CSE 753 Autumn 2017

Mid semester Exam Solutions

Instructor: Preethi Jyothi Date/Time: Sep 14, 2017, 3 to 5 pm

Name: _____ Roll number: _____

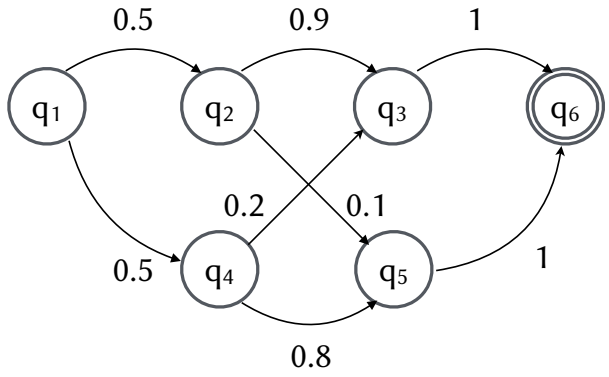
Instructions

- This is a closed book exam which should be completed individually. No form of collaboration or discussion is allowed.
- Write your name and roll number on the top of this page.
- Write your answers legibly in the space provided on the exam sheet. Provide derivations when they are asked for. If necessary, use the extra sheet on the back to work out your solutions.
- This exam consists of 4 questions with sub-parts. The maximum possible score is 100.
- Good luck!

Question	Score
HMMs	/25
Application of EM	/20
ASR for morphologically rich languages	/30
Mixed Bag	/25
Total	/100

Problem 1: HMMs (25 points)

(A) Consider the HMM shown in the figure below. (The transition probabilities are shown in the finite-state machine on the left and the observation probabilities corresponding to each state are shown on the right.) This model generates hidden state sequences and observation sequences of length 4. If S_1, S_2, S_3, S_4 represent the hidden states and O_1, O_2, O_3, O_4 represent the observations, then $S_i \in \{q_1, \ldots, q_6\}$ and $O_i \in \{a, b, c\}$. $\Pr(S_1 = q_1) = 1$ i.e. the state sequence starts in q_1.



	a	b	c
q_1	0.5	0.3	0.2
q_2	0.3	0.4	0.3
q_3	0.2	0.1	0.7
q_4	0.4	0.5	0.1
q_5	0.3	0.3	0.4
q_6	0.9	0	0.1

State whether the following three statements are true or false and justify your responses. If the statement is false, then state how the left expression is related to the right expression, using either $=, <$ or $>$ operators. (We use the following shorthand in the statements below: $\Pr(O = abbc)$ denotes $\Pr(O_1 = a, O_2 = b, O_3 = b, O_4 = c)$. **[15 points]**

(i) $\Pr(O = bbca, S_1 = q_1, S_4 = q_6) = \Pr(O = bbca | S_1 = q_1, S_4 = q_6)$
Ans: True, because $\Pr(S_1 = q_1, S_4 = q_6) = 1$.

(ii) $\Pr(O = acac, S_2 = q_2, S_3 = q_5) > \Pr(O = acac, S_2 = q_4, S_3 = q_3)$
Ans: True. $\Pr(O = acac, S_2 = q_2, S_3 = q_5) = K \times 0.09$ and $\Pr(O = acac, S_2 = q_4, S_3 = q_3) = K \times 0.04$.

(iii) $\Pr(O = cbcb | S_2 = q_2, S_3 = q_5) = \Pr(O = baac, S_2 = q_4, S_3 = q_5)$
Ans: False. $\Pr(O_4 = b | S_4 = q_6) = 0$. Therefore, $\Pr(O = cbcb | S_2 = q_2, S_3 = q_5) < \Pr(O = baac, S_2 = q_4, S_3 = q_5)$.

