Parameter estimation for discrete hidden Markov models

Junko Murakami [(1)] and Tomas Taylor [(2)]

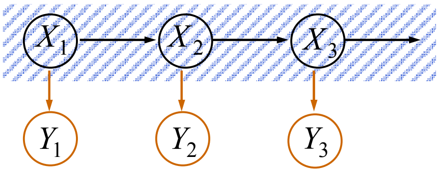
1. Victoria University of Wellington
2. Arizona State University

Outline

- Description of ‘simple’ hidden Markov models
- Maximum likelihood estimate (using Baum-Welch algorithm) – *mode*
- Bayes (or Least square error) estimate – *mean*
- Comparison of the *mode* and *mean*

‘simple’ HMMs?

‘simplest’ HMM (1)



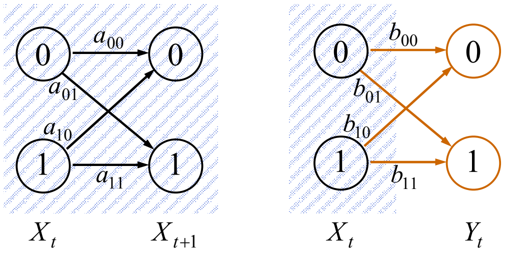
State sequence (Markov chain) $X^{1,n} = (X_1, X_2, \ldots, X_n)$

Observation sequence $Y^{1,n} = (Y_1, Y_2, \ldots, Y_n)$

$X^{1,n} \in \{0,1\}$ and $Y^{1,n} \in \{0,1\}$

‘simplest’ HMM (2)

Conditional Probabilities



Also, let $r_0 = P(X_1 = 0)$ and $r_1 = P(X_1 = 1)$.

Note : $a_{i1} = 1 - a_{i0}$, $b_{i1} = 1 - b_{i0}$, and $r_1 = 1 - r_0$ for $i = 0, 1$.

Baum-Welch Algorithm to Maximize the Log Likelihood

Baum-Welch Algorithm (1)

Consider the likelihood function $L(\theta)$,

$$L(\theta) = P(Y^{1,n}, X^{1,n} \mid \theta),$$

the probability of having sequences $Y^{1,n}$ and $X^{1,n}$, given the parameter set θ, where

$$\theta = \{ r_0, a_{00}, a_{11}, b_{00}, b_{11} \}.$$

Baum-Welch Algorithm (2)

Using the kth estimate, $\theta^{(k)}$, we want $\theta^{(k+1)}$ to be the θ - value that maximizes $Q(\theta, \theta^{(k)})$ defined below.

$$Q(\theta, \theta^{(k)}) = E(\log L(\theta) \mid Y^{1,n}, \theta^{(k)})$$

So, the goal is to maximize the expected value of the log likelihood function, given the observation sequence and the current estimate.

Baum-Welch Algorithm (3)

The algorithm finds two types of probabilities.
Let $i \in \{0, 1\}$.

- ***forward procedure***
  Recursively find $\alpha_t(i) = P(Y^{1,t}, X_t = i \mid \theta^{(k)})$,
  starting from time $t = 1$ up to $t = n$.
- ***backward procedure***
  Recursively find $\beta_t(i) = P(Y^{t+1,n} \mid X_t = i, \theta^{(k)})$,
  starting from time $t = n$ down to $t = 1$.

Then, uses $\alpha = \{\alpha_t(i)\}$, $\beta = \{\beta_t(i)\}$, and $\theta^{(k)}$ to compute $\theta^{(k+1)}$.

Baum-Welch Algorithm (4)

Characteristics

- An implementation of E-M algorithm.
- VERY widely used in various field.

Baum-Welch Algorithm (5)

Advantages

- Maximizes the likelihood the majority of times.
- The convergence is quick enough the majority of times.
- Still feasible when the state and observation space size is large.
- Implementation is easy.

Baum-Welch Algorithm (6)

Disadvantages

- Strong dependency on the initial estimate.
- Guaranteed only to find a local maximum.
- 'Overfitting' problem: not close to the true parameter set when the data size is small.
- Convergence is sometime very slow.
- Online computation is not possible.

Least Square Error (LSE) Estimate (or Bayes Estimate)

LSE (Bayes) Estimate (1)

Finds the expected value of the parameter set given an observation sequence; i.e.,

$$\hat{\theta} = E[\theta] = \int \theta \, P\left(\theta \mid Y^{1,n}\right) d\theta.$$

Assuming the uniform distribution of θ (i.e., letting $P(\theta) = 1$), and using Bayes' theorem, we have

$$E[\theta] = \frac{1}{P\left(Y^{1,n}\right)} \sum_{X^{1,n} \in \Omega_n} \int \theta \, P\left(Y^{1,n}, X^{1,n} \mid \theta\right) d\theta$$

where

$$P\left(Y^{1,n}\right) = \sum_{X^{1,n} \in \Omega_n} \int P\left(Y^{1,n}, X^{1,n} \mid \theta\right) d\theta.$$

LSE (Bayes) Estimate (2)

NOTE

The summation is over $\Omega_n \in \left\{\text{all the possible values of } X^{1,n}\right\}$, which has the size 2^n.

LSE (Bayes) Estimate (3)

First, we let

$$k_{ij} = \#\left(X_t = i \text{ and } X_{t+1} = j\right) \text{ and}$$
$$l_{iu} = \#\left(X_t = i \text{ and } Y_t = u\right)$$

for $i, j, u \in \{0, 1\}$, where # (event) means the total number of events over $t \in \{1, 2, \ldots, n\}$.

Let $K = \{k_{ij}\}$ and $L = \{l_{iu}\}$.

