UniS Centre for Vision Speech & Signal Processing
University of Surrey, Guildford GU2 7XH.

HMM tutorial 3

by Dr Philip Jackson

- Recap. of α, β and Viterbi
- Re-estimating models
  - Occupation
  - Transition
  - Baum-Welch formulae
- Gaussian pdf examples
  - Least squares
  - Maximum likelihood
  - B-W re-estimation
- Summary

http://www.ee.surrey.ac.uk/Personal/P.Jackson/tutorial/

Problem 1: Forward procedure

Consider $\alpha_t(i) = P(o_1, o_2, \ldots, o_t, x_t = i|\lambda)$:

1. Initially,
   $\alpha_1(i) = \pi_i \, b_i(o_1)$, for $1 \leq i \leq N$;

2. For $t = 2, 3, \ldots, T$,
   $\alpha_t(j) = \left[\sum_{i=1}^{N} \alpha_{t-1}(i) \, a_{ij}\right] b_j(o_t)$, for $1 \leq j \leq N$;

3. Finally,
   $P(\mathcal{O}|\lambda) = \sum_{i=1}^{N} \alpha_T(i)$.

Thus, we can solve Problem 1 efficiently by recursion.

Problem 1: Backward procedure

Define $\beta_t(i) = P(o_{t+1}, o_{t+2}, \ldots, o_T | x_t = i, \lambda)$:

1. Initially,
   $\beta_T(i) = 1,$ for $1 \leq i \leq N$;

2. For $t = T-1, T-2, \ldots, 1$,
   $\beta_t(i) = \sum_{j=1}^{N} a_{ij}\, b_j(o_{t+1})\, \beta_{t+1}(j),$ for $1 \leq i \leq N$;

3. Finally,
   $P(\mathcal{O}|\lambda) = \sum_{i=1}^{N} \pi_i\, b_i(o_1)\, \beta_1(i).$

We now have another efficient way of computing $P(\mathcal{O}|\lambda)$.

Problem 2: Viterbi algorithm

1. Initially,
   $\delta_1(i) = \pi_i b_i(o_1)$
   $\psi_1(i) = 0$ for $1 \leq i \leq N$;

2. For $t = 2, \ldots, T$,
   $\delta_t(j) = \max_i \left[\delta_{t-1}(i) a_{ij}\right] b_j(o_t)$
   $\psi_t(j) = \arg\max_i \left[\delta_{t-1}(i) a_{ij}\right]$ for $1 \leq j \leq N$;

3. Finally,
   $\Delta^* = \max_i [\delta_T(i)]$
   $x_T^* = \arg\max_i [\delta_T(i)]$;

4. Trace back, for $t = T-1, T-2, \ldots, 1$,

$$x_t^* = \psi_{t+1}\left(x_{t+1}^*\right), \quad \text{and} \quad X^* = \{x_1^*, x_2^*, \ldots, x_T^*\}. \quad (1)$$

