Web Search and Mining (CS 635)
Computer Science and Engineering
Indian Institute of Technology Bombay

Midterm Exam
2013-09-13 Friday
17:30–19:30 LCH03

NAME ______________________ ROLL ______________________

This exam has 6 printed page/s. Write your name and roll number on **EVERY SIDE (and not just sheet)**, because we may take apart your answer book and/or xerox it for correction. Write your answer clearly within the spaces provided and on any last blank page. Do not write inside the rectangles to be used for grading. **If you need more space than is provided, you probably made a mistake in interpreting the question.** Start with rough work elsewhere, but you need not attach rough work. Use the marks alongside each question for time management. **Illogical or incoherent answers are worse than wrong answers or even *no* answer, and may fetch negative credit.** You may not use any computing or communication device during the exam. You may use textbooks, class notes written by you, approved material downloaded **prior to the exam** from the course Web page, course news group, or the Internet, or notes made available by me for xeroxing. If you use class notes from other student/s, you must obtain them **prior to the exam** and **write down his/her/their name/s and roll number/s** here.

1. We will investigate a connection between locality preserving hash functions and metric spaces, and see that not all reasonable similarity scores have corresponding locality preserving hash functions.

1(a) Suppose there is a similarity function $\text{sim}(a, b)$ defined in some domain, and a locality preserving hash function family $\mathcal{F}$ is available such that

$$\Pr_{f \in \mathcal{F}}(f(a) = f(b)) = \text{sim}(a, b).$$

Let $\Delta_f(a, b)$ be 1 if $f(a) \neq f(b)$, and 0 otherwise. Complete the following by inserting one of $\leq, <, =, >, \geq$, and justify your choice.

$$\forall a, b, c: \qquad \Delta_f(a, b) + \Delta_f(b, c) \quad \_\_\_ \quad \Delta_f(a, c)$$

		1

Given $\Delta \in \{0, 1\}$, the lhs can be 0, 1, or 2 and the rhs can be 0 or 1. For lhs < rhs to be possible, lhs must be 0 and rhs must be 1. Can we have $\Delta_f(a, b) = \Delta_f(b, c) = 0$ but $\Delta_f(a, c) = 1$? This is equivalent to asking if $f(a) = f(b)$ and $f(b) = f(c)$ are simultaneously possible with $f(a) \neq f(c)$. Therefore,

$$\forall a, b, c: \qquad \Delta_f(a, b) + \Delta_f(b, c) \quad \geq \quad \Delta_f(a, c)$$

1(b) If $J(a,b) = |a \cap b|/|a \cup b|$ is the Jaccard similarity defined on sets a, b, using the above equality or inequality, either prove that the distance measure $1 - J(a,b)$ satisfies the triangle inequality $1 - J(a,b) + 1 - J(b,c) \geq 1 - J(a,c)$ for all a, b, c, or give a simple counter-example.

		2

It is easy to verify that $\Pr(\Delta_f(a,b) = 1) = \mathbb{E}_{f \in \mathcal{F}}(\Delta_f(a,b)) = 1 - J(a,b)$. Therefore,

$$\begin{aligned} \forall a, b, c: \quad 1 - J(a,b) + 1 - J(b,c) &= \mathbb{E}_{f \in \mathcal{F}}(\Delta_f(a,b)) + \mathbb{E}_{f \in \mathcal{F}}(\Delta_f(b,c)) \\ &= \mathbb{E}_{f \in \mathcal{F}}(\Delta_f(a,b) + \Delta_f(b,c)) \\ &\geq \mathbb{E}_{f \in \mathcal{F}}(\Delta_f(a,c)) = 1 - J(a,c). \end{aligned}$$

1(c) The Dice coefficient is similar to Jaccard, defined as

$$\text{Dice}(a,b) = \frac{2\,|a \cap b|}{|a| + |b|}$$

The overlap coefficient is defined as

$$\text{overlap}(a,b) = \frac{|a \cap b|}{\min\{|a|, |b|\}}$$

Using $a = \{1\}, b = \{2\}$, and a suitable (very simple) choice of c, argue that there can be no locality sensitive hash family for Dice and overlap coefficients.

		3

Choosing $c = \{1, 2\}$ establishes that neither $1 - \text{Dice}$ nor $1 - \text{overlap}$ satisfy triangle inequality. Therefore, they cannot have LSHF families.

2. We will extend minhash for Jaccard to the case of Jaccard similarity over weighted sets. This is strongly motivated by text applications and TFIDF weights.

