Final Exam

January 13, 2021 09:30 - 11:30

- Open lecture slides (printed version) and personal notes.
- You have to answer all the questions in CHINESE, but English terminologies are allowed.
- **Only copying mathematical terms without explanation may result to deduction on your score.**
- No electric devices except for a calculator are allowed.
- Total points: 100

---

1. (20 pts) Determine whether each of the statements below is true or false.(No explanation needed. No partial credit will be given.)

T (a) Hybrid system is an approach of integrating HMMs with Neural Networks in acoustic modeling. In this approach, the state posteriors are obtained with the output of Neural Networks.

F (b) The state transition probabilities in a Hybrid System are also modeled by the output of Neural Networks.

T (c) PCA is like an unsupervised training method, while LDA(Linear Discriminative Analysis) is more like a supervised training method.

F (d) In a speaker dependent speech recognition system, the model is trained with different utterances spoken by many different speakers.

T (e) In a transformer model, for an input sequence of words, the model itself can not obtain any information regarding the position/order for each word. Thus, we need positional encoding to solve this problem.

T (f) The difference between contextualized word embeddings(ELMo, BERT) and traditional word embeddings is that for the former the same word in different context may have different vector representations.

F (g) "Mismatch" in acoustic environment means that the speaker utterance and the corresponding transcript do not match, in other words, there are word errors in the transcript.

T (h) One of the possible approaches to cope with the acoustic environment mismatch problems is Cepstral Moment Normalization. When there is convolution noise in the MFCC features, subtracting the MFCC features $\bar{y}$ with its own mean $E[\bar{y}]$ yields CMS(Cepstral Mean Subtraction) features that are immune to convolutional noise.

T (i) Following the previous question (h), histogram equalization usually performs better than Cepstral Mean Subtraction (CMS) and Cepstral Mean and Variance Normalization (CMVN).

T (j) The EM(Expectation Maximization) algorithm is guaranteed to converge to a global optimal point.

2. Consider the following problem for Minimum Classification Error (MCE). Here $\{C_i, i = 1, \ldots, M\}$ are classes, $\Lambda = \{\lambda^{(i)}, i = 1, \ldots, M\}$ are models for the classes, and X is the set of observations.

$$d_i(X, \Lambda) = -g_i(X, \Lambda) + \left[\frac{1}{M-1}\sum_{j \neq i} g_j(X, \Lambda)^{\alpha}\right]^{\frac{1}{\alpha}},$$

where $x \in C_i$ and $g_i(X, \Lambda) = Prob(X|\lambda^{(i)})$, with the continuous loss function

$$l_i(X, \Lambda) = l(d_i(X, \Lambda)), \quad \text{where } l(d) = \frac{1}{1 + e^{-\gamma d}}.$$

(a) (5 pts) Explain the meaning of choosing the values 1 and ∞ for α.

(b) (5 pts) Now assume that $\alpha \to \infty$, and $\gamma \to \infty$. Given a set of observation-class pairs

$$\{(x_{11}, C_1), (x_{12}, C_1), (x_{21}, C_2), (x_{22}, C_2), (x_{31}, C_3), (x_{32}, C_3)\},$$

calculate $d_1(x_{11}, \Lambda)$ and $d_1(x_{12}, \Lambda)$ with the values of $g_i(X, \Lambda)$ in Table 1, and check whether there are classification errors for x_{11} and x_{12}.

$g_i(X, \Lambda)$	x_{11}	x_{12}	x_{21}	x_{22}	x_{31}	x_{32}
C_1	0.1	0.5	0.12	0.11	0.09	0.08
C_2	0.095	0.085	0.4	0.3	0.065	0.055
C_3	0.2	0.4	0.14	0.12	0.08	0.06

Table 1: The class-conditioned likelihood

(c) (5 pts) Now calculate the overall classification performance measure

$$L(\Lambda) = \sum_{i=1}^{3} \sum_{X \in C_i} l_i(X, \Lambda).$$

3. A class of students took an exam with integer scores from 0 to 7. The probability mass function(pmf)(similar to probability density function pdf except for discrete scores here) for the scores is shown in Table 2. The instructor wants to adjust the scores to the pmf in Table 3 by performing histogram equalization in the following way. He first calculates an "approximate integer cumulative distribution function(c.d.f) $s_k = T(x_k)$" as shown in eq. (1), where he multiplies the probabilities by 7 and rounds the results to the nearest integers for easier mapping.

