CSE 494/598 Fall 2002 Exam 2

Instructor: Subbarao Kambhampati
Given Saturday, 14th December, 2002 12 noon
Due Wednesday, 18th December, 2002 12 noon [You can drop them off at CSE office in my mailbox]

Must be submitted in hardcopy. You are encouraged to restrict your answer to the space provided. In cases where you absolutely must have more space, you can use the back of the sheets—but make a note so I won't miss the work.

If you have more than 2 exams during 14-18th time period, or if you have some other compelling circumstances, you may send me an email with request for deadline extension. [Only extensions that I grant in E-mail will be honored]

Name:____

The examination is going to be conducted under the following honor-code agreement. Please read and sign it:

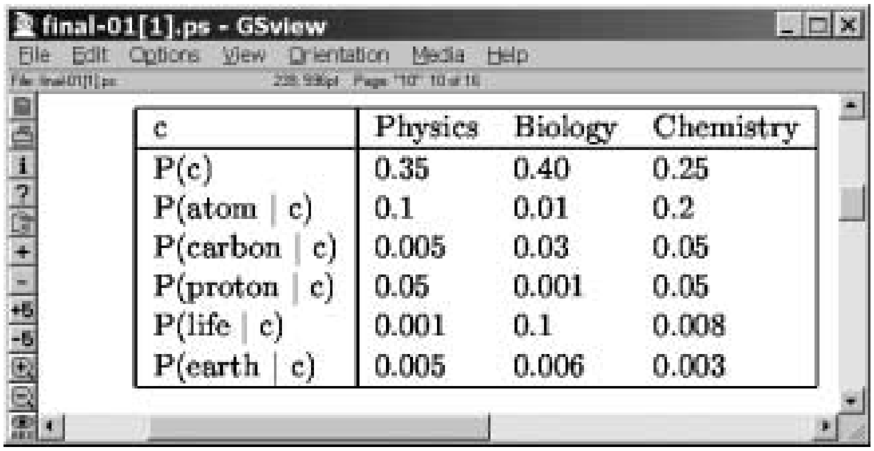
"I understand that this final take-home examination is supposed to be strictly individual effort. I certify, under the penalty of academic dishonesty, that I have not consulted with anyone other than the instructor in preparing my answers."

*Signature:*_________________________ *date:*_________________________

Qn I. Text classification using NBC [10]	
Qn II. Collaborative Filtering [10]	
Qn III. Integration [13]	
Qn IV. XML [10]	
Qn V. Coverage/Completeness [10]	
Qn VI. Short answer qns [35]	
Qn VII. Easy marks [12]	

Total (100pt)

Qn 1. [10pt] [Mooney] Assume we want to classify science texts into three categories—physics, biology and chemistry. The following probabilities have been estimated from analyzing a corpus of pre-classified web-pages gathered from Yahoo.

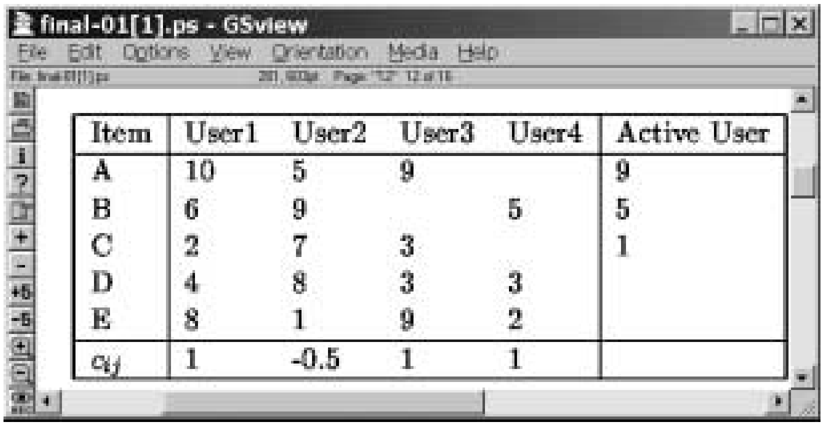

final-01[1].ps - GSview
File Edit Options View Orientation Media Help
File: final01[1].ps 238, 936pt Page: "10" 10 of 16

c	Physics	Biology	Chemistry
P(c)	0.35	0.40	0.25
P(atom | c)	0.1	0.01	0.2
P(carbon | c)	0.005	0.03	0.05
P(proton | c)	0.05	0.001	0.05
P(life | c)	0.001	0.1	0.008
P(earth | c)	0.005	0.006	0.003

