

# 050/520/600.666 Information Extraction from Speech and Text

Project # 1

Due February 24, 2006.

You will model letters of English text using hidden Markov models. Some ordinary text has been selected and, to keep matters simple, all numerals and punctuation have been purged, capital letters have been down-cased, and inter-word spacing, new lines and paragraph breaks, have all been normalized to single spaces. The alphabet of the resulting text is therefore the 26 lower-case English letters and the white-space, which is formally denoted as  $\mathcal{Y} = \{a, b, c, \dots, z, \#\} \equiv \{1, 2, 3, \dots, 26, 27\}$ , with  $\#$  denoting the white-space character.

The text is 35,000 characters long, and has been divided into a 30,000 character *training set*, named **A**, and a 5,000 character *test set*, named **B**.

1. Model the letter-sequence as the output of a fully connected 2-state HMM, whose states  $s_1$  and  $s_2$  *generate outputs* in  $\mathcal{Y}$  according to some unspecified probabilities  $q(y|s_i)$ ,  $i = 1, 2$ .

Let  $t_1$  denote the transition  $s_1 \rightarrow s_2$ , and  $t_2$  denote the self-loop on  $s_1$ . Similarly, let  $t_3$  denote the transition  $s_2 \rightarrow s_1$ , and  $t_4$  denote the self-loop on  $s_2$ , so that the transition probability matrix may be written as

$$\mathbf{p} = \begin{bmatrix} p(s_1|s_1) & p(s_2|s_1) \\ p(s_1|s_2) & p(s_2|s_2) \end{bmatrix} \equiv \begin{bmatrix} p(t_2) & p(t_1) \\ p(t_3) & p(t_4) \end{bmatrix}$$

Let the initial state of the hidden chain be either  $s_1$  or  $s_2$  with equal probability.

Use the Baum-Welch algorithm and the training text **A** to estimate the probabilities  $p(t_j)$ ,  $j = 1, 2, 3, 4$ , and the emission probabilities  $q(y|s_i)$ ,  $y \in \mathcal{Y}$  and  $i = 1, 2$ .

- (a) To get the reestimation going, initialize the transition probabilities as

$$p(t_1) = 0.51 = p(t_3) \quad \text{and} \quad p(t_2) = 0.49 = p(t_4),$$

and emission probabilities as

$$\begin{aligned} q(1|s_1) = q(2|s_1) = \dots = q(13|s_1) = 0.0370 &= q(14|s_2) = q(15|s_2) = \dots = q(26|s_2), \\ q(1|s_2) = q(2|s_2) = \dots = q(13|s_2) = 0.0371 &= q(14|s_1) = q(15|s_1) = \dots = q(26|s_1), \\ q(27|s_1) &= 0.0367 = q(27|s_2). \end{aligned}$$

What would happen if all probabilities were set to be uniform, i.e.  $\frac{1}{4}$  and  $\frac{1}{27}$ ?

(b) Plot the average log-probability of the training and test data after  $k$  iterations,

$$\frac{1}{|\mathbf{A}|} \log P_k(\mathbf{A}) \quad \text{and} \quad \frac{1}{|\mathbf{B}|} \log P_k(\mathbf{B}),$$

as a function of the number of iterations, for  $k = 1, 2, \dots, 600$ .

(c) Plot the emission probabilities of a few particular letters for each state, e.g.

$$q_k(a|s_1) \quad \text{and} \quad q_k(a|s_2), \quad \text{or} \quad q_k(n|s_1) \quad \text{and} \quad q_k(n|s_2),$$

as a function of the number of iterations, for  $k = 1, 2, \dots, 600$ .

(d) Study the emission probability distributions  $q_{600}(\cdot|s_1)$  and  $q_{600}(\cdot|s_2)$  to see where they differ the most, as well as how the transition probabilities differ from their initial values. Try to explain what the machine has learned about English text.

2. *Increasing Model Complexity:* Repeat the Exercises 1(a) through 1(d) with a fully connected 4-state HMM. Modify the initialization in 1(a) to account for 4 states.

3. *Alternate Initialization of Output Probabilities:* HMM estimation is sometimes sensitive to the initialization of the model parameters. You will now investigate an alternative to the initialization of Exercise 1(a).

(a) Compute the relative frequency  $q(y)$  of the letters in  $\mathcal{Y}$  from the entire text  $\mathbf{A}$ .

(b) Generate  $|\mathcal{Y}|$  random numbers  $r(y)$ , compute their average  $\bar{r} = \frac{1}{|\mathcal{Y}|} \sum_{y \in \mathcal{Y}} r(y)$ , and use it to create a *perturbation* vector  $\delta(y) = r(y) - \bar{r}$ .

(c) Choose a small  $\lambda > 0$ , though not too small, such that both

$$q(y|s_1) = q(y) - \lambda\delta(y) > 0 \quad \text{and} \quad q(y|s_2) = q(y) + \lambda\delta(y) > 0 \quad \forall y \in \mathcal{Y}.$$

Note:  $q(\cdot|s_1)$  and  $q(\cdot|s_2)$  are bona fide probability assignments on  $\mathcal{Y}$ . (Why?)

Use the two  $q(y|s_i)$  thus generated, along with the  $p(t_j)$  from Exercise 1(a), to initialize the Baum-Welch iteration. Compare the resulting plots of average log-probability versus  $k$  with those of 1(b), as well as the final values of the average log-probabilities.

**Caution:** Make sure you guard against numerical underflow problems when computing the forward and backward probabilities. Use the normalization described in Section 2.8 of Chapter 2 if needed.

**Submission:** Turn in all your plots and discussion. Do not turn in source code, but make sure it is well documented, in case it needs to be reviewed. You will be contacted for your source code, possibly asked to rerun your code on different training and test data or with a different initialization, if needed.