CSC 483/583: Final (275 pts)

May 10, 2017

Name (5 pts): KEY

Instructions

Read each question carefully before determining the best answer. If you don't know how to answer a question, you might want to skip it until you've done the questions you're confident about. **Show all work**, and indicate if any work for credit is included somewhere other than directly below the question.

You are allowed one 8½″ × 11″ page (double-sided) of notes and a simple, self-contained hand-calculator. If you do not own a hand-calculator you may use your cell phone strictly as a calculator. No devices of any kind are allowed to access the Internet!

Violation of this policy or of the student code of academic integrity results in a penalty greater than the value of this exam up to and including failing the course.

Graduate students have one additional question (marked GRAD STUDENTS ONLY). Because of this additional question, the credit for graduate students adds up to more than 275 points. Similarly to the homeworks, graduate students grades will be normalized at the end to be out of 275.

The exam also contains two bonus questions, which are counted in addition of the overall 275 points.

1. **(30 pts) Relevance feedback**

a. (15 pts) In the Rocchio relevance feedback algorithm (the SMART implementation), what weight setting for $\alpha/\beta/\gamma$ does a "Find pages like this one" search correspond to?

-3

$\alpha = 0$ or very small (there is no query!)

-5 each

$\beta = 1$, or large, but ≤ 1

$\gamma = 0$, or very small

b. (15 pts) Give two reasons why relevance feedback has been little used in web searches.

- Hard to explain
- Takes time
- Users reluctant to provide feedback

....

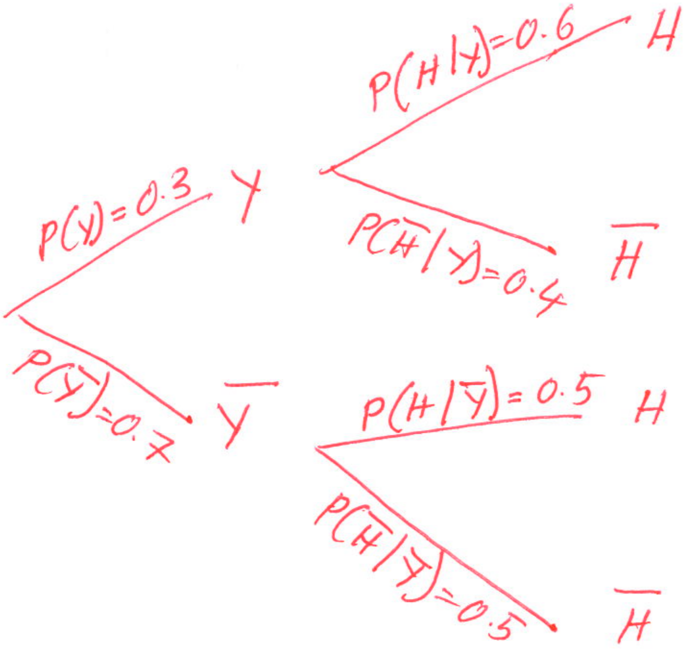
2. **(30 pts - GRAD STUDENTS ONLY) Probability theory**

According to exit polls at the last presidential election, 30% of voters in the presidential election were under 30. Of these young voters, 60% reported that they had voted for Hillary Clinton. On the other hand, the probability that a voter is 30 years old or older *and* voted for Hillary is 35%. Let Y represent the event that the voter is younger than 30, and let H represent the event that she is a Hillary voter. Let Y' and H' be the corresponding complements (i.e., *not* younger than 30, and has *not* voted for Hillary).

(a) (5 pts) Fill in the following using H, Y, their complements, and any appropriate conditionals or operators such as $\cap$ or $\cup$:

$$P(\ Y\ ) = 0.30$$
$$P(\ H/Y\ ) = 0.60$$
$$P(\ \overline{Y} \cap H\ ) = 0.35$$

(b) (10 pts) Draw the probability tree and fill in all the probability values for all branches. If you have to calculate some of them, show the expressions used for each.



$$P(\overline{Y} \cap H) = P(\overline{Y}) \cdot P(H/\overline{Y}) \Rightarrow P(H/\overline{Y}) = \frac{0.35}{0.7} = 0.5$$

