COMS E6998-3, Spring 2011, problem set 1

Due date: 5pm, 11th April 2011

Question 1 (15 points)

In this question we will build a log-linear model for language modeling. Assume we have a training sample (x_i, y_i) for $i = 1 \ldots n$, where each x_i is a prefix of a document (e.g., $x_i =$ "Yesterday, George Bush said") and y_i is the next word seen after this prefix (e.g., $y_i =$ "that"). As usual in log-linear models, we will define a function $\underline{\phi}(x, y)$ that maps any x, y pair to a vector in $\mathbb{R}^d$. Given parameter values $\underline{\theta} \in \mathbb{R}^d$, the model defines

$$P(y|x, \underline{\theta}) = \frac{e^{\underline{\theta} \cdot \underline{\phi}(x,y)}}{\sum_{y' \in \mathcal{V}} e^{\underline{\theta} \cdot \underline{\phi}(x,y')}}$$

where $\mathcal{V}$ is the *vocabulary*, i.e., the set of possible words; and $\underline{\theta} \cdot \underline{\phi}(x, y)$ is the inner product between the vectors $\underline{\theta}$ and $\underline{\phi}(x, y)$.

Given the training set, the training procedure returns parameters $\underline{\theta}^* = \arg\max_{\underline{\theta}} L(\underline{\theta})$, where

$$L(\underline{\theta}) = \sum_i \log P(y_i|x_i, \underline{\theta}) - C \sum_k \theta_k^2$$

and $C > 0$ is some constant.

We will make the (rather odd) choice of the first two features in the model:

$$\phi_1(x, y) = \begin{cases} 1 & \text{if } y = \texttt{model} \text{ and previous word in } x \text{ is } \texttt{the} \\ 0 & \text{otherwise} \end{cases}$$

$$\phi_2(x, y) = \begin{cases} 1 & \text{if } y = \texttt{model} \text{ and previous word in } x \text{ is } \texttt{the} \\ 0 & \text{otherwise} \end{cases}$$

So $\phi_1(x, y)$ and $\phi_2(x, y)$ are *identical features*.

Question (15 points): Show that for any training set, with ϕ_1 and ϕ_2 defined as above, the optimal parameters $\underline{\theta}^*$ satisfy the property that $\theta_1^* = \theta_2^*$.

Question 2 (15 points)

We now decide to build a bigram language model using log-linear models. We gather a training sample (x_i, y_i) for $i = 1 \ldots n$. Given a vocabulary of words $\mathcal{V}$, each x_i and each y_i is a member of $\mathcal{V}$. Each (x_i, y_i) pair is a *bigram* extracted from the corpus, where the word y_i is seen following x_i in the corpus.

The new model is similar to our previous model, except we choose the optimal parameters $\underline{\theta}^*$ to be $\arg\max L(\underline{\theta})$ where

$$L(\underline{\theta}) = \sum_i \log P(y_i|x_i, \underline{\theta})$$

The features in the model are of the following form:

$$\phi_i(x, y) = \begin{cases} 1 & \text{if } y = \texttt{model} \text{ and } x = \texttt{the} \\ 0 & \text{otherwise} \end{cases}$$

i.e., the features track pairs of words. To be more specific, we create one feature of the form

$$\phi_i(x, y) = \begin{cases} 1 & \text{if } y = w_2 \text{ and } x = w_1 \\ 0 & \text{otherwise} \end{cases}$$

for every (w_1, w_2) in $\mathcal{V} \times \mathcal{V}$.

Question (15 points): Assume that the training corpus contains all possible bigrams: i.e., for all $w_1, w_2 \in \mathcal{V}$ there is some i such that $x_i = w_1$ and $y_i = w_2$. The optimal parameter estimates $\underline{\theta}^*$ define a probability $P(y = w_2|x = w_1, \underline{\theta}^*)$ for any bigram w_1, w_2. Show that for any w_1, w_2 pair, we have

$$P(y = w_2|x = w_1, \underline{\theta}^*) = \frac{Count(w_1, w_2)}{Count(w_1)}$$

where $Count(w_1, w_2) =$ number of times $(x_i, y_i) = (w_1, w_2)$, and $Count(w_1) =$ number of times $x_i = w_1$.

Question 3 (40 points)

In conditional random fields (CRFs), a key idea was to define a "global" feature vector

$$\underline{\Phi}(\underline{x}, \underline{s})$$

that maps an input sequence $\underline{x}$ paired with an output sequence $\underline{s}$ to a d-dimensional feature vector. In *bigram* models for sequence modeling, we define

$$\underline{\Phi}(\underline{x}, \underline{s}) = \sum_{j=1}^{n} \underline{\phi}(\underline{x}, j, s_{j-1}, s_j)$$

where $\underline{\phi}$ is a *local* feature-vector definition. We described a decoding algorithm for CRFs that take this form, and also an algorithm for parameter estimation.

In this question we'll consider a *trigram* model for conditional random fields, where

$$\underline{\Phi}(\underline{x}, \underline{s}) = \sum_{j=1}^{n} \underline{\phi}(\underline{x}, j, s_{j-2}, s_{j-1}, s_j)$$

where $\underline{\phi}$ is again a *local* feature-vector definition, which can now consider sequences of three tags (s_{j-2}, s_{j-1}, s_j).

Question (15 points): Give pseudo-code for a dynamic-programming algorithm for decoding for the trigram model.

Question (15 points): In class we described a parameter estimation method for bigram CRF models. Describe an analogous parameter estimation method for trigram models. Your method should use analogous terms to the $q^i_j(a, b)$ terms employed for bigram models (see the notes on CRFs).

Question (10 points): Give pseudo code for a perceptron-based algorithm for parameter estimation for the trigram model.

Question 4 (20 points)

Consider a sequence modeling task where we have the following training data:

- 100 examples where $x_1 = \texttt{a}, x_2 = \texttt{b}$, and $s_1 = \texttt{A}, s_2 = \texttt{B}$.
- 100 examples where $x_1 = \texttt{a}, x_2 = \texttt{c}$, and $s_1 = \texttt{A}, s_2 = \texttt{C}$.
- 800 examples where $x_1 = \texttt{c}, x_2 = \texttt{d}$, and $s_1 = \texttt{B}, s_2 = \texttt{D}$.

Question (10 points): We first train a bigram HMM for the sequence modeling problem. List all non-zero parameters for the HMM. What is the output from the HMM on the three input sequences `a b`, `a c`, and `c d`?

Question (10 points): Describe features for a bigram CRF for the sequence modeling problem, which models the data correctly. (By "correctly" we mean that the output from the model on the input sequences `a b`, `a c`, and `c d` is `A B`, `A C`, and `B D` respectively.)