Questions for Flipped Classroom Session of COMS 4705 Week 10, Fall 2014. (Michael Collins)

Question 1 This question considers log-linear models. We'd like to build a model that estimates a distribution $p(tag|word)$ using a log-linear model. The variable tag can take any one of three values, *D, N, V*. The variable $word$ could potentially be any member of a set $\mathcal{V}$ of possible words. The set $\mathcal{V}$ contains the words *the, dog, sleeps*, as well as additional words (i.e., $|\mathcal{V}| > 3$). The distribution should give the following probabilities:

$$\begin{aligned}
p(D|the) &= 0.9 \\
p(N|dog) &= 0.9 \\
p(V|sleeps) &= 0.9 \\
p(D|word) &= 0.6 \text{ for any word other than } the, dog \text{ or } sleeps \\
p(N|word) &= 0.3 \text{ for any word other than } the, dog \text{ or } sleeps \\
p(V|word) &= 0.1 \text{ for any word other than } the, dog \text{ or } sleeps
\end{aligned}$$

Note that we have intentionally left the values for the following probabilities undefined: $p(N|the)$, $p(V|the)$, $p(D|dog)$, $p(V|dog)$, $p(D|sleeps)$, $p(N|sleeps)$. It is assumed that these probabilities could take any values such that $\sum_{tag} p(tag|word) = 1$ is satisfied for $word =$ *the, dog, sleeps*, or indeed any other word in $\mathcal{V}$.

Question 1a Define the features for a log-linear model that can model this distribution $p(tag|word)$ perfectly. Each feature should be an indicator function: i.e., each feature $f_j(x, y)$ can take only the values 0 or 1 depending on the values of x and y. Your model should make use of as few features as possible. We will give you 10 points for using 6 features, and will penalise you for using more than 6 features. (It may be possible to use only 5 features, but a model with 6 features may be more straightforward to analyse.)

Question 1b Write an expression for each of the probabilities

$$\begin{aligned}
&p(\text{D}|\text{cat}) \\
&p(\text{N}|\text{laughs}) \\
&p(\text{D}|\text{dog}) \\
&p(\text{V}|\text{sleeps})
\end{aligned}$$

as a function of the parameters in your model. (Assume that the words $laughs$ and cat are both members of the set $\mathcal{V}$.)

Question 1c What value do the parameters in your model take to give the distribution described above?

Question 2 Consider a log-linear model for translation alignments. For simplicity assume that English sentences and French sentences are always of length m. Our task is to model the conditional distribution

$$p(a_1 \dots a_m | e_1 \dots e_m)$$

where each alignment variable a_j can take any value in $\{1, 2, \dots m\}$ (for simplicity we will not allow the null word in the model).

We model this using a simple log-linear model, where

$$p(a_1 \dots a_m | e_1 \dots e_m) = \prod_{j=1}^{m} p(a_j | e_1 \dots e_m, j)$$

where

$$p(a_j | e_1 \dots e_m, j) = \frac{\exp\{v \cdot f(e_1 \dots e_m, j, a_j)}{\sum_{a \in \{1 \dots m\}} \exp\{v \cdot f(e_1 \dots e_m, j, a)}$$

Here $v \in \mathbb{R}^d$ is a parameter vector, and $f(e_1 \dots e_m, j, a_j)$ is a feature vector.

Assume that we know that the data always obeys the following constraints:

- If e_1 is equal to *the*, then $a_j = j$ for $j \in \{1 \dots m\}$ with probability 1.
- Conversely, if e_1 is not equal to *the*, then $p(a_1 \dots a_m | e_1 \dots e_m)$ is the uniform distribution over all m^m possible values for the alignment variables.

Give a definition of the feature-vector $f(e_1 \dots e_m, j, a)$ that will allow us to model the distribution described above.

Question 3 This question again concerns log-linear models. To recap the details from the lecture notes: we have a set $\mathcal{X}$ of possible inputs, and a finite set $\mathcal{Y}$ of possible labels. We have a feature vector $f(x, y) \in \mathbb{R}^d$ for any $x \in \mathcal{X}$ and $y \in \mathcal{Y}$. We have a parameter vector $v \in \mathbb{R}^d$. The log-linear model defines the conditional probability as

$$p(y|x; v) = \frac{\exp(v \cdot f(x, y))}{\sum_{y \in \mathcal{Y}} \exp(v \cdot f(x, y))}$$

To estimate the parameters of the model, we have a set of training examples $(x^{(i)}, y^{(i)})$ for $i \in \{1 \dots n\}$. The regularized log-likelihood function is

$$L(v) = \sum_{i=1}^{n} \log p(y^{(i)}|x^{(i)}; v) - \frac{\lambda}{2}||v||^2$$

where $\lambda > 0$ is a parameter. Recall that the derivatives of this function are

$$\frac{d}{dv_j} L(v) = \sum_{i=1}^{n} f_j(x^{(i)}, y^{(i)}) - \sum_{i=1}^{n} \sum_{y \in \mathcal{Y}} p(y|x^{(i)}; v) f_j(x^{(i)}, y) - \lambda v_j$$

The optimal parameters are

$$v^* = \arg\max_{v \in \mathbb{R}^d} L(v)$$

Question 3a Assume that for feature f_1, we have $f_1(x^{(i)}, y) = 0$ for all $i \in \{1 \dots n\}, y \in \mathcal{Y}$. What is the value of v_1^*? Make sure to justify your answer.

Question 3b (10 points) Assume that for feature f_2, we have $f_2(x^{(i)}, y) = 10$ for all $i \in \{1 \dots n\}, y \in \mathcal{Y}$. What is the value of v_2^*? Make sure to justify your answer.

Question 3c (10 points) Assume that for feature f_3, we have $f_3(x^{(i)}, y) = i$ for all $i \in \{1 \dots n\}, y \in \mathcal{Y}$. What is the value of v_3^*? Make sure to justify your answer.