HMM: The Learning Problem

Sorin Istrail

Department of Computer Science
Brown University, Providence
sorin@cs.brown.edu

March 31, 2020
Outline

1. Outline
2. The Basic Three HMM Problems
3. HMM: basic variables and probabilities
4. Solution to Problem 3: The Expectation-Maximization Algorithm
   - The Reestimation Equations
   - The EM Algorithm
   - Baum’s Q-function and the Baum-Welch Theorem
   - The Stochastic Constrains
   - Lagrangean multipliers for the solution of the max-optimization
• Observation sequence $O = o_1, ..., o_T$ and HMM model $\lambda = (A, B, \pi)$

**Problem 1:** The Evaluation Problem
Given: $O, \lambda$
Compute: $P(O | \lambda)$ the probability of the observation sequence given the HMM model

**Problem 2:** The Decoding Problem
Given: $O, \lambda$
Compute: A sequence of states $Q$ for the observation sequence $O$, $Q = q_1, ..., q_T$ which optimally “explains” the observation sequence.

**Problem 3:** The Learning Problem
Given: $O$
Compute: the parameters of an HMM model $\lambda$ that maximizes the probability $P(O | \lambda)$ of observing $O$ in the model $\lambda$
Problem 1

- Observation sequence $O = o_1, ..., o_T$
  and HMM model $\lambda = (A, B, \pi)$

- **Problem 1: The Evaluation Problem**
  Given: $O, \lambda$
  Compute: $P(O | \lambda)$ the probability of the observation sequence given the HMM model
Problem 2

- Observation sequence $\mathcal{O} = o_1, \ldots, o_T$ and HMM model $\lambda = (A, B, \pi)$

**Problem 2: The Decoding Problem**

Given: $\mathcal{O}, \lambda$

Compute: A sequence of states $Q$ for the observation sequence $\mathcal{O}$, $Q = q_1, \ldots, q_T$ which optimally “explains” the observation sequence.
Problem 3

- Observation sequence $O = o_1, \ldots, o_T$
  and HMM model $\lambda = (A, B, \pi)$

**Problem 3:** The Learning Problem
Given: $O$
Compute: the parameters of an HMM model $\lambda$ that maximizes the probability $P(O \mid \lambda)$ of observing $O$ in the model $\lambda$
Elements of an HMM

1. *N* is the number of **states** \( S = \{S_1, \ldots, S_N\} \).
   The HMM process proceeds in discrete units of time, \( t = 1, 2, 3, \ldots \).
   The state at time \( t \) is denoted by \( q_t \).

2. *M* is the number of distinct **observation symbols** per state \( V = v_1, \ldots, v_M \).

3. The **transition probability distribution** is given by \( A = \{a_{ij}\} \),
   where
   \[ a_{ij} = P[q_{t+1} = S_j \mid q_t = S_i], \quad 1 \leq i, j \leq N \]

4. The **observation symbols probability distribution** in state \( j \)
   is given by
   \[ B = \{b_j(k) = P[v_k \text{ at time } t \mid q_t = S_j], \quad 1 \leq j \leq N, 1 \leq k \leq M \}

5. The **initial state distribution** is given by
   \[ \pi = \{\pi_i = P[q_1 = S_i], \quad 1 \leq i \leq N \} \]
Basic variables and probabilities

- **a sequence of states** is \( Q = \{ q_1, q_2, \ldots, q_T \} \)
- The **probability of observing the sequence** \( O \) **in sequence of states** \( Q \) is

\[
P(O \mid Q) = \prod_{i=1}^{T} P(o_i \mid q_i)
\]

\[
P(O \mid Q) = b_{q_1}(o_1) \ldots b_{q_T}(o_T)
\]

\[
P(Q) = \pi q_1 a_{q_1 q_2} a_{q_2 q_3} \ldots a_{q_{T-1} q_T}
\]

\[
P(O, Q) = P(O \mid Q)P(Q)
\]
the probability of observing $O$ is

$$P(O) = \sum_{all\ Q} P(O \mid Q)P(Q)$$

$$= \sum_{q_1 \ldots q_T} \pi_{q_1} b_{q_1}(o_1) a_{q_1 q_2} b_{q_2}(o_2) \ldots a_{q_{T-1} q_T} b_{q_T}(o_T)$$
the Forward variable $\alpha_t(i)$

