WEB ANALYTICS

ANALISE OS SEUS DADOS PARA MELHORAR AS DECISÕES EMPRESARIAIS

João Mendes Moreira
jmoreira@fe.up.pt

E-AGRO INNOVATION
E-AGRO MARKETS
Cofinanciado por:
COMPETE 2020
PORTUGAL 2020
UNIÃO EUROPEIA
Fundo Europeu de Desenvolvimento Regional

METODOLGIA PARA PROJETOS DE ANÁLISE DE DADOS

INSERIR SUBTÍTULO

E-AGRO INNOVATION

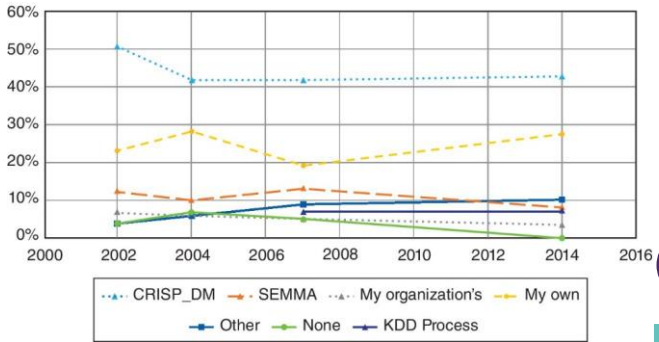
E-AGRO MARKETS

Cofinanciado por:
COMPETE 2020
PORTUGAL 2020
UNIÃO EUROPEIA
Fundo Europeu de Desenvolvimento Regional

Metodologia para projetos de análise de dados

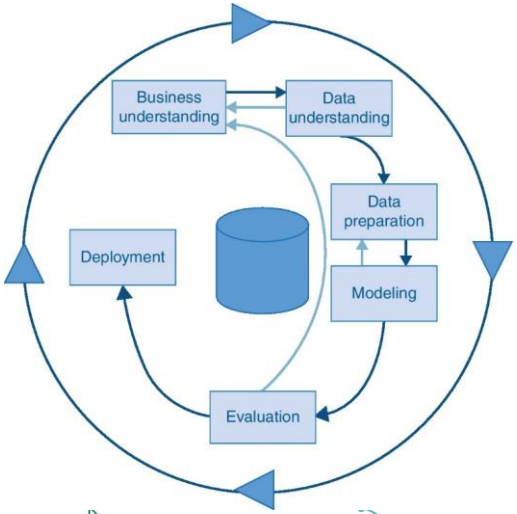
- The KDD process: metodologia com 9 passos
- The CRoss-Industry Standard Process for Data Mining (CRISP-DM): metodologia com 6 fases
- SEMMA - Sample, Explore, Modify, Model e Assess

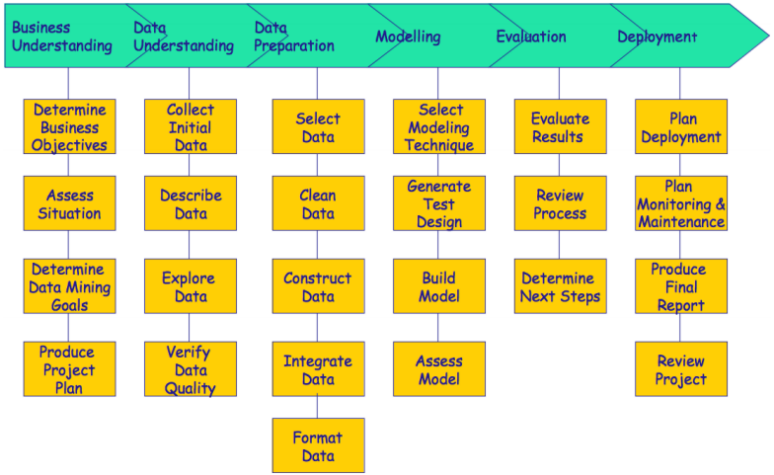
- Inquéritos realizados em 2002, 2004, 2007 e 2014 por kdnuggets sobre a utilização de metodologias de planeamento e desenvolvimento de projetos de análise de dados



Metodologia para projetos de análise de dados

- A metodologia CRISP-DM





AGROCLUSTER RIBATEJO PORTUGAL
E-AGRO INNOVATION
E-AGRO MARKETS

4

Projeto descritivo: 1 – Compreensão do negócio

- Vamos assumir que o hospital de Wisconsin, EUA, quer desenvolver um Sistema de apoio à decisão para ajudar no diagnóstico de um tipo de cancro (cancro da mama)
- Compreender os diferentes padrões que podem existir em tecidos mamários é um primeiro bjetivo
- Este objetivo remete-nos para técnicas de aglomeração de dados / clustering

Projeto descritivo : 2 – Compreensão dos dados

Attr	Atributo	Domínio	Tipo de dados	Em falta	Min/menor	Max/maior	Valores
1.	Sample code number	Id number	Polynomial	0	97519 (1)	1182404 (6)	1182404 (6), 127609 (5), … [643 more]
2.	Clump thickness	1 – 10	Integer	0	1	10	4.418
3.	Uniformity of cell size	1 – 10	Integer	0	1	10	3.134
4.	Uniformity of cell shape	1 – 10	Integer	0	1	10	3.207
5.	Marginal adhesion	1 – 10	Integer	0	1	10	2.807
6.	Single epithelial cell size	1 – 10	Integer	0	1	10	3.216
7.	Bare nuclei	1 – 10	Integer	16	1	10	3.545
8.	Bland chromatin	1 – 10	Integer	0	1	10	3.438
9.	Normal nucleoli	1 – 10	Integer	0	1	10	2.867
10.	Mitoses	1 - 10	Integer	0	1	10	1.589
11.	class	2 for benign, 4 for malignant	binomial	0	4 (241)	2 (458)	2 (458), 4 (241)

