# Assignment 2 Summaries for two chapters/sessions

## Sec 9.5 HMM training: Forward and Backward Algorithm.

Wenyue Liu 730028157

Typically, HMM model can be used to solve three types of questions.

1. Given model λ, how to calculate the probability of P(x| λ). In other words, how to evaluate the degree of matching between the model estimate and the observation?
2. Given λ and observation sequence O, how to find the best state sequence. In other words, how to find the hidden model?
3. Given O, how to train model?

The general idea for question 3 is by using EM algorithm, iteratively approximate the best model. For the EM algorithm, two hidden parameters are $\xi_t(i, j)$ and $\gamma_t(j)$. $\xi_t(i, j)$ is the probability of being in state i at time t and in state j at time t+1, this parameter is highly correlated to the A matrix (transmission matrix). The other probability $\gamma_t(j)$ is the probability of being in j at time t. This parameter is highly correlated to the emission matrix (B matrix).

$\xi_t(i, j) = P(q_t = i, q_{t+1} = j|O,\lambda)$, we can calculate it by $\xi_t(\mathrm{i}, \mathrm{j}) = \frac{P(q_t=i, q_{i+1}=j\ O|\lambda)}{P(O|\lambda)} = \frac{a_t(i)a_{ij}b_j(o_{t+1})\beta_{t+1}(j)}{\alpha_F(q_F)}$, and for $\gamma_t(j)$, the logic is similar, $\gamma_t(j) = \frac{\alpha_t(j)\beta_t(j)}{\alpha_F(q_F)}$.

The disadvantage for this method is EM algorithm provide generative result rather than discriminative result, the initial state has strong inference to the final answer. Therefore, for different tasks, the select of AB matrix initialization is an interesting topic.

## Sec 10.5 Maximum Entropy Markov Model (MEMM)

HMM is a generative model, however, a discriminative model (MEMM) is preferred in Part-of-Speech Tagging problem. In HMM, we estimate the tag through the likelihood, but in MEMM, we calculate the posterior directly. By using the discriminative model, we can introduce more features to enhance the prediction accuracy.

Rather than left to right one by one encoding, we use Viterbi encoding method as what we used in HMM to find the best sequence of tags, which is the optimal for the whole sentence.

For HMM, the Viterbi implication is

$$v_t(j) = \max v_{t-1}(i)P(s_j|s_i)P(o_t|s_j); 1 \le j \le N, 1 \le t \le T$$

For MEMM, rather than calculate the likelihood, we use posterior directly, so:

$$v_t(j) = \max v_{t-1}(i)P(s_j|s_i, o_t); 1 \le j \le N, 1 \le t \le T$$

In MEMM, we use logistical regression to calculate the weights on observations, features and hidden states. Logistical regression is a weak classifier, if we can introduce some powerful algorithm, it may provide better performance.