|  |  |  |  |
| --- | --- | --- | --- |
| Baum-Weltch Algorithm | | | |
| Mohammed Ahmed Mohammed Hossen. 17/12/2023 | | | |
|  | | | |
| Introduction   In this research I will go through the Baum Weltch algorithm.  It is the best time to learn this concept, since I was informed about both Forward / Backward algorithm and Viterbi algorithm.  I am now at this very moment using word by Microsoft and I have only got through the introduction part, and I was corrected during the writing process about 10 times or more  So, it's an interesting concept to learn how this thing works.   Some history   Since we are talking about Baum Weltch it's exciting to know a little thing about him and to waste sometimes so don't go through mathematics.  The Baum–Welch algorithm was named after its inventors Leonard E. Baum and Lloyd R. Welch. The algorithm and the Hidden Markov models were first described in a series of articles by Baum and his peers at the IDA Center for Communications Research, Princeton in the late 1960s and early 1970s. One of the first major applications of HMMs was speech processing. In the 1980s, HMMs were a useful tool in the analysis of biological systems and information, particularly genetic information. They have since become a crucial tool in the probabilistic modeling of genomic sequences. |  |  |  |
| Note that   This The Baum-Welch algorithm is a special case of the expectation-maximization algorithm used to find the unknown parameters of a hidden Markov model (HMM). It makes use of the forward-backward algorithm to compute the statistics for the expectation. |

### How does it work theoretically

Understanding the Baum-Welch Algorithm:

The Baum-Welch algorithm is an Expectation-Maximization (EM) algorithm used to estimate the parameters of an HMM based on a sequence of observations. Unlike the Forward-Backward and Viterbi algorithms, which deal with known parameters, Baum-Welch iteratively refines the HMM parameters to better fit the observed data.

Here's the basic idea:

Expectation Step (E-Step):

Using existing HMM parameters, calculate the forward probabilities (probability of observing the sequence up to a specific state) and backward probabilities (probability of observing the remaining sequence from a specific state).

Estimate the expected number of state transitions and the expected time spent in each state for each observation.

Maximization Step (M-Step):

Using the expectations from the E-Step, re-estimate the HMM parameters (transition probabilities and emission probabilities) to maximize the likelihood of the observations given the current model.

Repeat:

Continue iterating between the E-Step and M-Step until the parameters converge (change minimally) or a defined stopping criterion is met.

Mathematical Equations:

Let:

N be the number of states in the HMM.

T be the length of the observed sequence O = {o\_1, o\_2, ..., o\_T}.

A be the transition probability matrix (NxN).

B be the emission probability matrix (NxM), where M is the number of possible observations.

α\_t(i) be the forward probability at time t in state i.

β\_t(i) be the backward probability at time t in state i.

γ\_t(i) be the expected number of times in state i at time t.

ξ\_t(i,j) be the expected number of transitions from state i to state j at time t.

Then, the key equations are:

E-Step:

α\_t(i) = P(O\_1, ..., o\_t, i | A, B) (computed recursively using the Forward algorithm)

β\_t(i) = P(o\_t+1, ..., O\_T | i, A, B) (computed recursively using the Backward algorithm)

γ\_t(i) = P(i | O, A, B) = α\_t(i) \* β\_t(i) / P(O | A, B)

ξ\_t(i,j) = P(i, j | O, A, B) = γ\_t(i) \* A\_ij \* β\_(t+1)(j) / P(O | A, B)

M-Step:

A\_ij = Σ\_t=1^{T-1} ξ\_t(i,j) / Σ\_t=1^{T-1} γ\_t(i)

B\_i(k) = Σ\_(t=1:T, o\_t=k) γ\_t(i) / Σ\_t=1^T γ\_t(i)

## How does it work with an example

## Numeric Example of Baum-Welch Algorithm

Let's use a simple HMM with:

- 2 states: S1 and S2

- 3 possible observations: A, B, and C

- Transition probabilities:

- P (S1 -> S1) = 0.6, P (S1 -> S2) = 0.4

- P (S2 -> S1) = 0.5, P (S2 -> S2) = 0.5

- Emission probabilities:

