

CG168 Homework 4

Mark Johnson

due 9th October 2008

A Maximum Entropy model for parsing

Your homework assignment is to construct a maximum entropy parsing model for the dependency parsing framework we've been using in previous assignments.

The feature function vector \mathbf{f} consists of the following feature functions:

$$\begin{aligned}f_{y'}(y, \mathbf{x}) &= \delta_{y'}(y) \text{ for each } y' \in \mathcal{Y} \\f_{j,y',x'}(y, \mathbf{x}) &= \delta_{y'}(y)\delta_{x'}(x_j) \text{ for each } j \in 1, \dots, k, y' \in \mathcal{Y}, x' \in \mathcal{X}_j\end{aligned}$$

where $\mathbf{x} = (x_1, \dots, x_k)$ is a context vector, and \mathcal{X}_j is the j th component of \mathcal{X} .

Informally, f_y “fires” whenever y appears in the input, while $f_{j,y,x}$ fires whenever y appears together with x_j in the input. We let \mathbf{f} be the sequence containing all the f_y functions followed by all the $f_{j,y',x'}$ functions.

Count how often each feature fires on the training data, and prune the features that don't fire at least 5 times.

Then use a numerical technique such as Conjugate Gradient or L-BFGS to find the feature weights \mathbf{w} that minimize the following regularized negative log likelihood:

$$\begin{aligned}Q &= -\log L_D(\mathbf{w}) + R(\mathbf{w}) \\R(\mathbf{w}) &= 10 \sum_{j=1}^{|\mathbf{f}|} w_j^2 \\ \log L_D(\mathbf{w}) &= \sum_{i=1}^n \log P_{\mathbf{w}}(y_i | x_i) \\ P_{\mathbf{w}}(y|x) &= \frac{1}{Z_x(\mathbf{w})} \exp \mathbf{w} \cdot \mathbf{f}(y, x) \\ Z_x(\mathbf{w}) &= \sum_{y \in \mathcal{Y}} \exp \mathbf{w} \cdot \mathbf{f}(y, x) \\ \frac{\partial \log L_D}{\partial w_j} &= \sum_{i=1}^n (f_j(y_i, x_i) - E_{\mathbf{w}}[f_j | x_i]) \\ E_{\mathbf{w}}[f | x] &= \sum_{y \in \mathcal{Y}} f(y, x) P_{\mathbf{w}}(y | x)\end{aligned}$$

Questions:

1. *Specify the number of feature functions you found met the count threshold on the training data, the optimization method you used, the number of likelihood function evaluations required, and the value of the negative log likelihood of the training data L_D and the value of the regularizer R that you found at the optimal value of \mathbf{w} . (My program uses the OWLQN L-BFGS optimizer and required 298 function evaluations. At iteration 50, $-\log L_D = 477,140$ and $R = 23137.7$).*
2. *What is the accuracy of your classifier on the development set? I.e., how accurately does your classifier predict the y associated with each context x in the development set? (On the training set my classifier has an accuracy of 0.915).*
3. *How accurate is the dependency parser whose actions are chosen by the maximum entropy classifier you have just constructed? (On the training set my dependency parser has an accuracy of 0.80).*
4. *(200-level) What happens when one of the components of context, say x_j , takes a value not seen in the training data? Is this reasonable? Can you suggest a better way of dealing with “unknown words”? Implement your idea, and report its classification and parsing accuracies.*