(c) (5 pts) Suppose we are interested in the probability that the voter is under 30 *and* a Hillary supporter. Fill in the corresponding event, and compute its probability. Show work.

$$P(\,Y \cap H\,) = P(Y) \cdot P(H|Y) = 0.3 \cdot 0.6 = \boxed{0.18}$$

(d) (10 pts) Suppose we know that the chosen voter is a Hillary supporter. What is the probability that she is under 30? Fill in the event in parentheses, and compute the corresponding probability. Show all work.

$$P(\,Y|H\,) = \frac{P(H|Y) \cdot P(Y)}{P(H)} =$$

$$= \frac{0.3 \cdot 0.6}{0.3 \cdot 0.6 + 0.7 \cdot 0.5} = \frac{0.18}{0.18 + 0.35} = \boxed{0.34}$$

3. **(30 pts) Probabilistic IR**

a. (10 pts) Name two differences between the BM25 model and the vector space *tf.idf* model.

BM25 – based on probability theory
– explicit length normalization
– has hyper parameters

b. Consider the following document collection:

Doc1: the martian has landed on the latin pop sensation ricky martin → 11 unigrams 10 bigrams

Doc2: the martian is the science fiction movie of the year → 10 unigrams 9 bigrams

i. (10 pts) Under a unigram language model, what are $P(\text{the}|\text{Doc1})$ and $P(\text{martian}|\text{Doc1})$? Use Jelinek-Mercer smoothing with $\lambda = 0.5$. Do not remove stop words.

$$P(the/Doc1) = 0.5 \cdot \frac{2}{11} + 0.5\frac{5}{21} = 0.20$$

$$P(martian/Doc1) = 0.5\frac{1}{11} + 0.5\frac{2}{21} = 0.09$$

ii. (10 pts) Under a bigram language model, what are $P(\text{sensation}|\text{pop}, \text{Doc1})$ (interpreted as "sensation" follows "pop" in the text), and $P(\text{pop}|\text{the}, \text{Doc1})$? Do not remove stop words. Use Jelinek-Mercer smoothing, including backoff to unigram probabilities, and $\lambda_1 = \lambda_2 = \lambda_3 = \frac{1}{3}$.

$$P_{smoothed}(sensation/pop, D1) = \lambda_1 \cdot P(sensation/pop, doc1) + \lambda_2 \cdot P(sensation/doc1) + \lambda_3 \cdot P(sensation/collection) =$$

$$\frac{Count(\text{"sensation"}, \text{"pop"})}{Count(\text{"pop"})}$$

$$= \frac{1}{3} \cdot \frac{1}{1} + \frac{1}{3} \cdot \frac{1}{11} + \frac{1}{3} \cdot \frac{1}{21} = 0.37$$

$$P_{smoothed}(pop/the, doc1) = \lambda_1 \cdot P(pop/the, doc1) + \lambda_2 \cdot P(pop/doc1) + \lambda_3 \cdot P(pop/collection) =$$

$$= \frac{1}{3} \cdot \boxed{0} + \frac{1}{3} \cdot \frac{1}{11} + \frac{1}{3} \cdot \frac{1}{21} = 0.04$$

"pop" doesn't follow "the" in doc1

4. **(30 pts) Naive Bayes**

Based on the data in the table below: (i) (15 pts) how is the test document classified under a multinomial NB classifier?: and (ii) (15 pts) how is the test document classified under a multinomial NB classifier that uses only bigrams? You do not have to estimate parameters that you don't need for classifying the test document, but show all relevant work! For all classifiers use "add-one-smoothing".

	DocID	Words in document	in $c = China$
training set	1	Taipei Taiwan Taiwan	yes
	2	Macao Taiwain Shanghai	yes
	3	Sapporo Sapporo	no
	4	Japan Osaka Taiwan	no
test set	5	Taiwan Taiwan Sapporo	?