- P (A | S1) = 0.5, P (B | S1) = 0.3, P (C | S1) = 0.2

- P (A | S2) = 0.2, P (B | S2) = 0.7, P (C | S2) = 0.1

Now, imagine we observe the sequence: B, A, B, C

\*\*E-Step Iteration 1: \*\*

1. \*\*Forward Probabilities: \*\*

\*\*Time step 1 (Observation: B): \*\*

- α\_1(S1) = P (B | S1) \* pi\_S1 = 0.3 \* 0.5 = 0.15

- α\_1(S2) = P (B | S2) \* pi\_S2 = 0.7 \* 0.5 = 0.35

\*\*Time step 2 (Observation: A): \*\*

- α\_2(S1) = P (A | S1) \* [α\_1(S1) \* P (S1 -> S1) + α\_1(S2) \* P (S2 -> S1)] = 0.5 \* [0.15 \* 0.6 + 0.35 \* 0.5] = 0.0575

- α\_2(S2) = P (A | S2) \* [α\_1(S1) \* P (S1 -> S2) + α\_1(S2) \* P (S2 -> S2)] = 0.2 \* [0.15 \* 0.4 + 0.35 \* 0.5] = 0.035

\*\*Time step 3 (Observation: B): \*\*

- α\_3(S1) = P (B | S1) \* [α\_2(S1) \* P (S1 -> S1) + α\_2(S2) \* P (S2 -> S1)] = 0.3 \* [0.0575 \* 0.6 + 0.035 \* 0.5] = 0.010275

- α\_3(S2) = P (B | S2) \* [α\_2(S1) \* P (S1 -> S2) + α\_2(S2) \* P (S2 -> S2)] = 0.7 \* [0.0575 \* 0.4 + 0.035 \* 0.5] = 0.016225

\*\*Time step 4 (Observation: C): \*\*

- α\_4(S1) = P (C | S1) \* [α\_3(S1) \* P (S1 -> S1) + α\_3(S2) \* P (S2 -> S1)] = 0.2 \* [0.010275 \* 0.6 + 0.016225 \* 0.5] = 0.00178125

- α\_4(S2) = P (C | S2) \* [α\_3(S1) \* P (S1 -> S2) + α\_3(S2) \* P (S2 -> S2)] = 0.1 \* [0.010275 \* 0.4 + 0.016225 \* 0.5] = 0.00101125

So, the forward probabilities for all time steps are:

- α\_1(S1) = 0.15, α\_1(S2) = 0.35

- α\_2(S1) = 0.0575, α\_2(S2) = 0.035

- α\_3(S1) = 0.010275, α\_3(S2) = 0.016225

- α\_4(S1) = 0.00178125, α\_4(S2) = 0.00101125

2. \*\*Backward Probabilities: \*\*

\*\*Time step 4: \*\*

- β\_4(S1) = 1 (assuming no further observations)

- β\_4(S2) = 1

\*\*Time step 3 (Next Observation: C): \*\*

- β\_3(S1) = P (C | S1) \* β\_4(S1) \* P (S1 -> S1) + P (C | S2) \* β\_4(S2) \* P (S1 -> S2) = 0.2 \* 1 \* 0.6 + 0.1 \* 1 \* 0.4 = 0.14

- β\_3(S2) = P (C | S1) \* β\_4(S1) \* P (S2 -> S1) + P (C | S2) \* β\_4(S2) \* P (S2 -> S2) = 0.2 \* 1 \* 0.5 + 0.1 \* 1 \* 0.5 = 0.15

\*\*Time step 2 (Next Observation: B): \*\*

- β\_2(S1) = P (B | S1) \* β\_3(S1) \* P (S1 -> S1) + P (B | S2) \* β\_3(S2) \* P (S1 -> S2) = 0.3 \* 0.14 \* 0.6 + 0.7 \* 0.15 \* 0.4 = 0.0504

