Introduction to Digital Speech Processing (NTU, Autumn 2020) Instructor: Lin-shan Lee

Midterm Exam

November 11, 2020 09:30 - 11:30

- Open lecture slides (printed version) and personal notes. No electrical devices except for calculators are allowed.
- You have to answer all the questions in CHINESE, but English terminologies are allowed.
- Total points: 100

---

1. (8 pts) What is GMM? How is it usually used in HMMs for speech recognition?

2. (8 pts) What is K-means algorithm? How is it used in speech recognition?

3. In HMM, Viterbi Algorithm is used to **find the single best state sequence**. Variable $\delta_t(i)$ is defined as:

$$\delta_t(i) = \max_{q_1,q_2,\ldots,q_{t-1}} P\big[q_1, q_2, \ldots, q_{t-1}, q_t = i, o_1, o_2, \ldots, o_t|\lambda\big]$$

The induction step of the algorithm is:

$$\delta_{t+1}(j) = (\max_i \Big[\delta_t(i)a_{ij}\Big])b_j(o_{t+1})$$

(a) (3 pts) Can we change the induction step into the equation shown below? Please explain why.

$$\delta_{t+1}(j) = \max_i \Big[\delta_t(i)a_{ij}b_j(o_{t+1})\Big]$$

(b) (5 pts) Can we change the induction step into the equation shown below? Please explain why.

$$\delta_{t+1}(j) = (\max_i \Big[\delta_t(i)\Big])a_{\sigma(i)j}b_j(o_{t+1})$$

where $\sigma(i) = \underset{i}{\operatorname{argmax}}\, \delta_t(i)$

4. (9 pts) We wish to calculate the accuracy for some speech recognition results. Please list the insertions, deletions, substitutions, and calculate the accuracy with the formula taught in class(insertions, deletions, substitutions have the same penalty weight).

reference:	the dog sat on the mat
recognized:	the dogs on the mat are

5. Below is a dataset for training a bi-gram language model.

dataset
<sos> I am Sam <eos>
<sos> I am legend <eos>
<sos> Bob I am <eos>

(a) (4 pts) Calculate the probabilities: $P(\text{I}|\text{<sos>}), P(\text{am}|\text{I}), P(\text{Sam}|\text{am}), P(\text{<eos>}|\text{Sam})$

(b) (3 pts) Calculate the probability of $P(\text{<sos> I am Sam <eos>})$ using uni-gram plus bi-grams only.

(c) (3 pts) With the bi-grams trained above, for a given sentence "<sos> I am Bob <eos>", the probability $P(\text{"<sos> I am Bob <eos>"}) = 0$ (note that this given sentence is not in the training set). However, this sentence is a reasonable sentence and should not have zero probability. Propose a method to fix this problem (you do not have to explain your method in detail).

6. In language modeling, perplexity is a very useful parameter.

(a) (4 pts) What is perplexity of a language model with respect to a testing corpus?

(b) (3 pts) A training corpus consists of only a single sentence:

<sos> dsp so easy <eos>

The testing corpus also consists of only a single sentence:

<sos> so easy dsp <eos>

We use the training corpus to train a bi-gram language model(bi-grams that do not exist in the training corpus have probabilities equal to 0). What is the perplexity on the testing corpus?

(c) (3 pts) Following the previous question, what is the perplexity on the testing corpus if the testing corpus consists of only a single sentence:

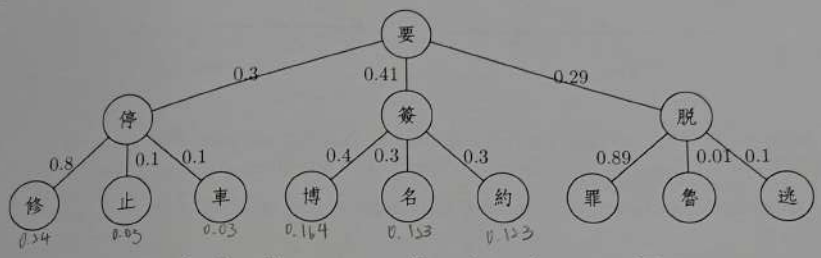
<sos> dsp so easy <eos>

7. (a) (5 pts) Speech signals are roughly categorized into voiced and unvoiced. Explain the distinction between the two.

(b) (5 pts) Explain how the derivatives of the 13 MFCC parameters ($14^{th} - 26^{th}$ and $27^{th} - 39^{th}$) are actually calculated.

