Introduction to Digital Speech Processing, Midterm Exam

Nov. 18, 2017, 15:30-17:30

- OPEN Lecture Power Point (Printed Version) and Personal Notes
- You have to use CHINESE sentences to answer all the questions, but you can use English terminologies
- Total points: 195

---

1. Alice and Bob are both interested in only four activities: playing baseball, going to the movies, watching television and studying. Their choices are influenced by the temperature of the cities they live, city A and city B, on a given day. One of their friends, Candy, has no definite information about the temperature in the cities, but she believes that they both operate as discrete Markov chains. Candy assumes that the weather conditions can be modeled as either "hot", "warm", or "cold", but she cannot observe them directly, that is, they are hidden because Candy lives in neither city A nor city B. Candy can see the blogs of Alice and Bob, where they post their daily activities. There are the observations Candy can get. The entire systems are like two hidden Markov models (HMMs).
   Candy set the following model setting:
   states = ' cold ', ' warm ', ' hot ' observations = 'baseball', 'movies', 'TV', 'study'
   Then Candy uses the following training algorithm to estimate the model parameters:

// Baum-Welch iterative training
Read in the observations (daily activities on Alices/Bobs blog)
Read in initial models
for iter =1 to iteration_num do
  Clean all accumulators
  for sample = 1 to num_of_samples do
    T ← length of the sample
    for t = T to 1 do
      calculate α_t (cold), α_t (warm) and α_t (hot)
      calculate β_t (cold), β_t (warm) and β_t (hot)
    end for
    calculate $\gamma_t(i)$, $\epsilon_t(i,j)$ iteratively where i, j = cold or warm or hot
    accumulate

$$\gamma_1(i),\ \sum_{t=1}^{T}\gamma_t(i),\ \sum_{t=1}^{T-1}\gamma_t(i),\ \sum_{o_t=baseball}\gamma_t(i),\ \sum_{o_t=movies}\gamma_t(i),\ \sum_{o_t=TV}\gamma_t(i),\ \sum_{o_t=study}\gamma_t(i),\ \sum_{t=1}^{T-1}\epsilon_t(i,j)$$

  end for
  update (A, B, π)
end for
Write out the new model

(a) (5) Are there any errors in the pseudo code above? Please point them out if yes.

(b) (15) Please use the following two models and **Viterbi algorithm** to classify (predict the author of) this collection: (TV, study, movie), and the corresponding state(weather).

Alice	Bob
start_probability = { 0.2, 0.2, 0.6}	start_probability = { 0.1, 0.4, 0.5}
transition_probability =	transition_probability =
{ 0.6, 0.4, 0.0	{ 0.3, 0.5 ,0.2
0.3, 0.5, 0.2	0.7, 0.3 ,0.0
0.1, 0.7, 0.2 }	0.0, 0.5 ,0.5 }
observation_probability =	observation_probability =
{ 0.0, 0.1, 0.3	{ 0.0, 0.1 ,0.2
0.5, 0.3, 0.4	0.1, 0.3 ,0.7
0.5, 0.5, 0.1	0.6, 0.5 ,0.1
0.0, 0.1, 0.2 }	0.3, 0.1 ,0.0 }

2. (a) (10) After you obtain the 13 MFCC parameters with the window. DFT/IDFT, filter bank and so on, how can you obtain the other 26 parameters? Explain what they are.

   (b) (10) Assume your cell phone receives your voice when you walk on the street and sends the signals to the cloud. The 39 MFCC parameters are then extracted in the cloud. Explain why the extra 26 parameters are useful here.

3. Given a HMM $\lambda = (A, B, \pi)$ with N states, an observation sequence $\bar{O} = o_1 o_2 \dots o_t \dots o_T$ (assuming o_i, o_j are mutually independent if $i \neq j$) and a state sequence $\bar{q} = q_1 q_2 \dots q_t \dots q_T$, define

$$\alpha_t(i) = \text{Prob}[o_1 o_2 \dots o_t, q_t = i|\lambda]$$
$$\beta_t(i) = \text{Prob}[o_{t+1} o_{t+2} \dots o_T | q_t = i, \lambda]$$

   (a) (5) Write down and explain the meaning of $\sum_{i=1}^{N} \alpha_t(i)\beta_t(i)$ in terms of probabilities.

   (b) (5) Write down and explain the meaning of $\frac{\alpha_t(i)\beta_t(i)}{\sum_{j=1}^{N} \alpha_t(j)\beta_t(j)}$ in terms of probabilities.

   (c) (5) Write down and explain the meaning of $\alpha_t(i)a_{ij}b_j(o_{t+1})\beta_{t+1}(j)$ in terms of probabilities.

   (d) (10) Formulate and describe the Viterbi algorithm to find the best state sequence $\bar{q}^* = q_1^* q_2^* \cdots q_t^* \cdots q_T^*$ giving the highest probability $\text{Prob}[\bar{O}, \bar{q}^*|\lambda]$. Why do we need backtracking?

4. (15) What is the perplexity of a language source? What is the perplexity of a language model with respect to a test corpus? How are they related to a "virtual vocabulary"?

5. (10) Explain why and how the unseen triphones can be trained using decision trees.

6. (10) Given a set of events $\{x_i$, i = 1, 2, ... M$\}$, $\{$p(x_i), i = 1, 2, ... M$\}$ and $\{$q(x_i), i = 1, 2, ... M$\}$ are two probability distributions. What is the Kullback-Leibler(KL) distance between p(x_i) and q(x_i) and what does it mean?

7. (10) Explain why and how **beam search** and **two-pass search** are useful in large vocabulary continuous speech recognition.

8. (a) (10) What are the voiced/unvoiced speech signals and their time domain waveform characteristics?

(b) (10) What is the pitch in speech signals? How is it related to the tones in Mandarin?

9. (20) Please briefly describe **LBG algorithm** and **K-means algorithm** respectively. Which one of the above two algorithms usually performs better? Explain your answer with descriptions, not just formula only.

10. (10) Explain the basic principles of back-off and interpolation to be used for language model smoothing.

11. (20) What is the maximum a posteriori (MAP) principle? How can it be used to integrate acoustic modeling and language modeling for large vocabulary speech recognition? Why and how this can be solved by a Viterbi algorithm over a series of lexicon trees?

12. Let p_1 be the central phoneme of a triphone model m_1, p_2 the central phoneme of another triphone model m_2, $p_1 \neq p_2$, and m_3 be another triphone model whoes central phoneme is also p_1(m_1, m_2, m_3 are 5 states models). Discuss: in a standard triphone training algorithm based on decision trees:

    (a) (5) Can the first state and the last state of m_1 share some training data? Why if not, or under what kind of conditions if yes?

    (b) (5) Can the first state of m_1 and the first state of m_2 share the same training data? Why if not, or under what kind of conditions if yes?

    (c) (5) Can the first state of m_1 and the first state of m_3 share the same training data? Why if not, or under what kind of conditions if yes?