Problem 2: Trellis diagram

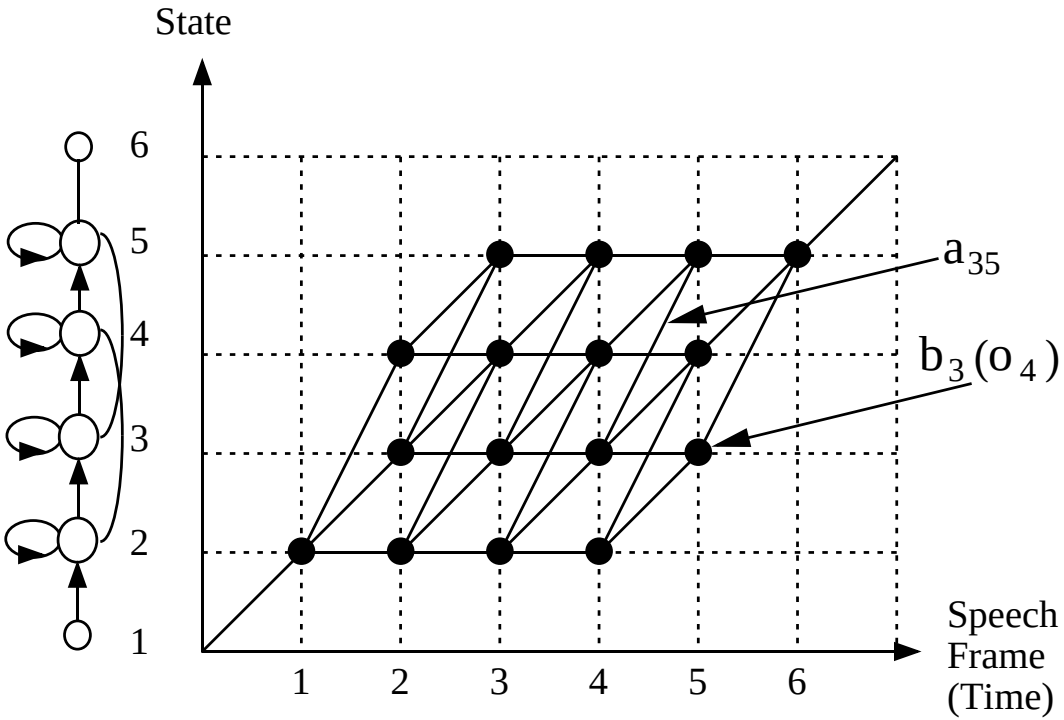

Figure 1.6 The Viterbi algorithm for Isolated Word Recognition.
From (Young et al. 1997), p. 10.

Problem 3: Re-estimation of λ

Parameters of the model $\lambda = \{\pi, A, B\}$ are adjusted to maximise $P(\mathcal{O}|\lambda)$, i.e., according to the ML criterion.

As before, in eq. 4 (Tut. 2),

$$P(\mathcal{O}|\lambda) = \sum_X P(\mathcal{O}, X|\lambda), \tag{2}$$

which means considering all possible state sequences.

We can perform this maximisation using the *Baum-Welch* formulae.

Baum-Welch re-estimation (occupation)

Consider

$$\gamma_t(i) \;=\; P\left(x_t = i | \mathcal{O}, \lambda\right). \tag{3}$$

By Bayes law,

$$\begin{aligned} \gamma_t(i) \;&=\; \frac{P\left(x_t = i, \mathcal{O}|\lambda\right)}{P(\mathcal{O}|\lambda)} \\ &=\; \frac{\alpha_t(i)\,\beta_t(i)}{P(\mathcal{O}|\lambda)}, \end{aligned} \tag{4}$$

where α_t and β_t are as in eqs. 5 & 6 previously, and $P(\mathcal{O}|\lambda) = \sum_{i=1}^{N} \alpha_T(i)$ is the solution to Problem 1.

Baum-Welch re-estimation (transition)

Define

$$\xi_t(i,j) \;=\; P\left(x_{t-1}=i, x_t=j|\mathcal{O},\lambda\right). \tag{5}$$

Similarly, by Bayes law,

$$\begin{aligned}
\xi_t(i,j) \;&=\; \frac{P\left(x_{t-1}=i, x_t=j, \mathcal{O}|\lambda\right)}{P(\mathcal{O}|\lambda)} \\
&=\; \left[\, P\left(x_{t-1}=i, o_1, \ldots, o_{t-1}|\lambda\right) \times \right. \\
&\qquad \left. P\left(x_t=j, o_t, \ldots, o_T|x_{t-1}=i, \lambda\right) \,\right] / P(\mathcal{O}|\lambda) \\
&=\; \left[\, \alpha_{t-1}(i)\; P\left(x_t=j, o_t|x_{t-1}=i, \lambda\right) \times \right. \\
&\qquad \left. P\left(o_{t+1}, \ldots, o_T|x_t=j, \lambda\right) \,\right] / P(\mathcal{O}|\lambda) \\
&=\; \frac{\alpha_{t-1}(i)\; a_{ij}\, b_j(o_t)\; \beta_t(j)}{P(\mathcal{O}|\lambda)}.
\end{aligned} \tag{6}$$

Baum-Welch re-estimation formulae

(a) Initial-state probabilities,

$$\hat{\pi}_i = \gamma_1(i) \qquad \text{for } 1 \leq i \leq N;$$

(b) State-transition probabilities,

$$\hat{a}_{ij} = \frac{\sum_{t=2}^{T} \xi_t(i,j)}{\sum_{t=2}^{T} \gamma_t(i)} \qquad \text{for } 1 \leq i, j \leq N;$$

(c) Discrete output probabilities,

$$\hat{b}_j(k) = \frac{\sum_{t=1}^{T}\Big|_{o_t=k} \gamma_t(j)}{\sum_{t=1}^{T} \gamma_t(j)} \qquad \text{for } 1 \leq j \leq N;$$

and $1 \leq k \leq K$.

For the new model $\hat{\lambda}$, it can be shown that,

$$P(\mathcal{O}|\hat{\lambda}) \geq P(\mathcal{O}|\lambda), \tag{7}$$

although it does not guarantee a *global* maximum.

