6.864, Fall 2010: Problem Set 4

Total points: 125 points
Due date: 5 PM, Friday, November 12, 2010
Submit writeup to Tahira Naseem in 32-G362 or email to tahira@csail.mit.edu.
Late policy: 5 allowed late days over the course of the semester.

Question 1 (25 points)

In this question we will develop an algorithm, based on the EM algorithm, for modeling of topics underlying documents. In this model, the training sample $x^1, x^2, \ldots x^m$ is a sequence of m documents. We will take each document x^i to consist of n words, $x^i_1, x^i_2, \ldots x^i_n$. The hidden variables y in the EM approach can take one of K values, $1, 2, \ldots K$. The model is defined as follows:

$$P(x, y|\Theta) = P(y) \prod_{j=1}^{n} P(x_j|y)$$

Thus if $\mathcal{V}$ is the vocabulary—the set of possible words in any document—the parameters in the model are:

- $P(y)$ for $y = 1 \ldots K$
- $P(w|y)$ for $y = 1 \ldots K$ and $w \in \mathcal{V}$

Our aim in this question will be to derive EM updates which optimize the log-likelihood of the data:

$$L(\Theta) = \sum_{i=1}^{m} \log P(x^i|\Theta) = \sum_{i=1}^{m} \log \sum_{y} P(x^i, y|\Theta)$$

Give pseudo-code showing how to derive an updated parameter vector Θ^t from a previous parameter vector Θ^{t-1}. I.e., show pseudo-code that takes as input parameter estimates $P^{t-1}(y)$ for all y and $P^{t-1}(w|y)$ for all w, y, and as output provides updated parameter estimates $P^t(y)$ and $P^t(w|y)$ using EM. Use the notation $C(w, x)$ to denote the number of times word w is seen in document x.

Question 2 (25 points)

In this question we'll derive an EM approach to word clustering. In this model, the training sample $x^1, x^2, \ldots x^m$ is a sequence of m bigrams of the following form: each x^i is of the form w^i_1, w^i_2 where w^i_1, w^i_2 are words, and w^i_2 is seen following w^i_1 in the corpus. The hidden variables y can take one of K values, $1, 2, \ldots K$. The model is defined as follows:

$$P(w_2, y|w_1, \Theta) = P(y|w_1)P(w_2|y)$$

Thus if $\mathcal{V}$ is the vocabulary—the set of possible words in any document—the parameters in the model are:

- $P(y|w)$ for $y = 1 \ldots K$, for $w \in \mathcal{V}$
- $P(w|y)$ for $y = 1 \ldots K$ and $w \in \mathcal{V}$

Our aim in this question will be to derive EM updates which optimize the log-likelihood of the data:

$$L(\Theta) = \sum_{i=1}^{m} \log P(w_2^i | w_1^i, \Theta) = \sum_{i=1}^{m} \log \sum_y P(w_2^i | y) P(y | w_1^i)$$

Give pseudo-code showing how to derive an updated parameter vector Θ^t from a previous parameter vector Θ^{t-1}. I.e., show pseudo-code that takes as input parameter estimates $P^{t-1}(y|w)$ for all y, w and $P^{t-1}(w|y)$ for all w, y, and as output provides updated parameter estimates $P^t(y|w)$ and $P^t(w|y)$ using EM.

Question 3 (25 points)

In lecture we saw how the forward-backward algorithm could be used to efficiently calculate probabilities of the following form for an HMM:

$$P(y_j = p | x, \Theta) = \sum_{y : y_j = p} P(y | x, \Theta)$$

and

$$P(y_j = p, y_{j+1} = q | x, \Theta) = \sum_{y : y_j = p, y_{j+1} = q} P(y | x, \Theta)$$

where x is some sequence of output symbols, and Θ are the parameters of the model (i.e., parameters of the form π_i, $a_{j,k}$ and $b_j(o)$ as defined in the lecture). Here y_j is the j'th state in a state sequence y, and p, q are integers in the range $1 \ldots N-1$ assuming an N state HMM.

(a) State how the following quantity can be calculated in terms of the forward-backward probabilities, and some of the parameters in the model:

$$P(y_2 = 1, y_3 = 2, y_4 = 1 | x, \Theta)$$

(we assume that the sequence x is of length at least 4)

(b) State how the following quantity can be calculated in terms of the forward-backward probabilities, and some of the parameters in the model:

$$P(y_2 = 1, y_5 = 1 | x, \Theta)$$

(we assume that the sequence x is of length at least 5. Don't worry too much about the efficiency of your solution: we **do** expect you to use forward and backward terms, but we **don't** expect you to calculate any other quantities using dynamic programming.)

