Organizing Web Information (CS 728) **Final Exam**
Computer Science and Engineering **2014-04-31 Wednesday**
Indian Institute of Technology Bombay **17:30–20:30 SIC201**

NAME _____ ROLL _____

This exam has 5 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. When discussing locality sensitive hash functions (LSHFs) for record linkage, we considered only data-agnostic hash functions such as minhash or the sign of the dot product with random vectors. Here we will develop data-driven LSHFs. N input instances, each a vector in $\mathbb{R}^D$, are together written as $X \in \mathbb{R}^{N \times D}$. We will learn K 1-bit hash functions

$$h_k(x_n) = \text{sign}(w_k \cdot x_n + b_k), \text{ where } k \in [1, K],\ n \in [1, N],$$

$w_k \in \mathbb{R}^D$ are suitably chosen to satisfy some desirable properties of LSHFs, and b_k is fixed afterward to ensure $\sum_n (w_k \cdot x_n + b_k) = 0$. We will assume the data is centered to zero mean, i.e., $\sum_n x_n = \vec{0}$, in which case each $b_k = 0$. A small fraction of instances are involved in pairwise supervision: some pairs (x_i, x_j) are labeled $y_{ij} = +1$ (respectively, -1), meaning hash functions should ideally map them to the same (respectively, opposite) value(s).

1.a Complete the following objective to maximize wrt $\{w_k\}$ the number of pairwise preferences that are respected:

$$J(\{w_k\}) = \sum_k \left[\sum_{i,j:y_{ij}=+1} h_k(x_i) \_\_\_\_\_ - \sum_{i,j:y_{ij}=-1} \_\_\_\_\_ h_k(x_j) \right].$$

		1

1.b Suppose, out of N instances, n_+ have $h_k(\cdot) = +1$ and $n_- = N - n_+$ have $h_k(\cdot) = -1$. Then the

$$\text{mean } \mu(h_k) = \frac{1}{N}\sum_n h_k(x_n) = \frac{\qquad - 2\qquad}{\qquad} \quad \text{and}$$

$$\text{variance } \sigma^2(h_k) = \frac{4}{\qquad}(\qquad N - \qquad).$$

		3

1.c Another desirable feature for each h_k is that it is *balanced*, i.e., splits the instances into two equal halves (assume N is even): $\sum_n h_k(x_n) = 0$. When h_k is balanced, what property does $\sigma^2(h_k)$ satisfy?

		1

1.d The above objective J is discontinuous, so we replace the sign function with signed magnitude. This gives us the relaxed objective:

$$\sum_k \left[\sum_{i,j:y_{ij}=+1} \underbrace{\qquad\qquad\qquad\qquad}_{(A)} - \sum_{i,j:y_{ij}=-1} \text{(same as expression A)} \right]$$

		2

1.e Is there a need to augment the above objective with a regularization term involving w_k, or can we simply force $\|w_k\|_2 = 1$ for all k? Justify.

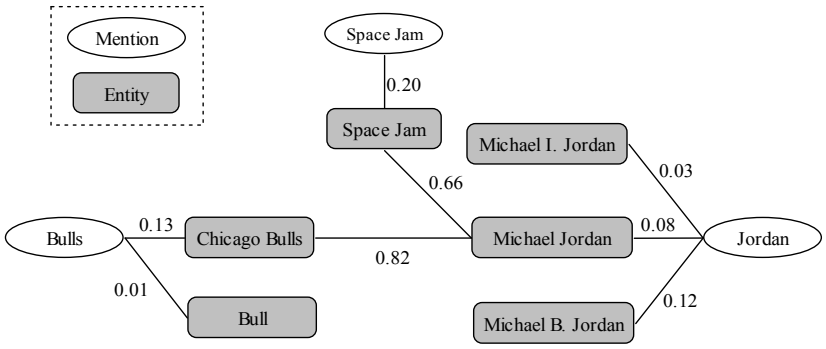
		2

1.f Suggest how to enhance the relaxed objective to encourage the balance constraint.