Parameter estimation examples

Example 0: LS estimate of the mean

We have a set of measurements $\mathcal{O} = \{o_1, o_2, \ldots, o_T\}$, from which we would like to estimate the mean, μ.

Starting with a least-squares approach, we can write an expression for the squared distance of the samples from the mean:

$$\begin{aligned} E &= \sum_{t=1}^{T} (o_t - \mu)^2 \\ &= \sum_{t=1}^{T} \left(o_t^2\right) - 2\mu \sum_{t=1}^{T} (o_t) + T\mu^2. \end{aligned} \tag{8}$$

Example 0 (continued)

To find the minimum, the derivative is set to zero:

$$\frac{\partial E}{\partial \mu} = 2T\hat{\mu} - 2\sum_{t=1}^{T}(o_t) = 0$$

$$\Rightarrow \quad \hat{\mu}_{\mathsf{LS}} = \frac{1}{T}\sum_{t=1}^{T}(o_t),$$

which gives the usual equation for evaluating the sample mean.

Example 1: ML estimate of the mean

Now, if we assume that observations are *continuous* and *normal*, of the form

$$o_t = \mu + n_t \quad \text{for } t \in \{1, 2, \ldots, T\},$$

where $n_t \sim \mathcal{N}(0, \Sigma)$ are independent Gaussian random variables with zero mean and variance Σ, estimate the value of μ from a set of T observations.

The likelihood function is of the form (scalar):

$$p(o|\mu) = \prod_{t=1}^{T} \frac{1}{\sqrt{2\pi\Sigma}} \exp\left[-\frac{(o_t - \mu)^2}{2\Sigma}\right].$$

Taking the logarithm and solving the ML equation, gives

$$\widehat{\mu}_{\mathsf{ML}} = \frac{1}{T} \sum_{t=1}^{T} o_t.$$

Example 2: ML estimate of the variance

Now estimate the variance Σ, assuming that μ is known.

It can be shown that the ML-estimated variance is

$$\hat{\Sigma}_{\mathsf{ML}} = \frac{1}{T} \sum_{t=1}^{T} (o_t - \mu)^2.$$

ML estimates for a multivariate Gaussian

Similarly, we can derive maximum likelihood estimates of the mean vector $\boldsymbol{\mu}$ and the covariance matrix $\boldsymbol{\Sigma}$ using their respective moments:

$$\hat{\boldsymbol{\mu}} = \frac{1}{T}\sum_{t=1}^{T}\mathbf{o}_t \tag{9}$$

and

$$\hat{\boldsymbol{\Sigma}} = \frac{1}{T}\sum_{t=1}^{T}(\mathbf{o}_t - \boldsymbol{\mu})(\mathbf{o}_t - \boldsymbol{\mu})'. \tag{10}$$

B-W re-estimation of Gaussian state parameters

Assuming that the observations come from an HMM with a *continuous* multivariate Gaussian distribution, i.e.:

$$b_j(\mathbf{o}_t) = \mathcal{N}\left(\mathbf{o}_t; \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j\right), \tag{11}$$

we can make a soft (i.e., probabilitistic) allocation of the observations to the states. Thus, if $\gamma_t(j)$ denotes the likelihood of being in state j at time t then eqs. 9 and 10 become weighted averages,

$$\widehat{\boldsymbol{\mu}}_j = \frac{\sum_{t=1}^{T} \gamma_t(j)\mathbf{o}_t}{\sum_{t=1}^{T} \gamma_t(j)} \tag{12}$$

and

$$\widehat{\boldsymbol{\Sigma}}_j = \frac{\sum_{t=1}^{T} \gamma_t(j)(\mathbf{o}_t - \boldsymbol{\mu}_j)(\mathbf{o}_t - \boldsymbol{\mu}_j)'}{\sum_{t=1}^{T} \gamma_t(j)}, \tag{13}$$

normalised by a denominator which is the total likelihood of all paths through node j.

Today's summary

- Recap. of likelihoods α_t and β_t
- Recap. of Viterbi algorithm
- Re-estimating models, $\Lambda = \{\lambda\}$
  - Occupation and transition
  - Baum-Welch formulae
- Gaussian pdf examples, $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$
  - Least squares
  - Maximum likelihood
  - B-W re-estimation

Next time

- Worked example illustrating Baum-Welch algorithm
- More on output pdfs: multivariate Gaussians, Gaussian mixtures and other types of pdf

Homework

1. Using the models you built last week and the forward-backward algorithm:
   - re-estimate model parameters for one state;
   - check for increased likelihood with new model.
2. Using Viterbi, test whether the updated parameters of your new model alter the state alignment.