AGROCLUSTER RIBATEJO PORTUGAL
E-AGRO INNOVATION
E-AGRO MARKETS

Projeto descritivo : 3 – Preparação dos dados

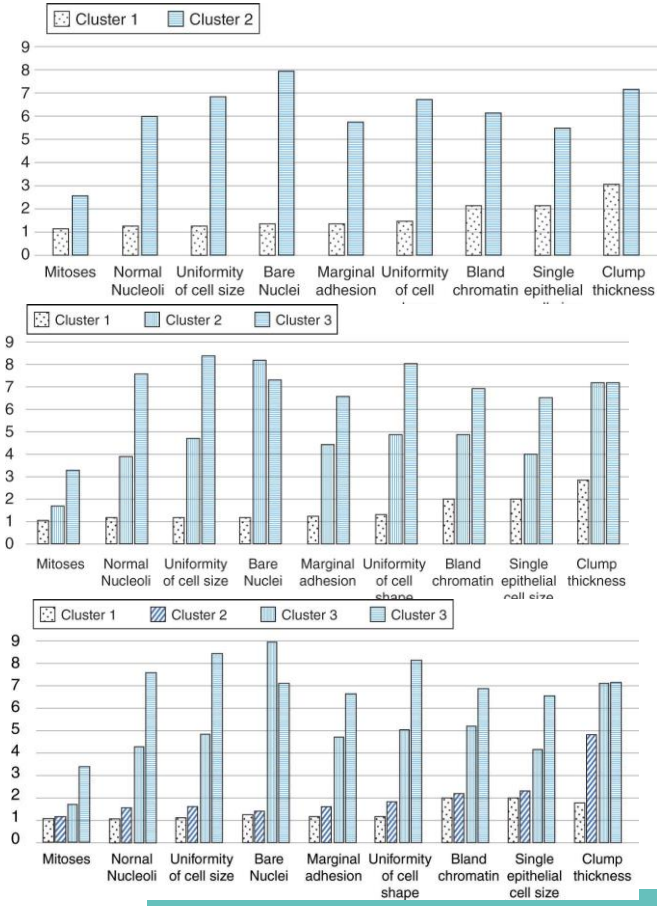
- Problemas com a qualidade dos dados:
  - Bare Nuclei tem 16 valores em falta
  - É o único problema de qualidade de dados deste conjunto de dados
  - Já que 16 instâncias de entre 699 não é muito significativo, opta-se por remover as instâncias com valores em falta
- Quais os atributos úteis para identificar padrões na massa mamária?
  - O código da imagem é irrelevante
  - Como só queremos identificar padões, a variável objetivo (ser ou não um tecido canceroso) não é relevante
    - Só queremos identificar padrões nas imagens
- Normalização: é necessária?
  - Neste caso não é necessária
    - Todos os atributos estão na mesma escala: 1 – 10

AGROCLUSTER RIBATEJO PORTUGAL E-AGRO INNOVATION E-AGRO MARKETS

7

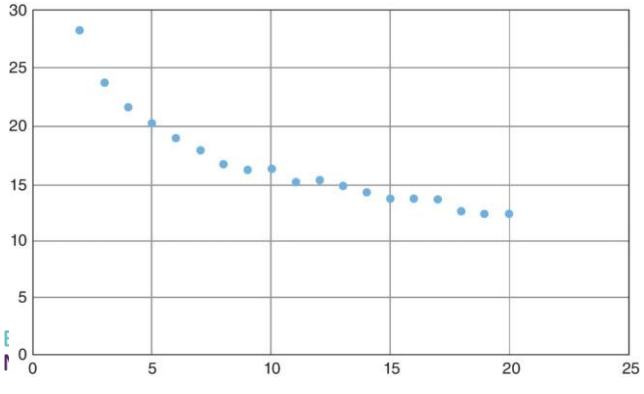
Projeto descritivo : 4 – Modelação

- Vamos utilizar o algoritmo de clustering mais popular, o k-medias«s
- Testam-se diferentes valores de k, é possível observer que entre $k = 2$ e $k = 4$, um dos clusters é razoavelmente estável e corresponde a tecidos benignos



Projeto descritivo : 4 – Modelação

- O correto número de clusters pode ser encontrado com o seguinte gráfico, apesar de neste caso, o cotovelo da curva não ser muito evidente
  - De qq forma, $k = 4$ foi o valor escolhido



Projeto descritivo : 5 – Avaliação

- Para o Hospital, a identificação dos quarto padrões de tecidos pode ser relevante em termos de análise

Projeto descritivo : 6 – Aplicação

- Esta fase não é, geralmente, feita por analistas de dados
- Como disponibilizar os métodos atrás vistos no dia a dia do hospital? É esse o objetivo
- Os resultados deste métodos podem ser usados na prática clínica?

Projeto preditivo: 1 – Compreensão do negócio

- Investidores, bancos e muitas outras instituições e acionistas têm interesse em prever o quão viável é uma empresa.
- O objetivo do negócio é o de prever se uma determinada empresa ficará insolvente nos próximos cinco anos.
- Este problema remete para técnicas preditivas de classificação binária

Projeto: 2a – Compreensão dos dados

Attribute	Description
X1	net profit / total assets
X2	total liabilities / total assets
X3	working capital / total assets
X4	current assets / short-term liabilities
X5	[(cash + short-term securities + receivables - short-term liabilities) / (operating expenses - depreciation)] * 365
X6	retained earnings / total assets
X7	EBIT / total assets
X8	book value of equity / total liabilities
X9	sales / total assets
X10	equity / total assets
X11	(gross profit + extraordinary items + financial expenses) / total assets
X12	gross profit / short-term liabilities
X13	(gross profit + depreciation) / sales
X14	(gross profit + interest) / total assets
X15	(total liabilities * 365) / (gross profit + depreciation)
X16	(gross profit + depreciation) / total liabilities

Atributos preditivos

Projeto: 2b – Compreensão dos dados

Attribute	Description
X17	total assets / total liabilities
X18	gross profit / total assets
X19	gross profit / sales
X20	(inventory * 365) / sales
X21	sales (n) / sales (n-1)
X22	profit on operating activities / total assets
X23	net profit / sales
X24	gross profit (in 3 years) / total assets
X25	(equity - share capital) / total assets
X26	(net profit + depreciation) / total liabilities
X27	profit on operating activities / financial expenses
X28	working capital / fixed assets
X29	logarithm of total assets
X30	(total liabilities - cash) / sales
X31	(gross profit + interest) / sales