2(a) First we review the unweighted Jaccard case. With a, b being sets, recall that we defined $\text{sketch}_\pi(a) = \min \pi(a)$. Instead of thinking about permutations, it may be easier in later parts of this exercise (and closer to actual code) to use a hash function f that is seeded with a random seed s (which can be an arbitrary bit sequence), and then maps any set element x to $[0,1]$ as a deterministic function $f_s(x)$. Over random choices of seed s, we want $f_s(x)$ to map to the uniform distribution $\mathcal{U}[0,1]$. We will ignore the possibility of ties in the real range $[0,1]$. Then we should define

$$\text{sketch}_s(a) = \arg\min_{x \in a} \underline{\qquad\qquad}$$

to ensure that $\Pr_s(\text{sketch}_s(a) = \text{sketch}_s(b))$ is the unweighted Jaccard similarity between a and b. (Complete with justification.)

		1

If there is no fear of collisions, hashing some number of items using $\mathcal{U}[0,1]$ is equivalent to a random permutation of those items on the number line. Therefore we have to define

$$\text{sketch}_s(a) = \arg\min_{x \in a} \underset{\sim\sim\sim}{f_s(x)}$$

2(b) Now let $a, b \in \mathbb{Z}_+^D$ be vectors with nonnegative integer elements. The *weighted Jaccard similarity* between a and b is defined as

$$J(a,b) = \frac{\sum_{d=1}^{D} \min\{a_d, b_d\}}{\sum_{d=1}^{D} \max\{a_d, b_d\}}.$$

Note that this generalizes the standard unweighted Jaccard similarity between sets, interpreted as the *characteristic vector* over sets, $d \in a \Leftrightarrow a_d = 1$ and $d \notin a \Leftrightarrow a_d = 0$. For the general case, we say the element d has *support* a_d in set a.

A crude way to generalize Jaccard is simply to make a_d copies of d, which we may call $(d,1),(d,2),\ldots,(d,a_d)$. We call this transformation of a (to effectively a multiset) as $M(a)$. Give a modified definition of $\text{sketch}_s(a)$ for this setting, so that $\Pr_s(\text{sketch}_s(a) = \text{sketch}_s(b))$ is now the *weighted* $J(a,b)$.

		2

Re-implement f_s to accept input d, t, where $d \in [1, D]$ and $t \in [0, a_d]$, and define

$$\text{sketch}_s(a) = \arg\min_{(d,t)\in M(a)} f_s(d,t)$$

2(c) Compare the number of invocations of f in the unweighted and weighted cases. Do you see a problem?

		1

To compute $\text{sketch}_s(a)$, f has to be invoked $\sum_d a_d$ times. This can be quite expensive if supports are large integers, because f must access pseudo-random functions.

2(d) Note that f can be any hash function that ensures these two key properties:

Uniformity: Over all random seeds s, the output of $\text{sketch}_s(a)$ is distributed uniformly over the epigraph $0 \le t \le a_d$. I.e., if a were drawn as a histogram over d as x-axis and a_d as y-axis, we sample a point uniformly from the area under the "curve" of a.

Consistency: Suppose b dominates a, i.e., $a_d \le b_d\ \forall d$. Given a seed s, suppose we draw $\text{sketch}_s(b)$ as (d,t), which satisfies not only $t \le b_d$ but also $t \le a_d$. Then $\text{sample}_s(a)$ would always return (d,t) as well.

Now think of all the replicates of d, i.e., $(d,1),(d,2),\ldots$ as being successively hashed by f. Suppose we get $f.(d,t) = r$. Each of $(d,t+1),(d,t+2),\ldots$ hashes to a value larger than r with probability $1-r$. Write down the distribution $g(\cdot)$ over the number of replicates that will hash to a value larger than r before a replicate hashes to a value smaller than r.

		2

g represents the geometric distribution corresponding to a Bernoulli trial with success probability r. If the (random) number of replicates giving hashes greater than r (i.e., failure) is K, then we have $\Pr(K = k) = (1 - r)^k r$ for $k = 0, 1, \ldots$

2(e) Using the above observation, complete the following pseudocode, with adequate justification, to find, for each d, that t for which $f_s(d, t)$ is smallest.

$i \leftarrow 0, r \leftarrow 1$
while $i \le a_d$ **do**
  seed distribution g using ______
  invoke g to get next *skip*
  answer $\leftarrow i$
  $i \leftarrow$ ______
  seed distribution $\mathcal{U}[0, 1]$ with ______
  invoke $\mathcal{U}[0, 1]$ to get *shrink*
  $r \leftarrow r$ ______
return (d, answer)

(Keep in mind that $f_s(d, t)$ is deterministic in s, d, t. Therefore, each invocation may need a different s to retain apparent randomness of output, while also ensuring consistency.)

4

The key is proper seeding to ensure the two required properties. Here is one solution, others are possible.

$i \leftarrow 0, r \leftarrow 1$
while $i \le a_d$ **do**
  seed distribution g using s, d, i
  invoke g to get next *skip*
  answer $\leftarrow i$
  $i \leftarrow i + \mathit{skip} + 1$
  seed distribution $\mathcal{U}[0, 1]$ with s, d, i
  invoke $\mathcal{U}[0, 1]$ to get *shrink*
  $r \leftarrow r \times \mathit{shrink}$
return (d, answer)

For the last item, observe that, given a uniform distribution has generated a value at most r, we can generate such as value by multiplying r with the result of invoking $\mathcal{U}[0, 1]$.

