Prova 2016, questões

Exploração de Dados (Universidade de Aveiro)

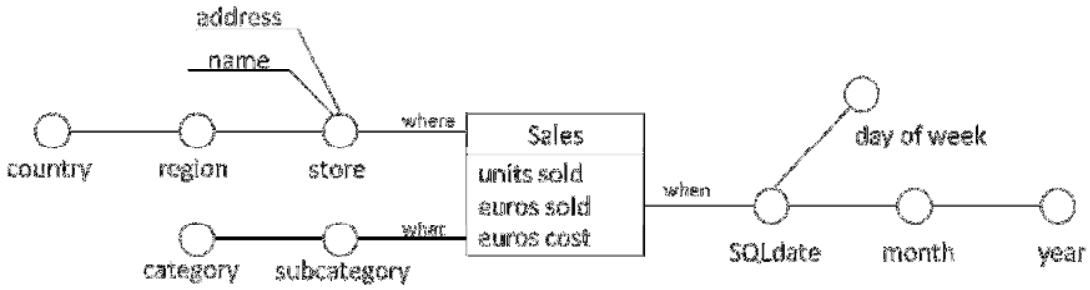
Departamento de Eletrónica, Telecomunicações e Informática

Exploração de dados & Data Mining

25 / 05 / 2016 : : 14h00 – 16h30

I *(folha 1)*

1. Consider the following multidimensional data model in DFM notation for a retail sales data mart:



   a. What is the implicit cardinality in the relationship between *name* and *address* ?

   b. What is the implicit cardinality in the relationship between *store* and *region* ?

   c. Write a MDX query to give the sum of units sold by store during January through April 2011.

   d. Write a MDX query to give the sum of profits (euros sold – euros cost) by store.

   e. Design a relational database model for data mart considering a star schema implementation. The design must include the names of the tables, the names of the attributes, and the primary key and the foreign keys of each table.

2. Comment the following sentence and give illustrative example to support your arguments: "The processing of a query using a range bitmap index may require a refinement step to filter out non-qualifying tuples"

3. An E-commerce company wants analyze the interactions (clickstream) between Web clients (browsers) and its Web server (website) to be able to determine which banners (advertisements) the visitors (costumers) should receive in future visits. The decision will be based on the visitor's interests or on previous interactions of the website visitors. For that purpose, the company wants to build a data mart to keep the full history of visitor's clickstream.

   To prevent the data mart from growing astronomically, the designers have chosen the grain to be one row for each completed customer session, and they want to be able to answer questions such as:

   - How many visitors consulted the product information before ordering?
   - How many customers looked at product information and never ordered?
   - How many customers began the ordering process but did not finish it? And where did they stop?

They have identified the following information requirements for each session:

- to know the duration in seconds and the number of pages visited;
- to know the session type (classified, unclassified, corrupted or inapplicable), the local content (such as requesting product information), the session context (such as ordering a product) and success status (whether the overall mission was achieved);
- to know which was the entry page: the page filename, the page source (static, dynamic, unknown), the page function (portal, search, product description or corporate information), the item type, the graphics type, the animation type and the sound type.
- to know how the visitor has arrived at the entry page (the website logs usually provide this information): the referral type (intranet, remote site, search engine or inapplicable), the referring URL and the referring website;
- to know the number of orders placed and the order amount in euros;
- to know the customer name, gender, marital status, education and country;
- to know the web server's date and time at the start of the session, including the SQL date and time, the weekday, the month, the year and the hour.

They also want analyze the Web profitability by keeping track of the activity costs and infrastructure cost to each sales transaction and so, they also want to know for each product sale:

- the quantity sold, the gross revenue, the gross profit and the following costs: manufacturing, storage, freight, special deal and other overhead;
- the product description, brand, package type, category and subcategory;
- the customer name, gender, marital status, education and country;
- the web server's date and time of sales transactions, including the SQL date, the weekday, the month, the year and the hour.

Design a multidimensional data model for this system using the DFM notation[1].

II *(folha 2)*

The following table shows the data for a two classes problem.

ID	atr1	atr2	label
1	0	2	A
2	1	2	A
3	-1	2	A
4	0	-3	B
5	1	0	B
6	-1	0	B

1. Plot the data and comment the following statement: 'the data is linearly separable'.
2. The following parameters correspond to the decision border (a line) computed by a linear SVM algorithm in the two dimensional space of the upper attributes. Represent the line graphically.

[1] This exercise was inspired on a case study presented in: Ralph Kimball and Margy Ross (2002). The Data Warehouse toolkit, the complete guide to dimensional modeling (Second edition). John Wiley and Sons, inc. ISBN: 0-471-20024-7.

Attribute	Parameters (weights)
atr1	0
atr2	-1.0
bias (w_0)	1.0

3. Find the parameters (w_0, w) of one of the many possible decision borders (lines) between the two classes. It has to be different from the solution at point 2.