$$B = 7$$

$$B_{bigram} = 7$$

$$P(China) = \frac{1}{2}, \quad P(\overline{China}) = \frac{1}{2}$$

(i)

$$P(Taiwan / China) = \frac{3+1}{6+7} = 0.30$$

$$P(Taiwan / \overline{China}) = \frac{1+1}{5+7} = 0.16$$

$$P(Sapporo / China) = \frac{0+1}{6+7} = 0.07$$

$$P(Sapporo / \overline{China}) = \frac{2+1}{5+7} = 0.25$$

$$\left.\begin{aligned} P(D5 / China) &= \frac{1}{2} \cdot \left(\frac{4}{13}\right)^2 \cdot \frac{1}{13} = 0.0036 \\ P(D5 / \overline{China}) &= \frac{1}{2} \cdot \left(\frac{2}{12}\right)^2 \cdot \frac{3}{12} = 0.0034 \end{aligned}\right) \Rightarrow China!$$

(ii)

$$P(Taiwan\ Taiwan / China) = \frac{1+1}{4+7} = \frac{2}{11}$$

$$P(Taiwan\ Taiwan / \overline{China}) = \frac{0+1}{3+7} = \frac{1}{10}$$

$$P(Taiwan\ Sapporo / China) = \frac{0+1}{4+7} = \frac{1}{11}$$

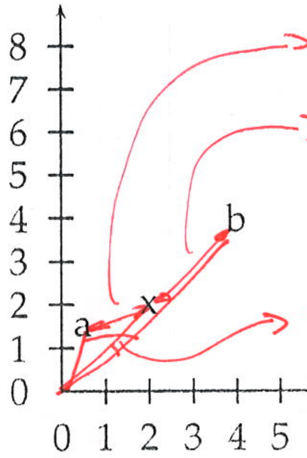
$$P(Taiwan\ Sapporo / \overline{China}) = \frac{0+1}{3+7} = \frac{1}{10}$$

$$\left.\begin{aligned} P(D5 / China) &= \frac{1}{2} \cdot \left(\frac{2}{11}\right)^2 \cdot \frac{1}{11} = 0.0015 \\ P(D5 / \overline{China}) &= \frac{1}{2} \left(\frac{1}{10}\right)^2 \cdot \frac{1}{10} = 0.0005 \end{aligned}\right) \Rightarrow China!$$

5. **(30 pts) Vector space classification**

In the figure below, which of the two vectors a and b is (i) (15 pts) most similar to x according to cosine similarity, and (ii) (15 pts) closest to x according to Euclidian distance? Show all work for computing these similarities and distances!

The vectors are: $a = (0.5 \;\; 1.5)$, $b = (4 \;\; 4)$, and $x = (2 \;\; 2)$.



$\to E(a,x)$

$\to E(x,b)$

visual explanation is OK!

$\to$ angles between x and a or b

$$||x|| = \sqrt{2^2+2^2} = 2.82$$

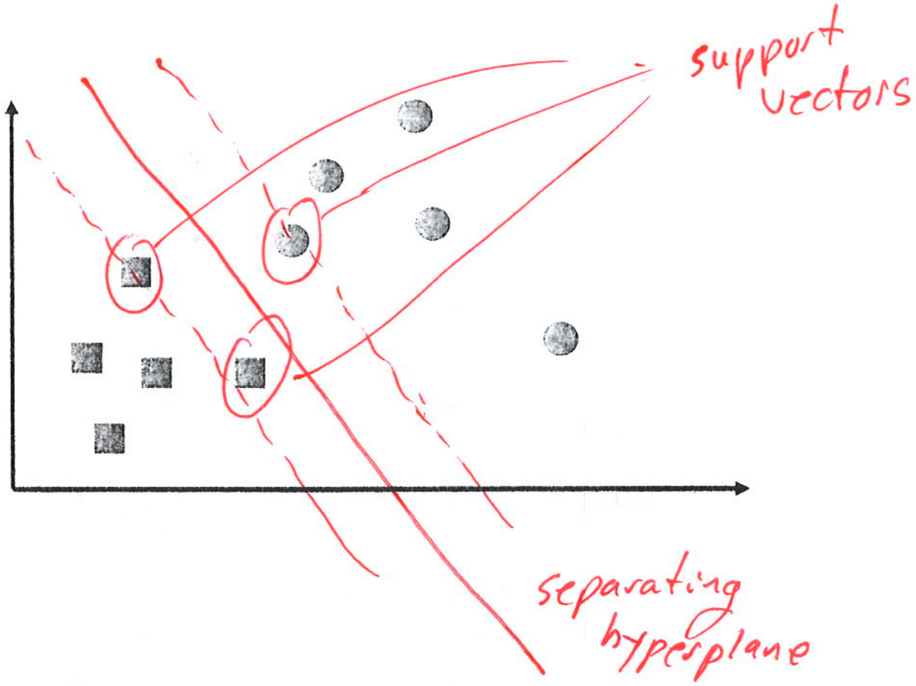
$$||a|| = \sqrt{0.5^2+1.5^2} = 1.58 \qquad ||b|| = \sqrt{4^2+4^2} = 5.65$$