x_k	$p_x(x_k)$	
$x_0 = 0$	0.19	0.19
$x_1 = 1$	0.25	0.44
$x_2 = 2$	0.21	0.65
$x_3 = 3$	0.16	0.81
$x_4 = 4$	0.08	0.89
$x_5 = 5$	0.06	0.95
$x_6 = 6$	0.03	0.98
$x_7 = 7$	0.02	1

Table 2: x_k are the scores, $p_x(x_k)$ is the probability mass function (pmf) for each original score.

y_j	$p_y(y_j)$	
$y_0 = 0$	0.00	0
$y_1 = 1$	0.00	0
$y_2 = 2$	0.00	0
$y_3 = 3$	0.15	1.05 1
$y_4 = 4$	0.20	1.35 2
$y_5 = 5$	0.30	0.65 5
$y_6 = 6$	0.20	0.85 6
$y_7 = 7$	0.15	1 7

Table 3: y_j are the scores, $p_y(y_j)$ is the probability mass function (pmf) for each adjusted score.

$$s_k = T(x_k) = round\Big[7\sum_{i=0}^{k} p_x(x_i)\Big], \qquad k = 0, 1, 2, \cdots, 7 \tag{1}$$

(a) (5 pts) Please calculate the values of s_k. (Note that you need to round the values to their nearest integers.)

(b) (10 pts) Now perform histogram equalization based on $s_k = T(x_k)$ in eq.(1) and $r_j = T(y_j)$. Please write down how the values of x_k are mapped to y_j.

4. (10 pts) Voice actors/actresses are well-known for their diverse voice characteristics. Assume the speaker features of an utterance can be represented as a 5-dim vector, and we have a data set of such vectors for the utterances produced by three voice actresses: *Minori*, *Kiyono*, and *Nao*. Now we perform Principal Component Analysis (PCA) on this data set, such that the dimensionality of these vectors can be reduced from 5 to 2. Mean of the set is $\mathbf{0}$, and the first two eigenvectors and the corresponding largest eigenvalues obtained by PCA are listed in Table 4. The obtained 2-dim plot for the utterances of the three voice actresses are in Figure 1, where the different voice actresses are labeled by different shapes of marks.

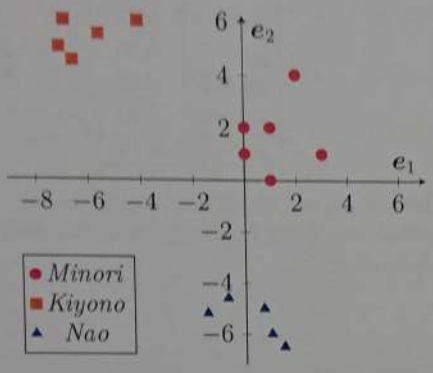

Figure 1: The 2-dim plot obtained with PCA.

Eigenvalue	Eigenvector
100	$\frac{1}{3}[1,0,2,-2,0]^T = \boldsymbol{e}_1$
50	$\frac{1}{5}[0,4,0,0,-3]^T = \boldsymbol{e}_2$

Table 4: Eigenvalues and eigenvectors

Index	Feature
1	$[4,1,1,0,-2]^T$
2	$[1,-2,0,2,4]^T$
3	$[3,6,-4,5,1]^T$
4	$[8,-6,0,1,2]^T$
5	$[1,5,0,-1,5]^T$

Table 5: Feature of unknown utterances

Now given the 5-dim feature vectors of some unknown utterances in Table 5, please perform dimensionality reduction and classify whom these utterances belong to based on the clusters for the speakers in Figure 1.

5. (a) (4 pts)Please explain what Matrix Factorization is.

   (b) (3 pts)Let A be an $M \times N$ matrix. Explain why matrix factorization can help in data compression(try to explain it with such parameters as M, N).

   (c) (5 pts)How can Matrix Factorization be used in a movie recommendation system?

6. (10 pts) Please explain how the EM-algorithm works as detailed as possible. What are the "E step" and the "M step"? Why do we have to use EM-algorithm in training HMMs?

7. (6 pts) There are three types of prosodic features mentioned in the lecture. List two of them and give two examples for each.

8. (7 pts) Pseudo relevance feedback(PRF) can be used to improve the retrieval performance. What is PRF and the basic principles for it? Why don't we simply use the first-pass results?

9. (5 pts) What is the difference between extractive and abstractive summarization?

transformer model