Atributos preditivos

Projeto: 2c – Compreensão dos dados

Attribute	Description
X33	operating expenses / short-term liabilities
X34	operating expenses / total liabilities
X35	profit on sales / total assets
X36	total sales / total assets
X37	(current assets - inventories) / long-term liabilities
X38	constant capital / total assets
X39	profit on sales / sales
X40	(current assets - inventory - receivables) / short-term liabilities
X41	total liabilities / ((profit on operating activities + depreciation) * (12/365))
X42	profit on operating activities / sales
X43	rotation receivables + inventory turnover in days
X44	(receivables * 365) / sales
X45	net profit / inventory
X46	(current assets - inventory) / short-term liabilities
X47	(inventory * 365) / cost of products sold

Atributos preditivos

Projeto: 2d – Compreensão dos dados

Attribute	Description
X49	EBITDA (profit on operating activities - depreciation) / sales
X50	current assets / total liabilities
X51	short-term liabilities / total assets
X52	(short-term liabilities * 365) / cost of products sold)
X53	equity / fixed assets
X54	constant capital / fixed assets
X55	working capital
X56	(sales - cost of products sold) / sales
X57	(current assets - inventory - short-term liabilities) / (sales - gross profit - depreciation)
X58	total costs /total sales
X59	long-term liabilities / equity
X60	sales / inventory
X61	sales / receivables
X62	(short-term liabilities *365) / sales
X63	sales / short-term liabilities

Atributos preditivos

Projeto: 2e – Compreensão dos dados

Attr	Type	Missing	Min/Least	Max/Most	Average/Values
X1	Real	3	-189.560	453.770	0.314
X2	Real	3	-141.410	1452.200	2.624
X3	Real	3	0	3876.100	5.553
X4	Real	30	-440.550	1099.500	1.826
X5	Real	8	-189.450	453.780	0.354
X6	Real	3	-23.207	331.460	0.800
X7	Real	3	-607.402	13315	2.093
X8	Real	25	-141.410	1452.200	2.624
X9	Real	1	0	3876.100	5.553
X10	Real	3	-440.550	1099.500	1.826
X11	Real	39	-189.450	453.780	0.354
X12	Real	30	-23.207	331.460	0.800
X13	Real	0	-607.420	13315	2.093
X14	Real	3	-189.560	453.770	0.314
X15	Real	2	-5611900	3599100	1802.696
X16	Real	25	-42.322	405.330	0.871

Estatísticas dos atributos

Projeto: 2f – Compreensão dos dados

Attr	Type	Missing	Min/Least	Max/Most	Average/Values
X17	Real	25	-0.413	1529.900	3.752
X18	Real	3	-189.560	453.770	0.314
X19	Real	0	-622.060	2156.800	0.562
X20	Real	0	0	7809200	1162.128
X21	Real	1622	-1325	27900	10.368
X22	Real	3	-216.800	454.640	0.288
X23	Real	0	-634.590	2156.800	0.424
X24	Real	124	-189.560	831.660	0.540
X25	Real	3	-459.560	1353.300	1.264
X26	Real	25	-21.793	612.880	0.831
X27	Real	311	-14790	2040800	1321.989
X28	Real	34	-490.090	1570	2.703
X29	Real	3	0.176	9.386	4.195
X30	Real	43	-149.070	152860	23.705
X31	Real	297	-622	2156.800	0.500
X32	Real	636	0	351630	237.064

Estatísticas dos atributos

Projeto: 2g – Compreensão dos dados

Attr	Type	Missing	Min/Least	Max/Most	Average/Values
X33	Real	763	0	884.200	7.473
X34	Real	818	-280.260	884.200	3.931
X35	Real	830	-169.470	445.470	0.356
X36	Real	841	0.000	3876.000	6.447
X37	Real	3265	-525.520	398920	190.201
X38	Real	841	-20.340	1099.500	2.188
X39	Real	839	-14.335	2156.500	0.440
X40	Real	869	-101.270	1014.600	0.883
X41	Real	914	-11.976	813.140	0.664
X42	Real	840	-35.214	2156.800	0.435
X43	Real	840	0	30393000	5034.468
X44	Real	840	0	22584000	3728.024
X45	Real	948	-0.599	5986.800	8.252
X46	Real	869	-101.260	1017.800	1.960
X47	Real	869	0	47794	80.008
X48	Real	843	-218.420	405.590	0.186

Estatísticas dos atributos

Projeto: 2h – Compreensão dos dados

Attr	Type	Missing	Min/Least	Max/Most	Average/Values
X49	Real	842	-9001	31.639	-1.408
X50	Real	865	0	261.500	2.161
X51	Real	846	0	21.261	0.385
X52	Real	872	0	354.360	0.462
X53	Real	875	-130.470	180440	107.276
X54	Real	875	-82.303	180440	108.245
X55	Real	844	-589300	4398400	10496.129
X56	Real	844	-1108300	1	-179.139
X57	Real	845	-15.813	71.053	0.291
X58	Real	844	-0.004	1108300	180.140
X59	Real	845	-256.990	119.580	0.290
X60	Real	953	0.000	361820	132.159
X61	Real	862	0.000	21110	16.433
X62	Real	844	0	25016000	4164.117
X63	Real	869	0.000	1042.200	8.635
X64	Real	875	0.000	294770	218.049
CLASS	Binomial	844	1 (194)	0 (5989)	0 (5989), 1 (194)

Estatísticas dos atributos

Projeto: 3 – Preparação dos dados

- Problemas identificados:
  - Valores em falta
    - 844 instâncias sem valor na variável objetivo e com valores em falta em quase todos os atributos
  - Atributos redundantes
    - 24 pares de atributos com valor de correlação absoluta maior que 0,8: 22 correlacionados positivamente e 2 negativamente.
  - Valores com ruído / valores extremos
    - Quase todos os atributos têm valores extremos / com ruído

