

IT655: Advanced topics in Data mining, EndSem exam

Sep 21, 2004.

2:00 PM - 5:00 PM

Roll: _____

Name: _____

Write all your answers in the space provided and use the last sheet for rough work. Do not spend time/space giving irrelevant details or details not asked for. Use the "marks" as a guideline for the amount of time you should spend on a question. You are allowed extra sheet only under special circumstances like total cancellation of a previously written answer. You are allowed to refer your class notes and the papers covered in the class. If your solution applies only if certain assumptions are made, state those assumptions clearly.

1. Given a trained HMM M , we studied how to compute the total probability $\Pr(\mathbf{x})$ of generating a sequence $\mathbf{x} = x_1, x_2, \dots, x_n$ from M using a recursive formula. Design a (recursive) formula for computing the probability $\Pr(s_i = q | \mathbf{x})$ of the state from which the i element (x_i) is generated is q

$$\begin{aligned}\Pr(s_i = q | \mathbf{x}) &= \frac{\Pr(s_i = q, \mathbf{x})}{\Pr(\mathbf{x})} \\ \Pr(s_i = q, \mathbf{x}) &= \Pr(s_i = q, \mathbf{x}_{1..i}) \Pr(\mathbf{x}_{i+1..n} | s_i = q) \\ &= \alpha(i, q) \beta(i+1, q)\end{aligned}$$

$$\beta(j, q) = \begin{cases} \sum_{q' \in \mathcal{S}} a_{qq'} b_{q'}(x_{j+1}) \beta(j+1, q') & \text{if } j < n \\ 1 & \text{if } j = n \end{cases}$$

2. List all the dictionary features activated in the Borthwick's model when you have a dictionary consisting of two entries "Useless bombastic architectures", "Useful", on each word of the phrase "Achitechtures can be bombastic yet not useful". Assume we have just two labels: "company_name" and "other" and dictionary match is case-insensitive.

Let,

$f_{F,l}(y_i, x, i, y_{i-1}) = 1$ if $l = y_i$ and x_i is the *first* word of a dictionary, = 0 otherwise.

$f_{C,l}(y_i, x, i, y_{i-1}) = 1$ if $l = y_i$ and x_i is the continuing word of a dictionary, = 0 otherwise.

$f_{E,l}(y_i, x, i, y_{i-1}) = 1$ if $l = y_i$ and x_i is the ending word of a dictionary, = 0 otherwise.

$f_{U,l}(y_i, x, i, y_{i-1}) = 1$ if $l = y_i$ and x_i is the unique word of a dictionary, = 0 otherwise.

The set of features that get a value of 1 are then $f_{E,c}(c, x, 0, y_{i-1})$, $f_{C,c}(c, x, 3, y_{i-1})$, $f_{U,c}(c, x, 6, y_{i-1})$ for $c =$ all possible labels from the set

{ "company_name_start", "company_name_continue", "company_name_end", "company_name_unique", otl

Note, $x_0 =$ "architectures" and so on.

3. In the Takeuchi et al paper, at each token a single class is chosen as label from the SVM ensemble. This is in contrast to a maximum entropy tagger where we estimate $\Pr(y_i|\mathbf{x}, i, y_{i-1})$ for each possible y_i and find the best sequence using Viterbi. Can you think of a way of combining the two methods?

Design a maximum entropy classifier with features as the score or label of the winning class from each one-vs-one classifier.

4. Assuming the same way of breaking up tags as in the Takeuchi paper, how many features does your model above have when the number of entity types is 3 (DATE, MONEY, PERSON)?

In this model there are $2*3+1=7$ labels. This means there are $7*6/2 = 21$ one-vs-one SVM outcomes. There is a feature for each of the outcomes on each label giving rise to $21*7=147$ features.

5. Assume after training your CRF model, you replicate one feature f to create another f' . Comment with reasons on the weights of f and f' on the following two scenarios?

- (a) Retrain using the CRF training algorithm?

The weights of the new feature will be equal with most common numerical optimization routines. However, there are infinite solutions possible corresponding to any assignment of weights as long as the sum of the two weights is the same.

- (b) Retrain using the Collins Perceptron algorithm, using the starting weights as those obtained from the first training for all except f' for which the starting weight is 0. Assume first training reached convergence with no error.

The weight of w' will remain at 0 since the previous solution converged with zero error and no update will be performed.

6. Suppose you have a VMM with states “A”, “AC”, “BC” and “B”, “ACC”. Factorize $\Pr(AACABB)$ in terms of the probabilities expressed in this VMM. For example, $\Pr(BCC) = \Pr(B|start) \Pr(C|B) \Pr(C|BC)$

$\Pr(A|start) \Pr(A|A) \Pr(C|A) \Pr(A|AC) \Pr(B|A) \Pr(A|A)$

7. Write an expression for the modified Viterbi algorithm for sequential CRFs that can capture a constraint that a particular label c cannot appear more than k times in a sequence.

$$\delta(i, y, l) = \begin{cases} \max_{y'} \delta(i-1, y', l - \llbracket y' == c \rrbracket) + \mathbf{W} \cdot \mathbf{f}(y, y', \mathbf{x}, i) & \text{if } i > 0 \\ 0 & \text{if } i = 0 \end{cases} \quad (1)$$

In the above equation $\llbracket a \rrbracket$ denotes a function that is 1 when the condition a is true and 0 otherwise. The best label then corresponds to the path traced by $\max_y \max_{l \leq k} \delta(|\mathbf{x}|, y, l)$.

8. What is the space required by the above algorithm?

$O(mnk)$ where m =number of labels, n = length of sequence.

9. In hierarchical text classification, you are given a tree of labels (like in Dmoz and Yahoo) and a given document is to be assigned to a leaf of the tree. Map this prediction task to a sequential labeling problem. You can assume a tree with all leaves at the same level. Clearly state the set of random variable(s) for a given document x , the probability distribution and the set of features.

Define a sequence where each level of the hierarchy is a sequence position. Thus, all sequences have the same length h = height of the tree. For any document we have a label sequence $\mathbf{y} = y_1 \dots y_h$ corresponding to labels at each position of the hierarchy. The correspond \mathbf{x} sequence is the same for each position. The set of values that each y_i can take is the set of labels at the i -th level of the hierarchy.

The features are as follows.

- For each (label, term) pair there is a feature.
- For each edge in the hierarchy tree define a feature.

Further, during Viterbi we can restrict to only those label sequences that confirm to the subsumption relationship conveyed by the hierarchy.

10. All three link-based classification papers ignore the textual similarity of a page with its neighbors. Can you design two features that can capture this useful information in the right way?

Let n be a neighbor of the i -th page.

- $f_s(y_i, y_n, \mathbf{x}) = \text{similarity}(x_i, x_n)$ if $y_i = y_n$
- $f_d(y_i, y_n, \mathbf{x}) = \text{similarity}(x_i, x_n)$ if $y_i \neq y_n$

11. Which of the papers covered so far did you like the most and why? (Be objective :))

The original CRF paper. This paper basically provided the last word on sequence labeling. It is a different matter that the writing sucks and therefore we read the “Shallow parsing” paper.: