Information Retrieval and Data Mining

Prof. Marco Tagliasacchi
Prof. Matteo Matteucci

February, 11 2015

1. **Question (8 pts)**:
   Consider a document collection that is represented by means of the following term-document matrix

$$A = \begin{bmatrix} 4 & 8 & 2 & 4 \\ 0 & 1 & 2 & 0 \\ 1 & 0 & 0 & 2 \\ 8 & 2 & 8 & 1 \\ 4 & 0 & 0 & 0 \end{bmatrix} \tag{1}$$

   Let $\mathbf{q} = [0, 0, 1, 1, 0]^T$ denote the query vector.
   (a) Compute document ranking adopting a *tf-idf* weighting scheme and cosine similarity (hint: use $\log_2$ to simplify your calculations)
   (b) Comment how the result would change if *idf* weighting is not adopted, justifying your answer.
   (c) Evaluate the result when retaining the top-k documents in terms of precision and recall. Assume that the set of relevant documents is $R = \{d_2, d_4\}$.
   (d) Evaluate the result using Kendall tau and footrule distance, assuming that correct ordering is $< d_4, d_2, d_1, d_3 >$.

   In case of ties, break the tie in favour of the document with the smallest index.
2. **Question (6 pts)**
   (a) Describe the PageRank centrality index.
   (b) Provide a small-scale example
   (c) Discuss the random surfer interpretation of PageRank

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

(a) Consider the following rankings (break the tie in favour of the object with the smallest index).

k	rank 1	rank 2	rank 3
1	a	a	b
2	b	c	a
3	c	d	c
4	d	e	f
5	e	b	e
6	f	f	d

- [T] [F] The aggregated ranking based on Borda's method is $< a, b, c, d, e, f >$
- [T] [F] $d_F(r_1, r_2) < d_F(r_1, r_3)$, where $d_F(\cdot, \cdot)$ denotes the footrule distance between two rankings.
- [T] [F] The median rank aggregation is $< a, b, c, d, e, f >$.
- [T] [F] In this case, aggregation based on footrule distance leads to the same results as median rank aggregation (hint: you don't need to compute aggregation based on footrule distance to answer to this question).

(b) Consider a system based on Latent Semantic Indexing, in which the term-document matrix $A = [d_1, \ldots, d_N]$ is approximated as follows $A_k = [d_{1,k}, \ldots, d_{N,k}] = U_k \Sigma_k V_k^T$, by retaining only the first k topics.

- [T] [F] If $\|d_{j_a} - q\|_2 < \|d_{j_b} - q\|_2$, then $\|d_{j_a,k} - q\|_2 < \|d_{j_b,k} - q\|_2$
- [T] [F] The value of k is chosen analytically based on the singular values $\sigma_1, \ldots, \sigma_k$.
- [T] [F] When new documents are added, the matrix U_k changes and V_k remains the same.

(c) Consider the following graph adjacency matrix, in which $E_{ij} = 1$ indicates that there is a link from node i to node j.

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

- [T] [F] The harmonic centrality index assigned to node 1 is equal to 2.
- [T] [F] The betweenness centrality index assigned to node 3 is equal to 0
- [T] [F] There are three paths of length 3 from node 4 to node 3 (hint: remind Katz centrality)

4. **Question (6 pts)**

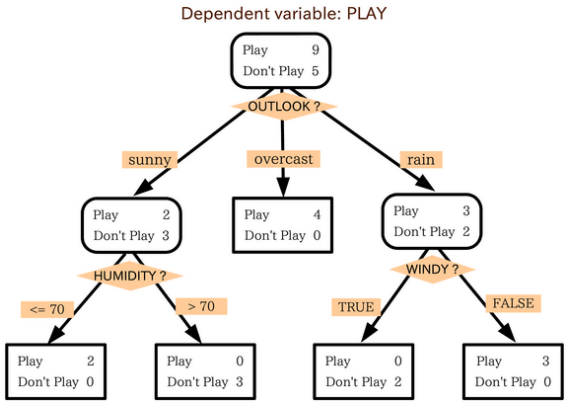
Answer the following questions about Decision Trees.

1. What is a *Decision Tree*? Provide a description of this kind of model.
2. Describe the learning algorithm for a Decision Tree. In particular describe the of splitting criterium, and the stopping conditions.
3. Describe in details *Information Gain*, *Information Gain Ratio*, and *Gini index*.

5. **Question (7 pts)**

Let assume we have learned the depicted Decision Tree from the dataset in the table.

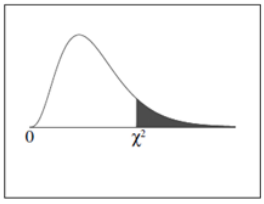
Outlook	Temp.	Hum.	Windy	Play
sunny	85	85	false	No
sunny	80	90	true	No
overcast	83	78	false	Yes
rain	70	96	false	Yes
rain	68	80	false	Yes
rain	65	70	true	No
overcast	64	65	true	Yes
sunny	72	95	false	No
sunny	69	70	false	Yes
rain	75	80	false	Yes
sunny	75	70	true	Yes
overcast	72	90	true	Yes
overcast	81	75	false	Yes
rain	71	80	true	No



Answer the following questions

1. Extract the rule set corresponding to the decision tree, assume `No` as default class
2. Prune one of the attributes from the rule corresponding to the rightmost path using the χ^2 test for independence with ($\alpha = 0.01$) (assume to have enough data).
3. Define coverage and accuracy for a rule, then provide coverage and accuracy for the rule before and after the pruning in the previous step.
4. Is the resulting rules set, i.e., the one obtained after the pruning, still a decision tree? Is it true in general? Why?

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