

Assignment 4: Markov Networks and Related Models

CS486/686 – Fall 2008

Out: November 18, 2008

Due: December 1, 2008

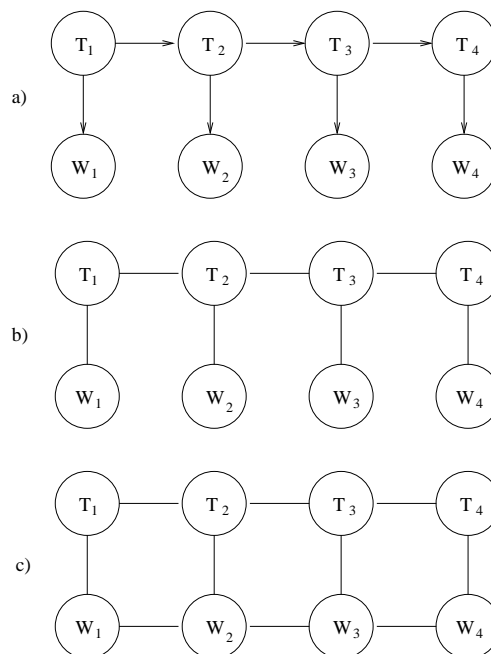
Be sure to include your name and student number with your assignment.

1. [64 pts] Part-of-speech tagging

Part-of-speech tagging is a basic task in natural language processing that consists of tagging each word in a text with a part of speech. Common parts of speech include noun, verb, adjective, adverb, article, preposition and conjunction. For example, the sentence below is tagged with corresponding parts of speech.

The mountain is high.
article noun verb adjective

Consider the three graphical models below that could be used for part-of-speech tagging. In each graph, the top nodes correspond to part-of-speech tags and the bottom nodes to words. Graph a) is a hidden Markov model, graph b) is a corresponding Markov network and graph c) is a slightly more complex Markov network. The task of part-of-speech tagging corresponds to computing $\max_{T_1, T_2, \dots, T_n} \Pr(T_1, T_2, \dots, T_n | W_1, W_2, \dots, W_n)$



- (a) [8 pts] Show the parameterization of each model. More precisely, indicate the form of the conditional distributions, potentials or features that must be defined to specify each model. Indicate how these conditional distributions, potentials or features are used to specify the joint distribution of each model.

- (b) **[8 pts]** Indicate how many distinct parameters are required to specify each model. Assume that the W_i variables can take W values (e.g., W words) and the T_i variables can take T values (e.g., T tags).
- (c) **[8 pts]** For each pair of graphical model indicate whether one model subsumes the other one. Here a graphical model subsumes a second graphical model when all the joint distributions that can be encoded by the second graphical model can also be encoded by the first graphical model.
- (d) **[8 pts]** Describe an advantage and disadvantage of each graphical model for part-of-speech tagging.
- (e) **[8 pts]** Since the sequence of words is always observed in part-of-speech tagging and we are looking for a function that maps words to part-of-speech tags, one could consider graphs b) and c) as conditional random fields and learn only the conditional distribution $\Pr(T_1, T_2, \dots, T_n | W_1, W_2, \dots, W_n)$. Describe the form of the potentials or features that are necessary to encode the conditional random fields in graphs b) and c). Indicate how these potentials or features are used to specify the conditional distribution of each model.
- (f) **[8 pts]** Describe an advantage of conditional random fields over their corresponding Markov networks for part-of-speech tagging.
- (g) **[8 pts]** The Markov networks in b) and c) can also be encoded as Markov logic networks. List the predicates and first-order formula that encode each Markov logic network using the Alchemy notation.
- (h) **[8 pts]** Describe an advantage of Markov logic networks over their corresponding Markov networks for part-of-speech tagging.

2. **[36 pts]** Collective text categorization

Text categorization is a common task in information retrieval. In its simplest form, text categorization may be done by using words as independent features. However, documents are much more rich and additional features may be used. To that effect, it is desirable to define richer graphical models that can exploit additional features. For instance, web pages link to each others and these links encode relations between documents that may be indicative of their category. Other information may include the author of each document and his/her affiliation (e.g., company, university, etc.).

Consider the following Markov logic network for text categorization encoded using the Alchemy notation. The first rule indicates that each word may influence the class of a page. The second rule indicates that documents joined by a link are likely to have the same topic.

- Predicates
 - `HasWord(word, page)`
 - `Topic(class, page)`
 - `LinkTo(linkid, page, page)`
- Rules
 - `HasWord(+w, p) => Topic(+c, p)`
 - `Topic(c, p1) ^ LinkTo(id, p1, p2) => Topic(c, p2)`

- (a) **[8 pts]** Suppose we have a vocabulary of W words, a corpus of P web pages, L links between these pages and C classes. How many nodes would the corresponding grounded Markov network have? In your opinion, is it worthwhile to use a Markov logic network to compactly encode this grounded Markov network?
- (b) **[8 pts]** Suppose that we use the Alchemy package to learn the weights of this Markov logic network. How many distinct weights would be learned?
- (c) **[8 pts]** Add another rule to the above Markov logic network to encode that two pages that have a link pointing to the same page are likely to have the same class. Similarly, add another rule to encode that two pages pointed to by links from the same page are likely to have the same topic.
- (d) **[12 pts]** Create additional rules to use the author of each page and the affiliation (e.g., university, company, etc.) of each author to improve the categorization of each page. This is an open question that has many good answers.