		1

1.g (Extra credit — answer on a separate sheet of paper.) The above optimization still misses a very important property we want $\{h_k\}$ to satisfy collectively — what is it? How would you further modify the optimization to encourage this property?

2. Consider a collective entity disambiguation task within one document. Each mention phrase is represented by a node m in a graph (shown as oval nodes in the accompanying diagram). A mention node m is connected to nodes representing candidate entities e (shown as shaded rectangles). The edge (m, e) is assigned a non-negative weight $w(m, e)$ based on the compatibility between e and the mention context of s. NA or $\perp$ is not explicitly represented as a node. A pair of entities are connected to each other with an edge (e_1, e_2), which has a non-negative symmetric weight $w(e_1, e_2)$ based on the semantic relatedness between e_1 and e_2. Edge weights may not be interpretable as probabilities.



Nodes have two states. A mention that occurs in the document sets the corresponding node to "on". (There might be some benefit to counting the number of times the mention occurs.) An entity node is "on" if it is mentioned in the document, and "off" otherwise. Entity e may have other possible mentions m' that did not occur in the document, which are represented by "off" mention nodes. The goal of inference is to estimate the state of all entity nodes, given the state of all mention nodes.

We can keep the problem combinatorial and hard by keeping entity node states binary. We can even enforce additional constraints, e.g., at most one entity neighbor of each mention node can be on. Here, we are interested in relaxing these constraints to get tractable approaches.

2.a Suppose we are interested in only *ranking* the entity nodes as against classifying them as "on" and "off". Propose and defend a method based on personalized PageRank. Be sure to define the teleport vector and explain why it will have the intended effect.

		1

2.b Another ranking approach is to associate a real score x_e, x_m with each entity and mention node, pin the mention scores x_m (how?), and solve for the entity scores x_e using the graph Laplacian. Recall that the Laplacian smoothness of a node score vector $\vec{x}$ wrt edge weights $w(i,j)$ is defined as $\sum_{(i,j)\in E} w(i,j)(x_i - x_j)^2$. Furnish a detailed objective and solution approach. (Note: what is the counterpart of teleport here?)

		4

2.c Instead of fixing $w(m,e)$ and $w(e,e')$, we can make them feature-dependent. Suppose $f(m,e)$ is a feature vector capturing various aspects of compatibility between e and the context of m. Then we can let $w(m,e) = \lambda \cdot f(m,e)$. Similarly, let $g(e,e')$ be a vector of features, each capturing some notion of relatedness between e and e', and let $w(e,e') = \mu \cdot g(e,e')$. Here λ, μ are model vectors to be trained. Supervision is provided in the form of documents, along with the values of all relevant x_e^*s. Furnish the following details to learn λ and μ (there may be no unique or best approach):

- Give a per-node (itemwise classification) and a node-pairwise (RankSVM-like) loss function.
- How would you regularize and possibly constrain λ and μ? Why?
- The overall optimization with objective and all constraints.

		5

3. Consider a correlation clustering (CC) problem instance (V, E, W, L) specified by a weighted, undirected graph (V, E) with $|V| = n$ nodes. Edge (u, v) has a *label* $\ell(u, v) \in \{-1, +1\}$. If $\ell(u, v) = +1$ (respectively, -1), that means we prefer nodes u and v to be in the same (respectively, different) cluster(s), and the strength of the preference is given by the edge weight $w(u, v) > 0$. If there is no edge between u and v, that means we have no preference about u and v being in the same or different clusters. Given a clustering, let $C(u)$ be the nodes in the same cluster as u. The objective of clustering is to minimize the sum of weights of mistaken or violated edges:

$$\sum_{e(u,v)=-1, u \in C(v)} w(u, v) + \sum_{e(u,v)=+1, u \notin C(v)} w(u, v).$$

A *perfect clustering* makes no mistakes.

3.a An *erroneous cycle* is a simple cycle (no repeated nodes) with exactly one edge labeled -1. If a graph has an erroneous cycle, can it have a perfect clustering? If a graph has a perfect clustering, can it have an erroneous cycle? Justify both answers.

