Statistical Model Training

Notes on Speech and Audio Processing

Chia-Ping Chen

Department of Computer Science and Engineering
National Sun Yat-Sen University
Kaohsiung, Taiwan ROC

Introduction

- We have seen how one can compute data-likelihood and posterior probability with HMM through the forward-backward algorithm.
- There is one problem left: In order to compute the likelihood, the parameters in the model must be known. How do we know their values?
- This is not like throwing dice or flipping coin that we can reasonably assign probabilities. In this case, the parameters must be learned from data.
- We have seen that maximum-likelihood criterion can be used in model training and the EM algorithm is one way to do it. Here we apply EM to the HMMs.

The $\mathcal{Q}$ Function for HMM

- When applying to HMM, the hidden variables are the sequence of states. Let Q denote a state sequence. Define the $\mathcal{Q}$ function as

$$\mathcal{Q} \triangleq \sum_{Q} p(Q|O,\Theta_o) \log p(Q,O|\Theta).$$

- We will show how to simply the $\mathcal{Q}$ function and relate it to quantities computable from the forward-backward algorithm.

Simplifying $\mathcal{Q}$ Function

- From the independence assumption of HMM,

$$p(Q,O) = p(Q)p(O|Q)$$
$$= p(q_1)\prod_{t=2}^{T} p(q_t|q_{t-1})\prod_{t=1}^{T} p(o_t|q_t).$$

- Taking the logarithm, we have

$$\log p(Q,O)$$
$$= \log p(q_1) + \sum_{t=2}^{T}\log p(q_t|q_{t-1}) + \sum_{t=1}^{T}\log p(o_t|q_t).$$

Simplifying $\mathcal{Q}$ Function II

- Putting it together,

$$
\begin{aligned}
\mathcal{Q} &\triangleq \sum_{Q} p(Q|O,\Theta_o) \log p(Q,O|\Theta) \\
&= \sum_{Q} p(Q|O,\Theta_o) \log p(q_1|\Theta) + \sum_{Q} p(Q|O,\Theta_o) \sum_{t=1}^{T} \log p(o_t|q_t,\Theta) \\
&\quad + \sum_{Q} p(Q|O,\Theta_o) \sum_{t=2}^{T} \log p(q_t|q_{t-1},\Theta) \\
&= \sum_{i=2}^{N-1} p(q_1 = i|O) \log \pi_i + \sum_{t=1}^{T} \sum_{i=2}^{N-1} p(q_t = i|O) \log b_i(o_t) \\
&\quad + \sum_{t=2}^{T} \sum_{i=2}^{N-1} \sum_{j=2}^{N-1} p(q_{t-1} = i, q_t = j|O) \log a_{ij}
\end{aligned}
$$

The Posterior Probabilities

- The posterior probabilities can be computed through forward-backward algorithm. Specifically

$$\gamma_i(t) = p(q_t = i|O) = \frac{\alpha_i(t)\beta_i(t)}{\sum_j \alpha_j(t)\beta_j(t)}$$

$$\xi_{ij}(t) = p(q_t = i, q_{t+1} = j|O) = \frac{p(q_t = i, q_{t+1} = j, O)}{p(O)}$$

where the joint probability of $p(q_t = i, q_{t+1} = j, O)$ is given by

$$p(q_t = i, q_{t+1} = j, O) = \alpha_i(t)a_{ij}b_j(o_{t+1})\beta_j(t+1).$$

Occupancy Numbers

- The expected number of transitions from state i to state j at time t is $\xi_{ij}(t)$. The expected number of transitions from state i to state j is

$$\sum_{t=1}^{T-1} \xi_{ij}(t).$$

- The occupancy number for state i is the expected number of times that $q_t = i$, and is given by

$$\sum_{t=1}^{T-1} \gamma_i(t).$$

Parameter Update Equations

The parameters are uncoupled in the Q function so the maximization can be carried out independently. The new set of parameters are

$$\begin{cases} \pi_i^* = \gamma_i(1) \\ a_{ij}^* = \frac{\sum_t \xi_{ij}(t)}{\sum_t \gamma_i(t)} \\ \mu_i^* = \frac{\sum_t \gamma_i(t) o_t}{\sum_t \gamma_i(t)} \\ {\sigma^2}_i^* = \frac{\sum_t \gamma_i(t)(o_t - \mu_i)(o_t - \mu_i)'}{\sum_t \gamma_i(t)} \end{cases}$$

One epoch of training finishes here and another starts. The learning continues until some stopping criterion is met.