CSE 250A Assignment 6

This assignment is due at the start of class on Thursday November 15, 2012. Instructions are the same as for previous assignments. You must work in partnership with one other student, but you may keep the same partner or change partners, as you wish. *Acknowledgment: The second question is an extended version of one written by Lawrence Saul.*

1. Language modeling with latent information

For this question, you will train language models using EM. Specifically, you will write code to estimate the parameters of the context model of Section 7 of the lecture notes, and, separately, the interpolation model of Section 9. For each part of this question, hand in the new code you write, and answer any questions that are asked explicitly.

(a) Download the corpus from `www.pxnguyen.com/files/hw6data.zip`. This file will be available on Friday morning. The corpus contains sentences from 4,531 articles published in the magazine *Slate*, obtained from `http://www.americannationalcorpus.org/OANC/index.html`. Phuc Nguyen has preprocessed the data to make all words lower case, remove most punctuation, and identify sentence boundaries. The corpus has been randomly divided into 90% training articles and 10% tuning articles. (For this problem we do not need a test set.)

(b) Write code to identify the 10,000 most common words in the training set. Include the four pseudo words $\langle START \rangle$, $\langle END \rangle$, $\langle NUMBER \rangle$, and $\langle OTHER \rangle$. Use these 10,000 words in the following parts.

(c) Reuse code that you wrote for the fourth assignment to learn a unigram model p_1, a bigram model p_2, and a trigram model p_3 from the training set. Write code implementing EM to learn, using the tuning set, the three parameters in the interpolated model

$$p(w_3|w_1, w_2) = \lambda_1 p_1(w_3) + \lambda_2 p_2(w_3|w_2) + \lambda_3 p_3(w_3|w_1, w_2).$$

How should you initialize the λ_i values for EM training? What are the final λ_i values after convergence? What do these values mean?

(d) Write code that learns a context model from the training set for any fixed number c of contexts up to 30. Plot the log likelihood of the tuning set as a function of c, for c between 1 and 30. Which value of c is best?

(e) For $c = 6$, try different initializations for the parameters of the context model. How much does the final log likelihood achieved on the training set vary?

(f) For $c = 6$ with a good initialization, investigate the top ten words triggering each context. Discuss whether it is more informative to find the words w for which $p(w|z)$ is highest, or for which $p(z|w)$ is highest. What is the difference in meaning between these two probabilities? Similarly, investigate the top ten words w' suggested by each context. Is it more informative to maximize $p(w'|z)$ or $p(z|w')$?

(g) The context model and the interpolated model are quite different. Explain why it is, or is not, meaningful to compare the numerical log likelihood achieved by each model. Compute these log likelihoods for the λ_i obtained in part (c) and for the optimal c chosen in part (d). Discuss the results.

(h) Given your findings from part (g), do you expect it to be worthwhile to train on the tuning set a model that interpolates between the unigram, bigram, and context models? What λ_i values do you expect to get for this model? Train this model. Do you get the results that you said you expected?

(i) For the two models from part (g), and for the model from part (h), compute the sequence of most likely words starting with the word "She." That is, compute

$$\begin{aligned} w_2 &= \text{argmax}_w\ p(w|\langle START\rangle, \text{“She”}) \\ w_3 &= \text{argmax}_w\ p(w|\text{“She”}, w_2) \end{aligned}$$

and so on. Stop when you reach w_{10}. For each of the three models, show the sequence of words that you obtain. Discuss what these sequences reveal about the strengths and weaknesses of the models.

(j) For a given model, consider the word sequence

$$\text{argmax}_{w_2,\ldots,w_{10}}\ p(w_2|\langle S\rangle, \text{“She”})p(w_3|\text{“She”}, w_2)\cdots p(w_{10}|w_8, w_9).$$

Is this sequence the same as the sequence obtained as described in part (i), or not? Explain.

2. A Gaussian mixture model

Consider a multivariate Gaussian mixture model with two components. The model has a binary random variable Y with values $y \in \{0, 1\}$ and a vector-valued random variable X with values $x \in \mathbb{R}^d$. The graphical model is $Y \rightarrow X$. The CPT for Y is $p(Y = i) = \rho_i$. The CPT for X is

$$p(x|Y = i) = (2\pi)^{\frac{d}{2}}|C|^{-\frac{1}{2}} \exp -\frac{1}{2}(x - \mu_i)^T C^{-1}(x - \mu_i).$$

The parameters of the model are two prior probabilities ρ_0 and ρ_1, two mean vectors μ_0 and μ_1, and one covariance matrix C.

(a) Compute the posterior distribution $p(Y = 1|x)$ as a function of the five parameters of the model described above.

(b) Given independent, identically distributed, fully observed training examples, state the maximum likelihood estimates of the five parameters.

(c) Show that

$$p(Y = 1|x) = \sigma(w \cdot x + b) \text{ where } \sigma(z) = \frac{1}{1 + \exp -z}.$$

As part of your answer, express the weights $w \in \mathbb{R}^d$ and $b \in \mathbb{R}$ explicitly in terms of the original parameters of the model.

(d) The expression for $p(Y = 1|x)$ in part (c) is a logistic regression model. Consider training this model using the same data described in part (b), by maximizing the conditional likelihood $p(y|x)$. Prove that, in general, the parameter estimates obtained in this way are not the same as the parameter estimates obtained in part (b).

(e) Explain intuitively why the trained parameter estimates are different. Which estimates are preferable under what circumstances?

(f) Now suppose that the training examples are independent and identically distributed as before, but all the y values are unobserved. Derive an EM algorithm to learn the five parameters of the original model.

(g) Resolve the following paradox: With zero labeled training examples, it is impossible to train a classifier, but the EM algorithm of part (f) seems to do just that.