Information Retrieval and Data Mining

Prof. Marco Tagliasacchi
Prof. Matteo Matteucci

September, 15 2015

Very Important Notes

- Answers to questions 1, 2, and 3 should be delivered on a different sheet with respect to 4 and 5
- If you need a calculator this should not be to any extent programmable or network connected

1. **Question (8 pts)**: Consider a graph that is described by the following adjacency matrix

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

   (a) Compute the in-degree centrality index of each node

   (b) Compute the closeness centrality index of each node

   (c) Compute the betweenness centrality index of node 2

   (d) Construct the HITS authority matrix **A** and compute the first iteration, including the normalization step.

2. **Question (6 pts)**

   Illustrate the construction of a kd-tree. It might helpful to use a small-scale example with at least 8 points in a 2-dimensional space.

3. **Questions (5 pts - each statement can be either TRUE or FALSE)**

(a) Consider three rankings $r = [a, b, c, d]$, $s = [a, c, d, b]$ and $p = [a, b, d, c]$.

- T F According to Spearman's footrule distance, ranking r is closer to p than to s.
- T F Kendall's tau distance between r and s is equal to 1/6.
- T F Kendall's tau distance is always smaller than Spearman's footrule distance
- T F The maximum (unnormalized) Kendall tau distance is $N(N-1)/2$, where N is the number of objects.

(b) Consider the outcome of three queries, for which ground truth (GT) relevance judgements are available

k	q_1	GT	q_2	GT	q_3	GT
1	a	1	a	1	f	0
2	b	1	c	1	b	1
3	c	0	f	1	c	0
4	d	0	g	0	a	1
5	e	0	d	0	h	0

- T F MAP is equal to 5/6
- T F Precision always increases when increasing k
- T F The average recall at $k = 3$ is equal to 5/6
- T F Recall never decreases when increasing k

(c) Consider a term-document matrix A.

$$A = \begin{bmatrix} 1 & 2 & 4 & 4 & 2 \\ 1 & 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix} \tag{2}$$

- T F The idf of term t_1 is equal to zero.
- T F The cosine similarity between document 1 and document 2 is equal to $\sqrt{10}/5$
- T F The L_1 distance between document 1 and document 2 is equal to 3

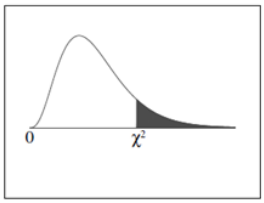
4. **Question (5 pts)** Starting from the dataset hereafter

Size	Colour	Shape	Weight	Expensive
Big	Red	Square	Heavy	Yes
Small	Blue	Triangle	Light	Yes
Small	Blue	Square	Light	No
Big	Green	Triangle	Heavy	No
Big	Blue	Square	Light	No
Big	Green	Square	Heavy	Yes
Small	Red	Triangle	Light	Yes

1. Extract a rule set for the class *Expensive=Yes* out of it using *Sequential Covening* (2 points)

2. Use the the χ^2 test for independence with ($\alpha = 0.01$) to test for pruning the *Color* attribute from the most general rule, i.e., the one with less antecedents. (2 points)
3. Define coverage and accuracy for a rule, then provide coverage and accuracy for all the rules previously extracted. (1 point)

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

5. **Question (8 pts)**

Answer the following questions:

1. What is the goal of frequent pattern mining (a.k.a. association rules mining)? (1 point)
2. What does the a-priori principle states? (1 point)
3. Describe the a-priori algorithm for *Frequent Itemset Generation*. (1 points)
4. Describe the algorithm for *Rule Generation* out of an itemset. (1 point)
5. Starting from the table find all the itemset with minimum support 50%. (2 points)
6. Out of the biggest itemsets extract the association rules with at least 80% confidence. (2 points)

TID	A	B	C	D	E
T1	0	1	0	0	0
T2	1	1	1	1	1
T3	0	0	1	1	0
T4	1	0	0	0	1
T5	1	1	1	1	0