(B) Modify the recurrence for $v_t(j)$ (i.e. the probability of being in state j after seeing the first t observations) in the original Viterbi algorithm such that the best state sequence satisfies the following condition: For any three consecutive states i, j, k in the sequence, $i \neq k$. (Here, j may or may not be equal to i.) **[10 points]**
Ans:

Modified Viterbi recursion becomes:

$$v_t(k) = \max_j v_t(j,k)$$

$$v_t(j,k) = \max_{i \neq k} v_{t-1}(i,j) a_{jk} b_k(O_t)$$

$$v_1(j,k) = \begin{cases} 1 & \text{if } k = \text{start state} \\ 0 & \text{otherwise} \end{cases}$$

Problem 2: Application of EM (20 points)

Consider a first order HMM with two states 0 and 1, and output alphabet $\{a, b\}$. Suppose the output distribution at state 0 is uniform, and the output distribution in state 1 assigns probabilities 3/4 and 1/4 respectively for a and b. The only transition from state 1 is to itself (self-loop, with probability 1), whereas from state 0, the HMM transitions to state 1 with probability p (and stays at state 0 with probability $1 - p$), where p is a parameter of the HMM.

Consider an observation $x \in \{a, b\}^2$ produced by this HMM starting at state 0, for two time steps. Suppose we are given n such observations $(x_1, \ldots, x_n)$. Describe the EM algorithm to estimate p that maximizes the probability of the observed output. Explicitly state both the E step and the M step. You may state the algorithm in terms of quantities $N_{\alpha,\beta} = |\{i \mid x_i = (\alpha, \beta)\}|$, where $(\alpha, \beta) \in \{a, b\}^2$, and $N_{a,a} + N_{a,b} + N_{b,a} + N_{b,b} = n$
[20 points]

Ans:

The auxiliary function, $Q(p, p^{t-1})$ is defined as follows:

$$Q(p, p^{t-1}) = \sum_i \sum_z \Pr(z|x_i; p^{t-1}) \log \Pr(x_i, z; p)$$

where x_i refers to the ith observed variable and z refers to the hidden variable.

In this problem, x_i is one of $\{(a,a), (a,b), (b,a), (b,b)\}$ and $z \in \{0, 1\}$ corresponds to the second state traversed on encountering an x_i.

$$Q(p, p^{t-1}) = \sum_{\substack{x_i=(a,a)\\x_i=(a,b)}} \big(\Pr(0|(a,a);p^{t-1})\log\Pr((a,a),0;p) + \Pr(1|(a,a);p^{t-1})\log\Pr((a,a),1;p)$$
$$+ \Pr(0|(a,b);p^{t-1})\log\Pr((a,b),0;p) + \Pr(1|(a,b);p^{t-1})\log\Pr((a,b),1;p)\big)$$
$$+ \sum_{\substack{x_i=(b,a)\\x_i=(b,b)}} \big(\Pr(0|(b,a);p^{t-1})\log\Pr((b,a),0;p) + \Pr(1|(b,a);p^{t-1})\log\Pr((b,a),1;p)$$
$$+ \Pr(0|(b,b);p^{t-1})\log\Pr((b,b),0;p) + \Pr(1|(b,b);p^{t-1})\log\Pr((b,b),1;p)\big)$$

E step:

$$\gamma_1 = \Pr(z=0|x_i=(a,a);p^{t-1}) = \tfrac{\frac{1}{4}(1-p)}{\frac{1}{4}(1-p)+\frac{3}{8}p} = \tfrac{2-2p}{2+p}\ ;\ \gamma_2 = 1-\gamma_1 = \tfrac{3p}{2+p}$$
$$\gamma_3 = \Pr(z=0|x_i=(a,b);p^{t-1}) = \tfrac{\frac{1}{4}(1-p)}{\frac{1}{4}(1-p)+\frac{1}{8}p} = \tfrac{2-2p}{2-p}\ ;\ \gamma_4 = 1-\gamma_1 = \tfrac{p}{2-p}$$
$$\gamma_5 = \Pr(z=0|x_i=(b,a);p^{t-1}) = \tfrac{\frac{1}{4}(1-p)}{\frac{1}{4}(1-p)+\frac{3}{8}p} = \tfrac{2-2p}{2+p}\ ;\ \gamma_6 = 1-\gamma_5 = \tfrac{3p}{2+p}$$
$$\gamma_7 = \Pr(z=0|x_i=(b,b);p^{t-1}) = \tfrac{\frac{1}{4}(1-p)}{\frac{1}{4}(1-p)+\frac{1}{8}p} = \tfrac{2-2p}{2-p}\ ;\ \gamma_8 = 1-\gamma_7 = \tfrac{p}{2-p}$$

M step:

