

050/520/600.666 Information Extraction from Speech and Text

Homework # 3

Due February 25, 2005.

1. *HMMs with State Durations*: Consider an HMM with outputs produced by states. Let \mathcal{S} denote the state-space and \mathcal{Y} the discrete and finite output alphabet, as usual, and let the transition and output probabilities be denoted by

$$a_{ij} = P(s_t = j | s_{t-1} = i) \quad \text{and} \quad b_j(k) = P(y_t = k | s_t = j) \quad i, j \in \mathcal{S}, \quad k \in \mathcal{Y}.$$

Modify the HMM to explicitly model state durations as before. Set $a_{ii} = 0$ for all states i . For some initial state s_0 and probability distributions $\delta_i(\tau)$ of the duration for each state i , where $\tau = 1, 2, \dots, T$ and $\sum_{\tau=1}^T \delta_i(\tau) = 1$, assume that the state process,

- *makes a transition* to reach state $s_1 \neq s_0$ with probability $a_{s_0 s_1}$,
- *decides* to stay in the state s_1 for a duration τ with probability $\delta_{s_1}(\tau)$,
- *produces* the first τ outputs y_1, \dots, y_τ from s_1 with probability $\prod_{t=1}^{\tau} b_{s_1}(y_t)$,
- *makes a transition* to reach state $s_{\tau+1} \neq s_1$ with probability $a_{s_1 s_{\tau+1}}$,
- *decides* to stay in the state $s_{\tau+1}$ for a duration τ' with probability $\delta_{s_{\tau+1}}(\tau')$,
- *produces* the next τ' outputs $y_{\tau+1}, \dots, y_{\tau+\tau'}$ from state $s_{\tau+1}$ with probability $\prod_{t=1}^{\tau'} b_{s_{\tau+1}}(y_{\tau+t})$,

and so on. Observe carefully the notational difference in time-indices from Homework 2.

Let $\alpha_t(j)$ denote the probability that the state-process has *just reached* state j after having produced the observations y_1, \dots, y_{t-1} .

- (a) Rederive the *forward* probabilities $\alpha_t(j)$ for $t > 1$ in terms of $\alpha_{t'}(i)$ for $t' < t$.
- (b) How should the forward recursion be initialized? (Hint: Start with $t = 1$.)

Let $\beta_t(i)$ denote the probability that the state-process, having *just reached* state i , will produce the observations y_t, \dots, y_n .

- (c) Derive the *backward* probabilities $\beta_t(i)$ for $t < n$ in terms of $\beta_{t'}(j)$ for $t' > t$.
- (d) How should the backward recursion be initialized?

Derive the re-estimation formulae for the HMM parameters a_{ij} , $b_j(k)$ and $\delta_i(\tau)$.

2. *Levenshtein Distance*: Given two strings $\mathbf{A} = a_1a_2 \dots a_k$ and $\mathbf{B} = b_1b_2 \dots b_l$ made up of symbols from a common alphabet \mathcal{X} , define the Levenshtein (or string-edit) distance $L(\mathbf{A}, \mathbf{B})$ between them to be the minimum number of insertions, deletions and substitutions of letters required to transform \mathbf{A} into \mathbf{B} .

(a) Show that $L(\cdot, \cdot)$ is a bona fide *distance*. In other words, argue why

i. $L(\mathbf{A}, \mathbf{A}) = 0$ for all strings \mathbf{A} ,

ii. $L(\mathbf{A}, \mathbf{B}) = L(\mathbf{B}, \mathbf{A})$ for all strings \mathbf{A} and \mathbf{B} , and

iii. $L(\mathbf{A}, \mathbf{C}) \leq L(\mathbf{A}, \mathbf{B}) + L(\mathbf{B}, \mathbf{C})$ for all strings \mathbf{A} , \mathbf{B} and \mathbf{C} .

(b) For each symbol $a \in \mathcal{X}$, consider designing an *elementary* weighted finite state machine with unique start- and end-states, and arcs labeled $\langle x, c \rangle$ to denote the output $x \in \mathcal{X} \cup \{\epsilon\}$ and cost $c \in \{0, 1\}$, where ϵ represents the null symbol. Design the elementary machines such that in the machine obtained by *concatenating* the elementary machines of the symbols of \mathbf{A} , the minimum cost of producing \mathbf{B} , from the start-state of a_1 to the end-state of a_k , is exactly the Levenshtein distance $L(\mathbf{A}, \mathbf{B})$. Draw the elementary machine and clearly label the output (e.g. $a, \epsilon, x \neq a$) and cost of each arc.

(c) Modify the Viterbi algorithm of p22 in the textbook to construct an algorithm that computes $L(\mathbf{A}, \mathbf{B})$ from the concatenated machine for \mathbf{A} as described above.

3. Work on Project 1.