$$\left.\begin{array}{l} \cos(a,x) = \frac{a \cdot x}{||a|| \cdot ||x||} = \frac{0.5 \cdot 2 + 1.5 \cdot 2}{||x|| \cdot 1.58} = \frac{2.53}{||x||} \\ \cos(a,b) = \frac{b \cdot x}{||b|| \cdot ||x||} = \frac{8+8}{5.65 \cdot ||x||} = \frac{2.81}{||x||} \end{array}\right) \Rightarrow b \text{ is closer!}$$

$$\left.\begin{array}{l} E(a,x) = \sqrt{(2-0.5)^2 + (2-1.5)^2} = 1.58 \\ E(b,x) = \sqrt{(4-2)^2 + (4-2)^2} = 2.82 \end{array}\right) \Rightarrow a \text{ is closer!}$$

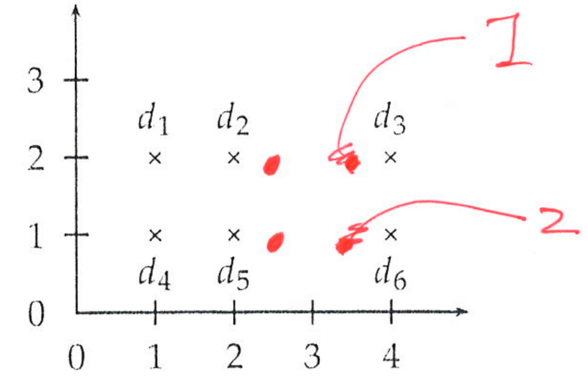
6. (30 pts) **Support Vector Machines**

In the figure below, please: (i) (15 pts) circle the support vectors for the two classes (squares and circles), and (ii) (15 pts) draw the optimal separating hyperplane as set by support vector machines between the two classes, as well as all necessary evidence to justify it.



7. **(30 pts) Clustering**

a. (15 pts) Compute the K-means centroids for the points in the figure below after one iteration, assuming that the initial centroids (randomly assigned) are d_3 and d_6. Show all work. Draw the resulting centroids in the figure.



$$\text{Centroid 1} = \frac{d_1+d_2+d_3}{3} = \left(\frac{4+2+1}{3}, \frac{2+2+2}{3}\right) = \left(\overset{2.3}{\cancel{\text{[illegible]}}}, 3\right)$$

$$\text{Centroid 2} = \frac{d_4+d_5+d_6}{3} = \left(\frac{4+2+1}{3}, 1\right) = \left(\overset{2.3}{\cancel{\text{[illegible]}}}, 1\right)$$

b. (5 pts) What are the centroids after the second K-means iteration? Explain your solution.

The centroids stay the same because no points are reassigned.

c. (10 pts) The K-means++ algorithm works by initializing the regular K-means algorithm with centroids that are assigned such that each is placed as far away as possible from the previously assigned one. How would you improve upon this idea?

Assign the next centroid as far away as possible from <u>all</u> previous centroids.

8. **(30 pts) Crawling**

a. (15 pts) Why is it better to partition hosts (rather than individual URLs) between the nodes of a distributed crawl system?

if you partition URLs, you risk having multiple crawlers connect to the same host at the same time. Violates the politeness requirement.

b. (15 pts) Which of these are must-have properties of a crawler (circle the correct answer(s)):

[✓ i.] Robustness to misleading sites such as spider traps.

ii. Language detection: the crawler must prioritize pages in English.

[✓ iii.] Politeness: a crawler should respect policies indicating the rate at which the crawler can visit pages.

iv. Variety: a crawler must fetch pages in multiple media: text, video, images, etc.