Setting $\partial Q/\partial p$ to 0 and solving for p, we get:

$$p = \frac{N_{aa}\gamma_2 + N_{ab}\gamma_4 + N_{ba}\gamma_6 + N_{bb}\gamma_8}{n}$$

Problem 3: ASR for morphologically rich languages (30 points)

Words in a language can be composed of sub-word units called *morphemes*. For simplicity, in this problem, we consider there to be three sets of morphemes, $V_{\text{pre}}, V_{\text{stem}}$ and V_{suf} – corresponding to prefixes, stems and suffixes. Further, we will assume that every word consists of a single stem, and zero or more prefixes and suffixes. That is, a word is of the form $w = p_1 \cdots p_k \sigma s_1 \cdots s_\ell$ where $k, \ell \geq 0$, and $p_i \in V_{\text{pre}}$, $s_i \in V_{\text{suf}}$ and $\sigma \in V_{\text{stem}}$.

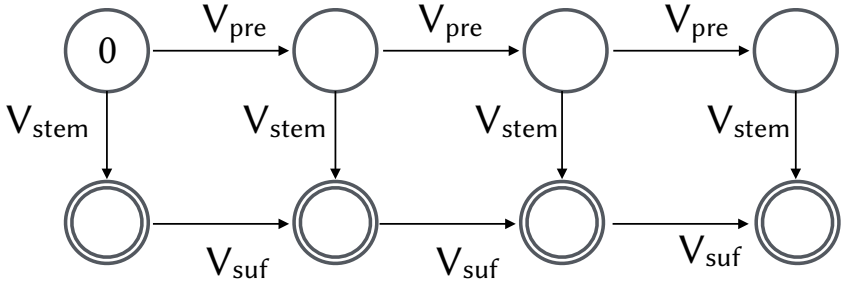
For example, a word like FAIR consists of a single morpheme (a stem), where as the word UNFAIRNESS is composed of three morphemes, UN + FAIR + NESS, which are a prefix, a stem and a suffix, respectively.

(A) Suppose we want to build an ASR system for a language using morphemes instead of words as the basic units of language. Which WFST(s) in the $H \circ C \circ L \circ G$ framework should

be modified in order to utilize morphemes? Briefly describe what the inputs and outputs of the modified WFST(s) should be. [**4 points**]

Ans: Modify L and G. L: Inputs are monophones, outputs are morphemes. G: Acceptor over morphemes.

(B) Recall that a word is of the form $w = p_1 \cdots p_k \sigma s_1 \cdots s_\ell$ where $k, \ell \geq 0$, and $p_i \in V_{\text{pre}}$, $s_i \in V_{\text{suf}}$ and $\sigma \in V_{\text{stem}}$. Draw a finite-state acceptor over morphemes ($V_{\text{pre}} \cup V_{\text{stem}} \cup V_{\text{suf}}$) that accepts only words with at most four morphemes. Your FSA should not have more than 15 states. You may draw a single arc labeled with a set to indicate a collection of arcs, each labeled with an element in the set. [**12 points**]



(C) In this part, your goal is to define a smoothed unigram language model. For a word w, let $[w]$ denote its stem and let $\langle w \rangle$ denote the rest of the word (the modifiers), i.e. the ordered set of prefixes and of suffixes. E.g. if the word $w =$ UNFAIRNESSLESSNESS, then $[w] =$ FAIR and $\langle w \rangle = ((\text{UN}), (\text{NESS}, \text{LESS}, \text{NESS}))$.

Suppose you are given models P_{stem} and P_{mod} such that $P_{\text{stem}}(\sigma) = \Pr[[W] = \sigma]$ and $P_{\text{mod}}(\mu) = \Pr[\langle W \rangle = \mu]$, where W denotes a word drawn according to the unigram probability. Also suppose that you are given a dataset, with unigram count $\pi(w)$ for each word w, and also a discounted count $\pi^*(w)$ (using some discounting rule). You may assume that all stems in V_{stem} occur in the data.

(i) Define a smoothed backoff model for $\Pr[\langle W \rangle = \mu | [W] = \sigma]$. The smoothing should ensure that for every stem σ, the discounted count corresponding to σ is redistributed among unseen words w with $[w] = \sigma$, such that w gets a mass proportionate to $P_{\text{mod}}(\langle w \rangle)$.