- The Forward variable is defined by

$$\alpha_t(i) = P(o_1 o_2 ... o_t, q_t = S_i)$$

- i.e., the probability of the prefix of the sequence of observations $o_1 ... o_t$ until time $t$ and being in state $S_i$ at time $t$
The Backward variable $\beta_t(i)$

- The Backward variable is defined by

$$\beta_t(i) = P(o_{t+1}o_{t+2}...o_T, q_t = S_i)$$

- i.e., the probability of the suffix of the sequence of observations $o_{t+1}o_{t+2}...o_T$ until end of sequence $t$ and being in state $S_i$ at time $t$
The \( \delta_t(i) \) variable is defined by

\[
\delta_t(i) = \max_{q_1...q_{t-1}} P(q_1...q_{t-1}q_t, o_1...o_{t-1}o_t)
\]

i.e., the best score (highest probability) along the single path, at time \( t \) which accounts for the first \( t \) observations and ends in state \( S_i \).
By far the most difficult of the three problems.

We want to adjust the parameters of the model \( \lambda = (A, B, \pi) \) to maximize the probability of observing the sequence in the model.

There is no exact analytical solution to this problem.

Both Problem 1 and Problem 2 have solutions given by algorithms that we presented in CS 1810. Those algorithms are exact and having computing time \( O(N^2 T) \).
We can choose $\lambda' = (A', B', \pi')$ such that $P(O | \lambda')$ is locally maximal.

We use the **Baum-Welch Algorithm**. This is an iterative algorithm. We iterate until no improvement is possible. At that point we reached a local maxima. For this iteration we are using a method for reestimation of the HMM parameters.
We first define a new variable $\xi$

$\xi_t(i,j) = \text{the probability of being in state } S_i \text{ at time } t \text{ and state } S_j \text{ at time } t + 1, \text{ given the model and the observation sequence}$

$\xi_t(i,j) = P(q_t = S_i, q_{t+1} = S_j \mid \mathcal{O}, \lambda)$
From the definition of $\alpha$ and $\beta$ variables we can write $\xi$ as follows:

$$\xi_t(i, j) = \frac{\alpha_t(i) a_{ij} b_j(o_{t+1}) \beta_{t+1}(j)}{P(O \mid \lambda)}$$

$$= \frac{\alpha_t(i) a_{ij} b_j(o_{t+1}) \beta_{t+1}(j)}{\sum_{i'=1}^{N} \sum_{j'=1}^{N} \alpha_t(i') a_{i'j'} b_{j'}(o_{t+1}) \beta_{t+1}(j')}$$
The numerator is:

\[ P(q_t = S_i, q_{t+1} = S_j, O | \lambda) \]

and the denominator is the normalization factor to give the probability:

\[
P(O | \lambda) = \sum_{i' = 1}^{N} \sum_{j' = 1}^{N} \alpha_t(i') a_{i'j'} b_{j'}(o_{t+1}) \beta(j')
\]
the Gamma variable $\gamma_t(i)$

- As $\gamma_t(i)$ is the probability of being in state $S_i$ at time $t$ given the observation sequence and model we have

$$\gamma_t(i) = \sum_{j=1}^{N} \xi_t(i,j)$$
The expected number of times state $S_i$ is visited or equivalently the expected number of transitions made from $S_i$ is

\[
\sum_{t=1}^{T-1} \gamma_t(i) = \text{the expected number of transitions from } S_i
\]

Similarly,

\[
\sum_{t=1}^{T-1} \xi_t(i, j) = \text{the expected number of transitions from } S_i \text{ to } S_j
\]
Reestimating $\pi_i$