Projeto: 3 – Preparação dos dados

- Os seguintes passos foram realizados:
  - 844 instâncias sem valor na variável objetivo foram removidas
    - Todas da classe maioritária (sem bancarrota)
  - Todos os atributos preditivos foram normalizados coma a norma z
  - Remoção dos valores extremos,
    - 167 (17 da classe 1 e 150 da classe 0) valores normalzados com valor absolute maior do que 5
  - Todos os valoes em falta foram preenchidos com a media dos valores das instâncias da mesma classe

Projeto: 4 – Modelação

- Utilizaram-se 3 algoritmos:
  - K-NN (k=15, sem afinação de parâmetros e com a distância Euclideana)
    - Técnica Hold-out: 70% para treino do modelo e 30% para avaliação da técnica de seleção de atributos *wrapper*
      - Os atributos selecionados foram X6, X11, X24, X27 e X60
  - C4.5 (25% de confiança para poda e no mínimo 2 instâncias na folha)
  - Florestas aleatórias (500 árvores)
- Todas as 3 técnicas utilizaram as instâncias da partição de 70% para validação cruzada com 10 partições

Projeto: 5 – Avaliação

K-NN matriz de confusão
5 variáveis predtivas

	True 0	True 1	Class Precision
Predicted 0	5812	106	98.21%
Predicted 1	14	84	85.71%
Class Recall	99.76%	44.21%	

C4.5 matriz de confusão
5 variáveis preditivas

	True 0	True 1	Class Precision
Predicted 0	5799	100	98.30%
Predicted 1	27	90	76.92%
Class Recall	99.54%	47.37%	

Florestas aleatórias matriz de confusão
Todos os atributos peditivos

	True 0	True 1	Class Precision
Predicted 0	5794	90	98.47%
Predicted 1	32	100	75.76%
Class Recall	99.45%	52.63%	

Projeto: 6 – Aplicação

- Este tipo de resultados podiam ser disponibilizados de várias formas
  - Por exemplo, numa página web
- Se fosse esse o caso, a página web:
  - A fase de aplicação deveria envolver a construção do site web usando os resultados preditivos obtidos pelos analistas de dados

AGROCLUSTER
RIBATEJO

Web mining

INSERIR SUBTÍTULO

E-AGRO INNOVATION

E-AGRO MARKETS

Cofinanciado por:
COMPETE 2020
PORTUGAL 2020
UNIÃO EUROPEIA
Fundo Europeu de Desenvolvimento Regional

Summary

- Text Mining
  - Feature extraction
- Recommender Systems
  - Knowledge based
  - Content based
  - Model based
- Social Networks

Working with Texts

Text mining tasks

- Descriptive
  - Group similar documents
  - Look for texts about similar issues and words that frequently appear together
- Predictive
  - Classification of documents into one or more topics
  - Sentiment analysis and opinion mining
- Similar to data mining tasks
  - After transformation of texts into tabular, attribute-value format

Text mining tasks

- Main phases
  - Data acquisition
    - Convert the text into a sequence of characters
    - Cleaning of unnecessary information
  - Feature extraction
    - Extract relevant features from the original data
  - Data preprocessing
    - Improve data quality
  - Model induction
    - Standard machine learning
  - Evaluation and interpretation of the results

Feature extraction

- Main steps to extract features from text
  - Tokenization
  - Stemming
  - Lemmatization
  - Removal of stop words
  - Conversion to structured data

Feature extraction

Message received	Class
I like my sister's birthday party	Family
I liked the company party	Work
I am not bringing them from school	Family
I will talk and bring the contract	Work
I talked to other companies	Work
My wife is having contractions	Family

- Tokenization
  - Extract, for each text, a sequence of words from the stream of characters
    - Look at white spaces and punctuation characters
    - Each word in the sequence is called a lexical token
- if a word appears more than once in the text:
  - Its token will appear more than once in the sequence of tokens
  - bag-of-words

Stemming

Original words	Stems
studied, studying, student, studies, study	studi, studi, student, studi, studi
miner, mining, mine	miner, mine, mine
vegetable, vegetarian, vegetate	veget, vegetarian, veget, veget
eating, ate, eats, eater	eat, ate, eat, eater

- Look for a common base form able to represent many token variations
- Avoid having a very large number of tokens
  - Can result in a very sparse data set
- Convert each token to its stem
  - Stemming algorithms
    - Porter stemmer is the most common
    - Stem of “studied”, “studying”, “student”, “studies”, “study”: **studi**
    - Stem of “student”: **student**

Stemming

Message received after stemming	Class
I, like, my, sister, birthday, parti	Family
I, like, the, compani, parti	Work
I, am, not, bring, them, from, school	Family
I, will, talk, and, bring, the, contract	Work
I, talk, to, other, compani	Work
My, wife, is, have, contract	Family

- Look for a common base form able to represent many token variations
- Avoid having a very large number of tokens
  - Can result in a very sparse data set
- Convert each token to its stem
  - Stemming algorithms
    - Porter stemmer is the most common
    - Stem of “studied”, “studying”, “student”, “studies”, “study”: **studi**
    - Stem of “student”: **student**

AGROCLUSTER RIBATEJO PORTUGAL
E-AGRO INNOVATION
E-AGRO MARKETS

34

Lemmatization

Message received	Class
i like my sister birthday party	Family
i like the company party	Work
i be not bring them from school	Family
i will talk and bring the contract	Work
i talk to other company	Work
my wife be have contraction	Family

- More sophisticated variation of stemming
- Uses a vocabulary and takes grammatical aspects of language into account
  - Performing a morphological analysis
- Returns the dictionary form of a word, called lemma

Removal of stop words

Stems after removal of stop words	Class
sister, birthday, parti	Family
compani, parti	Work
bring, school	Family
talk, bring, contract	Work
talk, compani	Work
wife, contract	Family

- Further reduce the number of stems by removing stop words
  - Adjectives (good, bad, large...)
  - Adverbs (fast, nicely, not...)
  - Articles (a, an, the)
  - Negations (none, not, never...)
  - Pronouns (I, he, my, his, yours, ours …)
  - Prepositions (at, by, for, from, in, to...)
  - Conjunctions (and, but, or, with. . .)
  - Frequent verbs (are, be, is, was, has...)
  - Qualifiers (a little, less, more, very, yet...)

