An Erratum for ``A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition''

Send bugs to: Ali Rahimi (ali@mit.edu)

Introduction

Rabiner's excellent tutorial on hidden markov models [1] contains a few subtle mistakes which can result in flawed HMM implementations. This note is intended as a companion to the tutorial and addresses subtle mistakes which appear the sections on ``scaling'' and ``multiple observations sequences.'' Following is a summary of the terms introduced in the tutorial and corrections to some of the equations.

Definitions and Notation

Section III.A of Rabiner introduces in eq (r18) the *forward variable* α:

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

and the *backward variable* β:

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

Equation (r37) in section III.C defines the posterior probability of going from state i to state j at time t and shows that it can be computed in terms of the forward and backward variables:

$$\begin{aligned} \xi_t(i,j) &= P(q_t = S_i, q_{t+1} = S_j|O,\lambda) \\ &= \frac{\alpha_t(i) a_{ij} b_j(O_{t+1}) \beta_{t+1}(j)}{P(O|\lambda)} \end{aligned} \quad (1)$$

Finally, in eqs (r38) and (r27), it is observed that $\gamma_t(i)$, the probability of being in state i at time t, is related to ξ_t and can be written in two different forms:

$$\begin{aligned} \gamma_t(i) &= \sum_{j=1}^{N} \xi_t(i,j) \\ &= \frac{\alpha_t(i)\beta_t(i)}{P(O|\lambda)} \end{aligned} \tag{2}$$

These terms rely on the forward and backward variables, but modern floating point engines do not have the necessary precision to compute α and β according to the recursions provided by equations (r20) and (r25).

Instead, section V.A of Rabiner introduces a way to compute γ and ξ using alternative entities.

Scaling

Section V.A introduces the *scaled forward variable* $\hat{\alpha}$ and the *scaled backward variable* $\hat{\beta}$. These variables are easy to compute on modern machines and will not result in underflows. Section V.A also describes how to use the unscaled variables to compute γ and ξ.

Rabiner's eqs (r91-r92b) for computing $\hat{\alpha}$ are misleading, and no recursion is provided for computing $\hat{\beta}$. This section derives recursions for both $\hat{\alpha}$ and $\hat{\beta}$.

The scaled forward variable

We are looking for a recursion to calculate a variable $\hat{\alpha}$ such that

$$\hat{\alpha}_t(i) = \frac{\alpha_t(i)}{\sum_{j=1}^{N} \alpha_t(j)} = C_t \alpha_t(i) \tag{3}$$

The following is a corrected version of the recursion of eqs (r91-r92b)

$$\begin{aligned} \bar{\alpha}_1(i) &= \alpha_1(i) \\ \bar{\alpha}_{t+1}(j) &= \sum_{i=1}^{N} \hat{\alpha}_t(i) a_{ij} b_j(O_{t+1}) \\ c_{t+1} &= \frac{1}{\sum_i \bar{\alpha}_{t+1}(i)} \end{aligned}$$

$$\hat{\alpha}_{t+1}(i) = c_{t+1}\bar{\alpha}_{t+1}(i)$$

The proof that this recursion results in the criterion of eq (3) is by induction:

Base case. According to the recursion, we get

$$\bar{\alpha}_1(i) = \alpha_1(i), \quad \hat{\alpha}_1(i) = \frac{\alpha_1(i)}{\sum_j \alpha_1(j)}$$

which satisfies the condition of eq (3) with $C_1 = c_1$.