4. Given x=[atr1=0.5; atr2= 0.7]. Indicate the class for this new example, according to the decision border of point 2 and point 3 respectively.

5. Given the following data set of the names of movies divided in two classes (new and old). What is the probability of a new movie with the name "Top" to belong the old or new class of movies ?

OLD	NEW
Top Gun	Top Gear
Shy People	Gun Shy
Top Hat	

III *(folha 2)*

1. For evaluating the performance of the classifier, the data sets must be divided into training and testing. Describe (briefly) the methodology of cross-validation.

2. The performance of the classifiers is presented as a matrix, called the 'confusion matrix'. Assuming that a problem has three classes:

a) What is the number of rows and columns of the confusion matrix for this case and what is the meaning of the entries in the matrix ?

b) Which are the measures displayed in the matrix?

3. What is the principal difference between classification and regression?

4. How to apply KNN (K nearest neighbour) to a data set with qualitative attributes?

5. Comment: "the KNN classifier needs the training set during the test phase while Naive Bayes do not need the training set during test phase".

6. Comment the following statement: 'The SVM classifier is a maximum margin classifier'.

Data Mining - Exame 25/05/2012

Parte I

1-

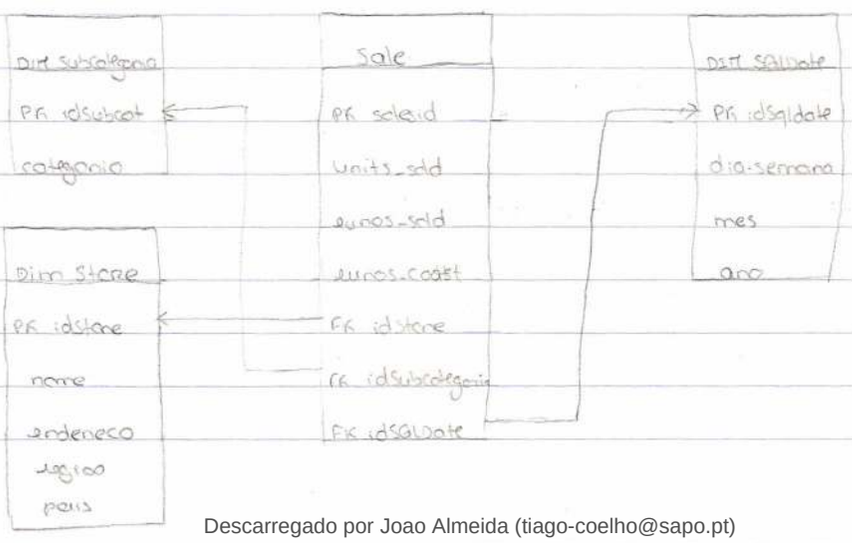
a) Um-para-um (1 .. 1)

b) nsgeo (1) — (*) store (1 .. *)

c)
```
Select {
    [Measures].[Units_Sold],
    [DIM Store].[Name].[Name]
    } ON COLUMNS
From Sales
where [DIM SalDate].[Year].&[2011] AND ([DIM SQLdate].[Month].&[1]:[DIM SQLdate].
[Month].&[4])
```

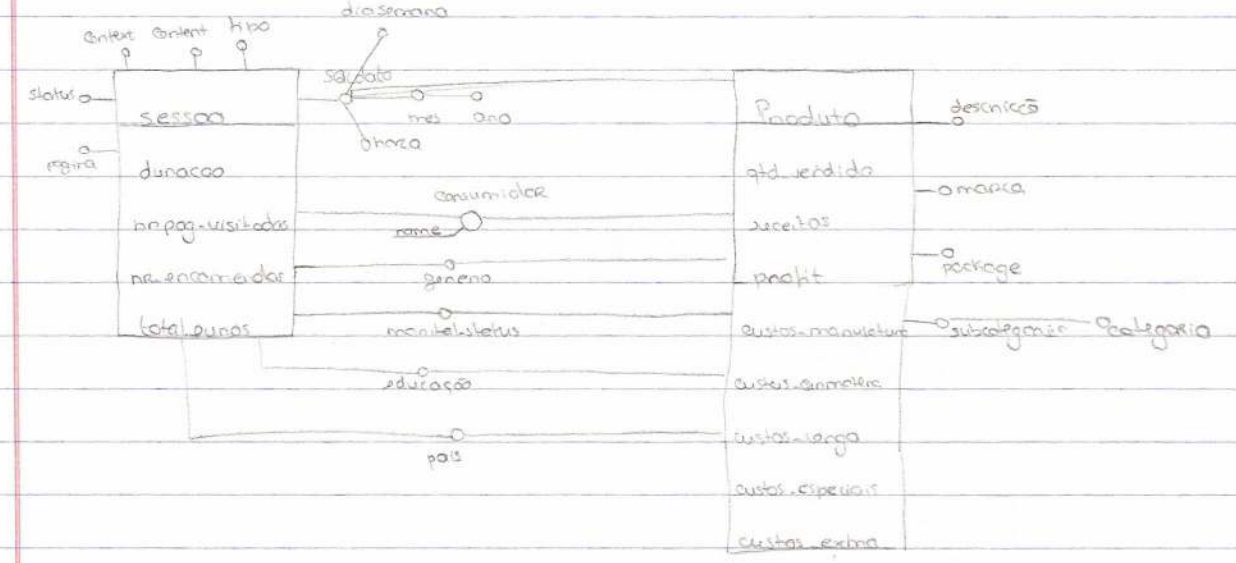
d)
```
With member
    Measure.[Profits] as [Measures].[Euros Sold] - [Measure].[Euros Cost]
Select {
    [Measure].[Profits],
    [DIM Store].[Name].[Name]
} on columns
From Sales
```