Removal of stop words

Stems after removal of stop words	Class
sister, birthday, parti	Family
compani, parti	Work
bring, school	Family
talk, bring, contract	Work
talk, compani	Work
wife, contract	Family

- Decision of which stop words to remove depends on the application
  - E.g. presence of adjectives and negations is important when mining opinions
  - Number of stems removed from 23 to 9
  - Reduce the sparsity of the resulting table

38

Conversion to structured data

- Creates a table with binary (presence of a stem in the text) or quantitative (frequency of a stem in the text) values

birthday	bring	compani	contract	parti	school	sister	talk	wife	Class
1	0	0	0	1	0	1	0	0	Family
0	0	1	0	1	0	0	0	0	Work
0	1	0	0	0	1	0	0	0	Family
0	1	0	1	0	0	0	1	0	Work
0	0	1	0	0	0	0	1	0	Work
0	0	0	1	0	0	0	0	1	Family

Recommender systems

Recommendation task

- Explicit feedback
  - The user is asked by the system to directly express preference on items
    - E.g.: rating them on a scale or ranking them
- Implicit feedback
  - Information about a user's preference is obtained by watching the user interaction with the system
    - E.g.: recording which items were viewed, listened to, scrolled past, bookmarked, saved, purchased, linked to, copied, …

Recommendation task

- Given a matrix with
  - Set of users
  - Set of items
  - Recorded feedbacks
- Returns
  - Predictive model
- Depending on the feedback, there are two tasks:
  - Rating prediction (explicit feedback)
  - Item recommendation (implicit feedback)

$$\{u_1, u_2, \ldots, u_n\}$$

$$\{i_1, i_2, \ldots, i_m\}$$

$$\{r_{ui} \mid u \in users, i \in items\}$$

42

Examples

- Rating prediction
  - What rating would James give to Titanic and Forrest Gump?

	Titanic	Pulp Fiction	Iron Man	Forrest Gump	The Mummy
Eve	1	4	5		3
Fred	5	1		5	2
Irene	4	1	2	5	
James	?	3	4	2	4

Item recommendation

- What is the likelihood of positive feedback from James for Titanic and Forrest Gump?

	Titanic	Pulp Fiction	Iron Man	Forrest Gump	The Mummy
Eve	1	1	1		1
Fred	1	1		1	1
Irene	1	1	1	1	
James	?	1	1	?	1

AGROCLUSTER RIBATEJO PORTUGAL E-AGRO INNOVATION E-AGRO MARKETS

43

Knowledge based techniques

- Recommendations are based on
  - Item attributes
    - E.g.: price and type of a car, number of airbags, trunk size, …
  - User requirements
    - E.g.: “the maximum acceptable price of a car is $8,000” and “the car should be safe and suitable for a family”
  - And domain knowledge describing some dependencies between user requirements and item properties
    - E.g.: “a family car should have a large trunk”
  - Or between user requirements
    - E.g.: “if a safe family car is required, the maximum acceptable price must be more than $2,000”

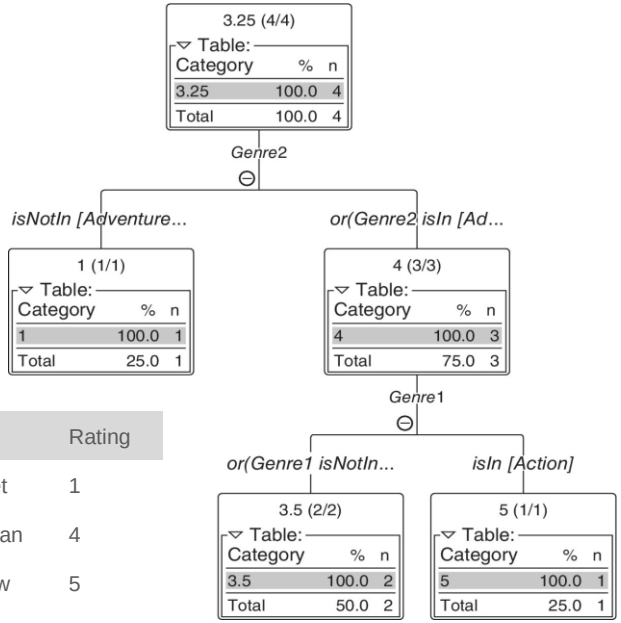
Knowledge based techniques

- Recommendation is an interactive and an iterative process
  - User specifies requirements according to the items recommended in the given state of the “conversation” with the system
- High cost of preparing the underlying knowledge base, which is domain-dependent

Content based techniques

- User interests are learned by machine learning techniques
  - Learns a model of the feedback (the target attribute) of a given user from the attributes (the explanatory variables) of items rated or ranked by the user in the past
  - Use the induced predictive model to predict ratings or rankings for items not seen by the user

Content based techniques



ID	$Genre_1$	$Genre_2$	Year	Country	Length	Director	$Actor_1$	$Actor_2$	Rating
Titanic	Drama	Romance	1997	USA	194	J. Cameron	L. DiCaprio	K. Winslet	1
Pulp Fiction	Drama	Crime	1994	USA	154	Q. Tarantino	J. Travolta	U. Thurman	4
Iron Man	Action	Adventure	2008	USA	126	J. Favreau	R. Downey Jr.	G. Paltrow	5
The Mummy	Fantasy	Adventure	1999	USA	125	S. Sommers	B. Fraser	R. Weisz	3
Forrest Gump	Drama	Romance	1994	USA	142	R. Zmeckis	T. Hanks	R. Wright	?