LSE (Bayes) Estimate (4)

If we fix r_0 as $1/2$ for simplicity, $P\left(Y^{1,n}, X^{1,n} \mid \theta\right)$ is in the form

$$\frac{1}{2} a_{00}^{k_{00}} \left(1 - a_{00}\right)^{k_{01}} a_{11}^{k_{11}} \left(1 - a_{11}\right)^{k_{10}} b_{00}^{l_{00}} \left(1 - b_{00}\right)^{l_{01}} b_{11}^{l_{11}} \left(1 - b_{11}\right)^{l_{10}},$$

and so both $\int \theta \, P\left(Y^{1,n}, X^{1,n} \mid \theta\right) d\theta$ and $\int P\left(Y^{1,n}, X^{1,n} \mid \theta\right) d\theta$ are functions of K and L, $\theta = \{r_0, a_{00}, a_{11}, b_{00}, b_{11}\}$.

NOTE: Because of the symmetry in the probability distribution, the integration should be under some restriction; e.g., $a_{00} \geq a_{11}$.

LSE (Bayes) Estimate (5)

Fact

1. To evaluate the integrals, all we need to know are the values of $\{K, L\}$.
2. $\{K, L\}$ can be expressed as a function of $\{k_1, k_{11}, l_{11}, X_1, X_n\}$ instead, where k_1 is the number of 1's in $X^{1,n}$.

→ All we need is $\omega_n = \{k_1, k_{11}, l_{11}, X_1, X_n\}$

LSE (Bayes) Estimate (6)

Fact

Given a particular $Y^{1,n}$, different state sequences $X^{1,n}$ can produce the same value of $\omega_n = \{k_1, k_{11}, l_{11}, X_1, X_n\}$.

→

Let $h_n(\omega_n)$ be the number of $X^{1,n}$ values that corresponds to the ω_n given.

LSE (Bayes) Estimate (7)

If we find the values of $h_n(\omega_n)$ for all ω_n, then the summations can be done over ω_n such that $h_n(\omega_n) > 0$, instead of over all possible values of $X^{1,n} \in \Omega_n$.

The algorithm shows that the number of ω_n values such that $h_n(\omega_n) > 0$ are polynomial of n.

LSE (Bayes) Estimate (7)

If we find the values of $h_n(\omega_n)$ for all ω_n, then the summations can be done over ω_n such that $h_n(\omega_n) > 0$, instead of over all possible values of $X^{1,n} \in \Omega_n$.

The algorithm shows that the number of ω_n values such that $h_n(\omega_n) > 0$ are polynomial of n.

NOTE : The observation state space size can be extended from $m = 2$ (this example) to any integer m in general.

LSE (Bayes) Estimate (8)

Let $h_1(0, 0, 0, 0, 0) = 1$
for t from 1 to $n-1$
    with all $\omega_t = (k_1, k_{11}, l_{11}, 0, X_t)$ such that $h_t(\omega_t) > 0$
        increment $h_{t+1}(k_1, k_{11}, l_{11}, 0, 0)$ and
        $h_{t+1}(k_1 + 1, k_{11} + X_t, l_{11} + X_{t+1}, 0, 1)$
        by the value $h_t(\omega_t)$
end for

(Because of the symmetry, we can find $h_n(\omega_n)$ for $X_1 = 1$ once the ones for $X_1 = 0$ is obtained.)

LSE (Bayes) Estimate (9)

Advantages

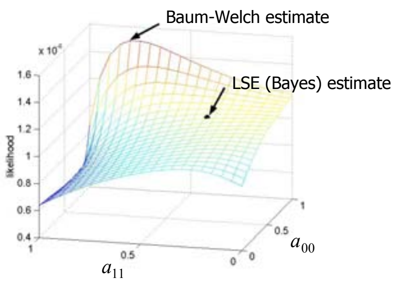
- Closer than B-W estimates to the true parameters when the data size is small.
- Online computation is possible.
- Finds the exact expected values (unbiased).
- One-time computation.

LSE (Bayes) Estimate (10)

Disadvantage

- Computational complexity grows still exponentially in the state space size.

Example 1



Example 2: B-W and LSE estimates with a small data set (1)

Outline:

Generate 200 θ - values, randomly with respect to the determinant of $A = \{a_{ij}\}$ and to the difference $b_{00} - b_{10}$.

↓

For each θ, generate a set of $\{X^{1,n}, Y^{1,n}\}$, $n = 100$, and obtain the estimates.

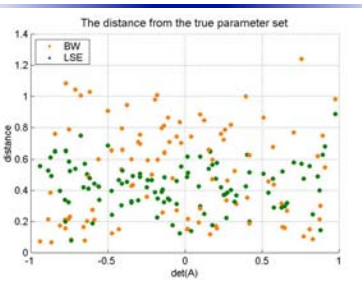
As for B-W estimates:

Find 10 estimates using 10 randomly picked initial estimates.

↓

Pick the one with the largest basin.

Example 2: B-W and LSE estimates with a small data set (2)



The first 100 are plotted. On the average, the B-W estimates (orange dots) were farther away from the true parameters than LSE ones (green dots) by 0.073 and less stable.

Referenes

- J. A. Bilms, "A gentle tutorial of the EM algorithm and its application to parameter estimation for Gaussian mixture and hidden Markov models," International Computer Science Institute, Tech. Rep. ICSI-TR-97-021, April 1998.
- L. E. Baum, "An inequality and associated maximization technique in statistical estimation for probabilistic functions of Markov processes," *Inequalities*, vol. 3, pp. 1-8, 1972.
- J. Murakami, "Parameter estimate of a hidden Markov chain," Unpublished Ph.D. Dissertation, Arizona State University, Tempe, AZ, USA, May 2005.

Acknowledgement

The presenter (J. Murakami) would like to thank Keio University and NZIMA for their support.