8. There are many different strategies for search or decoding.
   - **Exhaustive Search**: Exhaustively enumerate all possible output sequences with their probabilities, then output the one with the highest probability.
   - **Beam Search(beam width k)**: At the first time index, we select k tokens with the highest probabilities. At each subsequent time index, we continue to select k tokens with the highest probabilities.
   - **Greedy Search**: At any time index, we search for and output the token with the highest probability. (You can view this as beam search with $k = 1$.)

   Given a tree with tokens as nodes and the edge weights representing the bi-gram probabilities (e.g. $P($ 停 $|$ 要 $) = 0.3$):



   (a) (3 pts) What is the decoding output with **exhaustive search**?

   (b) (3 pts) What is the decoding output with **greedy search**?

   (c) (3 pts) What is the decoding output of **beam search with $k = 2$** ?

9. (7 pts) Bob is a hard-working student. There are many courses for the new semester. He made a table as below listing the attributes of the courses and then decided whether to take a course or not as listed on the rightmost column in the table. You are to analyze how he made the decision using a decision tree.

Course Name (課程名稱)	compulsory (是否為必修)	credit (幾學分)	tests (是否有考試)	whether to take (是否要修)
DSP	No	3	Yes	**TAKE**
ADA	Yes	3	No	**NOT TAKE**
ML	No	4	No	**TAKE**
LA	No	3	No	**NOT TAKE**
DLHLP	Yes	4	No	**TAKE**

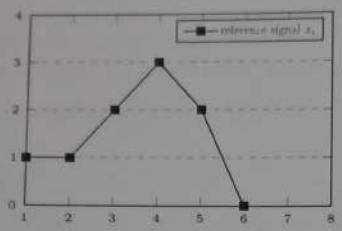
   Construct a decision tree so that each leaf node of the tree clearly indicates he decided to take a course or not. (It is fine not to use all the attributes, and you just only have to provide one solution if there are multiple solutions.)

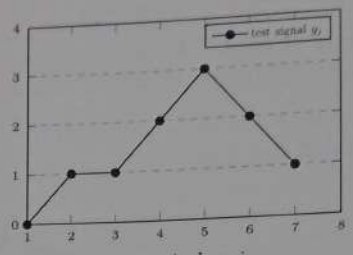
10. (8 pts) What is the context dependency when we try to train HMMs for small sound units?

11. Below are two signals:
the reference signal $[x_i, i = 1, 2, \ldots, 6]$ and test signal $[y_j, j = 1, 2, \ldots, 7]$, respectively.

i	1	2	3	4	5	6
x_i	1	1	2	3	2	0

j	1	2	3	4	5	6	7
y_j	0	1	1	2	3	2	1





We want to find an optimal path for matching two signals with **Dynamic Time Warping(DTW)**. Define $D(i, j)$ to be the accumulated minimum distance up to (i, j).

- endpoint constraints: the optimal path must begin at $(i, j) = (1, 1)$ and end at $(i, j) = (6, 7)$.
- local constraints: only the three moves shown in Fig. 1 are allowed.
- recursive relationship:

$$D(i,j) = \min\left(D(i,j-1)+d(i,j),\ D(i-1,j-1)+\frac{1}{2}d(i,j),\ D(i-1,j)+d(i,j)\right) \quad (1)$$

for $i = \{2, 3, \ldots, 6\}$, $j = \{2, 3, \ldots, 7\}$, where $d(i, j) = (x_i - y_j)^2$.

(a) (9 pts) Finish the dynamic programming table($D(i, j)$) shown in Fig. 2.
(The first row and column are done for you.)

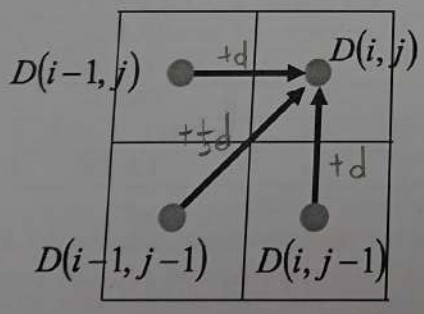


Figure 1: Illustration of the recursive relationship for Eq. (1).

7	1	7	5	3	4	2	1.5
6	2	7	5	3	2	1	4
5	3	6	4	2	1	2	4.5
4	2	2	1.5	1	2	3	4
3	1	1	1	1.5	4	5	6
2	1	1	1	2	6	7	8
1	0	1	2	6	15	19	19
j	y_j / x_i	1	1	2	3	2	0
	i	1	2	3	4	5	6

Figure 2: dynamic programming table

(b) (4 pts) Find an optimal path for matching the two signals (remember that this path should begin at $(1, 1)$ and end at $(6, 7)$).