(520|600).666
Information Extraction from Speech and Text

Programming Assignment # 1

Due March 7, 2024.

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, upper case letters have been lower-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, and is formally denoted as $\mathcal{Y} = \{\mathtt{a}, \mathtt{b}, \mathtt{c}, \ldots, \mathtt{z}, \#\}$, 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 this text with a fully connected 2-state HMM, with states ① and ②.

   Let t_1 denote the transition ① → ②, and t_2 denote the self-loop ① → ①. Similarly, let t_3 denote the transition ② → ①, and t_4 denote the self-loop ② → ②, so that the transition probability matrix may be written as

   $$\mathbf{p} = \begin{bmatrix} p(①|①) & p(②|①) \\ p(①|②) & p(②|②) \end{bmatrix} \equiv \begin{bmatrix} p(t_2) & p(t_1) \\ p(t_3) & p(t_4) \end{bmatrix}$$

   Let the initial state of the Markov chain be either ① or ② with *equal probability*.

   Let the emission probabilities be associated with the states, i.e. let

   $$q(y\,|\,t_2) \;\equiv\; q(y\,|\,t_3) \;\equiv\; q(y\,|\,①) \quad \text{and} \quad q(y\,|\,t_1) \;\equiv\; q(y\,|\,t_4) \;\equiv\; q(y\,|\,②).$$

   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 $q(y\,|\,s)$, $y \in \mathcal{Y}$ and $s \in \{①, ②\}$.

   (a) Initialize the transition probabilities to be *slightly different from uniform*, as

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

   Initialize the emission probabilities to also be *slightly different from uniform*, as

   $$q(\mathtt{a}|①) = q(\mathtt{b}|①) = \ldots = q(\mathtt{m}|①) = 0.0370 = q(\mathtt{n}|②) = q(\mathtt{o}|②) = \ldots = q(\mathtt{z}|②),$$

$$q(\mathtt{a}|②) = q(\mathtt{b}|②) = \ldots = q(\mathtt{m}|②) = 0.0371 = q(\mathtt{n}|①) = q(\mathtt{o}|①) = \ldots = q(\mathtt{z}|①),$$
$$q(\#|①) = 0.0367 = q(\#|②).$$

What would happen if all probabilities were set to be uniform, i.e $\frac{1}{2}$ and $\frac{1}{27}$?

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

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

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

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

$$q_k(\mathtt{a}|①) \text{ versus } q_k(\mathtt{a}|②) \qquad \text{and} \qquad q_k(\mathtt{n}|①) \text{ versus } q_k(\mathtt{n}|②),$$

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

(d) Study the emission probability distributions $q_{600}(\cdot|①)$ and $q_{600}(\cdot|②)$ 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 a vector of *random* numbers $r(y)$, compute the average $\overline{r} = \frac{1}{|\mathcal{Y}|}\sum_{y\in\mathcal{Y}} r(y)$, and use it to create a zero-mean *perturbation* vector $\delta(y) = r(y) - \overline{r}$.

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

$$q(y|①) = q(y) - \lambda\delta(y) > 0 \qquad \text{and} \qquad q(y|②) = q(y) + \lambda\delta(y) > 0 \qquad \forall\, y \in \mathcal{Y}.$$

Note: $q(\cdot|①)$ and $q(\cdot|②)$ are bona fide probability assignments on $\mathcal{Y}$. (Why?)

Use the two $q(y|s)$ 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 mitigate numerical underflow problems when computing the forward and backward probabilities. Use the normalization described in §2.8 if needed.

Submission: Turn in all your plots and discussion, and your source code, via GradeScope; make sure your code is well documented. Points may be deducted for incomprehensible code. Your code may be rerun on different training and test data or with a different initialization to check its correctness; make sure it runs on a linux machine with standard compilers/libraries/environments.