reparou-se que o valor da classe é sempre o mesmo assim conclui-se que não existe uma relação pois o valor é sempre o mesmo para diferentes valores dos atributos.

## Feature Reduction

For both datasets:

1. Compute the PCA decomposition. Plot the corresponding eigenvalues sorted in decreasing order. What would be an adequate number of reduced dimensions, m, for this dataset?

Uma imagem com texto, file, diagrama, captura de ecrã

Descrição gerada automaticamente

Figura 10- Eigenvalues for both datasets

O número de dimensões adequado para os dados será aquele que seja m < d sendo d o número de atributos de cada dataset sendo por isso 7 para o Pima dataset(8 atributos) e 18 para o Lisbon dataset(24 atributos)

1. Compute the SVD decomposition. Plot the corresponding singular values sorted in decreasing order. What would be an adequate number of reduced dimensions, m, for this dataset?

Uma imagem com texto, file, diagrama, Tipo de letra

Descrição gerada automaticamente

Figura 12- Singular Values Pima Dataset

O número de dimensões adequado para os dados “Pima” seria 7 e para o “Lisbon” é

1. Using the decomposition results of (a) and (b), compute the dimensionality-reduced versions of both datasets. Explain how you perform the dimensionality reduction. State the number of features of the reduced datasets.

Para realizar a redução da dimensionalidade decidiu-se selecionar os componentes da decomposição PCA que explicassem 90% da variação dos dados, para o dataset Pima reduziu para 7 e no Lisbon foi para 9. Para a decomposição SVD selecionou-se os valores que expliquem 90% da variação de dados

## Feature Discretization

For one dataset of your choice, compute a discretized version with one unsupervised technique and with one supervised technique, at your choice. State the chosen discretization technique as well as the number of discretization intervals for each feature. Comment on the results.

Usou-se o dataset “pimas” para realizar a discretização dos atributos.

Usou-se o método sem supervisão(unsupervised) *equal width discretization* para cada um dos atributos tivemos os seguintes intervalos:

* Age

"(20.9,33]" "(33,45]" "(45,57]" "(57,69]" "(69,81.1]"

* Glucose

"(-0.199,39.8]" "(39.8,79.6]" "(79.6,119]" "(119,159]" "(159,199]"

* Insulin

"(-0.846,169]" "(169,338]" "(338,508]" "(508,677]" "(677,847]"

* BMI

"(-0.0671,13.4]" "(13.4,26.8]" "(26.8,40.3]" "(40.3,53.7]" "(53.7,67.2]"

* DiabetesPedigreeFunction

"(0.0757,0.546]" "(0.546,1.01]" "(1.01,1.48]" "(1.48,1.95]" "(1.95,2.42]"

* Outcome

"(-0.001,0.2]" "(0.2,0.4]" "(0.4,0.6]" "(0.6,0.8]" "(0.8,1]"

* Pregnancies

"(-0.017,3.4]" "(3.4,6.8]" "(6.8,10.2]" "(10.2,13.6]" "(13.6,17]"

* BloodPressure

"(-0.122,24.4]" "(24.4,48.8]" "(48.8,73.2]" "(73.2,97.6]" "(97.6,122]"

* SkinThickness

"(-0.099,19.8]" "(19.8,39.6]" "(39.6,59.4]" "(59.4,79.2]" "(79.2,99.1]"

Como método de supervisão usou-se o método ChiMerge e tivemos os seguintes intervalos por atributos:

* Pregnancies

0.5 2.5 6.5

* Glucose

22.0 99.5 114.5 115.5 123.5 125.5 127.5 152.5 154.5 156.5 157.5 166.5

* BloodPressure

42.0 61.5 75.5

* SkinThickness

7.5 23.5 31.5

* Insulin

14.5 57.5 58.5 78.5 80.0 87.5 91.5 95.5 99.5 113.0 137.5 143.0

* BMI

9.10 22.80 22.95 23.25 23.55 26.35 27.65 27.85 28.45 28.85 30.25 30.60 32.25 32.35 32.85 33.00 33.25 36.65 36.95 37.15

37.85 38.05 40.85 41.25 41.65 42.65 42.85 43.45 43.55 44.55 45.35 45.95 48.10

* DiabetesPedigreeFunction

0.1265 0.1295 0.1655 0.1775 0.1845 0.1950 0.2385 0.2420 0.2715 0.2785 0.2955 0.2980 0.3010 0.3025 0.3135 0.3435 0.3465 0.3840 0.4015 0.4035 0.4465 0.4515 0.5365 0.5440 1.1075 1.1565 1.1780 1.2375 1.2750 1.3970 1.8370

* Age

24.5 30.5 42.5 54.5

Na discretização não supervisionada dividiu-se para cada atributo 5 intervalos em que cada intervalo tem aproximadamente a mesma largura. Na discretização supervisionada cada atributo têm o seu número de intervalos, pois estes são os que otimizam algum critério relacionado à tarefa de aprendizado supervisionado, como a maximização da separabilidade entre as classes. O que é possível observar no algoritmo chiMerge é os cut-points de cada variável, estes representam os limites dos intervalos resultantes da discretização.

A diferença entre os dois métodos da discretização (supervisionada ou não supervisionada) é que a não supervisionada é focada apenas nos valores de cada atributo enquanto a supervisionada considera os valores da classe.