(c) Say that we now wanted to calculate probabilities for an HMM such as the following:

$$\max_{y : y_j = p} P(y | x, \Theta)$$

so this is the maximum probability of any state sequence underlying x, with the constraint that the j'th state y_j is equal to p.

How would you modify the definition of the forward and backward terms—i.e., the recursive method for calculating them—to support this kind of calculation? How would you then calculate

$$\max_{y : y_3 = 1} P(y | x, \Theta)$$

assuming that the input sequence x is of length at least 3?

Question 3 (50 points)

In this problem we will apply the EM algorithm to learn parameters of a *probabilistic edit distance* function. The training set consists of pairs of words $\{(w_1, w'_1), \ldots (w_m, w'_m)\}$. Each word is composed of a sequence of characters drawn from an alphabet Σ.

The hidden variables are a sequence of *operations* that produce a pair of words (w, w'). Each operation, denoted by z, can be one of the following:

- (c, c'): append c to w and c' to w'
- $(c, \boldsymbol{\epsilon})$: append c to w
- $(\boldsymbol{\epsilon}, c')$: append c' to w'
- #: stop producing characters

For example, for the pair of words *cat* and *fast*, two valid operation sequences are $(c, f), (a, a), (\boldsymbol{\epsilon}, s), (t, t), \#$ and $(c, \boldsymbol{\epsilon}), (\boldsymbol{\epsilon}, f), (a, a), (t, s), (\boldsymbol{\epsilon}, t), \#$.

The parameters of the model are the probabilities of each operation z:

- $P(c, c')$ for all $c, c' \in \Sigma$
- $P(c, \boldsymbol{\epsilon})$ for all $c \in \Sigma$
- $P(\boldsymbol{\epsilon}, c')$ for all $c' \in \Sigma$
- $P(\#)$

Let $\mathbf{z}$ be a sequence $(z_1, \ldots, z_m)$ of operations, $v_1(\mathbf{z})$ be the first word induced by sequence $\mathbf{z}$, and $v_2(\mathbf{z})$ be the second word. The probability of a pair of words and an operation sequence is then:

$$P(w, w', \mathbf{z} \mid \Theta) = \begin{cases} \prod_{i=1}^{|\mathbf{z}|} P(z_i) & \text{if } v_1(\mathbf{z}) = w \text{ and } v_2(\mathbf{z}) = w', \\ 0 & \text{otherwise.} \end{cases}$$

As usual, your goal is to find a maximum likelihood estimate of the parameters Θ.

(a) [5 points]

Assume that you have access to a function $f(w, w')$ that returns the set of all possible operation sequences $\mathbf{z}$ that produce the pair (w, w'). Write down the expression for the log-likelihood of the parameters $L(\Theta)$ in terms of $P(w, w', \mathbf{z} \mid \Theta)$ and your training sample $\{(w_1, w'_1), \ldots (w_m, w'_m)\}$.

(b) [10 points]

Again using function $f(w, w')$, write expressions for the expected counts of each of the parameters $\text{Count}^t(c, c')$, $\text{Count}^t(c, \boldsymbol{\epsilon})$, $\text{Count}^t(\boldsymbol{\epsilon}, c)$, and $\text{Count}^t(\#)$ in terms of previous parameter estimates $P^{t-1}(z)$.

If we naïvely applied the EM algorithm of the previous section, we would have to sum over exponentially many operation sequences for each word pair. For that reason, we will use a dynamic program with tables similar to the forward and backward tables for HMM to more efficiently compute expected counts.

For a pair of words $w = c_1 \dots c_p$ and $w' = c'_1 \dots c'_q$, define the following α (forward) and β (backward) probabilities:

$$\alpha(k, \ell) \equiv P(c_1 \dots c_k, c'_1 \dots c'_\ell)$$
$$\beta(k, \ell) \equiv P(c_k \dots c_p, c'_\ell \dots c'_q)$$

We will now proceed to define these dynamic programs.

(c) [5 points]

Specify the base cases $\alpha(0, 0)$ and $\beta(p + 1, q + 1)$.

(d) [10 points]

Specify the recursive case for $\alpha(k, \ell)$ (the case for β is very similar).

(e) [5 points]

In terms of p and q, what is the time complexity for filling in the α table?

(f) [10 points]

Using the α and β values, write expressions for the expected counts of the parameters $\text{Count}^t(c, c')$, $\text{Count}^t(c, \epsilon)$, and $\text{Count}^t(\#)$ in terms of previous estimates $P^{t-1}(z)$. ($\text{Count}^t(\epsilon, c)$ is similar to $\text{Count}^t(c, \epsilon)$.)

(g) [5 points]

Write the new EM estimates for parameter $P^t(c, c')$ in terms of the counts in the previous question.