Hidden Markov Model - II

Parameter Estimation for HMM (Training)

The *maximum likelihood estimate* method maximizes the data likelihood to decide the parameter value. That is,

$$\lambda^* = \arg\max_{\lambda} p(\mathbf{o}|\lambda) \tag{1}$$

It would be great if the above equation yields closed-form solution for λ^*. In the case that such is not possible, we have the following iterative algorithm for parameter re-estimation.

Auxiliary Function

Consider the expectation value of the joint probability $\log p(\mathbf{S}, \mathbf{o}|\lambda)$, i.e.,

$$E \log p(\mathbf{S}, \mathbf{o}|\lambda) = \sum_{\mathbf{s}} p(\mathbf{s}|\mathbf{o}, \lambda) \log p(\mathbf{s}, \mathbf{o}|\lambda) \tag{2}$$

where $\mathbf{S}$ denotes the random state sequence. To maximize (2) with respect to λ iteratively, we first define an auxiliary function

$$\mathcal{Q}(\lambda, \lambda_o) = \sum_{\mathbf{s}} p(\mathbf{s}|\mathbf{o}, \lambda_o) \log p(\mathbf{s}, \mathbf{o}|\lambda) \tag{3}$$

Note the posterior probability $p(\mathbf{s}|\mathbf{o}, \lambda_o)$ is computed according to λ_o (known), while $\log p(\mathbf{s}, \mathbf{o}|\lambda)$ depends on the variable λ.

Data Likelihood and $\mathcal{Q}(\lambda, \lambda_o)$

$\mathcal{Q}(\lambda, \lambda_o)$ and the data likelihood $p(\mathbf{o}|\lambda)$ are related by

$$\begin{aligned}
&\mathcal{Q}(\lambda, \lambda_o) - \mathcal{Q}(\lambda_o, \lambda_o) \\
&= \sum_{\mathbf{s}'} [p(\mathbf{s}'|\mathbf{o}, \lambda_o) \log p(\mathbf{s}', \mathbf{o}|\lambda) - p(\mathbf{s}'|\mathbf{o}, \lambda_o) \log p(\mathbf{s}', \mathbf{o}|\lambda_o)] \\
&= \sum_{\mathbf{s}'} p(\mathbf{s}'|\mathbf{o}, \lambda_o) [\log p(\mathbf{o}|\lambda) + \log p(\mathbf{s}'|\mathbf{o}, \lambda)] - \\
&\quad \sum_{\mathbf{s}'} p(\mathbf{s}'|\mathbf{o}, \lambda_o) [\log p(\mathbf{o}|\lambda_o) + \log p(\mathbf{s}'|\mathbf{o}, \lambda_o)] \\
&= \log p(\mathbf{o}|\lambda) - \log p(\mathbf{o}|\lambda_o) - \sum_{\mathbf{s}'} p(\mathbf{s}'|\mathbf{o}, \lambda_o) \log \frac{p(\mathbf{s}'|\mathbf{o}, \lambda_o)}{p(\mathbf{s}'|\mathbf{o}, \lambda)} \\
&= \log p(\mathbf{o}|\lambda) - \log p(\mathbf{o}|\lambda_o) - D(p_o||p).
\end{aligned} \tag{4}$$

It follows that

$$\begin{aligned}
\log p(\mathbf{o}|\lambda^*) - \log p(\mathbf{o}|\lambda_o) &= \mathcal{Q}(\lambda^*, \lambda_o) - \mathcal{Q}(\lambda_o, \lambda_o) + D(p_o||p) \\
&\geq \mathcal{Q}(\lambda^*, \lambda_o) - \mathcal{Q}(\lambda_o, \lambda_o)
\end{aligned} \tag{5}$$

since $D(p_o||p)$, the *KL-distance* between distributions p_o and p, is always non-negative. Suppose λ^* maximizes $\mathcal{Q}(\lambda, \lambda_o)$. The data likelihood $p(\mathbf{o}|\lambda)$ is *non-decreasing* from λ_o to λ^*, and eventually converges to a local maximum.

$\mathcal{Q}$ Function with HMM

From the *conditional independence assumptions* of HMM, we have

$$p(\mathbf{s}, \mathbf{o}) = p(\mathbf{s})p(\mathbf{o}|\mathbf{s}) = p(s_1) \prod_{t=2}^{T} p(s_t|s_{t-1}) \prod_{t=1}^{T} p(o_t|s_t). \tag{6}$$

Taking logarithm, we have

$$\log p(\mathbf{s}, \mathbf{o}) = \log p(s_1) + \sum_{t=2}^{T} \log p(s_t|s_{t-1}) + \sum_{t=1}^{T} \log p(o_t|s_t). \tag{7}$$