Ans:

$$\Pr[\langle w \rangle = \mu | [w] = \sigma] = \begin{cases} \frac{\pi^*(w)}{\sum_{[w]=\sigma} \pi(w)} & \text{if } \pi(w) > 0 \text{ where } [w] = \sigma, \langle w \rangle = \mu \\ \alpha(\sigma) \cdot P_{\text{mod}}(\mu) & \text{otherwise} \end{cases}$$

where

$$\alpha(\sigma) = \left(1 - \frac{\sum_{[w]=\sigma} \pi^*(w)}{\sum_{[w]=\sigma} \pi(w)}\right) \cdot \frac{1}{\mathcal{K}} \text{ and } \mathcal{K} = \sum_{\substack{\langle w \rangle = \mu \text{ s.t.} \\ \pi(w)=0}} P_{\text{mod}}(\mu)$$

(ii) Next, use the expression from above to define a unigram language model $P(w)$.

Ans: $P(w) = P_{\text{stem}}([w]) \Pr[\langle w \rangle | [w]]$

Problem 4: Mixed Bag (25 points)

(A) The recommended textbook for CS 753 is in the main library in one of several shelves. It is known *a priori* that the book is in shelf i with probability $p_i < 1$. However, all the books are so chaotically ordered that even if someone correctly guesses that the book is in shelf i, the probability of finding it there is only q_i. A student searches for the book in a particular shelf, say shelf i, and does not find the book. Conditioned on this event, what is the probability that the book is in a shelf j? Consider both cases $j = i$ and $j \neq i$.

[**7 points**]

Ans: Let E_i refer to the book being in shelf i and F_i refer to finding the book in shelf i. Then, $\Pr(E_i) = p_i$ and $\Pr(F_i|E_i) = q_i$. We need to find:

$$\Pr(E_j|\overline{F_i}) = \frac{\Pr((\overline{F_i})|E_j) \Pr(E_j)}{\Pr(\overline{F_i})} = \begin{cases} \frac{p_j}{1-p_i q_i} & \text{if } i \neq j \\ \frac{(1-q_i)p_i}{1-p_i q_i} & \text{if } i = j \end{cases}$$

(B) In a hybrid system, how are DNNs used within HMM acoustic models? Briefly explain what the DNN computes and how it is used to compute the observation probabilities in the HMM. [**4 points**]

Ans: In a hybrid system, DNNs provide observation probabilities to the HMMs. The DNNs compute posterior probabilities over triphone states. These DNN posteriors are scaled by priors over the triphone states to give scaled likelihoods, which in turn are used as observation probability densities in the HMM.

Describe one advantage and one disadvantage of hybrid HMM-DNN systems over HMM-GMM systems. [**4 points**]

Ans:

Advantage: Can exploit power of DNN-based acoustic models, DNN training can use frame-level acoustic features which are correlated.

Disadvantage: Computationally more intensive.

(C) Calculate the total number of parameters in the following components of an ASR system.

Acoustic model: Inputs are given as d-dimensional acoustic vectors. Suppose there are N distinct triphones and we use a K-state HMM to model each of them. Let us assume each HMM has exactly τ non-zero transition probabilities and uses m-component GMMs to model the observation probability densities. Also assume full covariance matrices for all the GMM probability densities. [**6 points**]

Ans: $\tau N + mdNK + m\frac{d(d+1)}{2}NK + (m-1)NK$

Language model: We use a trigram language model. Assume that the word vocabulary is of size V. The language model should define a probability distribution over sentences which start and end with special markers (which are not part of the vocabulary). The parameters of the language model are simply all the conditional probabilities necessary to define such a distribution. [**4 points**]

Ans: $(V-1) + V^2(V+1)$ parameters

$V^2(V+1)$ trigram parameters: $P(z|x,y)$ where z is any word $\in \{V \cup \langle/s\rangle\}$, x is any word $\in \{V \cup \langle s\rangle\}$, y is any word in V.

V or $V-1$ bigram parameters at the beginning of a sentence: $P(x|\langle s\rangle)$ where $x \in \{V \cup \langle/s\rangle\}$ or $x \in V$ (both answers will be awarded full points).