Model based collaborative filtering

- Recognize similarities between users according to their feedback
  - Recommends items preferred by like-minded users
    - k-nearest neighbors of the user *u* which rated an item *i*
    - According to some vector similarity measure *sim(u,v)* of the feedback vectors of user *u* and other users
- Can produce good results even without user/item attributes

Item recommendation

$$\hat{r}_{ui} = \frac{\sum_{v \in \mathcal{N}_i^{u,k}} sim(u,v)}{k}$$

Rating prediction

$$\hat{r}_{ui} = \bar{r}_u + \frac{\sum_{v \in \mathcal{N}_i^{u,k}} sim(u,v) \cdot (\phi_{vi} - \bar{r}_v)}{\sum_{v \in \mathcal{N}_i^{u,k}} |sim(u,v)|}$$

AGROCLUSTER RIBATEJO PORTUGAL

E-AGRO INNOVATION

E-AGRO MARKETS

48

Cosine vector similarity

$sim^{cv}(\mathbf{x}, \mathbf{y})$	Eve	Fred	Irene	James
Eve	1.00	0.75	0.75	0.87
Fred		1.00	0.75	0.58
Irene			1.00	0.58
James				1.00

$$sim^{cv}(\mathbf{x}, \mathbf{y}) = \frac{\sum_{i=1}^{m} x_i y_i}{\left(\sum_{i=1}^{m} x_i^2 \sum_{i=1}^{m} y_i^2\right)^{\frac{1}{2}}}$$

$$\mathcal{N}_{Titanic}^{James,2} = \{Eve, Fred\}$$

$$\mathcal{N}_{ForrestGump}^{James,2} = \{Fred, Irene\}$$

$$\hat{r}_{James\ Titanic} = \frac{sim^{cv}(James,Eve)+sim^{cv}(James,Fred)}{2} = \frac{0.87+0.58}{2} = 0.725$$

$$\hat{r}_{James\ ForrestGump} = \frac{sim^{cv}(James,Fred)+sim^{cv}(James,Irene)}{2} = \frac{0.58+0.58}{2} = 0.58$$

AGROCLUSTER RIBATEJO PORTUGAL

E-AGRO INNOVATION

E-AGRO MARKETS

Pearson correlation similarity

$sim^{pc}(\mathbf{x}, \mathbf{y})$	Eve	Fred	Irene	James
Eve	1.000	-0.716	-0.762	-0.005
Fred		1.000	0.972	0.565
Irene			1.000	0.600
James				1.000

$$\mathcal{N}_{Titanic}^{J,2} = \{I, F\}$$

$$\bar{r}_J = \frac{3+4+4}{3} = 3.67,$$

$$\bar{r}_I = \frac{4+1+2+5}{4} = 3,$$

$$\bar{r}_F = \frac{5+1+5+2}{4} = 3.25$$

$$\hat{r}_{J\,Titanic} = \bar{r}_J + \frac{sim^{pc}(J,I)\cdot(r_{I\,Titanic} - \bar{r}_I) + sim^{pc}(J,F)\cdot(r_{F\,Titanic} - \bar{r}_F)}{|sim^{pc}(J,I)| + |sim^{pc}(J,F)|}$$

$$= 3.67 + \frac{0.6\cdot(4-3) + 0.565\cdot(5-3.25)}{0.6+0.565} = 1.36$$

Model based collaborative filtering

- Use Machine Learning algorithms to predict the rating given by users to unrated items
  - Look for models able to map the users and the items into a common latent space
- The dimensions of the latent space are named factors
  - The important factors are named latent factors
    - Important, implicit, interests or properties present in an item or an user
  - Latent factors are found by applying matrix factorization algorithms to the matrix with user-item ratings

Model based collaborative filtering

- There are several matrix factorization techniques
  - E.g.: Low-rank matrix factorization
- Low-rank matrix factorization
  - Given user-item matrix *R* with *n* rows and *m* columns
    - Decompose R into a product of two matrix with lower dimensions:
      - User matrix *W* of *n* rows and *k* columns
      - Item matrix *H* of *m* columns and *k* rows
        - Where K is the dimension of the latent space
      - Such that

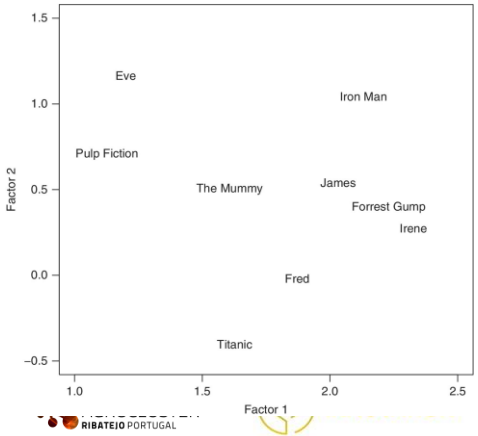
$$W \cdot H = \hat{R}$$

52

Matrix factorization example

R	Titanic	Pulp Fiction	Iron Man	Forrest Gump	The Mummy
Eve	1	4	5		3
Fred	5	1		5	2
Irene	4	1	2	5	
James	?	3	4	?	4

H	Titanic	Pulp Fiction	Iron Man	Forrest Gump	The Mummy
$factor_1$	1.626	1.126	2.131	2.229	1.607
$factor_2$	-0.406	0.706	1.041	0.394	0.497



W	$factor_1$	$factor_2$
Eve	1.200	1.164
Fred	1.871	-0.023
Irene	2.327	0.276
James	2.034	0.539

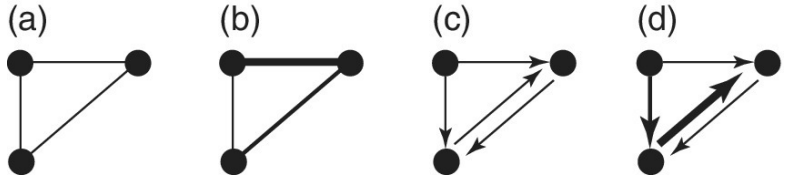
$\hat{R}$	Titanic	Pulp Fiction	Iron Man	Forrest Gump	The Mummy
Eve	1.478	2.171	3.767	3.132	2.507
Fred	3.052	2.091	3.965	4.162	2.997
Irene	3.671	2.814	5.246	5.294	3.877
James	3.088	2.671	4.896	4.745	3.537