		1

3.b Complete the following with informal justification: The total weight of edges violated by an optimal clustering is the minimal weight set of edges whose removal will eliminate _____ in the graph.

		1

3.c In graph theory, the multicut problem (V', E', W', P) is specified as follows. We are given a weighted, undirected graph (V', E', W') and a collection of $K = |P|$ pairs of distinct source and sink nodes $P = \{(s_k, t_k), k = 1, \ldots, K\}$. (Note that edges do not have labels.) The goal is to find a minimum weight set of edges whose removal disconnects every s_k from the corresponding t_k. The problem is hard for $K \geq 3$ and the best-known approximation factor for general graphs is $O(\log K)$.

Give a simple polynomial-time and -space transformation of a CC instance (V, E, W, L) to a multicut instance (V', E', W', P), to be solved by a multicut algorithm. Show that the CC instance can thereby be solved to within a $O(\log n)$ factor of optimal. (Hint: The multicut instance cannot have edge labels, so something must be done about either positive or negative edges.)

		4

3.d The above transformation employs a multicut algorithm to solve CC, but leaves open the possibility that CC may be easier than multicut. Given a multicut instance (V, E, W, P), give a polynomial-time and -space transformation to a CC instance $(\widehat{V}, \widehat{E}, \widehat{W}, L)$ such that

- If a CC solver returns with objective o, then there is a multicut with weight at most o.
- An optimal CC solution can be mapped back to an optimal multicut solution efficiently.

		4

4. We will develop a model for querying and clicking on Web pages based on the query words and the location of the user (but ignoring page contents). There are K topics. Globally, there is a Dirichlet hyper-prior over topics with parameters $\alpha \in \mathbb{R}^K$. Suppose we model D pages. For each page $d \in [1, D]$, first $\text{Dir}(\alpha)$ is invoked to generate a multinomial distribution over topics with parameters $\theta \in \mathbb{R}^K$. On each page there are F query+click events. The fth such event is characterized by exactly one topic z, starts with a query having Q words $w_1, \ldots, w_Q$, and then a click c that is characterized by the coordinates (latitude and longitude) of the user. All Q words are generated iid from a per-topic multinomial distribution with parameters $\beta \in \mathbb{R}^{K \times W}$. Location distributions are captured by L means (centers) $\mu_\ell \in \mathbb{R}^2$ and covariances (spreads) $\Sigma_\ell \in \mathbb{R}^{2\times 2}$, for $\ell \in [1, L]$. Denote μ_ℓ, Σ_ℓ together as ϕ_ℓ. A topic (like string theory) may have multiple centers. A center may have diverse spread (e.g., IPL vs. string theory). On choosing the topic z, a per-topic multinomial prior over locations, $\pi \in \mathbb{R}^{K \times L}$, is used to choose a center and spread, which is then used to generate c.

4.a Draw a complete plate diagram for the generative story, labeling all distribution names, dimensions of parameters, etc.

		3

4.b Complete the following probability expression for the observed data given the parameters, giving brief explanation:

$$\Pr\left(\left\{(w_{df}, c_{df})_{f=1}^{F}\right\}_{d=1}^{D} \middle| \alpha, \beta, \pi\right) = \prod_{d=1}^{D}\left[\sim\sim \prod_{f=1}^{F}\left(\sim\sim \left[\prod_{q=1}^{Q} \sim\sim\right] \Pr(\ell_{df}|\pi_{z_{df}}) \sim\sim \right)\right]$$

		4

4.c Given a new query, write down the high-level roadmap and expressions to estimate the posterior distribution over the user location. If the meaning is clear, there is no need to simplify the expressions.

		2

4.d Instead of mapping a hard topic decision z to a prior over locations, we could extend the dimensionality of α and θ to $W + L$, and draw two multinomials from it (for each f), one generating the query words and the other generating the click. What would be a potential limitation of this more symmetric approach?

		1

Total: 40