- β\_2(S2) = P (B | S1) \* β\_3(S1) \* P (S2 -> S1) + P (B | S2) \* β\_3(S2) \* P (S2 -> S2) = 0.3 \* 0.14 \* 0.5 + 0.7 \* 0.15 \* 0.5 = 0.063

\*\*Time step 1 (Next Observation: A): \*\*

- β\_1(S1) = P (A | S1) \* β\_2(S1) \* P (S1 -> S1) + P (A | S2) \* β\_2(S2) \* P (S1 -> S2) = 0.5 \* 0.0504 \* 0.6 + 0.2 \* 0.063 \* 0.4 = 0.018144

- β\_1(S2) = P (A | S1) \* β\_2(S1) \* P (S2 -> S1) + P (A | S2) \* β\_2(S2) \* P (S2 -> S2) = 0.5 \* 0.0504 \* 0.5 + 0.2 \* 0.063 \* 0.5 = 0.01428

So, the backward probabilities for all time steps are:

- β\_4(S1) = 1 β\_4(S2) = 1

- β\_3(S1) = 0.14 β\_3(S2) = 0.15

- β\_2(S1) = 0.0504 β\_2(S2) = 0.063

- β\_1(S1) = 0.018144 β\_1(S2) = 0.01428

3. \*\*Expected Counts: \*\*

γ\_1(S1) = α\_1(S1) \* β\_1(S1) / P(B) = 0.15 \* 0.018144 / 0.5 = 0.0054432

γ\_1(S2) = α\_1(S2) \* β\_1(S2) / P(B) = 0.35 \* 0.01428 / 0.5 = 0.009996

γ\_2(S1) = α\_2(S1) \* β\_2(S1) / P(A) = 0.0575 \* 0.0504 / 0.25 = 0.01158

γ\_2(S2) = α\_2(S2) \* β\_2(S2) / P(A) = 0.035 \* 0.063 / 0.25 = 0.00882

γ\_3(S1) = α\_3(S1) \* β\_3(S1) / P(B) = 0.010275 \* 0.14 / 0.5 = 0.002877

γ\_3(S2) = α\_3(S2) \* β\_3(S2) / P(B) = 0.016225 \* 0.15 / 0.5 = 0.0048675

γ\_4(S1) = α\_4(S1) \* β\_4(S1) / P(C) = 0.00178125 \* 1 / 0.25 = 0.007125

γ\_4(S2) = α\_4(S2) \* β\_4(S2) / P(C) = 0.00101125 \* 1 / 0.25 = 0.004045

4. \*\*Expected Transitions: \*\*

ξ\_1(S1, S1) = π\_S1 \* γ\_1(S1) \* P\_S1\_S1 \* β\_2(S1) / α\_2(S1) = 0.5 \* 0.3 \*0.6 \* 0.0504 / 0.018144

= 0.009792

ξ\_1(S1, S2) = π\_S1 \* γ\_1(S1) \* P\_S1\_S2 \* β\_2(S2) / α\_2(S2) = 0.0054432 \* 0.5 \*0.4 \*.063 / 0.035 = 0.011168

ξ\_1(S2, S1) =0

ξ\_1(S2, S2) =0

ξ\_2(S1, S1) = γ\_2(S1) \* P (A | S1) \* P\_S1\_S1 \* β\_3(S1) / α\_3(S1) ==.0575 \* 0.5 \* 0.6 \* 0.14 / 0.010275= 0.14997

ξ\_2(S1, S2) = γ\_2(S1) \* P (A | S1) \* P\_S1\_S2 \* β\_3(S2) / α\_3(S1) = 0.0575 \* 0.5 \* 0.4 \* 0.15 / 0.01027 5 = 0.15524

ξ\_2(S2, S1) = γ\_2(S2) \* P (A | S2) \* P\_S2\_S1 \* β\_3(S1) / α\_3(S1) = 0.00882 \* 0.2 \* 0.5 \* 0.14 / 0.010275 = 0.02505

ξ\_2(S2, S2) = γ\_2(S2) \* P (A | S2) \* P\_S2\_S2 \* β\_3(S2) / α\_3(S2) = 0.00882 \* 0.2 \* 0.5 \* 0.15 /