Assuming that the probability of each evidence word is independent of other word occurrences given the category of the text, compute the (posterior) probability for each of the possible categories each of the following short texts; and based on that, their most likely classification. Assume that the categories are disjoint and exhaustive (i.e., every text is either physics, or biology or chemistry and no text can be more than one). Assume that words are first stemmed to reduce them to their base form (atoms→atom) and ignore any words that are not in the table:

A: ***the carbon atom is the foundation of life on earth.***
B. the carbon atom contains 12 protons.

Qn II (10pt) [Mooney] Consider the following ratings matrix for five users and five items—with the 5th user—in the last column—being the active user. The pearson correlation coefficients of the first four users with respect to the active user are already calculated for you and given in the row titled “Cij”. Using the technique for collaborative filtering discussed in the class and in the homework, compute the ratings for D and E for the active user. Use two closest neighbors in the calculation. Assume that the significance weights are uniformly set to 1.



Item	User1	User2	User3	User4	Active User
A	10	5	9		9
B	6	9		5	5
C	2	7	3		1
D	4	8	3	3	
E	8	1	9	2	
c_{ij}	1	-0.5	1	1	

Qn III (13pt) Suppose we are interested in building a mediator for integrating a bunch of data sources that export information about CS courses. We would like the mediator to be able to provide the information about the course name, the institution at which it is being offered, the term of offering and the average enrollment in the course. We also want to offer information about the instructors that teach these courses, and for each instructor the institution at which the instructor teaches and his/her average teaching evaluation numbers. (*for the following questions, you can either use SQL style syntax or the datalog-style syntax. I will be forgiving of syntax problems).*

[2]Design a mediator schema for this application (note that a schema may have more than one relation).

[2]Suppose we got a source—called ASU-CS-Underground---which exports, for a bunch of CS instructors at ASU---their teaching evaluation numbers. Using LAV approach, write this source as a (materialized) view on the mediator schema.

[2]We have got another source called ASU-CS-S02-Catalog—which exports the set of CS courses being taught in Spring 2003, the instructor who will be teaching them, and the rooms in which the classes will be taught. Write this source too as a materialized view on the mediator schema.

[2]] If a student in Arizona has the following question: What are the CS courses being taught next term (S 2003) by instructors whose average evaluations are higher than 4.3. Show this as a query on the mediator schema

[5]Show—using the bucket algorithm---how the previous query is reformulated as queries on the data sources.

Qn IV [10] I am interested in exporting the ASU CSE course catalog—available at http://www.asu.edu/aad/catalogs/courses/cse.html in XML form. Here is a typical entry from that catalog:

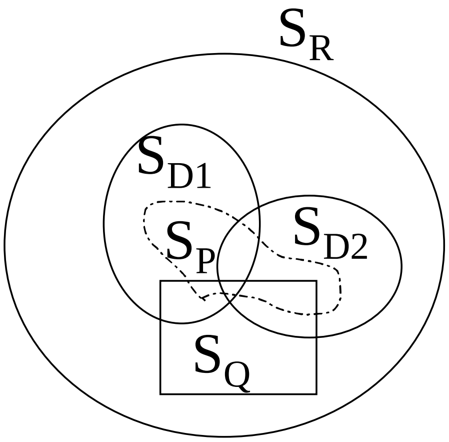
> **CSE 471 Introduction to Artificial Intelligence. (3)**
> *fall and spring*
> State space search, heuristic search, games, knowledge representation techniques, expert systems, and automated reasoning. Prerequisites: CSE 240, 310.

[4] Decide on a set of tags, and show their use by converting this course into XML format using your tags. [It is okay if your tag set is tailored just for this course]

[3]Write a DTD for your catalog. [It is okay if your DTD is tailored just for this course]

[3] Suppose someone is interested in finding out all courses typically offered in summer and need 310 as a pre-requisite. Write an Xquery query that will answer their question on your XML file.

Qn V [10pt] Consider a relation R whose complete extension (i.e., the total set of tuples in it) is S_R. There are two data sources D1 and D2 which have stored some parts of R and their extensions are sets S_{D1} and S_{D2} (each of these sets are subsets of S_R). We have a query Q that is a selection on the relation R, and the set of tuples satisfying Q in R is S_Q. Finally, we have a query plan P, to answer the query Q. When we run P on D1 and D2, we get set S_P of tuples. The query plan P may or may not be completely correct—i.e., it may give wrong answers, or miss answers. See the Venn diagram to the right which illustrates a possible relation between these sets.



