MP5: Hidden Markov Model

Group:Siheng Pan, Yangge Li

1. Introduction

In this mp, we are taking both audio and visual features and using hidden markov model to do speech recognition.

The audio feature vectors are the 24-dimensional MFCCs and visual feature vector is the distance between mouth corners and the distance between the upper/lower lips and the line connecting the mouth corners. Using these features can effectively reduce the dimension of the feature vector and increase the accuracy of classification.

In this mp, we train Hidden Markov Models for different classes and using the model for classification. The model learn the observation mean, observation covariance, initial state probabilities and state transition probabilities from the training data by using the forward-backward algorithm.

2. Methods

HMM Training:

For this MP, we are using forward-backward algorithm to train the Gaussian Hidden Markov model and classify the testing data set.

The forward step of the algorithm is to calculate

$$\alpha_t(i) = P(O_1 O_2 .....O_t, q_t = S_i|\lambda)$$

which is the probability of the partial observation sequence $O_1 O_2 .....O_t$ (until time t) and state S_i at time t, given the model λ. $\alpha_t(i)$ can be solved inductively as follows:

(1) Initialization:

$$\alpha_1(i) = \pi_1 b_i(O_1),\ 1 \le i \le N$$

(2) Induction:

$$\alpha_{t+1}(j) = [\sum_{i=1}^{N} \alpha_t(i) a_{ij}] b_j(O_{t+1}),\ 1 \le t \le T-1,\ 1 \le j \le N$$

In a similar manner, the backward variable $\beta_t(i)$ can be defined as

$$\beta_t(i) = P(O_{t+1} O_{t+2} .....O_T|q_t = S_i, \lambda)$$

which is the probability of the partial observation sequence from $t+1$ to the end. given state S_i at time t and the model λ. $\beta_t(i)$ can also be solved inductively:

(1) Initialization:

$$\beta_T(i) = 1,\ 1 \le i \le N$$

(2) Induction:

$$\beta_t(i) = \sum_{j=1}^{N} a_{ij} b_j(O_{t+1}) \beta_{t+1}(j),\ t = T-1, T-2, ..., 1,\ 1 \le i \le N$$

After getting α and β, the next step is to calculate the probability of being in state S_i at time t, given the observation sequence O, and the model λ

$$\gamma_t(i) = P(q_t = S_i|0,\lambda)$$

$\gamma_t(i)$ can be expressed simply in terms of the forward-backward variables α and β:

$$\gamma_t(i) = \frac{\alpha_t(i)\beta_t(i)}{P(O|\lambda)} = \frac{\alpha_t(i)\beta_t(i)}{\sum_{i=1}^{N} \alpha_t(i)\beta_t(i)}$$

In order to describe the procedure for reestmation of HMM parameters, $\xi_t(i,j)$, the probability of being in state S_i at time t, and state S_j at time $t+1$, given the model adn the observation sequence is defined as:

$$\xi_t(i,j) = P(q_t = S_i, q_{t+1} = S_j|0,\lambda)$$

$\xi_t(i,j)$ cam ne written by the forward and backward variables in the form:

$$\xi_t(i,j) = \frac{\alpha_t(i)a_{ij}b_j(O_{t+1})\beta_{t+1}(j)}{P(O|\lambda)} = \frac{\alpha_t(i)a_{ij}b_j(O_{t+1})\beta_{t+1}(j)}{\sum_{i=1}^{N}\sum_{j=1}^{N} \alpha_t(i)a_{ij}b_j(O_{t+1})\beta_{t+1}(j)}$$

Then, we will calculate the new observation mean and covariance using the above calculated values.

After calculating all these components, we can reestimate state transition probability by doing

$$\bar{a}_{ij} = \frac{expected\ number\ of\ transitions\ from\ state\ S_i\ to\ state\ S_j}{expected\ number\ of\ transitions\ from\ state\ S_i} = \frac{\sum_{t=1}^{T-1} \xi_t(i,j)}{\sum_{t=1}^{T-1} \gamma_t(i)}$$

We will repeat this procedure for 25 times to get a better result.

HMM Evaluation:

We will evaluate $P(O|\lambda)$ as

$$P(O|\lambda) = \sum_{i=1}^{N} \alpha_T(i)$$

We will get $log(P(O|\lambda))$ and compare the result between different classes

3. Results

---------- Accuracy: Forward Backward ----------

	Audio_Recognition	Video_Recognition	AV_Recognition
2	1	0.6	1
5	1	1	1
Overall	1	0.8	1

---

4. Discussion

From the result above, the HMM model seems to give the best result while using only the audio feature. The result is worst while using only the video features. The result is showing that HMM works better for audio data than video data.

5. Extra credit

Introduction:

We implement first extra credit assignment for this MP.

Method:

Instead of using forward-backward algorithm to calculate $log(P(O|\lambda))$, for this part we are required to use the Viterbi algorithm to calculate the best state sequence and compute the likelihood given that state sequence.

For the Viterbi algorithm,

$$\delta_t(i) = \max_{q_1,q_2,\ldots\ldots,q_{t-1}} P[q_1 q_2 \ldots\ldots q_t = i, O_1 O_2 \ldots O_t|\lambda]$$

is defined to represent the best score (highest probability) along a single path, at time t, which accounts for the first t observations and ends in state S_i. $\delta_t(i)$ can be solved recursively following the steps below

(1) Initialization:

$$\delta_1(i) = \pi_i b_i(O_1),\ 1 \le i \le N$$

$$\phi_1(i) = 0 \text{ where } \phi \text{ is an array that maximize } \delta_t(i)$$

(2) Recursion:

$$\delta_t(j) = \max_{1\le i\le N}[\delta_{t-1}(i)a_{ij}]b_j(O_t),\ 2 \le t \le T,\ 1 \le j \le N$$

$$\phi_t(j) = \arg\max_{1\le i\le N}[\delta_{t-1}(i)a_{ij}],\ 2 \le t \le T,\ 1 \le j \le N$$

After the two steps above, we get the log likelihood by doing

$$log(P^*) = log(\max_{1\le i\le N}[\delta_T(i)])$$

Result:

---------- Accuracy: Viterbi ----------

	Audio_Recognition	Video_Recognition	AV_Recognition
2	1	0.6	1
5	1	1	1
Overall	1	0.8	1

Discussion:

From the result, we can see that the Viterbi algorithm and the forward-backward algorithm give us the exactly classification accuracy and classification result.