CSE 250B Quiz 5

Tuesday February 11, 2014

Instructions. Do this quiz in partnership with exactly one other student. Write both your names at the top of this page. **Circle one name that we will call out when we return the quiz. Choose a first name or last name that is likely to be unique in the class.**

Discuss the answer to the question with each other, and then write your joint answer below the question. Use the back of the page if necessary. It is fine if you overhear what other students say, because you still need to decide if they are right or wrong. You have seven minutes. The maximum score is three points.

Question. Consider the contrastive divergence method for training a linear chain CRF. Assume that each tag is re-assigned once starting from the training label $\bar{y}$, and ignore sparsity. Notation: J is the number of low-level feature functions, m is the number of alternative tags, n is the length of each training sequence (assume that all sequences have the same length), S is the number of training examples, and T is the number of epochs of training. **Explain the mistake in the lemma and proof below.**

Lemma: The time complexity of this training method is $O(nm(ST + mJ)$.
Proof: The total time needed is ST multiplied by the time needed for one training example $\langle \bar{x}, \bar{y} \rangle$. Because each tag is re-assigned once starting from the training label $\bar{y}$, there are n conditional probability computations. Each of these involves computing the probability of each of m tag values, which is a constant time operation, when the g_i matrices for $\bar{x}$ have been computed explicitly. For a given $\bar{x}$, precomputing the g_i matrices requires $O(nm^2J)$ time, ignoring sparsity. Hence, the total time needed is $O(STnm + nm^2J) = O(nm(ST + mJ))$.

Answer. The g_i matrices depend not only on $\bar{x}$ but also on the current weights w_j. Hence, they must be recomputed each time a training example is processed; they cannot be precomputed. The total time needed is $O(ST(nm + nm^2J)) = O(STnm^2J)$.

Additional notes. In the hardcopy quiz, there was also a typographical error, a missing parenthesis in the last big-O expression. A more subtle observation is that

a factor m can be removed from the time complexity. The relevant equation is

$$p(v|x, \bar{y}_{-i}; w) = \frac{[\exp g_i(y_{i-1}, v)][\exp g_{i+1}(v, y_{i+1})]}{\sum_{v'}[\exp g_i(y_{i-1}, v')][\exp g_{i+1}(v', y_{i+1})]}.$$

For all m alternative possible values of v in position i, we need to compute only one row of the g_i matrix, that is the row that involves y_{i-1}, and only one column of the g_{i+1} matrix, the column that involves y_{i+1}.