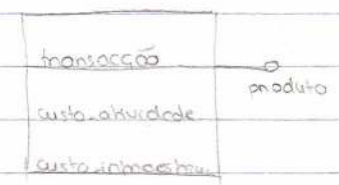
e)



2- As tabelas de factos são por norma muito grande e muito usadas nas queries. Para isto, podemos criar indices separados para cada chave da dimensão, criar indices de conjuntos de chaves que são muito utilizadas em conjunto ou criar indices para tabelas de dimensão. Existem 3 tipos de indices sendo que um deles é o indice bitmap baseado em intervalos. Aqui, os valores dos atributos são particionados num pequeno nº de valores. Se a distribuição dos valores for desconhecida, os intervalos podem ser preenchidos de forma irregular o que leva a tempos de acesso às queries desconhecidos. Assim, requer um passo de refinamento para filtrar registos não qualificados este passo leva a que os intervalos sejam mais uniformes para uma melhor distribuição.
(exemplo do intervalo de idades (slide OLAP_Queries pg. 12)

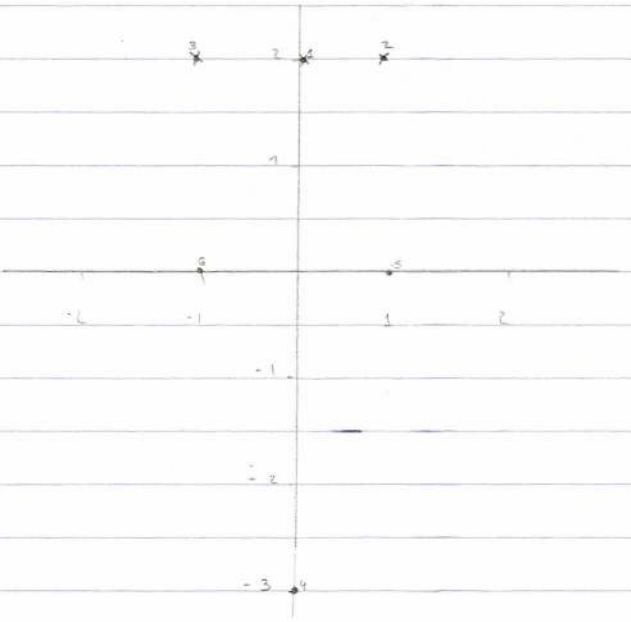
3.





Parte II

1.

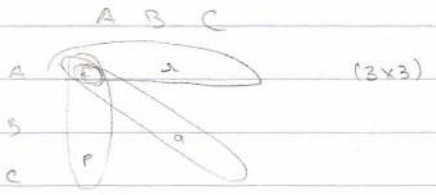


Sim, os dados são linearmente separáveis pois é possível traçar uma reta que separe os dois conjuntos. (p.e: $y=1$)

Parte III

1. Dado um conjunto de dados (conjunto de treino) cada registo tem:
   - conjunto de atributos(x)
   - nome da classe (label)
   - na fase de treino tentamos encontrar o modelo $f(w,x)$
   - dado um conjunto de teste: estudar a performance do classificador
   - O cross validation permite avaliar a performance, em que divide o conjunto de dados em K subconjuntos disjuntos. Conjunto de dados de tamanho (N) dividido em K partições em que treina com (K-1) e testa com 1 partição

2.

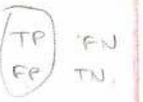


(3x3)

a) Nº de linha: 3 (3x3)

Nº de colunas: 3

O significado dos enlaces na matrix é a intercepção das classes que existem eventualmente



com as classes previstas pelo classificador

b) As medidas que podemos ver através da matrix é: precisão, recall, accuracy
(vertical) (horizontal) (diagonal)

3.