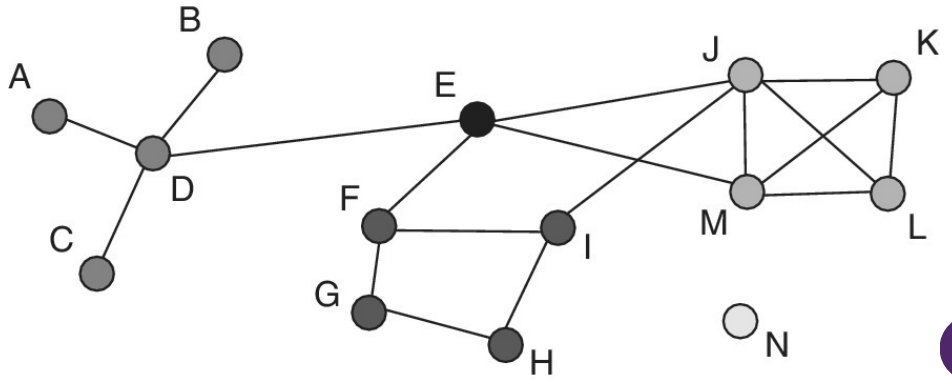
E-AGRO MARKETS

Social Network Analysis

Representing Social Networks

- Use graphs, which can be:
  - Undirected
    - Weightless (a)
    - Weighted (b)
  - Directed
    - Weightless (c)
    - Weighted (d)





55

Adjacency matrix

- Rows and columns represent nodes
  - Sparse matrix

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
A	0	0	0	1	0	0	0	0	0	0	0	0	0	0
B	0	0	0	1	0	0	0	0	0	0	0	0	0	0
C	0	0	0	1	0	0	0	0	0	0	0	0	0	0
D	1	1	1	0	1	0	0	0	0	0	0	0	0	0
E	0	0	0	1	0	1	0	0	0	1	0	0	1	0
F	0	0	0	0	1	0	1	0	1	0	0	0	0	0
G	0	0	0	0	0	1	0	1	0	0	0	0	0	0
H	0	0	0	0	0	0	1	0	1	0	0	0	0	0
I	0	0	0	0	0	1	0	1	0	1	0	0	0	0
J	0	0	0	0	1	0	0	0	1	0	1	1	1	0
K	0	0	0	0	0	0	0	0	0	1	0	1	1	0
L	0	0	0	0	0	0	0	0	0	1	1	0	1	0
M	0	0	0	0	1	0	0	0	0	1	1	1	0	0
N	0	0	0	0	0	0	0	0	0	0	0	0	0	0

- Adjacency matrix in the 2nd power
  - Number of paths (sequence of edges) of length 2 between pairs of nodes

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
A	1	1	1	0	1	0	0	0	0	0	0	0	0	0
B	1	1	1	0	1	0	0	0	0	0	0	0	0	0
C	1	1	1	0	1	0	0	0	0	0	0	0	0	0
D	0	0	0	4	0	1	0	0	0	1	0	0	1	0
E	1	1	1	0	4	0	1	0	2	1	2	2	1	0
F	0	0	0	1	0	3	0	2	0	2	0	0	1	0
G	0	0	0	0	1	0	2	0	2	0	0	0	0	0
H	0	0	0	0	0	2	0	2	0	1	0	0	0	0
I	0	0	0	0	2	0	2	0	3	0	1	1	1	0
J	0	0	0	1	1	2	0	1	0	5	2	2	3	0
K	0	0	0	0	2	0	0	0	1	2	3	2	2	0
L	0	0	0	0	2	0	0	0	1	2	2	3	2	0
M	0	0	0	1	1	1	0	0	1	3	2	2	4	0
N	0	0	0	0	0	0	0	0	0	0	0	0	0	0

56

Basic properties of nodes

- Degree
  - Number of connections of the node
  - In-degree (directed graphs)
    - Sum of corresponding columns of adjacency matrix
  - Out-degree (directed graphs)
    - Sum of corresponding rows of adjacency matrix

Node	A	B	C	D	E	F	G	H	I	J	K	L	M	N
Degree	1	1	1	4	4	3	2	2	3	5	3	3	4	0

57

Basic properties of nodes

- Distance
  - The minimum number of edges the information has to take from one node to the other

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
A	0	2	2	1	2	3	4	5	4	3	4	4	3	∞
B	2	0	2	1	2	3	4	5	4	3	4	4	3	∞
C	2	2	0	1	2	3	4	5	4	3	4	4	3	∞
D	1	1	1	0	1	2	3	4	3	2	3	3	2	∞
E	2	2	2	1	0	1	2	3	2	1	2	2	1	∞
F	3	3	3	2	1	0	1	2	1	2	3	3	2	∞
G	4	4	4	3	2	1	0	1	2	3	4	4	3	∞
H	5	5	5	4	3	2	1	0	1	2	3	3	3	∞
I	4	5	4	3	2	1	2	1	0	1	2	2	2	∞
J	3	3	6	2	1	2	3	2	1	0	1	1	1	∞
K	4	4	4	3	2	3	4	3	2	1	0	1	1	∞
L	4	4	4	3	2	3	4	3	2	1	1	0	1	∞
M	3	3	3	2	1	2	3	3	2	1	1	1	0	∞
N	∞	∞	∞	∞	∞	∞	∞	∞	∞	∞	∞	∞	∞	0

58

Basic properties of nodes

- Closeness
  - Reflects how accessible a node is in the network
  - Sensitive to and decreases with the size of the network
    - Instead of an infinite value, the number of nodes in the network is substituted

$$closeness(v) = \frac{1}{\sum_{u \neq v} distance(u,v)}$$

Node	A	B	C	D	E	F	G	H	I	J	K	L	M	N
Closeness	0.0196	0.0196	0.0196	0.025	0.0286	0.025	0.0204	0.0196	0.0238	0.0270	0.0217	0.0217	0.0256	0.0054

59

Basic properties of nodes

- Betweenness
  - Assesses how important the position of a node in the network is
    - *nsp (u,t)*: the number of shortest paths from node *u* to node *t*
    - $nsp_v(u,t)$: the number of shortest paths from node *u* to node *t* that go through node *v*

$$betweenness(v) = \sum_{u \neq v \neq t} \frac{nsp_v(u,t)}{nsp(u,t)}$$

Node	A	B	C	D	E	F	G	H	I	J	K	L	M	N
Betweenness	0	0	0	30	37.17	14.50	2.17	1.33	10.67	18	0	0	6.17	0