Induction. If $\hat{\alpha}_t = C_t \alpha_t$, then

$$\begin{aligned}
\bar{\alpha}_{t+1}(j) &= C_t \sum_{i=1}^{N} \alpha_t(i) a_{ij} b_j(O_{t+1}) \\
&= C_t \alpha_{t+1}(j) \\
c_{t+1} &= \frac{1}{\sum \bar{\alpha}_{t+1}(i)} = \frac{1}{C_t \sum \alpha_{t+1}(j)} \qquad (4) \\
\hat{\alpha}_{t+1}(i) &= c_{t+1}\bar{\alpha}_{t+1}(i) \\
&= \frac{C_t \alpha_{t+1}(i)}{C_t \sum \alpha_{t+1}(i)} \\
&= \frac{\alpha_{t+1}(i)}{\sum \alpha_{t+1}(j)}
\end{aligned}$$

which was what we wanted to show. Note that as a consequence of eq (4) and the definition of C_t in eq (3), we obtain a useful expression for C_t in terms of c_t:

$$\begin{aligned}
C_t &= \frac{1}{c_{t+1} \sum \alpha_{t+1}(j)} = \frac{C_{t+1}}{c_{t+1}} \\
C_t &= C_{t-1} c_t = \prod_{\tau=1}^{t} c_\tau \qquad (5)
\end{aligned}$$

we also define the term D_t which we will use to scale β, and observe its realtionship to C_t:

$$D_t = \prod_{\tau=t}^{T} c_\tau \tag{6}$$

$$C_t D_{t+1} = \prod_{\tau=1}^{t} c_\tau \prod_{\tau=t+1}^{T} c_\tau = \prod_{\tau=1}^{T} c_\tau = C_T \tag{7}$$

The scaled backward variable

The following recursion produces desired values of $\hat{\beta}$ if we satisfy ourselves with defining $\hat{\beta}_t(i) = D_t \beta_t(i)$:

$$\begin{aligned} \bar{\beta}_T(i) &= \beta_T(i) \\ \bar{\beta}_t(j) &= \sum_{i=1}^{N} a_{ij} b_j(O_{t+1}) \hat{\beta}_{t+1}(i) \\ \hat{\beta}_t(i) &= c_t \bar{\beta}_t(i) \end{aligned}$$

Note that defining $\hat{\beta}_t(i) = D_t \beta_t(i)$ is *not* the same as imposing the requirement

$$\hat{\beta}_t(i) = \frac{\beta_t(i)}{\sum_{j=1}^{N} \alpha_t(j)} \text{ (not true!)}$$

The proof that the recursion produces the desired result is again inductive:

Base case.

$$\bar{\beta}_T(i) = \beta_T(i), \quad \hat{\beta}_T(i) = D_T \beta_T$$

which satisfies the condition of the backward scaling with $D_T = c_T$.

Induction. If $\hat{\beta}_{t+1} = D_{t+1} \beta_{t+1}$, then

$$\begin{aligned}\bar{\beta}_t(j) &= D_{t+1}\sum_{i=1}^{N} a_{ij}b_j(O_{t+1})\beta_{t+1}(i) \\ &= D_{t+1}\beta_t(j) \\ \hat{\beta}_t(j) &= c_t\bar{\beta}_t(j) = c_t D_{t+1}\beta_t = D_t\beta_t.\end{aligned} \tag{8}$$

The last step uses the definition of D_t from eq (6) and produces a result in agreement with the scaling requirement.

We have shown recursions for computing $\hat{\alpha}_t = C_t\alpha_t$ and $\hat{\beta}_t = D_t\beta_t$, with C_t and D_t defined by eqs (5,6). The next section uses provides alternative ways of computing ξ and γ using these variables.

Using $\hat{\alpha}$ and $\hat{\beta}$

Substituting the scaled variables in the definition for ξ, we get:

$$\begin{aligned}\xi_t(i,j) &= \frac{1}{P(O|\lambda)}\alpha_t(i)a_{ij}b_j(O_{t+1})\beta_{t+1}(j) \\ &= \hat{\alpha}_t(i)a_{ij}b_j(O_{t+1})\hat{\beta}_{t+1}(j)\frac{1}{P(O|\lambda)}\frac{1}{C_t}\frac{1}{D_{t+1}}\end{aligned} \tag{9}$$