Answer the following questions (use |S| for size of set S; you can talk about unions and intersections of sets).

Coverage of D1 w.r.t R =	
Overlap of D1 with D2 w.r.t R =	
Precision of P =	
Recall of P =	
Conditions when P is complete =	
Conditions when P is sound =	
Conditions when P is "source-complete" in that provides all answers to Q that are present in D1 and D2 =	

Qn VI [Short answer questions—each question carries 3 points; except the last one which has 5points]

1. Consider a query Q: Q(x,y):- M(x,y);

   And a query plan P: Q'(u,v) :- M(u,w) & R(w,v).

   Using ideas of containment mapping, answer (a) if P a Sound query plan for the query Q and (b) Is P a complete query plan?

2. Here is a query Q, defined in the deductive database terminology. Is this query expressible in relational algebra? If so give a relational algebra query (you can write in SQL, and I will be forgiving of errors in SQL syntax as long as you show you got the idea):

   Q(x,z) :- P(x,y) & R(y,z)
   Q(x,z) :- M(x,z)

3. Don was married to Mary for 20 years, and they have seen over 300 movies together. Mary however seems to agree with Don on only 80% of the movies (in terms of whether they were good or bad). Of late, Don has been carrying on with Roxy over at the office and they have seen some 20 movies together. Roxy seems to agree with Don a 100%. Now, there is a new movie in town—called "*Gulf Wars: Episode 2—Clone of the Attack*". He hasn't yet seen it, but both Roxy and Mary have—and they have differing opinions. What does collaborative filtering method say he should do in this kind of scenario?

4. For the Buck-shot clustering algorithm that we discussed in class, explain (a) what is the advantage gained by using the hierarchical clustering as a first phase before K-means? (b) what is special about the number Sqrt(n)---the number of samples used by buckshot in the first phase?

5. Explain—very briefly---what is “screen-scraping” and how XML standard is supposed to help alleviate it.

6. A small company has decided that it needs to provide a uniform interface to its databases. It decided to investigate the mediator technology to do this. The company has, for the past seven years, been using three separate databases which, among themselves store the employee and product data. Should this company go with the GAV or LAV model for its mediator? Justify your answer.

7.. In database query optimization, query plans that involve joins are considered better than query plans that involve Cartesian products. What is the reason for this?

8. Explain as succinctly and clearly as possible what is meant by the statement that XML is “semi-structured” (in contrast to “unstructured text” and “structured (relational) data”).

9. State, as clearly as possible, what is the main advantage of the Naïve bayes assumption? What may we be losing because of the assumption?

10. What is the advantage of keeping statistics on overlap between sources, over and above the statistics on source coverage?

11.[5pt] One of the main problems faced in clustering high-dimensional text data is that the cosine similarities between points in high dimensions tend to be 0 (two random high dimensional vectors have a high probability of being orthogonal to each other). A way out of this is to use latent semantic indexing ideas to reduce dimensionality. Assuming the matrix T-D gives the original term-document matrix, explain, as clearly as possible, the steps that will be involved in this idea. Pay special attention to what are the data points that are clustered.

Qn VII Fun points [6+6]

[6][Participation question] Metaphors you (were forced to) live by. From time to time I (Rao) used various metaphors to make certain points in the class. Here is one example. List 3 additional examples, and for each example, briefly explain the point (from the course) that I was trying to illustrate using that metaphor

The story of the guy who prays to his God, and the God shows up and gives him 3 "boons"---with the caveat that whatever the guy gets, all his villagers get too.	*Used to motivate the inverse document frequency weights.*

[6] Suppose you were to list the topics we covered in this course in the order of their interestingness to you ranked from 1 to 1000—with the most interesting topics at the top of the list, while least interesting topics in terms of the time spent on them at the bottom. Write down the top 5 and bottom five topics from the list. Note that I fully realize that you may have liked this course so much that even the least liked topic is something you liked quite a bit; so please don’t tell me that there are no “not liked” topics…

Most liked topics	*Least liked topics*
1.	996.
2.	997.
3.	998.
4.	999.
5.	1000.