9. **(30 pts) Anchor text, for web and citation analysis**

a. (15 pts) Anchor text on the web: Given the collection of anchor-text phrases for a web page x, suggest a heuristic for choosing one term from this collection that is most descriptive of x.

Use a variant of tfidf:

↳ tf: most common term in anchortext

└ idf: computed on a large index, to identify most distinct words

rank by: tf x idf

b. (15 pts) Anchor text for citation analysis: Let's design a citation analysis similar to Google Scholar. The system will process all papers in a given field of Computer Science, say Algorithms. What are the nodes and the arcs in this citation graph? What does it mean that there is a directed arc from node a to node b? How would you extract the anchor text for this problem?

- nodes = papers
- arcs = citations
- directed arc from a to b = a cites b
- anchor text: text in the same clause ~~with the~~ with the actual citation

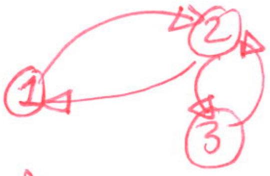
10. **(30 pts) Link analysis**

Consider a web graph with three nodes 1, 2 and 3. The links are as follows: 1 → 2, 3 → 2, 2 → 1, 2 → 3.

(i) (20 pts) Write down the transition probability matrices for the surfer's walk with teleporting, for the following three values of the teleport probability mass α: (a) $\alpha = 0$ (i.e., no teleporting); (b) $\alpha = 0.6$ (i.e., 60% of probability mass is reserved for teleportation probabilities), and (c) $\alpha = 1$.

(ii) (10 pts) For the configuration with $\alpha = 0.6$, compute the steady state vector after one iteration using the power method, assuming that at the beginning of the random walk the surfer is on node 2 with probability 1.0.

(i)



$\alpha = 0$

$$\begin{array}{c|c|c|c} & 1 & 2 & 3 \\ \hline 1 & 0 & 1 & 0 \\ 2 & 0.5 & 0 & 0.5 \\ 3 & 0 & 1 & 0 \end{array}$$

$\alpha = 0.6$

$$\begin{array}{c|c|c|c} & 1 & 2 & 3 \\ \hline 1 & 0.2+ & 0.2+0.4=0.6 & 0.2 \\ 2 & 0.2+0.2 & 0.2 & 0.2+0.2 \\ 3 & 0.2 & 0.2+0.4 & 0.2 \end{array}$$

$\alpha = 1$

$$\begin{array}{|c|c|c|} \hline 1/3 & 1/3 & 1/3 \\ \hline 1/3 & 1/3 & 1/3 \\ \hline 1/3 & 1/3 & 1/3 \\ \hline \end{array}$$

(ii) $x_0 = (0, 1, 0)$

$x_1 = x_0 \cdot P = (0.4, 0.2, 0.4)$

11. **(10 pts) Bonus question 1:** Modeling the *back button* in a browser

A user of a browser can, in addition to clicking a hyperlink on the page x he is currently browsing, use the *back button* to go back to the page from which he arrived at x. Can such a user of back buttons be modeled as a Markov chain? How would we model repeated invocations of the back button?

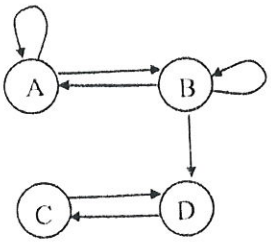
- add aprobability of going back, β.
- Must maintain an additional data structure, to keep track of the previous page visited.
- So: α – teleportation prob. mass
  β – "back button" probability
  $1-\alpha-\beta$ – mass for regular edges

→ implemented by the transpose of the original P matrix

12. **(10 pts) Bonus question 2 (recommended for grad students): Ergodic Markov chains**

Which of the two graphs below are ergodic? Justify your answer. For example, if a graph is not ergodic, indicate which property is violated, and where in the graph. If ergodic, indicate which properties are supported and why.

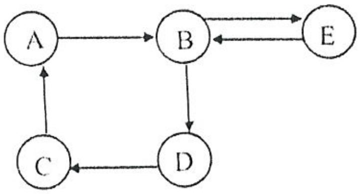
(i) (5 pts) Graph 1:



Not ergodic: no path going from D to B

not irreducible

(ii) (5 pts) Graph 2:



Not ergodic: the sets {B, C} and {A, D, E} are cyclically visited.

not aperiodic