- A set of reasonable reestimations for the parameters $\pi, A, B$ are given as follows:

  $$\bar{\pi} = \gamma_1(i), 1 \leq i \leq N$$

- i.e., the expected frequency (number of times) in state $S_i$ at time $(t = 1)$ is $= \gamma_1(i)$
Reestimating $A = \{a_{ij}\}$

$$\bar{a}_{ij} = \frac{\sum_{t=1}^{T-1} \xi_t(i,j)}{\sum_{t=1}^{T-1} \gamma_t(i)}$$

i.e., (expected number of transitions from $S_i$ to $S_j$)/ (exected number of transitions from $S_i$)
Reestimating $B = b_j(k)$

\[ \bar{b}_j(k) = \frac{\sum_{t=1,o_t=v_k}^T \gamma_t(j)}{\sum_{t=1}^T \gamma_t(j)} \]

i.e., (expected number of times in state $S_j$ observing observation symbol $v_k$) / (expected number of times in state $S_j$)
Let the current model $\lambda = (A, B, \pi)$

Compute the above reestimation to get a new model $\bar{\lambda} = (\bar{A}, \bar{B}, \bar{\pi})$

Then

1. $\lambda$ is a local optimum, i.e., $\lambda = \bar{\lambda}$, or
2. $\bar{\lambda}$ is more likely that $\lambda$ in the sense that

$$P(O | \bar{\lambda}) > P(O | \lambda)$$

, i.e., we have found a new model $\bar{\lambda}$ from which the observation sequence is more likely to have been produced.
The maximum likelihood HMM estimate

- If we consider this reestimation, the final result of this reestimation procedure is called a maximum likelihood estimate of the HMM.
- The Forward-Backward algorithm leads to a local maxima.
Baum’s Q-function and the Baum-Welch Theorem

The reestimation formulas can be derived directly by maximization (using constrained optimization) of Baum’s auxiliary function:

\[
\max_{\lambda} Q(\lambda, \lambda) = \sum_Q P(Q | O, \lambda) \log(P(O, Q | \lambda))
\]

Baum-Welch Theorem:

\[
\max_{\lambda} (Q(\lambda, \lambda))
\]

implies that

\[
P(O | \lambda) \geq P(O | \lambda)
\]
The reestimation procedure can be implemented as the **Expectation-Maximization (EM) Algorithm** due to Dempster, Laird and Rubin (1977).

- **E-step** (Expectation) is the calculation of the Baum’s auxiliary function $Q(\lambda, \bar{\lambda})$
- **M-step** (Maximization) is the maximization of $\bar{\lambda}$
The stochastic contraints

- The stochastic contraints for the model are automatically satisfied at each iteration:

  \[ \sum_{i=1}^{N} \bar{\pi}_i = 1, 1 \leq j \leq N \]

  \[ \sum_{i=1}^{N} \bar{a}_{ij} = 1, 1 \leq j \leq N \]

  \[ \sum_{k=1}^{M} \bar{b}_j(k) = 1 \]
We can solve the parameter estimation problem as a constraint optimization problem for

$$P(O \mid \lambda)$$

under the stochastic constraints by using the Lagrangean multipliers method. It shows that $P$ is maximized when the following hold:
The Basic Three HMM Problems
HMM: basic variables and probabilities
Solution to Problem 3: The Expectation-Maximization Algorithm

Lagrangean multipliers for the solution of the optimization

\[ \pi_i = \frac{\pi_i \frac{\partial P}{\partial \pi_i}}{\sum_{k=1}^{N} \pi_k \frac{\partial P}{\partial \pi_k}} \]

\[ a_{ij} = \frac{a_{ij} \frac{\partial P}{\partial a_{ij}}}{\sum_{k=1}^{N} a_{ik} \frac{\partial P}{\partial a_{ik}}} \]

\[ b_i(k) = \frac{b_i(k) \frac{\partial P}{\partial b_i(k)}}{\sum_{l=1}^{M} b_i(l) \frac{\partial P}{\partial b_i(l)}} \]
By appropriate manipulation of those formulas the right-hand sides of each equation can be readily converted to be identical to the right-sides of the EM algorithm reestimations.

This shows that the reestimation formulas are indeed exactly correct at local optimal points of $P(O \mid \lambda)$. 