But $C_t D_{t+1} = C_T$ according to eq (7), and $P(O|\lambda) = \sum_i \alpha_T(i)$, according to eq (r21) of Rabiner, so

$$\begin{aligned}P(O|\lambda) &= \sum_i \alpha_T(i) = 1/C_T \\ P(O|\lambda)C_T &= 1.\end{aligned}$$

Therefore eq (10) simplifies to

$$\xi_t(i,j) = \hat{\alpha}_t(i)a_{ij}b_j(O_{t+1})\hat{\beta}_{t+1}(j) \tag{10}$$

which is a simple way of computing ξ from the scaled variables.

γ can be computed from ξ using eq (2):

$$\begin{aligned}\gamma_t(i) &= \sum_{j=1}^{N}\xi_t(i,j) = \frac{1}{P(O|\lambda)}\alpha_t(i)\beta_t(i) \\ &= \hat{\alpha}_t(i)\hat{\beta}_t(i)\frac{1}{P(O|\lambda)}\frac{1}{C_t}\frac{1}{D_t} \\ &= \hat{\alpha}_t(i)\hat{\beta}_t(i)\frac{1}{c_t} \end{aligned} \quad (11)$$

These two entities can be used as-is in the Baum-Welch and Viterbi algorithms.

Multiple observations sequences

Section V.B of Rabiner explains how to use multiple observations sequences for training. In the M step of Baum-Welch, a new state transition matrix is computed according to eq (r109):

$$\bar{a_{ij}} = \frac{\sum_{k=1}^{K}\sum_{t=1}^{T_k-1}\xi_t(i,j)}{\sum_{k=1}^{K}\sum_{t=1}^{T_k-1}\gamma_t(i)}$$

and the observation matrix is updated according to eq (r110):

$$\bar{b_i}(l) = \frac{\sum_{k=1}^{K}\sum_{t=1,s.t.O_t=v_l}^{T_k-1}\gamma_t(i)}{\sum_{k=1}^{K}\sum_{t=1}^{T_k-1}\gamma_t(i)}$$

Onces ξ and γ have been computed, these equations can be used directly to update the state transition matrix and the emission probabilities. Equation (r111) incorrectly substitutes eqs (2) and (11) into (r109). Equations (13) and (14) are easy to use and should be used for computing the updates. However, for the sake of completeness, equations analogous to (r111) with the correct substitutions are included here:

$$\begin{aligned}\bar{a_{ij}} &= \frac{\sum_{k=1}^{K}\sum_{t=1}^{T_k-1}\hat{\alpha}_t(i)a_{ij}b_j(O_{t+1})\hat{\beta}_{t+1}(j)}{\sum_{k=1}^{K}\sum_{t=1}^{T_k-1}\hat{\alpha}_t(i)\hat{\beta}_t(i)\frac{1}{c_t}} \\ \bar{b_i}(l) &= \frac{\sum_{k=1}^{K}\sum_{t=1,s.t.O_t=v_l}^{T_k-1}\hat{\alpha}_t(i)\hat{\beta}_t(i)\frac{1}{c_t}}{\sum_{k=1}^{K}\sum_{t=1}^{T_k-1}\hat{\alpha}_t(i)\hat{\beta}_t(i)\frac{1}{c_t}}\end{aligned}$$

Conclusion

There are two salient corrections proposed in the paper: the first corrects Rabiner's notation for computing the scaled variables. The second correction is in the way the HMM parameters are updated in the M step under multiple observation sequences. This note also provides an inductive proof that the recursions provide the desired results.

If you notice bugs in this note, please inform the author.

Bibliography

1

L. R. Rabiner.
A tutorial on hidden markov models and selected apllications in speech recognition.
In A. Waibel and K.-F. Lee, editors, *Readings in Speech Recognition*, pages 267-296. Kaufmann, San Mateo, CA, 1990.

Next Group Up Previous

Ali Rahimi 2000-12-30