Homework 7 | CMSC 723

Maximum Entropy Models

Due: November 18$^{\text{th}}$, 2013

1 MaxEnt Math (40 pts)

Suppose we have an unregularized maximum entropy model with input x and output y. The output is an n-dimensional binary vector of the form $y = < y_1, y_2, \ldots, y_n >$, where $y_i \in \{0, 1\}$. Our model has the following n features:

$$f_1(x,y) = \begin{cases} 1, & \text{if } y_1 = 1 \\ 0, & \text{otherwise} \end{cases}$$

$$f_2(x,y) = \begin{cases} 1, & \text{if } y_2 = 1 \\ 0, & \text{otherwise} \end{cases}$$

$$\vdots$$

$$f_n(x,y) = \begin{cases} 1, & \text{if } y_n = 1 \\ 0, & \text{otherwise} \end{cases}$$

The probability distribution for this model is given below, where λ is a vector of feature weights and the denominator normalizes over all possible n-dimensional binary vectors:

$$P\left(y\,|\,x\right) = \frac{e^{\lambda_1 f_1(x,y) + \lambda_2 f_2(x,y) + \cdots + \lambda_n f_n(x,y)}}{\sum_{y'} e^{\lambda_1 f_1(x,y') + \lambda_2 f_2(x,y') + \cdots + \lambda_n f_n(x,y')}} \tag{1}$$

Rewrite the right-hand side of (**??**) to show that

$$P\left(y\,|\,x\right) = \prod_{i=1}^{n} P_i\left(y_i\,|\,x\right)$$

where each P_i is the probability distribution specified by a maximum entropy model with a single feature.

Hint: If we define the single feature f_i' for P_i as below, then $f_i' = f_i(x, y)$.

$$f_i'(x, y_i) = \begin{cases} 1, & \text{if } y_i = 1 \\ 0, & \text{otherwise} \end{cases}$$

2 Word Root Identification (60 pts)

We're going to design a maximum entropy model to identify the root of a given word. The root may be either a prefix of the given word (*acrobat* comes from the Greek root *acro*, which has meanings such as *height* and *top*), or a suffix (the root of *inspect* is *spect*, which means *to look*). For this problem, we'll ignore cases in which the root occurs in the middle of the word (e.g. *vert* in *advertisement*).

The probability of root r given word w is as follows:

$$P\left(r \,|\, w\right) = \frac{e^{\lambda f(w,r)}}{\displaystyle\sum_{r'} e^{\lambda f(w,r')}}$$

where the denominator normalizes over all possible prefixes and suffixes of w. As an example, for the word *lemon*, we have to consider the set of prefixes (*lemo*, *lem*, *le*, *l*), the set of suffixes (*emon*, *mon*, *on*, *n*), and *lemon* itself.

We'd like this distribution to give us the following probabilities:

$$\begin{aligned}
&P\left(anti \,|\, antipathy\right) = 0.9 \\
&P\left(hyper \,|\, hypersonic\right) = 0.7 \\
&P\left(homeo \,|\, homeopathy\right) = 0.8 \\
&P\left(sect \,|\, intersect\right) = 0.5 \\
&P\left(super \,|\, supersonic\right) = 0.7 \\
&P\left(sect \,|\, sector\right) = 0.6 \\
&P\left(insect \,|\, insect\right) = 0.2
\end{aligned}$$

2.1 Feature Design (40 pts)

Your task is to design a set of indicator features (binary features whose only possible values are 0 and 1) that can represent this distribution. While many possible solutions exist, you'll be penalized if you use more than four features. Below is an example feature:

$$f_1(w, r) = \begin{cases} 1, & \text{if } w = \text{“}\textit{example}\text{”} \\ 0, & \text{otherwise} \end{cases}$$

In devising your features, you should pay attention to the number of **distinct** probabilities, not the exact probabilities.

2.2 Using Your Features to Predict Roots (20 pts)

Let's say you're given the parameter vector λ, where $\lambda_1, \lambda_2, \ldots, \lambda_n$ are weights for each of your n features. Write expressions for the following probabilities:

$$P\left(a \,|\, apathy\right)$$
$$P\left(sect \,|\, bisect\right)$$
$$P\left(mason \,|\, masonic\right)$$
$$P\left(plutocracy \,|\, plutocracy\right)$$