0.016225 = 0.008477

ξ\_3(S1, S1) = γ\_3(S1) \* P (B | S1) \* P\_S1\_S2 \* β\_4(S1) / α\_4(S1) = 0.002877 \* 0.3 \* 0.4 \* 1 /

0.00178125= 0.255977

ξ\_3(S1, S2) = γ\_3(S1) \* P (B | S1) \* P\_S1\_S2 \* β\_4(S1) / α\_4(S1) = 0.002877 \* 0.3 \* 0.4 \* 1 /

0.00178125= 0.255977  
 ξ\_3(S2, S1) = γ\_3(S2) \* P (B | S2) \* P\_S2\_S1 \* β\_4(S1) / α\_4(S1) = 0.0048675 \* 0.7 \* 0.5 \* 1 /

0.00178125 = 0.03937

ξ\_3(S2, S2) = γ\_3(S2) \* P (C | S2) \* P\_S2\_S2 \* β\_4(S2) / α\_4(S2) = 0.0048675 \* 0.1 \* 0.5 \* 1 / 0.00101125 = 0.240625

Please note that ξ\_4(S1, S1) and ξ\_4(S1, S2) are not defined because there is no next state after the 4th time step

\*\*M-Step Iteration 1: \*\*

1. Update transition probabilities:

P\_S1\_S1 = Σ\_t ξ\_t(S1, S1) / Σ\_t γ\_t(S1) = 0.9

P\_S1\_S2 = Σ\_t ξ\_t(S1, S2) / Σ\_t γ\_t(S1) = 0.1

P\_S2\_S1 = Σ\_t ξ\_t(S2, S1) / Σ\_t γ\_t(S2) =0

P\_S2\_S2 = Σ\_t ξ\_t(S2, S2) / Σ\_t γ\_t(S2) =1

2. Update emission probabilities:

B\_S1(A) = 0 (There are no observations of A in state S1 with the BABC sequence.)

B\_S1(B) = γ\_1(S1) / Σ\_o γ\_t(S1) = 0.0054432 / (0.0054432 + 0.01158 + 0.002877 + 0.007125) = 0.292

B\_S1(C) = 1 - B\_S1(A) - B\_S1(B) = 1 - 0 - 0.292 = 0.708

B\_S2(A) = 0 (There are no observations of A in state S2 with the BABC sequence.)

B\_S2(B) = γ\_1(S2) / Σ\_o γ\_t(S2) = 0.009996 / (0.009996 + 0.00882 + 0.0048675 + 0.004045) = 0.543

B\_S2(C) = 1 - B\_S2(A) - B\_S2(B) = 1 - 0 - 0.543 = 0.457

The updated probabilities still indicate that state S1 is more likely to emit C (70.8%) compared to B (29.2%), while state S2 remains more likely to emit B (54.3%) compared to C (45.7%)..

Repeat the E-Step and M-Step until the parameters converge (the changes become negligible). This might take multiple iterations.

This is a simplified example, but it demonstrates the core principles of the Baum-Welch algorithm. With each iteration, the expected counts and transition probabilities refine the HMM to better fit the observed sequence.

## Some life uses of the algorithm

Speech Recognition

The algorithm helps in training models to decipher human speech patterns more accurately. This is crucial in voice recognition software used in virtual assistants like Siri and Alexa, enhancing their ability to understand and respond to spoken commands or queries.

Computational Biology

In genetics and molecular biology, the Baum-Welch algorithm aids in predicting gene sequences from observed data. This application contributes significantly to genetic research, enabling better understanding and potentially aiding in the development of personalized medicine.

Finance

In finance, particularly in stock market analysis, the Baum-Welch algorithm assists in predicting future stock prices. By modeling the stock market as a hidden Markov model, it estimates probabilities associated with different market states, aiding investors in making more informed investment decisions.

The algorithm's adaptability in learning from sequential data, its capacity to extract meaningful patterns, and its ability to estimate hidden parameters make it a valuable tool across diverse fields, improving prediction accuracy and decision-making processes.