

CG168 Homework 3

Mark Johnson

due 9th October 2008

Maximum Entropy models

This homework is about Maximum Entropy models of the form:

$$P(u) = \frac{1}{Z(\mathbf{w})} \exp \mathbf{w} \cdot \mathbf{f}(u), \text{ where}$$
$$Z(\mathbf{w}) = \sum_{u' \in \mathcal{U}} \exp \mathbf{w} \cdot \mathbf{f}(u')$$

Homework:

1. Suppose we wanted to calculate the partition function for a MaxEnt model over (y, \mathbf{x}) pairs for the CONLL data from earlier homework. What set do we have to sum over to compute the partition function Z ? Suppose we can enumerate one million (y, \mathbf{x}) pairs a second; how long would it take to compute the partition function? (You can assume there are 45 POS tags and approximately 40,000 words in the English language, and MacKay (2003) says that there are about 3×10^3 seconds in an hour, 3×10^7 seconds in a year and 3×10^{17} seconds since the big bang).
2. We saw in class that the MLE $\hat{\mathbf{w}}$ of a MaxEnt model is the \mathbf{w} such that the expected value $E_D[f_j]$ of each feature f_j in the data D equals the expected value $E_{\hat{\mathbf{w}}}[f_j]$ of f_j under the model $P_{\hat{\mathbf{w}}}$. But intuitively, the data D might contain more information than just the expected number of times each feature occurs. This question investigates whether a MaxEnt model can capture arbitrary dependencies between pairs of features. Specifically, suppose $|\mathcal{U}| = 4$ and we have two binary features f_1 and f_2 that uniquely identify each $u \in \mathcal{U}$. Further, suppose $P_D(u) = 0.33$ except when $f_1(u) = f_2(u) = 1$; then $P_D(u) = 0.01$. Is there a MaxEnt model with features f_1 and f_2 that has this distribution? If not, what would you have to do to obtain a MaxEnt model that equals P_D ?