2(f) Roughly how many times will the loop execute for each d? Give a heuristic argument.

1

In expectation, r is halved every iteration. So the expected skip doubles. Therefore, the loop executes about $\log a_d$ times, which is exponentially better than the first attempt.

3. Text documents generally contain mentions of entities. E.g., in the sentence "Albert played the violin", *Albert* may be a mention of Einstein, the Physicist, or John Albert, the violin maker from Philadelphia. Some tokens like *Albert* and perhaps *violin* are mention tokens, and the rest are content tokens. (The classification depends on the universe of entities known to us.)

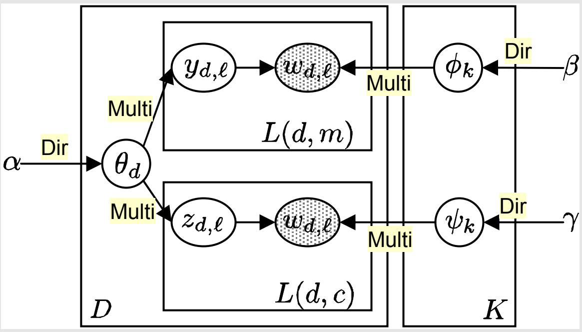
3(a) Initially, we will assume that the segmentation of documents $d = 1, \ldots, D$ into mention and content tokens is known. Suppose document d has $L(d, m)$ mention tokens and $L(d, c)$ content tokens.

Each entity $k = 1, \ldots, K$ in our entity catalog (say, Wikipedia) will be regarded as a "topic". Associated with each topic will be two multinomial models for words, one generating mention words (parameters $\phi_{k,w}$), and the other generating content words (parameters $\psi_{k,w}$). These will be generated from global Dirichlet priors with parameters β and γ respectively. Given the topic, we assume that mention and content words are independent of each other.

Naturally, given a document may mention many entities, it must be modeled as a multi-topic document. Each document will have an associated multinomial distribution over topics, with parameters θ_d generated from a Dirichlet prior having global parameters α. Each word position in the document will have an associated hidden topic $y_{d,\ell}$ or $z_{d,\ell}$ according as the position has a mention or context word. The word itself is denoted $w_{d,\ell}$.

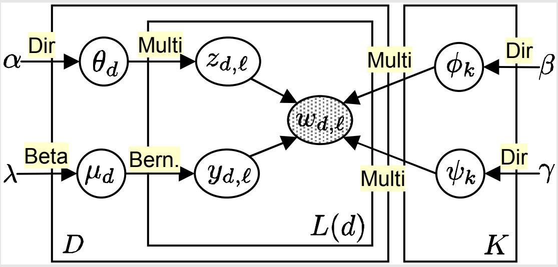
Draw a plate diagram for the generative process of the whole corpus, and label plates, nodes and edges completely with all necessary information.

| | | 2 |



3(b) Next we will remove the unrealistic assumption that the segmentation is known. Let document d have length $L(d)$. Each token makes a binary decision $y_{d,\ell}$ to be a mention or content token by tossing a coin with mention probability μ_d, generated for each document using a global beta prior with parameters λ. $z_{d,\ell}$ is the hidden topic used to generate the word at position ℓ of document d. Other symbols retain their earlier meanings. Draw a modified plate diagram for the new setting.

| | | 3 |



3(c) Based on the second plate diagram, write down the probability of generating a document $w_1, \ldots, w_L$ using the second model. Only $\alpha, \beta, \gamma, \lambda$ should be free in your final expression, and all other variables should be suitably marginalized or aggregated. Specifically, write out the full form of $\Pr(w_{d,\ell} | z_{d,\ell}, y_{d,\ell}, \_\_\_\_)$. But you need not expand known forms for beta, Dirichlet or other distributions.

		3

$$\begin{aligned}
&\Pr(w_1, \ldots, w_L | \alpha, \beta, \gamma, \lambda) \\
&\quad = \int_\theta \int_\mu \int_{\vec{\phi}} \int_{\vec{\psi}} \Pr(\theta|\alpha) \Pr(\mu|\lambda) \left[\prod_{k=1}^{K} \Pr(\phi_k|\beta) \Pr(\psi_k|\gamma) \right] \\
&\qquad \left[\prod_\ell \sum_{y_\ell} \sum_{z_\ell=1}^{K} \Pr(z_\ell|\theta) \Pr(y_\ell|\mu) \Pr(w_\ell|z_\ell, y_\ell, \phi, \psi) \right] d\theta \, d\mu \, d\vec{\phi} \, d\vec{\psi}
\end{aligned}$$

$$\Pr(w|z, y, \phi, \psi) = \begin{cases} \phi_{z,w} & y = \text{mention} \\ \psi_{z,w} & y = \text{content} \end{cases}$$

Total: 25