Using (7) in (2), we have

$$\begin{aligned}
\mathcal{Q}(\lambda, \lambda_o) &= \sum_{\mathbf{s}'} p(\mathbf{s}'|\mathbf{o}, \lambda_o) \log p(\mathbf{s}', \mathbf{o}|\lambda) \\
&= \sum_{\mathbf{s}'} p(\mathbf{s}'|\mathbf{o}, \lambda_o) \log p(s_1|\lambda) + \sum_{\mathbf{s}'} p(\mathbf{s}'|\mathbf{o}, \lambda_o) \sum_{t=1}^{T} \log p(o_t|s_t, \lambda) \\
&\quad + \sum_{\mathbf{s}'} p(\mathbf{s}'|\mathbf{o}, \lambda_o) \sum_{t=2}^{T} \log p(s_t|s_{t-1}, \lambda) \\
&= \sum_{i=2}^{N-1} p(S_1 = i|\mathbf{o}) \log \pi_i + \sum_{t=1}^{T} \sum_{i=2}^{N-1} p(S_t = i|\mathbf{o}) \log b_i(o_t) \\
&\quad + \sum_{t=2}^{T} \sum_{i=2}^{N-1} \sum_{j=2}^{N-1} p(S_{t-1} = i, S_t = j|\mathbf{o}) \log a_{ij}
\end{aligned} \tag{8}$$

Posterior Probabilities

In (8), the posterior probability of state i at time t, and the posterior probability of states i, j at consecutive times $t, t+1$ can be computed as follows

$$\begin{aligned}
\gamma_i(t) &\triangleq p(S_t = i|\mathbf{o}) \\
&= \frac{p(S_t = i, \mathbf{o})}{p(\mathbf{o})} \\
&= \frac{\alpha_i(t)\beta_i(t)}{\sum_j \alpha_j(t)\beta_j(t)}
\end{aligned} \tag{9}$$

$$\begin{aligned}
\xi_{ij}(t) &\triangleq p(S_t = i, S_{t+1} = j|\mathbf{o}) \\
&= \frac{p(S_t = i, S_{t+1} = j, \mathbf{o})}{p(\mathbf{o})} \\
&= \frac{\alpha_i(t) a_{ij} b_j(o_{t+1}) \beta_j(t+1)}{p(\mathbf{o})}
\end{aligned} \tag{10}$$

State Occupancy

Let $I(S_t = i|\mathbf{o})$ be the indicator function of the event that $S_t = i$. It is a random variable with value 0 or 1. The total number of occupancy for state i is

$$\sum_{t=1}^{T} I(S_t = i|\mathbf{o}) \tag{11}$$

with the expectation value of

$$C(i|\mathbf{o}) = E\left(\sum_{t=1}^{T} I(S_t = i|\mathbf{o})\right) = \sum_{t=1}^{T} E(I(S_t = i|\mathbf{o})) = \sum_{t=1}^{T} \gamma_i(t). \tag{12}$$

State Transition

Let $I(S_t = i, S_{t+1} = j|\mathbf{o})$ be the indicator function of the event that $S_t = i$ and $S_{t+1} = j$. The expectation value of the total number of transitions from state i to state j is

$$\sum_{t=1}^{T-1} \xi_{ij}(t). \tag{13}$$

Parameter Update

The parameter set is updated according to

$$\begin{aligned}
\pi_i^* &= \gamma_i(1) \\
a_{ij}^* &= \frac{\sum_t \xi_{ij}(t)}{\sum_t \gamma_i(t)} \\
b_j^*(k) &= \frac{\sum_{t \in \{t|o_{t+1}=k\}} \sum_i \xi_{ij}(t)}{\sum_t \sum_i \xi_{ij}(t)}
\end{aligned} \tag{14}$$

where the denominators and the numerators are the probability counts.

Speech as HMMs

- Each phone (or other acoustic unit) is an HMM with a number of states depending on the length.
- It follows all words, sentences are HMMs as well, since they are concatenation of the phone HMMs.

Common Practices

- State emitting probability is often modelled by the Gaussian mixture model (GMM).
- The GMMs can be initialized by k-means clustering or a global mean and covariance.
- The number of mixtures can be increased incrementally via splitting.
- The initial parameters of new mixtures are dependent on the parent mixtures.
- The HMM *state transition diagram* is often *left-to-right*, sometimes allowing *state-skipping*.

Parameters and Data

- The model complexity is often measured in terms of the total number of parameters.
- This number is closely related to the amount of training data, to avoid *over-training* and *under-training*.
- We also apply *parameter-tying schemes* to strike a balance between reliable estimates and the refinements of the models.

Decoding Speech

The basic problem of ASR is to find an “optimal” word sequence given acoustic observations. That is,

$$\hat{W} = \arg\max_{W} p(W|\mathbf{o}). \tag{15}$$

This is the same as

$$\hat{W} = \arg\max_{W} \frac{p(\mathbf{o}|W)p(W)}{p(\mathbf{o})} = \arg\max_{W} p(\mathbf{o}|W)p(W) \tag{16}$$

where $p(\mathbf{o}|W)$ is called the *acoustic model score* and $p(W)$ is called the *language model score*.

Evaluation Measure: Word Error Rate

$$\text{WER} = \frac{S + D + I}{N} \times 100\% \tag{17}$$

where N is the number of tokens in the reference, S is the number of substitution errors, D is the number of deletion errors, and I is the number of insertion errors. Note that S, D, I is determined by a minimal-editorial-distance (MED) alignment between the recognition hypothesis and the reference.