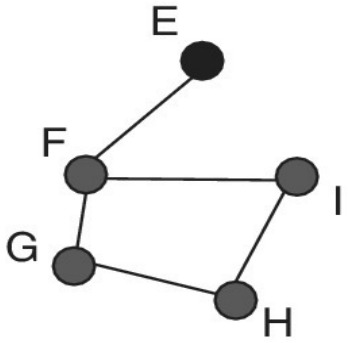
60

Basic properties of nodes

- Betweenness example for node G
  - Given:
    - nsp(F,H) = nsp(E,H) = 2
    - $nsp_G(F,H) = nsp_G(E,H) = 1$
    - nsp(E,F) = nsp(E,I) = nsp(F,I) = nsp(H,I) = 1
    - $nsp_G(E,F) = nsp_G(E,I) = nsp_G(F,I) = nsp_G(H,I) = 0$
  - Betweenness (G)

    $= nsp_G(E,F)/nsp(E,F) + nsp_G(E,H)/nsp(E,H) + nsp_G(E,I)/nsp(E,I) +$
    $nsp_G(F,H)/nsp(F,H) + nsp_G(F,I)/nsp(F,I) + nsp_G(H,I)/nsp(H,I)$

    = 0/1+1/2+0/1+1/2+0/1+0/1 = 1



61

Properties of networks

- Clustering coefficient
  - Measures the tendency of a node v to be included in a triad
    - *triangle(u,v,t)* =1 if the nodes *u,v* and *t* form a triangle and *triangle(u,v,t)=0* otherwise
    - *triple(u,v,t) =1* if the nodes *u* and *t* are both connected to the node *v*, otherwise, *triple(u,v,t)* =0
    - If *degree(v)* <2, the clustering coefficient is either equal to zero or not defined

$$clust\_coef(v) = \frac{\sum_{u \neq v \neq t} triangle(u,v,t)}{\sum_{u \neq v \neq t} triple(u,v,t)}$$

Node	A	B	C	D	E	F	G	H	I	J	K	L	M	N
Clust_coef	-	-	-	0	0.17	0	0	0	0	0.40	1	1	0.67	-

AGROCLUSTER RIBATEJO PORTUGAL

E-AGRO INNOVATION

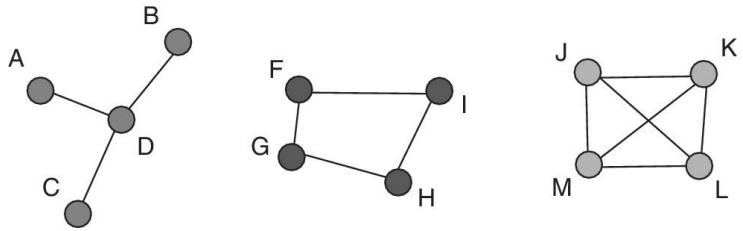
E-AGRO MARKETS

Properties of networks

- Diameter
  - The longest of all the distances between the nodes of the network
  - Diameter of the example network: 5
- Cliques
  - Subset of nodes such that every two nodes in these subsets are connected
  - Cliques of size 3 are the following subsets: {E, J, M }, {J, K, L}, {J, K, M }, {J, L, M } and {K, L, M }
  - Clique of size four is the subset {J, K, L, M }

Properties of networks

- Clustering coefficient
  - Expresses the probability that the triples in the network are connected to form a triangle
  - Clustering coefficient of the example network: 0.357
  - Clustering coefficients of the star and ring networks: 0.0
  - Clustering coefficient of the fully connected network: 1.0



64

Properties of nodes and networks

- Centralization
  - *C*: the centrality score
    - Degree
    - Closeness
    - Betweenness

$$C(N) = \sum_{v}(\max_{u} c(u) - c(v))$$

  - *max c(u)*: the maximum centrality score from all nodes u of the network (including node *v*)
  - *c(v)*: the centrality score of node *v*
- Modularity
  - Expresses the degree to which a network displays cluster structures
    - Often called communities

Properties of nodes and networks

- For the example network
  - $C^{degree}(N) = 0.187$
  - $C^{closeness}(N) = 0.202$
  - $C^{betweenness}(N) = 0.395$
  - $C^{closeness}$(star network) = (0.33-0.2) + (0.33-0.2) + (0.33-0.2) + (0.33-0.33)= 0.4
  - $C^{closeness}$(ring network) = 4 × (0.25-0.25) = 0
  - $C^{closeness}$(fully connected) = 4 × (0.33-0.33) = 0
  - Modularity(example nework) = 0.44

Final remarks

- A good pre-processing is very important in text mining
- Trend to combine text mining with natural language processing to get better results.
- Measuring the performance of a recommender system is not easy
  - Coverage, scalability, robustness, novelty, serendipity,
- Cold start problem in recommender systems
- Context-based and group recommendations
- Basic node/network properties can be used as features in machine learning applications.
  - E.g. link prediction, community detection, etc.

Final remarks and Literature

- Most important in text mining is the good pre-processing of the text.
- A trend is to combine text mining with natural language processing to get better results.
- Measuring the performance of a recommender system is not so easy.
  - coverage, scalability, robustness, novelty, serendipity,
- Cold start problem arises when new user or item arrive to the system.
- Context-based and group recommendations
- Basic node/network properties can be used as features in machine learning applications.
  - e.g. link prediction, community detection, etc.

Literature

- Weiss, S.M., Indurkhya, N., and Zhang, T. (2015) Fundamentals of Predictive Text Mining, Springer-Verlag.
- Ricci, F., Rokach, L., Shapira, B., and Kantor, P. (2010). Recommender Systems Handbook, Springer-Verlag.
- Zafarani, R., Abbasi, M., and Liu, H. (2014) Socia Media Mining: An introduction, Cambridge University Press.

AGROCLUSTER RIBATEJO PORTUGAL

E-AGRO INNOVATION

E-AGRO MARKETS

AGROCLUSTER
RIBATEJO

E-AGRO INNOVATION

E-AGRO MARKETS

Cofinanciado por:

COMPETE 2020

PORTUGAL 2020

UNIÃO EUROPEIA
Fundo Europeu de Desenvolvimento Regional