Information Retrieval and Data Mining

Prof. Matteo Matteucci, Ing. Luca Bondi

February, 22 2016

Very Important Notes

- Answers to questions 1, 2, and 3 should be delivered on a different sheet with respect to 4 and 5
- State cleary with number and letter which part of each exercise you are answering
- If you need a calculator this should not be to any extent programmable or network connected

1. **Question (8 pts)**:
   Imagine we have a data set where each record is a list of categorical weather conditions on a randomly selected number of days, and the labels correspond to whether a girl named Arya went for a horse ride on that day.

Sky	Temperature	Humidity	Wind	Horse Ride
Cloudy	Warm	Low	Low	Yes
Rainy	Cold	Medium	Low	No
Sunny	Warm	Medium	Low	Yes
Sunny	Hot	High	High	No
Snow	Cold	Low	High	No
Rainy	Warm	High	Low	Yes

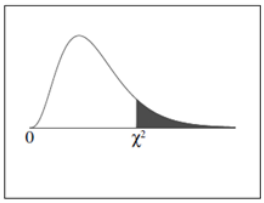
   (a) **Describe in details** what a rule is and how a model based on a set of rules works

   (b) **Describe in details** the *Sequential Covering* algorithm to learn a rule set out of a dataset

   (c) **Learn a set of rules** for *Horse Ride=Yes* out of the previous dataset by *Sequential Covering*

   (d) Use the the χ^2 test for independence with ($\alpha = 0.05$) to **prune the rule** with the highest coverage from the previous point (assume you have enough examples to apply the χ^2 statistics). What can you derive after the pruning?

Chi-Square Distribution Table



The shaded area is equal to α for $\chi^2 = \chi^2_\alpha$.

df	$\chi^2_{.995}$	$\chi^2_{.990}$	$\chi^2_{.975}$	$\chi^2_{.950}$	$\chi^2_{.900}$	$\chi^2_{.100}$	$\chi^2_{.050}$	$\chi^2_{.025}$	$\chi^2_{.010}$	$\chi^2_{.005}$
1	0.000	0.000	0.001	0.004	0.016	2.706	3.841	5.024	6.635	7.879
2	0.010	0.020	0.051	0.103	0.211	4.605	5.991	7.378	9.210	10.597
3	0.072	0.115	0.216	0.352	0.584	6.251	7.815	9.348	11.345	12.838
4	0.207	0.297	0.484	0.711	1.064	7.779	9.488	11.143	13.277	14.860
5	0.412	0.554	0.831	1.145	1.610	9.236	11.070	12.833	15.086	16.750
6	0.676	0.872	1.237	1.635	2.204	10.645	12.592	14.449	16.812	18.548
7	0.989	1.239	1.690	2.167	2.833	12.017	14.067	16.013	18.475	20.278
8	1.344	1.646	2.180	2.733	3.490	13.362	15.507	17.535	20.090	21.955

2. **Question (5 pts)**:
   Consider the Markov Chain represented by the following transition probability

$$E = \begin{bmatrix} 0 & 1/2 & 1/2 \\ 1/2 & 1/2 & 0 \\ 1/2 & 1/2 & 0 \end{bmatrix}$$

   (a) Compute the probability that starting from a random state you will end on state 3 in 2 steps.

   (b) Is the process represented by the previous transition matrix an ergothic one? Why?

   (c) Let assume the process is ergothic, this means you will arrive at some point in any state; starting on state 1 how many steps you will need on average to get in state 3?

   (d) Consider the adjacency matrix associated to the previous chain and compute the PageRank score associated to each node in the corresponding graph, setting $\mu = 0$ (Seeley index). In practice: (i) write the PageRank/Seeley linear system, (ii) find a possible solution setting one of the variables equal to α, (iii) find the solution corresponding to the unit norm for the resulting vector.

3. **Question (6 pts)**:
   Let consider the Knowledge Discovery Process, which starts from **data** to generate **knowledge**, reported in Figure 1

   (a) Discuss the 5 steps from `data` to `knowledge` in the figure and the reason for the corresponding feedback arrow

   (b) The outcome of this process is a pattern or a set of patterns. A pattern, to be good, should be `novel`, `useful`, and `understandable`; can you explain each of the previous terms and discuss how each of them is related/affected to/by the previous steps? (e.g., is the usefulness of a pattern mainly affected by the transformation step or the selection step?)

4. **Question (8 pts)**:
   An IR system produces the following rankings in answer to queries q_1, q_2 and q_3. The underscored documents are the ones relevant to the user.

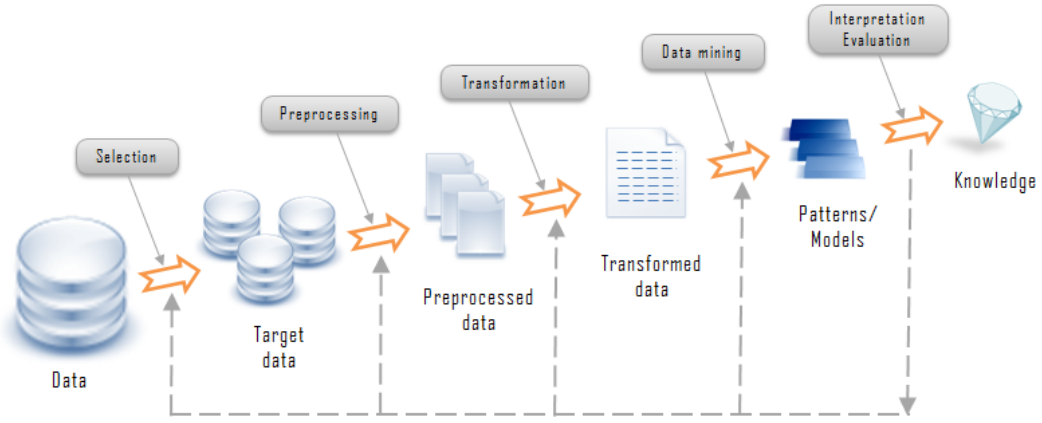


Figure 1: The Knowledge Discovery Process

R	q_1	q_2	q_3
1	A	G	H
2	L	E	G
3	G	D	L
4	F	F	C
5	D	L	D
6	E	I	A
7	B	B	E
8	H	H	B
9	I	C	F
10	C	A	I

(a) Draw the precision-recall curve and the interpolated precision-recall curve

(b) Compute the Mean Average Precision

(c) Compute the R-precision

(d) Draw the Receiver-Operating-Characteristic

5. **Question (5 pts)**:

(a) Describe how to build an inverted index.

(b) Illustrate a way to process an intersection query such as $q = t_a \wedge t_b$ on top of an inverted index.